Recursive estimation of conditional spatial medians and conditional quantiles

Citation for published version (APA):
Belitser, E. N., & de Andrade Serra, P. J. (2014). Recursive estimation of conditional spatial medians and conditional quantiles. *Sequential Analysis*, *33*(4), 519–538. https://doi.org/10.1080/07474946.2014.961856

DOI:
10.1080/07474946.2014.961856

Document status and date:
Published: 01/01/2014

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Sequential Analysis, 33: 519–538, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 0747-4946 print/1532-4176 online
DOI: 10.1080/07474946.2014.961856

Recursive Estimation of Conditional Spatial Medians and Conditional Quantiles

Eduard Belitser[1] **and Paulo Serra**[2]
[1]Department of Mathematics, VU University Amsterdam, Amsterdam, The Netherlands
[2]Institut für Mathematische Stochastik, Georg-August-Universität Göttingen, Göttingen, Germany

Abstract: We consider the problem of constructing an on-line (recursive) algorithm for tracking a conditional spatial median, a center of a multivariate distribution. In the one-dimensional case we also track conditional quantiles of arbitrary level. We establish a nonasymptotic upper bound for the L_p-risk of the algorithm, which is then minimized under different assumptions on the magnitude of the variation of the spatial median or quantile. We derive convergence rates for the examples we consider.

Keywords: Conditional spatial median; Convergence rate; Half-space symmetric distributions; Recursive algorithm; Sequential estimation.

Subject Classifications: 62L12; 62G20; 62L20.

1. INTRODUCTION

Often, in applications, one wishes to recover a functional dependence between different parameters of the underlying distribution on the basis of observations from that distribution. Nonparametric model regression is one common approach to this problem. Strictly speaking, regression is the conditional expectation of one random variable (vector) given another one. Thus, certain moment conditions in regression models are unavoidable. Moreover, in additive regression stronger structural conditions are usually imposed on the noises, for example, normality of the errors and moment restrictions. However, sometimes it is desirable to minimize the conditions on the moments (and the form) of the distribution of the noises. If, for example, we only assume that the error at each moment has a zero quantile

Received October 6, 2013, Revised April 24, 2014, Accepted May 18, 2014
Recommended by Nitis Mukhopadhyay

Address correspondence to P. Serra, Institut für Mathematische Stochastik, Georg-August-Universität Göttingen, Goldschmidtstrasse 7, D-37077 Göttingen, Germany; E-mail: pdeandr@uni-goettingen.de

of certain fixed level (for identifiability purposes), we obtain the so-called quantile regression model first introduced into the literature in Bassett and Koenker (1978); see Koenker (2005) for a nice account on this topic.

The quantile regression model is quite important in fields such as econometrics, social sciences, and ecology. In these areas, one often studies response variables whose relation with its measured predictors is complex. In such cases, the conditional expectation of the response variable might simply be insensitive to these relations and will provide a poor description of the underlying phenomenon. Error bounds on certain regression estimates can be viewed as crude quantile regressions (cf. Takeuchi et al., 2006) but the levels of these quantilels can be estimated directly. In such situations, we get a more comprehensive and robust description of the data by estimating conditional quantiles of different levels, rather than the conditional mean. This seems to be of particular relevance in applications—for example, in ecology—where data often display heterogeneous variances; cf. Cade and Noon (2003).

However, a notion of multivariate quantile is not straightforward. Several attempts have been made to define a multidimensional analogue of the median, as some form of *center* of a distribution. These are usually based on the notion of depth of a point, first proposed by Tukey (1974a,b, 1975). Roughly speaking, given a distribution $\mathbb{P}$ with support in $\mathbb{R}^d$, the depth of a point $x \in \mathbb{R}^d$ with respect to $\mathbb{P}$—call it $\text{depth}(x, \mathbb{P})$—measures the *centrality* of x in $\mathbb{P}$. Depth functions should be chosen in such a way that they provide a center-outward ordering of the points $x \in \mathbb{R}^d$ via contours of the function $x \mapsto \text{depth}(x, \mathbb{P})$. Tukey's depth, the *half-space depth*, is introduced in the next section. In the one-dimensional case, points of maximum half-space depth are medians, making this a consistent generalization of the concept median for multidimensional distributions. Besides, the half-space depth behaves well and has many attractive properties (cf. Zuo and Serfling, 2000); other notions of depth function can also be found in the cited article. Points of center defined based on the half-space depth function can also be shown to have high breakdown points as shown by Donoho and Gasko (1992)—at least $1/(d+1)$ in d dimensions.

In this article, we work with distributions that have a natural notion of center. These are *half-space symmetric distributions*—d-dimensional distributions $\mathbb{P}$ for which there exists a point $\theta \in \mathbb{R}^d$ such that any half-space H that contains θ verifies $\mathbb{P}(H) \geq 1/2$. This point θ is the center of the distribution in that it has maximal half-space depth. For absolutely continuous distributions, this point is unique and we refer to it as the *spatial median*.

In this article, we are concerned with the problem of recovering a d-variate, conditional, spatial median function. More precisely, suppose that at each time moment $k \in \mathbb{N}$ we observe a random vector X_k and the problem is to recover its spatial median by using observations available by this time moment. The spatial median may evolve with time so that we are dealing with nonparametric situation. Another important complicating factor is that we do not assume the traditional independence of the observations. In fact, the observations can be arbitrarily dependent and at time moment $k+1$ we would like to recover the conditional spatial median of X_{k+1} given $\boldsymbol{X}_k = (X_1, \ldots, X_k)$. This necessarily puts our problem into the sequential estimation framework. It is desirable to design a sequential estimation procedure such that the estimate of the evolving conditional spatial median at the current time moment is based on the estimate of the conditional spatial median at the previous time moment and a small correction based on the current observation. A procedure of such kind was first proposed by Robbins and Monro (1951), which gave rise to the area of stochastic approximation algorithms. There is an enormous body of literature on this topic by now. We also propose a recursive procedure for estimating the evolving conditional spatial median and derive a general upper bound for its quality in terms of L_p-risk. Clearly, if the the underlying conditional spatial median oscillates uncontrollably, no estimation procedure can provide good accuracy. However, if oscillations of the time varying conditional spatial median can be controlled in one way or other, estimation should be possible. Informally, one could say that some sort of accumulation of information occurs. For example, either variation in values of conditional spatial median is slowing down with time or observations are made more frequently. We demonstrate that this leads to useful non-void upper bounds by choosing appropriate settings. Depending on how accumulation of information in time occurs, we derive different asymptotic regimes for estimation quality by applying our general nonasymptotic upper bound. In the one-dimensional case we can estimate sequentially time varying conditional quantiles of arbitrary level $\alpha_k \in (0, 1)$, $k \in \mathbb{N}$, not only a drifting median. A related problem for independent one dimensional observations was treated in Belitser and Serra (2013).

This article is structured as follows. In Section 2 we present some basic definitions, explain the model in detail, specify our assumptions on the distribution of the observations, and define the sequential estimation procedure. Section 3 contains some auxiliary lemmas and our main results. Different variational setups for the drifting spatial median (quantiles of changing levels in one dimensional case) are treated in Section 4, including estimation of a static median, a stabilizing spatial median, and a spatial median that varies as a Lipschitz function. Finally, the proofs of the main results from Section 3 are in Appendix A and auxiliary technical results that are used in these proofs are in Appendix B.

2. PRELIMINARIES

Throughout, for $d \in \mathbb{N}$, $x, y \in \mathbb{R}^d$ (columns by default), $\|x\|_p$ is the l_p norm (with $p \geq 1$) on $\mathbb{R}^d$; $\|x\| = \|x\|_2$ and $x^T y$ are the usual Euclidean norm and inner product in $\mathbb{R}^d$; $I\{A\}$ is the indicator function of a set A. We use bold symbols (both upper- and lowercase) to represent matrices and families of vectors. Denote $\boldsymbol{x}_k = (x_1, \ldots, x_k)$, $x_i \in \mathbb{R}^d$, $k \in \mathbb{N}_0$, with the convention that $\boldsymbol{x}_0$ is an empty vector. For a $(d \times d)$-matrix $\boldsymbol{M}$, let $\lambda_{(i)}(\boldsymbol{M})$, $i = 1, \ldots, d$, denote the ith largest eigenvalue of $\boldsymbol{M}$. Denote by $\boldsymbol{O}$ the zero matrix and by $\boldsymbol{I}$ the identity matrix, whose dimensions will be determined by the context. We adopt the convention that $\sum_{i \in \varnothing} \boldsymbol{A}_i = \boldsymbol{O}$ and $\prod_{i \in \varnothing} \boldsymbol{B}_i = \boldsymbol{I}$ for matrices $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ of appropriate dimensions. When applied to matrices, $\| \cdot \|_p$ represents the matrix norm induced by the l_p vector norm: $\|\boldsymbol{M}\|_p = \max_{\|x\|_p = 1} \|\boldsymbol{M}x\|_p$. Given $x \in \mathbb{R}^d$ and a unit vector w in $\mathbb{R}^d$, define

$$H(x, w) = \{y \in \mathbb{R}^d : w^T y \geq w^T x\},$$

the closed half-space that contains the point $x + w$ and is delimited by the hyperplane containing x which is orthogonal to w.

For any distribution $\mathbb{P}$ with support in $\mathbb{R}^d$, the *half-space depth* (cf. Tukey, 1974a,b, 1975) of $x \in \mathbb{R}^d$ with respect to $\mathbb{P}$ is defined as

$$\operatorname{depth}(x, \mathbb{P}) = \inf\big\{\mathbb{P}(H) : H \text{ is a closed half-space, } x \in H\big\}. \tag{2.1}$$

The fundamental property of depth functions is that they induce an ordering of the elements in the support of the distribution $\mathbb{P}$ from a "center"— a point of maximal depth—outwards via the contours of the depth function. In case $d = 1$, the center is a point θ such that $\operatorname{depth}(\theta, \mathbb{P}) \geq 1/2)$ and this is always a median of $\mathbb{P}$. Besides, for $d = 1$ the quantiles of level α and $1 - \alpha$, $\alpha \in [0, 1/2)$, can be associated with the depth function as follows:

$$\theta(\alpha) = \inf\big\{x \in \mathbb{R} : \operatorname{depth}(x, \mathbb{P}) \geq \alpha\big\} \quad \text{and} \quad \theta(1-\alpha) = \sup\big\{x \in \mathbb{R} : \operatorname{depth}(x, \mathbb{P}) \geq \alpha\big\},$$

which are always well defined. Define $\theta(1/2) = \inf\big\{x \in \mathbb{R} : \operatorname{depth}(x, \mathbb{P}) \geq 1/2\big\}$.

When $d \geq 2$, it is not automatic that a given distribution has a "natural" center θ in that there might not be a unique point that maximizes a given depth function. In this article, we work with a particular family of distributions with a proper notion of center, the so-called *half-space symmetric* distributions; cf. Zuo and Serfling (2000). A distribution $\mathbb{P}$ is said to be half-space symmetric about some θ in the support of $\mathbb{P}$, if $\mathbb{P}(X \in H) \geq 1/2$ for every closed half-space H containing θ. Note that in one dimension, all distributions are half-space symmetrical about their medians. In the general, multidimensional case, this center is unique for absolutely continuous, half-space symmetrical distributions and we will refer to it as the *spatial median* of $\mathbb{P}$. It is straightforward to check that for such distributions the half-space depth function has a unique point of maximum at θ and its maximal value is $1/2$; that is, $\theta = \theta(1/2)$ as defined before.

Suppose then that at each time moment $k \in \mathbb{N}$ we observe a random vector X_k, so that by time n we have n observations $\boldsymbol{X}_n = (X_1, \ldots, X_n)$. Let $\mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1})$ denote the conditional distribution of X_k given $\boldsymbol{X}_{k-1} = \boldsymbol{x}_{k-1}$; that is, $\mathbb{P}_k(A|\boldsymbol{x}_{k-1}) = \mathbb{P}(X_k \in A|\boldsymbol{X}_{k-1} = \boldsymbol{x}_{k-1})$ for a measurable $A \subseteq \mathbb{R}^d$. Further, let $\mathscr{X} \subset \mathbb{R}^d$ represent the (common) support of each observation so that $\mathbb{P}(\boldsymbol{X}_k \in \mathscr{X}^k) = 1$. We assume that for each $\boldsymbol{x}_{k-1} \in \mathscr{X}^{k-1}$, $k \in \mathbb{N}$, the conditional distributions $\mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1})$ are absolutely continuous (with respect to the Lebesgue measure) and half-space symmetrical about some $\theta_k = \theta_k(1/2) = \theta_k(\boldsymbol{x}_{k-1}, 1/2)$, which is the only point in $\mathscr{X}$ of maximal depth satisfying

$$\operatorname{depth}\big(\theta_k(\boldsymbol{x}_{k-1}, 1/2), \mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1})\big) \geq 1/2. \tag{2.2}$$

The goal is to sequentially estimate a predictable process $\{\theta_k, k \in \mathbb{N}\}$; that is, at each time moment $k \in \mathbb{N}$, $\theta_k = \theta_k(\boldsymbol{X}_{k-1})$ is estimated by using the observations $\boldsymbol{X}_k = (X_1, \ldots, X_k)$ available by that moment. When $d = 1$, $\theta_k = \theta_k(\boldsymbol{X}_{k-1}, \alpha_k)$ will be a sequence of quantiles of level $\alpha_k \in (0, 1)$, $k \in \mathbb{N}$, which are fixed in advance. When $d \geq 2$, $\theta_k = \theta_k(\boldsymbol{X}_{k-1}) = \theta_k(\boldsymbol{X}_{k-1}, 1/2)$ will be the spatial median as defined before. Ideally, we want our procedure to approach θ_k as time progresses. If, however, this is impossible, then the procedure should at least stay in proximity of θ_k as close as possible. Until now we imposed few assumptions on the observations $X_1, X_2, \ldots$. In fact, the observations can be arbitrarily distributed and can have an arbitrary dependence structure. Clearly, the stated problem in its full generality has no feasible solution. Thus, in order to come up with some nonvoid results, we need to impose some assumptions on the conditional distributions of X_k given $\boldsymbol{X}_{k-1} = \boldsymbol{x}_{k-1}$, $k \in \mathbb{N}$, at the same time trying to keep these conditions as weak as possible. From now on, we will call θ_k spatial median keeping in mind that the same notation is used for a quantile of level $\alpha_k \in (0, 1)$ in case $d = 1$.

We impose the following conditions on the conditional distributions $\mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1})$, $k \in \mathbb{N}$.

(A) The distributions $\mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1})$ are absolutely continuous and half-space symmetrical and for some positive b, B, δ, the following inequalities hold for any $\epsilon \in (0, \delta]$ and any unit vectors $v, w \in \mathbb{R}^d$: almost surely

$$\epsilon b(v^T w) \leq (v^T w)\big[\mathbb{P}(X_k \in H(\theta_k - \epsilon v, w)|\boldsymbol{X}_{k-1}) - 1/2\big] \leq \epsilon B(v^T w), \quad k \in \mathbb{N},$$

where $\theta_k = \theta_k(\boldsymbol{X}_{k-1}, 1/2)$ is the spatial median of $\mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1})$.

(B) The support $\mathscr{X}$ is bounded so that $\sup_{x\in\mathscr{X}} \|x\| \leq C_{\mathscr{X}}$ and the spatial median θ_k takes values in some compact subset $\Theta \subseteq \mathscr{X}$ so that $\sup_{\theta\in\Theta} \|\theta\| \leq C_\Theta$, for some $0 \leq C_\Theta \leq C_{\mathscr{X}}$.

Remark 2.1. The requirement of bounded support of observations seems to be restrictive but it is reasonable from a practical perspective. One can think of X_k as truncated versions of some Y_k (with an unbounded support): $X_k = Y_k I\{\|Y_k\| \leq C_{\mathscr{X}}\} + Y_k(C_{\mathscr{X}}/\|Y_k\|)I\{\|Y_k\| > C_{\mathscr{X}}\}$.

Note that under the assumption that the distributions $\mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1})$ are half-space symmetrical, the fraction in (A) is trivially positive. This is because any half-space containing θ_k will contain at least half of the mass of the distribution. Furthermore, if such a hyperspace containing the conditional spatial median θ_k is moved along a (fixed) direction $-\epsilon v$, then the mass captured by this hyperspace changes lineally in ϵ.

When $d = 1$, condition (A) may be replaced with the following assumption:

(C) For some positive b, B, δ and a sequence $\alpha_k \in (0, 1)$, $k \in \mathbb{N}$, the following inequalities hold for any $\epsilon \in [-\delta, \delta]$: almost surely

$$\epsilon b \leq \mathbb{P}(X_k \in H(\theta_k + \epsilon, -1)|\boldsymbol{X}_{k-1}) - \alpha_k \leq \epsilon B, \quad k \in \mathbb{N},$$

where $\theta_k = \theta(\boldsymbol{X}_{k-1}, \alpha_k)$ is the conditional quantile of level α_k of $\mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1})$.

Note that the probability in the previous display is simply $\mathbb{P}(X_k \leq \theta_k + \epsilon|\boldsymbol{X}_{k-1})$. Condition (C) is appropriate in the case when we are interested in sequential estimation of arbitrary quantiles of a one-dimensional distribution. When $\alpha_k = 1/2$, $k \in \mathbb{N}$, condition (C) reduces to condition (A) with $d = 1$.

Condition (A) (or (C)) is fulfilled if, for example, the conditional distributions $\mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1})$ are absolutely continuous with conditional densities $f_k(\cdot|\boldsymbol{x}_{k-1})$ such that for some positive b, B, δ,

$$0 < b \leq f_k(x|\boldsymbol{x}_{k-1}) \leq B < \infty, \quad \boldsymbol{x}_{k-1} \in \mathscr{X}^{k-1}, \; k \in \mathbb{N},$$

for almost all (with respect to the Lebesgue measure) $x \in \mathscr{X}$ such that $\|x - \theta_k\| \leq \delta$.

Remark 2.2. Notice that, even under the above conditions, we deal with a rather general framework: the observations can be dependent and not identically distributed. Besides, our problem is stated in a robust setting in the sense that we do not assume anything about the moments of the observations X_k, they simply may not exist.

Condition (A) is rather natural for the important particular case of independent observations X_k, $k \in \mathbb{N}$. In this case, the conditional spatial median θ_k becomes unconditional (θ_k does not depend on $\boldsymbol{X}_{k-1}$) and bounded uniformly in k under our assumption that $\mathscr{X}$ is a bounded set. The observations can then be expressed in the form $X_k = \theta_k + \xi_k$, $k \in \mathbb{N}$, with independent noises ξ_k. Condition (A) means that the noises ξ_k have zero spatial median and their probability distribution behaves regularly in the neighborhood of zero in the sense that they degenerate neither into zero nor into delta function. This condition does not seem too restrictive for another important case: when the observations come from a Markov model. In this case, the conditional distributions $\mathbb{P}_k$ depends only on two arguments, x_k and x_{k-1}.

Now let $\boldsymbol{b} = \{b_1, \dots, b_d\}$ be any orthonormal basis for $\mathbb{R}^d$, which will be fixed through the remainder of this article. Consider a random vector D such that $\mathbb{P}(D = b_i) = 1/d$, $i = 1, \dots, d$. We call such a random vector a *random direction in* $\mathbb{R}^d$.

Introduce the *shift function*

$$S(u, v, w) = \big(I\{u \in H(v, w)\} - 1/2\big)w, \quad u \in \mathscr{X} \subset \mathbb{R}^d, \ v \in \mathbb{R}^d, \tag{2.3}$$

where w is a unit vector in $\mathbb{R}^d$. Note that this vector valued function takes the values $w/2$ or $-w/2$ depending on whether the argument u belongs to the half-space $H(v, w)$ or not. For estimating arbitrary quantiles of level α in the case $d = 1$, we use a different shift function, namely,

$$R(u, v, \alpha) = \alpha - I\{u \leq v\}, \quad u \in \mathscr{X} \subset \mathbb{R}, \ v \in \mathbb{R}, \tag{2.4}$$

where $\alpha \in (0, 1)$. This shift function only takes the values $\alpha - 1$ and α depending respectively on whether $u \leq v$ or not. Note that this is simply the shift function (2.3) with $d = 1$, $w = -1$ and with $1/2$ replaced with α.

Let $\{\gamma_k, \ k \in \mathbb{N}\}$ be a nonnegative sequence bounded by some constant Γ:

$$0 \leq \gamma_k \leq \Gamma, \quad k \in \mathbb{N}. \tag{2.5}$$

Define the recursive algorithm for estimating a conditional spatial median in $\mathscr{X} \subset \mathbb{R}^d$, $d \geq 2$:

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \gamma_k S(X_k, \hat{\theta}_k, D_k), \quad k \in \mathbb{N}, \tag{2.6}$$

where D_k is an independent sequence of random directions. An algorithm for sequential estimating an α_k-level quantiles in $\mathscr{X} \subset \mathbb{R}$ is as follows:

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \gamma_k R(X_k, \hat{\theta}_k, \alpha_k), \quad k \in \mathbb{N}, \tag{2.7}$$

where $\alpha_k \in (0, 1)$, $k \in \mathbb{N}$. In both cases the sequence of *step sizes* $\{\gamma_k, \ k \in \mathbb{N}\}$ satisfies (2.5) and $\hat{\theta}_1 \in \Theta$ is some fixed initial value from Θ.

3. MAIN RESULTS

In this section we formulate the main results of this article. We start with two technical lemmas that are needed in the proofs of the main theorems. All proofs can be found in Appendix A. For the sake of brevity, we denote by θ_k the parameter value to be estimated at the time moment k: for $d \geq 2$ this will be a spatial median $\theta_k = \theta_k(\boldsymbol{X}_{k-1}) = \theta_k(\boldsymbol{X}_{k-1}, 1/2)$ and for $d = 1$ an α_k-level quantile $\theta_k = \theta_k(\boldsymbol{X}_{k-1}, \alpha_k)$. It will be clear from the context which case we are referring to. Introduce $\mathscr{F}_k = \sigma(\boldsymbol{X}_k)$, $k \in \mathbb{N}$, the σ-algebra generated by $\boldsymbol{X}_k = (X_1, \ldots, X_k)$, and let $\rhd_0$ be the trivial σ-algebra.

Lemma 3.1. *Let* $\{\hat{\theta}_k, k \in \mathbb{N}\}$ *be defined by* (2.6). *Then* $\|\hat{\theta}_k\| \leq \sqrt{d}(C_{\mathscr{X}} + \Gamma/2)$, $k \in \mathbb{N}$. *This implies that* $\|\hat{\theta}_k - \theta_k\| \leq C = C_\Theta + \sqrt{d}(C_{\mathscr{X}} + \Gamma/2)$, $k \in \mathbb{N}$.

Lemma 3.2. *For a fixed sequence* $\alpha_k \in (0, 1)$, $k \in \mathbb{N}$, *let* $\{\hat{\theta}_k, k \in \mathbb{N}\}$ *be defined by* (2.7). *Then* $|\hat{\theta}_k| \leq C_{\mathscr{X}} + \Gamma$, $k \in \mathbb{N}$. *This implies that* $|\hat{\theta}_k - \theta_k| \leq c = C_\Theta + C_{\mathscr{X}} + \Gamma$, $k \in \mathbb{N}$.

These lemmas are used in the proof of Lemma 3.3 and of Lemma 3.4, respectively, and can be seen as preliminary rough versions of our main theorems (Theorems 3.1 and 3.2 below), stating that the approximations given by algorithms (2.6) and (2.7) do not stray to infinity and live on a compact set. The following lemmas give information about the average behavior of the shift functions (2.3) and (2.4).

Lemma 3.3. *Let the function* $S(u, v, w)$ *be defined by* (2.3) *and the sequence* $\{\hat{\theta}_k, k \in \mathbb{N}\}$ *be defined by* (2.6). *Then* $\|S(u, v, w)\| \leq 1/2$ *uniformly over* $u \in \mathscr{X}$, $v \in \mathbb{R}^d$ *and over all unit vectors* w *in* $\mathbb{R}^d$. *Moreover, if conditions* (*A*) *and* (*B*) *are fulfilled, then*

$$\mathbb{E}\big[S(X_k, \hat{\theta}_k, D_k)|\mathscr{F}_{k-1}\big] = -\boldsymbol{M}_k(\hat{\theta}_k - \theta_k), \quad k \in \mathbb{N},$$

for some sequence of $\mathscr{F}_{k-1}$*-measurable random matrices* $\boldsymbol{M}_k = \boldsymbol{M}_k(\boldsymbol{X}_{k-1})$ *such that almost surely* $\lambda_1 \leq \lambda_{(1)}(\boldsymbol{M}_k) \leq \lambda_{(d)}(\boldsymbol{M}_k) \leq \lambda_2$, $k \in \mathbb{N}$, *for constants* $0 < \lambda_1 \leq \lambda_2 < \infty$. (*The constants* λ_1 *and* λ_2 *depend on* b, B, δ *from* (*A*), *and on* C_Θ *and* $C_{\mathscr{X}}$ *from* (*B*) *via* C *from Lemma* 3.1.)

Lemma 3.4. *Let the function* $R(u, v, \alpha)$ *be defined by* (2.4), *the sequence* $\{\hat{\theta}_k, k \in \mathbb{N}\}$ *be defined by* (2.7) *and* $\alpha_k \in (0, 1)$. *Then* $|R(u, v, \alpha)| \leq \max\{\alpha, 1 - \alpha\}$ *uniformly over* $u \in \mathscr{X}$, $v \in \mathbb{R}$ *and* $\alpha \in (0, 1)$. *Moreover, if conditions* (*C*) *and* (*B*) *are fulfilled, then*

$$\mathbb{E}\big[R(X_k, \hat{\theta}_k, \alpha_k)|\mathscr{F}_{k-1}\big] = -M_k(\hat{\theta}_k - \theta_k), \quad k \in \mathbb{N},$$

for some sequence of $\mathscr{F}_{k-1}$*-measurable random variables* $M_k = M_k(\boldsymbol{X}_{k-1})$ *such that almost surely* $\lambda_1 \leq M_k \leq \lambda_2$, $k \in \mathbb{N}$, *for constants* $0 < \lambda_1 \leq \lambda_2 < \infty$. (*The constants* λ_1 *and* λ_2 *depend on* b, B, δ *from* (*C*), *and on* C_Θ *and* $C_{\mathscr{X}}$ *from* (*B*) *via* c *from Lemma* 3.2.)

An informal interpretation of Lemma 3.3 is as follows. Firstly, since the matrices $\boldsymbol{M}_k$ are almost surely positive definite, then the shift function $S(X_k, \hat{\theta}_k, D_k)$ gives the "right average direction" from $\hat{\theta}_k$ toward the conditional spatial median

θ_k. Secondly, since the eigenvalues of $\boldsymbol{M}_k$ are bounded from zero and from infinity, the "average length" of the shift $S(X_k, \hat{\theta}_k, D_k)$ is a controlled multiple of the distance between $\hat{\theta}_k$ and the conditional spatial median θ_k. The same interpretation can be given for Lemma 3.4.

We are ready to state our main results. Theorem 3.1 makes a statement about algorithm (2.6) as a sequential estimation procedure for conditional spatial medians in the case where $d \geq 2$ and Theorem 3.2 makes a statement about algorithm (2.7) as a sequential estimation procedure for conditional quantiles of level α_k in the one-dimensional case $d = 1$.

Theorem 3.1. *Let assumptions* (*A*) *and* (*B*) *hold, the sequence* $\hat{\theta}_k$ *be defined by* (2.6), $\delta_k = \delta_k(\boldsymbol{X}_{k-1}) = \hat{\theta}_k - \theta_k$, $k \in \mathbb{N}$. *Then for any* $k_0, k \in \mathbb{N}$, $k_0 \leq k$, *a sequence* $\{\gamma_k, k \in \mathbb{N}\}$ *as in* (2.5) *such that* $\gamma_i \lambda_2 \leq 1$ (λ_2 *as in Lemma* 3.3) *for all* $i \in \{k_0, \ldots, k\}$ *and any* $p \geq 1$, *the following relation holds*:

$$\mathbb{E}\|\delta_{k+1}\|_p^p \leq C_1 \exp\left\{-p\lambda_1 \sum_{i=k_0}^{k} \gamma_i\right\} + C_2 \left[\sum_{i=k_0}^{k} \gamma_i^2\right]^{p/2} + C_3 \max_{k_0 \leq i \leq k} \mathbb{E}\|\theta_{i+1} - \theta_{k_0}\|_p^p, \qquad (3.1)$$

where $C_1 = 3^{p-1} d^{p/2} K_p^p C^p$, $C_2 = 3^{p-1} d B_p \big(1 + K_p^2 \lambda_2/\lambda_1\big)^p$, $C_3 = 3^{p-1}\big(1 + K_p^2 \lambda_2/\lambda_1\big)^p$, *where* C *is the constant from Lemma* 3.1, λ_1 *and* λ_2 *from Lemma* 3.3, *and* K_p *is the constant from Lemma B*.2.

Theorem 3.2. *Let assumptions* (*C*) *and* (*B*) *hold and the sequence* $\hat{\theta}_k$ *be defined* (*for a fixed sequence* $\alpha_k \in (0, 1)$, $k \in \mathbb{N}$), *by* (2.7). *Define* $\delta_k = \delta_k(\boldsymbol{X}_{k-1}) = \hat{\theta}_k - \theta_k$, $k \in \mathbb{N}$. *Then for any* $k_0, k \in \mathbb{N}$, $k_0 \leq k$, *a sequence* $\{\gamma_k, k \in \mathbb{N}\}$ *as in* (2.5) *such that* $\gamma_i \lambda_2 \leq 1$ (λ_2 *as in Lemma* 3.4) *for all* $i \in \{k_0, \ldots, k\}$ *and any* $p \geq 1$, *the following relation holds*:

$$\mathbb{E}|\delta_{k+1}|^p \leq C_1 \exp\left\{-p\lambda_1 \sum_{i=k_0}^{k} \gamma_i\right\} + C_2 \left[\sum_{i=k_0}^{k} \gamma_i^2\right]^{p/2} + C_3 \max_{k_0 \leq i \leq k} \mathbb{E}|\theta_{i+1} - \theta_{k_0}|^p, \qquad (3.2)$$

where $C_1 = 3^{p-1} D^p$, $C_2 = 3^{p-1} 2^p B_p \big(1 + \lambda_2/\lambda_1\big)^p$, $C_3 = 3^{p-1}\big(1 + \lambda_2/\lambda_1\big)^p$, *where* D *is the constant from Lemma* 3.2, λ_1 *and* λ_2 *from Lemma* 3.4, *and* K_p *is the constant form Lemma B*.2.

Remark 3.1. The right-hand side of (3.2) depends on the quantile levels α_k, $k \in \mathbb{N}$, via the constants b and δ from (C). Also, the closer α_k it to one (resp. zero), the larger (resp. smaller) the value $|\theta_k|$, and therefore the constant C_Θ becomes larger and then also D from Lemma 3.2. Besides, there is too little probability mass in the neighborhood of extreme quantiles, so that condition (C) is more difficult to fulfill as α_k gets closer to one (or zero). This makes constants δ and b smaller, which in turn makes constant λ_1 smaller and constant λ_2 bigger. All of these changes lead to an increase of the constants C_1, C_2, and C_3 featured in inequality (3.2).

4. VARIATIONAL SETUPS FOR THE DRIFTING QUANTILES

Theorem 3.1 and Theorem 3.2 deliver (up to the values of the proportionality constants) the same explicit, nonasymptotic bound for the L_p-risk of the estimating recursive procedures (2.6) and (2.7). This bound depends mainly on three quantities: k_0, the sequence $\{\gamma_k, k \in \mathbb{N}\}$ and the variation of the spatial median θ_k.

Since the algorithms (2.6) and (2.7) are initiated with an arbitrary value $\hat{\theta}_1$ and the shift functions (2.3) and (2.4) are a.s. bounded, there is a minimal number of iterations which are needed for the estimating sequence to reach a neighborhood of the drifting quantile of interest—this is the so-called *burn-in period* of the algorithm. The length of the burn-in period is controlled by making an appropriate choice for k_0.

The step size sequence $\{\gamma_k, k \in \mathbb{N}\}$ induces averaging of the iterates of the algorithms (2.6) and (2.7), and its influence on the bound given by the theorems is explicit. If we are to minimize the right-hand side of (3.1) and (3.2), it should be clear that $\{\gamma_k, k \in \mathbb{N}\}$ must be chosen such that $\sum_{i=1}^k \gamma_i$ diverges as $k \to \infty$ and such that $\sum_{i=1}^k \gamma_i^2$ converges as $k \to \infty$. These are the classical conditions for the step sizes of Robbins-Monro type procedures and well known in the literature. Intuitively, if the sum $\sum_{i=k_0}^k \gamma_i^2$ is small, then the algorithm can "approach" θ_k arbitrarily close, and if the sum $\sum_{i=k_0}^k \gamma_i$ is large, then algorithm can "reach" any point $\theta \in \Theta$.

The variation of the drifting conditional spatial median also has a nonnegligible contribution to the accuracy of the sequential estimating procedures (2.6) and (2.7). This is reasonable since, if the median changes arbitrarily in-between observations, we should not expect it to the "estimable". In the following subsections we specify, for different assumptions on the variation of the spatial median, an appropriate value for k_0 and an appropriate sequence γ_k that minimizes the bound in (3.1) and (3.2).

4.1. Static Parameter

We assume in this section that $\theta_k = \theta_0$, $k \in \mathbb{N}$, a.s., for some unknown $\theta_0 \in \Theta$, this corresponds to a parametric setup. Clearly, in this case the third term on the right-hand side of (3.1) and (3.2) vanishes.

Take $\gamma_j = C_\gamma j^{-1} \log j$ and for $q \in (0, 1)$, $n_0 = \lfloor qn \rfloor$, where $\lfloor a \rfloor$ is the whole part of $a \in \mathbb{R}$. Let $n \geq 2/q = N_q$ such that $n_0 \geq 2$. For large enough C_γ and all $n \geq N_q$ we have

$$\sum_{j=n_0}^{n} \gamma_j \geq c_\gamma \log n_0 \sum_{j=n_0}^{n} \frac{1}{k} \geq \frac{\log n}{2\lambda_1},$$

from where for all $p \geq 1$,

$$\exp\left\{-p\lambda_1 \sum_{j=n_0}^{n} \gamma_j\right\} \leq n^{-p/2}.$$

Using the fact that $\sum_{j=n_0}^n \gamma_j^2 \leq c(\log n)^2 n^{-1}$ for some constant $c > 0$ we have

$$\left(\sum_{j=n_0}^{n} \gamma_j^2\right)^{p/2} \leq (n^{-1/2} \log n)^p.$$

We conclude that we can rewrite (3.1) as

$$\max_{n\geq N_q} \mathbb{E}\left(\frac{\sqrt{n}}{\log n}\|\delta_n\|_p\right)^p \leq C, \tag{4.1}$$

and (3.2) as

$$\max_{n\geq N_q} \mathbb{E}\left(\frac{\sqrt{n}}{\log n}|\delta_n|\right)^p \leq C, \tag{4.2}$$

both representations holding for all $p \geq 1$. The logarithmic term in the rate cannot be avoided and is a consequence of the recursiveness of the algorithm.

Note that by taking $p > \epsilon^{-1}$, using Markov's inequality and (4.1), we derive that

$$\begin{aligned}\sum_{n=1}^{\infty} P\big(n^{1/2-\epsilon}\|\hat{\theta}_n - \theta_0\|_1 > c\big) &\leq \sum_{n=1}^{\infty} P\big(d^{\frac{p-1}{p}} n^{1/2-\epsilon}\|\hat{\theta}_n - \theta_0\|_p > c\big)\\ &\leq \sum_{n=0}^{\infty} \frac{d^{p-1}n^{p/2-p\epsilon}\mathbb{E}\|\delta_n\|_p^p}{c^p} \leq C\sum_{n=1}^{\infty}\frac{(d\log n)^p}{n^{p\epsilon}} < \infty. \end{aligned}\tag{4.3}$$

By application of the Borel-Cantelli Lemma, we conclude that $\|\hat{\theta}_n - \theta_0\|_1 \to 0$ as $n \to 0$ with probability 1 at a rate $n^{1/2-\epsilon}$ for all $\epsilon > 0$. The same can be shown to be true in the one-dimensional case for algorithm (2.7) using (4.2).

The particular setup presented in this section, where the parameter is fixed, might seem atypical since we are, mainly concerned with estimating drifting parameters. The algorithms (2.6) and (2.7) are however recursive and easy to implement, whereas direct estimation might be more involved. Note that we only require the spatial median being estimated to remain fixed; the conditional distribution of the observations is allowed to change.

4.2. Stabilizing Parameter

Suppose now that the spatial median is stabilizing. Such a situation arises if, for example, the expectation of the oscillations of the median converges to zero with a certain rate. More specifically, assume that $\Delta\theta_i = \theta_i(\boldsymbol{X}_{i-1}) - \theta_{i+1}(\boldsymbol{X}_i)$ verifies

$$\mathbb{E}\|\Delta\theta_i\|_p^p \leq \rho_i^p, \quad i \in \mathbb{N},$$

for $p \geq 1$ and some decreasing sequence ρ_i. Assume then that we have $\rho_i = c_\rho i^{-\beta}$ for some constants $c_\rho > 0$ and $\beta \geq 0$.

Consider first the case $\beta \geq 3/2$. In this case the variation of the parameter vanishes so quickly that we are essentially in the setup of the previous section. Indeed, take γ_i and n_0 as in the previous section. The first and second terms on the right-hand side of (3.1) can be bounded in the same way as in the previous section. As for the third term, by using the Hölder inequality,

$$\begin{aligned}\mathbb{E}\left(\sum_{i=n_0}^{n}\|\Delta\theta_i\|_p\right)^p &\leq (n-n_0)^{p-1}\sum_{i=n_0}^{n}\mathbb{E}\|\Delta\theta_i\|_p^p \leq C(n-n_0)^p\rho_{n_0}^p\\ &\leq c\big((n-n_0)n_0^{-\beta}\big)^p \leq Cn^{-(\beta-1)p} \leq Cn^{-p/2},\end{aligned}\tag{4.4}$$

leading to the same bounds as in the previous section.

Consider now the case $0 < \beta < 3/2$. Let $\gamma_i = C_\gamma(\log i)^{1/3} i^{-2\beta/3}$, $n_0 = n - n^{2\beta/3}(\log n)^{2/3}$. By using the elementary inequality $(1+x)^\alpha \le 1 + \alpha x$ for $0 < \alpha < 1$ and $x \ge -1$, we obtain that for sufficiently large n (that is, $n \ge N_1 = N_1(\beta)$) and sufficiently large constant C_γ

$$\begin{aligned}\sum_{i=n_0}^{n} \gamma_i &\ge C_\gamma(\log n_0)^{1/3} \sum_{i=n_0}^{n} \frac{1}{i^{2\beta/3}} \ge C_\gamma(\log n_0)^{1/3} \int_{n_0}^{n} \frac{dx}{x^{2\beta/3}} \\ &= \frac{C_\gamma(\log n_0)^{1/3}}{1-2\beta/3}\left[n^{1-2\beta/3} - n^{1-2\beta/3}\left(1 - n^{2\beta/3-1}(\log n)^{2/3}\right)^{1-2\beta/3}\right] \\ &\ge \frac{C_\gamma(\log n_0)^{1/3}}{1-2\beta/3}\left[n^{1-2\beta/3} - n^{1-2\beta/3}\left(1 - n^{2\beta/3-1}(\log n)^{2/3}(1-2\beta/3)\right)\right] \\ &= C_\gamma(\log n_0)^{1/3}(\log n)^{2/3} \ge \frac{\log n}{2h}.\end{aligned}$$

This yields the same bound for the first term on the right-hand side of (3.1): for $n \ge N_1$ and sufficiently large constant C_γ,

$$\exp\left\{-ph \sum_{i=n_0}^{n} \gamma_i\right\} \le n^{-p/2}.$$

Now we bound the second term on the right-hand side of (3.1): for $n \ge N_2 = N_2(\beta)$,

$$\left(\sum_{i=n_0}^{n} \gamma_i^2\right)^{p/2} \le C\left((\log n)^{2/3} n_0^{-4\beta/3}(n-n_0)\right)^{p/2} \le c\left((\log n)^{2/3} n^{-\beta/3}\right)^p.$$

Similarly to (4.4), the third term on the right-hand side of (3.1) is bounded by

$$\mathbb{E}\left(\sum_{i=n_0}^{n} \|\Delta\theta_i\|_p\right)^p \le c\left((n-n_0)n_0^{-\beta}\right)^p \le C\left((\log n)^{2/3} n^{-\beta/3}\right)^p,$$

for sufficiently large n (that is, $n \ge N_3 = N_3(\beta)$). Finally we obtain that for $0 < \beta < 3/2$ and sufficiently large constant C_γ in the algorithm step $\gamma_i = C_\gamma(\log i)^{1/3} i^{-2\beta/3}$, (3.1) implies that

$$\max_{n \ge N_\beta} \mathbb{E}\left(\frac{n^{\beta/3}}{(\log n)^{2/3}}\|\delta_n\|_p\right)^p \le C,$$

where $N_\beta = \max(N_1, N_2, N_3)$ is the burn-in period of the algorithm.

Remark 4.1. If we choose $\gamma_i = C_\gamma(\log i)^{\alpha_1} i^{-\alpha}$ and $n_0 = n - n^\alpha(\log n)^{\alpha_2}$, $0 < \alpha < 1$, $\alpha_1, \alpha_2 \ge 0$, $\alpha_1 + \alpha_2 \ge 1$ in case $0 < \beta < 3/2$, then we get the following bound of the convergence rate: for sufficiently large n and sufficiently large constant C_γ

$$\mathbb{E}\|\delta_n\|_p^p \le C\left(n^{-\min\{\beta-\alpha, \alpha/2\}}(\log n)^{\max\{\alpha_2, \alpha_1+\alpha_2/2\}}\right)^p.$$

Thus, the choice $\alpha = 2\beta/3$, $\alpha_1 = 1/3$, $\alpha_2 = 2/3$ is optimal in the sense of the minimum of the right-hand side of the above inequality.

Remark 4.2. Much in the same way as for (4.3), we can establish that for any $\epsilon > 0$,

$$\lim_{n\to\infty} n^{\beta/3-\epsilon}\|\delta_n\|_1 = 0 \quad \text{with probability 1.}$$

Finally, consider the case $\beta = 0$; that is, we assume the following weak requirement: $\mathbb{E}\|\Delta\theta_i\|_p^p \le c$, $i \in \mathbb{N}$, for some uniform constant c. Take $n - n_0 = N$, $\gamma_i = \gamma$ for some $N \in \mathbb{N}$, $\gamma > 0$. Then Theorem 3.1 implies that

$$\max_{n\ge N} \mathbb{E}\|\delta_n\|_p^p \le C_1 e^{-phN\gamma} + C_2 N^{p/2}\gamma^p + C_3 N^p c = D.$$

We thus have that the algorithm will track the spatial median in the proximity of size D, which we can try to minimize by choosing appropriate constants N and γ.

The exact same computations will give the respective bounds for the right-hand side of (3.2) as for (3.1) by simply replacing everywhere $\|\cdot\|_p$ with $|\cdot|$.

4.3. Lipschitz Varying Median with Asymptotics in the Sampling Frequency

We consider now a slightly different setup where we assume that the spatial median is changing, on average, like a Lipschitz function. In this setup we assume that the data are sampled from a continuous-time process X_t, $t \in [0, 1]$, which we observe with frequency n. This means that for each $n \in \mathbb{N}$ we have a different model, namely,

$$X_0^n \sim P_0^n, \quad X_k^n | \boldsymbol{X}_{k-1}^n \sim P_k^n(\cdot|\boldsymbol{X}_{k-1}^n), \quad k \le n \in \mathbb{N}, \tag{4.5}$$

where the spatial median $\theta_k^n = \theta_k^n(\boldsymbol{X}_{k-1}^n)$ verifies, for some $p \in \mathbb{N}$, $\kappa = \kappa(d, p) < \infty$

$$\mathbb{E}\|\theta_k^n(\boldsymbol{X}_{k-1}^n) - \theta_{k_0}^n(\boldsymbol{X}_k^n)\|_p^p \le \kappa^p\Big(\frac{k-k_0}{n}\Big)^{\beta p}.$$

We could have, for example, that $\theta_k^n(\boldsymbol{X}_{k-1}^n) = \vartheta(k/n)$, almost surely, where $\vartheta(\cdot) \in \mathscr{L}(L,\beta) = \{g(\cdot) : \|g(t_1) - g(t_2)\|_1 \le L|t_1 - t_2|^\beta, t_1, t_2 \in [0,1]\}$ for some $0 < \beta \le 1$ and $L > 0$, a space of vector valued Lipschitz functions. The nonparametric median estimation problem ($d = 1$, $\alpha_k = 1/2$, $k \in \mathbb{N}$) has been studied in Belitser and Korostelev (1992) and Belitser and van de Geer (2000) for such an asymptotic regime.

Let $\gamma_k \equiv C_\gamma(\log n)^{(2\beta-1)/(2\beta+1)} n^{-2\beta/(2\beta+1)}$, (constant in k) for $k = 1, \dots, n$, and

$$k_0 = k_0(n) = k - (\log n)^{2/(2\beta+1)} n^{2\beta/(2\beta+1)},$$

for $k \ge K_n = (\log n)^{2/(2\beta+1)} n^{2\beta/(2\beta+1)}$. Note that for $K_n/n \to 0$ as $n \to \infty$ for any $0 < \beta \le 1$.

For sufficiently large C_γ,

$$\sum_{i=k_0}^{k} \gamma_i = C_\gamma(\log n)^{(2\beta-1)/(2\beta+1)} n^{2\beta/(2\beta+1)}(k - k_0) \ge C_\gamma \log n \ge \frac{\log n}{3\lambda_1},$$

leading to

$$\exp\left\{-p\lambda_1 \sum_{i=k_0}^{k} \gamma_i\right\} \leq cn^{-p/3}.$$

In much the same way, we have

$$\left(\sum_{i=k_0}^{k} \gamma_i^2\right)^{p/2} \leq C\Big((\log n)^{\frac{2\beta-1}{2\beta+1}} n^{-\frac{2\beta}{2\beta+1}} (k-k_0)^{1/2}\Big)^p = C\Big((\log n)^{\frac{2\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}}\Big)^p.$$

From our assumption on the variation of the parameter, we have

$$\max_{i=k_0,\ldots,k} \mathbb{E}\|\theta_{i+1}^n - \theta_{k_0}^n\|_p^p \leq c\left(\frac{k-k_0}{n}\right)^{-p\beta} \leq C\Big((\log n)^{\frac{2\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}}\Big)^p.$$

Finally, combining the three bounds with Theorem 3.1, we obtain

$$\sup_{\vartheta\in\mathscr{L}(L,\beta)} \max_{i\geq K_n} \mathbb{E}\|\delta_i\|_p^p \leq C\Big((\log n)^{\frac{2\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}}\Big)^p.$$

The exact same computations hold in case $d = 1$ and give the same bound for the right-hand side of (3.2) as for (3.1) by simply replacing everywhere $\|\cdot\|_p$ with $|\cdot|$.

APPENDIX A: PROOFS OF THE RESULTS FROM SECTION 3

In this appendix we present the proofs to our main results from Section 3. The technical lemmas used here are collected in Appendix B. If index k is involved in one of the relation below, then the corresponding relation holds for all $k \in \mathbb{N}$.

Proof of Lemma 3.1. Recall that $\boldsymbol{b} = \{b_1, \ldots, b_d\}$ is an orthonormal basis where random directions D_k take their values. Introduce a hypercube of size $h > 0$ with respect to this basis:

$$C(h) = \left\{v \in \mathbb{R}^d : |b_i^T v| \leq h, i = 1, \ldots, d\right\}.$$

We will prove a slightly stronger assertion, namely, that $\hat{\theta}_k \in C(C_{\mathscr{X}} + \Gamma/2)$ for all $k \in \mathbb{N}$, which, of course, implies the claim of the lemma. Recall that $\hat{\theta}_1 \in \Theta \subseteq \mathscr{X}$ so that $\hat{\theta}_1 \in C(C_{\mathscr{X}}) \subset C(C_{\mathscr{X}} + \Gamma/2)$. By induction, it is enough to show that if $\hat{\theta}_k \in C(C_{\mathscr{X}} + \Gamma/2)$, then $\hat{\theta}_{k+1} \in C(C_{\mathscr{X}} + \Gamma/2)$.

Assume that $\hat{\theta}_k \in C(C_{\mathscr{X}} + \Gamma/2)$ and suppose $D_{k+1} = b \in \boldsymbol{b}$. Then $\hat{\theta}_{k+1} = \hat{\theta}_k \pm \gamma_k b/2$ depending on whether $X_k \in H(\hat{\theta}_k, b)$ or not. As compared with $\hat{\theta}_k$, the only change in $\hat{\theta}_{k+1}$ is the lth coordinate $b^T\hat{\theta}_{k+1}$ with respect to the basis $\boldsymbol{b}$. Thus, we only need to show that $|b^T\hat{\theta}_{k+1}| \leq C_{\mathscr{X}} + \Gamma/2$.

Consider the case $X_k \in H(\hat{\theta}_k, b)$, then $b^T X_k \geq b^T\hat{\theta}_k$ and $\hat{\theta}_{k+1} = \hat{\theta}_k + \gamma_k b/2$. Recall also that $X_k \in \mathscr{X}$, which implies that $|b^T X_k| \leq C_{\mathscr{X}}$. If $b^T\hat{\theta}_k \geq 0$, then $b^T X_k \geq b^T\hat{\theta}_k \geq 0$ and $|b^T\hat{\theta}_{k+1}| = b^T\hat{\theta}_k + \gamma_k/2 \leq b^T X_k + \gamma_k/2 \leq C_{\mathscr{X}} + \Gamma/2$. If $b^T\hat{\theta}_k \leq 0$, then again $|b^T\hat{\theta}_{k+1}| = |b^T\hat{\theta}_k + \gamma_k/2| \leq |b^T\hat{\theta}_k| \leq C_{\mathscr{X}} + \Gamma/2$ since $\hat{\theta}_k \in C(C_{\mathscr{X}} + \Gamma/2)$.

Similarly, consider the case $X_k \notin H(\hat{\theta}_k, b)$. Then $b^T X_k < b^T \hat{\theta}_k$ and $\hat{\theta}_{k+1} = \hat{\theta}_k - \gamma_k b/2$. If $b^T \hat{\theta}_k \leq 0$, then $b^T X_k < b^T \hat{\theta}_k \leq 0$ and $|b^T \hat{\theta}_{k+1}| = |b^T \hat{\theta}_k| + \gamma_k/2 \leq |b^T X_k| + \gamma_k/2 \leq C_{\mathscr{X}} + \Gamma/2$. If $b^T \hat{\theta}_k > 0$, then $|b^T \hat{\theta}_{k+1}| = |b^T \hat{\theta}_k - \gamma_k/2| \leq |b^T \hat{\theta}_k| \leq C_{\mathscr{X}} + \Gamma/2$.

We established that $|b^T \hat{\theta}_{k+1}| \leq C_{\mathscr{X}} + \Gamma/2$ and this completes the proof of the lemma. □

Proof of Lemma 3.2. As compared to the proof of the previous lemma, we now have $\hat{\theta}_{k+1} = \hat{\theta}_k + \gamma_k \alpha_k$ if $X_k \geq \hat{\theta}_k$, and $\hat{\theta}_{k+1} = \hat{\theta}_k - \gamma_k(1 - \alpha_k)$ if $X_k < \hat{\theta}_k$. We repeat the same arguments as in the proof of Lemma 3.1 with Γ instead of $\Gamma/2$ (since $\alpha_k \in (0, 1)$) and $b = 1$. □

Proof of Lemma 3.3. It is easily seen that by construction $\|S(u, v, w)\| \leq 1/2$ uniformly over $u \in \mathscr{X}$, $v \in \mathbb{R}^d$ and over all unit vectors w in $\mathbb{R}^d$. For a random variable Y, let $\mathbb{E}_Y$ represent expectation with respect to the law of Y. Recall that $\theta_k = \theta_k(\boldsymbol{X}_{k-1})$ is a predictable process with respect to the filtration $\mathscr{F}_0 \cup \{\mathscr{F}_k, k \in \mathbb{N}\}$. Clearly, the lemma holds true if $\|\hat{\theta}_k - \theta_k\| = 0$ with, for example, $\boldsymbol{M}_k = \boldsymbol{I}$. Assume therefore that $\|\hat{\theta}_k - \theta_k\| > 0$.

Denote for brevity $\epsilon_k = \|\theta_k - \hat{\theta}_k\|$ and $e_k = (\theta_k - \hat{\theta}_k)/\epsilon_k$, the unit vector in the direction of $\hat{\theta}_k - \theta_k$. First note that, since D_k is independent of $\mathscr{F}_{k-1}$,

$$\begin{aligned} e_k^T \mathbb{E}[S(X_k, \hat{\theta}_k, D_k)|\mathscr{F}_{k-1}] &= e_k^T \mathbb{E}\Big[D_k\big(I\{X_k \in H(\hat{\theta}_k, D_k)\} - 1/2\big)|\mathscr{F}_{k-1}\Big] \\ &= \mathbb{E}_{D_k}\Big[(e_k^T D_k)\big(\mathbb{P}(X_k \in H(\hat{\theta}_k, D_k)|\boldsymbol{X}_{k-1}) - 1/2\big)\Big]. \quad \text{(A.1)} \end{aligned}$$

Since the conditional distributions $\mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1})$ are half-space symmetrical about θ_k,

$$e_k^T \mathbb{E}[S(X_k, \hat{\theta}_k, D_k)|\mathscr{F}_{k-1}] = (e_k^T D_k)\big(\mathbb{P}\big(X_k \in H(\hat{\theta}_k, D_k)|\boldsymbol{X}_{k-1}\big) - 1/2\big) \geq 0 \quad \text{a.s.}, \quad \text{(A.2)}$$

so that (A.1) is almost surely positive. By Lemma 3.1, $\epsilon_k = \|\hat{\theta}_k - \theta_k\| \leq C$, with C as defined in this lemma. We consider two cases depending on the value of $\epsilon_k \in (0, C]$.

First consider the case $0 < \epsilon_k < \delta$, with δ defined in assumption (A). Note that we can write

$$\hat{\theta}_k = \theta_k - \epsilon_k e_k, \quad \text{(A.3)}$$

so that, by using representation (A.1) and assumption (A),

$$\epsilon_k b \mathbb{E}_{D_k}\big[e_k^T D_k\big]^2 \leq e_k^T \mathbb{E}\big[S(X_k, \hat{\theta}_k, D_k)\big|\mathscr{F}_{k-1}\big] \leq \epsilon_k B \mathbb{E}_{D_k}\big[e_k^T D_k\big]^2.$$

Since D_k takes values in an orthonormal basis, $\mathbb{E}[v^T D_k]^2 = 1/d$ for any unit vector v in $\mathbb{R}^d$. The last two relations imply that almost surely

$$(bd^{-1})\epsilon_k \leq e_k^T \mathbb{E}\big[S(X_k, \hat{\theta}_k, D_k)\big|\mathscr{F}_{k-1}\big] \leq (Bd^{-1})\epsilon_k.$$

Consider now the case $\delta \leq \epsilon_k \leq C$. Using this assumption, the Cauchy-Schwarz inequality, and (A.2), it follows from (A.1) that

$$e_k^T \mathbb{E}\big[S(X_k, \hat{\theta}_k, D_k)\big|\mathscr{F}_{k-1}\big] \leq \frac{1}{2} \leq \frac{\epsilon_k}{2\delta}.$$

To derive a lower bound for the last display note that for $\epsilon \in [0, \delta]$ and any unit vectors $v, w \in \mathbb{R}^d$,

$$\epsilon b(v^T w) \leq (v^T w)\big[\mathbb{P}(X_k \in H(\theta_k - \epsilon v, w)|\boldsymbol{X}_{k-1}) - 1/2\big],$$

by assumption (A). Note also that since the distributions $\mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1})$ are half-space symmetric about θ_k, then the right-hand side of the previous display is monotonically increasing in ϵ. We then obtain that for any two unit vectors $v, w \in \mathbb{R}^d$ and any $\epsilon \geq \delta$,

$$\delta b(v^T w) \leq (v^T w)\big[\mathbb{P}(X_k \in H(\theta_k - \epsilon v, w)|\boldsymbol{X}_{k-1}) - 1/2\big].$$

Since $D_k \in \boldsymbol{b}$ (an orthonormal basis in $\mathbb{R}^d$), then, for any unit vector v, with probability at least $1/d$ we must have that $v^T D_k \geq 1/\sqrt{d}$. From this fact and (A.1) a lower bound follows for the case $\delta \leq \epsilon_k \leq C$:

$$e_k^T \mathbb{E}\big[S(X_k, \hat{\theta}_k, D_k)|\mathcal{F}_{k-1}\big] \geq \frac{\delta b}{d^{3/2}} \geq \frac{\epsilon_k \delta b}{C d^{3/2}}.$$

Summarizing, we established that almost surely

$$C_1 \|\theta_k - \hat{\theta}_k\|^2 \leq (\theta_k - \hat{\theta}_k)^T \mathbb{E}\big[S(X_k, \hat{\theta}_k, D_k)\big|\mathcal{F}_{k-1}\big] \leq C_2 \|\theta_k - \hat{\theta}_k\|^2, \tag{A.4}$$

where $C_1 = \min\{b/d, b\delta/(Cd^{3/2})\}$ and $C_2 = \max\{B/d, 1/(2\delta)\}$.

From assumption (A) we have that, for any unit vectors v, w in $\mathbb{R}^d$ and $\epsilon \in (0, \delta]$,

$$\big\|w\big(\mathbb{P}(X_k \in H(\theta_k - \epsilon v, w)|\boldsymbol{X}_{k-1}) - 1/2\big)\big\| \leq \big|\mathbb{P}(X_k \in H(\theta_k - \epsilon v, w)|\boldsymbol{X}_{k-1}) - 1/2\big| \leq B\epsilon,$$

by the Cauchy-Schwarz inequality. Thus, if $\|\hat{\theta}_k - \theta_k\| < \delta$, then, by using the previous display and (A.3),

$$\big\|\mathbb{E}[S(X_k, \hat{\theta}_k, D_k)|\mathcal{F}_{k-1}]\big\| = \big\|\mathbb{E}_{D_k}\big[D_k \mathbb{P}\big(X_k \in H(\hat{\theta}_k, D_k)\big|\boldsymbol{X}_{k-1}\big) - 1/2\big]\big\| \leq B\|\theta_k - \hat{\theta}_k\|.$$

In case $\|\hat{\theta}_k - \theta_k\| \geq \delta$, we trivially have

$$\big\|\mathbb{E}\big[S(X_k, \hat{\theta}_k, D_k)\big|\mathcal{F}_{k-1}\big]\big\| \leq \frac{1}{2} \leq \frac{\|\theta_k - \hat{\theta}_k\|}{2\delta}.$$

We conclude that almost surely

$$\big\|\mathbb{E}\big[S(X_k, \hat{\theta}_k, D_k)\big|\mathcal{F}_{k-1}\big]\big\| \leq C_3 \|\theta_k - \hat{\theta}_k\|. \tag{A.5}$$

with $C_3 = \max\{B, 1/(2\delta)\}$.

The statement of the lemma follows from (A.4) and (A.5) by applying Lemma B.1 below. □

Proof of Lemma 3.4. The proof of Lemma 3.4 is the same as that of Lemma 3.3 by making some minor modifications for the quantities involved in the proof. We take $d = 1$, replace everywhere $1/2$ with α_k (which is then bounded by 1), set $D_k = -1$, replace the constant C from Lemma 3.1 with the constant c from Lemma 3.2, $\theta_k = \theta_k(\boldsymbol{X}_{k-1}, \alpha_k) =$ is the conditional α_k-level quantile of $\mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1})$, take constants b, B, and δ from assumption (C), and replace $\|\cdot\|$ with $|\cdot|$. □

Proof of Theorem 3.1. For the sake of brevity, denote $\theta_k = \theta_k(\boldsymbol{X}_{k-1})$, $S_k = S(X_k, \hat{\theta}_k, D_k)$, $s_k = \mathbb{E}[S(X_k, \hat{\theta}_k, D_k)|\mathscr{F}_{k-1}]$ and $T_k = S_k - s_k$, $k \in \mathbb{N}$. We have

$$\mathbb{E}[T_k|\mathscr{F}_{k-1}] = \mathbb{E}[S_k - s_k|\mathscr{F}_{k-1}] = s_k - s_k = 0, \quad k \in \mathbb{N}.$$

It follows that $\{T_k, k \in \mathbb{N}\}$, is a (vector) martingale difference sequence with respect to the filtration $\mathscr{F}_0 \cup \{\mathscr{F}_k\}_{k\in\mathbb{N}}$.

Rewrite the algorithm equation (2.6) as

$$\delta_{k+1} = \delta_k + \Delta\theta_k + \gamma_k T_k + \gamma_k s_k, \quad k \in \mathbb{N}.$$

In view of (A) and Lemma 3.3 the decomposition $s_k = -\boldsymbol{M}_k \delta_k$ holds almost surely for an $\mathscr{F}_{k-1}$-measurable, symmetric, positive definite matrix $\boldsymbol{M}_k$ such that $0 < \lambda_1 \leq \lambda_{(1)}(\boldsymbol{M}_k) \leq \lambda_{(d)}(\boldsymbol{M}_k) \leq \lambda_2 \leq \infty$ almost surely. We have

$$\delta_{k+1} = \Delta\theta_k + \gamma_k T_k + (\boldsymbol{I} - \gamma_k \boldsymbol{M}_k)\delta_k, \quad k \in \mathbb{N}.$$

By iterating the relation from above, we obtain that for any $k_0 = 1, \ldots, k$

$$\begin{aligned}\delta_{k+1} &= (\boldsymbol{I} - \gamma_k \boldsymbol{M}_k)(\boldsymbol{I} - \gamma_{k-1}\boldsymbol{M}_{k-1})\delta_{k-1} + \Delta\theta_k + \gamma_k T_k \\ &\quad + (\boldsymbol{I} - \gamma_k \boldsymbol{M}_k)(\Delta\theta_{k-1} + \gamma_{k-1}T_{k-1}) \\ &= \left[\prod_{i=k_0}^{k} (\boldsymbol{I} - \gamma_i \boldsymbol{M}_i)\right] \delta_{k_0} + \sum_{i=k_0}^{k} \left[\prod_{j=i+1}^{k} (\boldsymbol{I} - \gamma_j \boldsymbol{M}_j)\right] (\Delta\theta_i + \gamma_i T_i). \qquad \text{(A.6)}\end{aligned}$$

Denote $A_i = \sum_{j=k_0}^{i} \gamma_j T_j$, $B_i = \sum_{j=k_0}^{i} \Delta\theta_j$ and $H_i = A_i + B_i$. Applying the Abel transformation (Lemma B.3) to the second term of the right-hand side of (A.6) yields

$$\sum_{i=k_0}^{k} \left[\prod_{j=i+1}^{k} (\boldsymbol{I} - \gamma_j \boldsymbol{M}_j)\right] (\Delta\theta_i + \gamma_i T_i) = H_k - \sum_{i=k_0}^{k-1} \gamma_{i+1} M_{i+1} \left[\prod_{j=i+2}^{k} (\boldsymbol{I} - \gamma_j \boldsymbol{M}_j)\right] H_i. \quad \text{(A.7)}$$

In particular, note that if we take $d = 1$, $\boldsymbol{M}_j = \lambda_1$, $\Delta\theta_j = 0$ for $j = k_0, \ldots, k$, $T_{k_0} = 1$, $T_j = 0$ for $j = k_0 + 1, \ldots, k$, we have that (since $0 \leq \gamma_j \lambda_1 \leq 1$ for $j = k_0, \ldots, k$)

$$\sum_{i=k_0}^{k-1} \lambda_1 \gamma_{i+1} \prod_{j=i+2}^{k} (1 - \gamma_j \lambda_1) = 1 - \prod_{j=k_0+1}^{k} (1 - \gamma_j \lambda_1) \leq 1, \qquad \text{(A.8)}$$

which we will use later.

Using (A.7), we can rewrite our expansion of δ_{k+1} in (A.6) as follows:

$$\delta_{k+1} = \left[\prod_{i=k_0}^{k} (\boldsymbol{I} - \gamma_i \boldsymbol{M}_i)\right] \delta_{k_0} + H_k - \sum_{i=k_0}^{k-1} \gamma_{i+1} M_{i+1} \left[\prod_{j=i+2}^{k} (\boldsymbol{I} - \gamma_j \boldsymbol{M}_j)\right] H_i.$$

The previous display, the Minkowski inequality, and the submultiplicative property of the operator norm ($\|\boldsymbol{AB}\|_p \leq \|\boldsymbol{A}\|_p \|\boldsymbol{B}\|_p$) imply that

$$\begin{aligned}\|\delta_{k+1}\|_p &\leq \|\delta_{k_0}\|_p \left\| \prod_{i=k_0}^{k} (\boldsymbol{I} - \gamma_i \boldsymbol{M}_i) \right\|_p + \|H_k\|_p \\ &\quad + \sum_{i=k_0}^{k-1} \gamma_{i+1} \|M_{i+1}\|_p \|H_i\|_p \left\| \prod_{j=i+2}^{k} (\boldsymbol{I} - \gamma_j \boldsymbol{M}_j) \right\|_p. \qquad \text{(A.9)}\end{aligned}$$

By using Lemma 3.3, Lemma B.2, (A.8), and the elementary inequality $1 - x \le e^{-x}$,

$$\|\delta_{k+1}\|_p \le K_p\|\delta_{k_0}\|_p \prod_{i=k_0}^{k}(1-\gamma_i\lambda_1) + \max_{k_0\le i\le k}\|H_i\|_p\left[1 + K_p^2\sum_{i=k_0}^{k-1}\gamma_{i+1}\lambda_2\prod_{j=i+2}^{k}(1-\gamma_i\lambda_1)\right]$$
$$\le K_p\|\delta_{k_0}\|_p \exp\left\{-\lambda_1\sum_{i=k_0}^{k}\gamma_i\right\} + \left(1 + \frac{K_p^2\lambda_2}{\lambda_1}\right)\left(\max_{k_0\le i\le k}\|A_i\|_p + \max_{k_0\le i\le k}\|B_i\|_p\right)$$

almost surely, where the constant $K_p = K_p(d)$ is from Lemma B.2.

Take now the pth power of both sides of the last relation and apply the Hölder inequality $|\sum_{i=1}^{m} a_i|^p \le m^{p-1}\sum_{i=1}^{m}|a_i|^p$ for $m = 3$ to get

$$\|\delta_{k+1}\|_p^p \le 3^{p-1}K_p^p\|\delta_{k_0}\|_p^p \exp\left\{-p\lambda_1\sum_{i=k_0}^{k}\gamma_i\right\}$$
$$+ 3^{p-1}\left(1 + \frac{K_p^2\lambda_2}{\lambda_1}\right)^p\left(\max_{k_0\le i\le k}\|A_i\|_p^p + \max_{k_0\le i\le k}\|B_i\|_p^p\right).$$

Recall that the sequence $\left\{\sum_{j=k_0}^{i}\gamma_jT_j,\ i \ge k_0\right\}$ is a martingale with respect to the filtration $\{\mathcal{F}_i, i \ge k_0\}$ and that the coordinates of T_j verify $|T_{jl}| \le 2\|S_j\| \le 1$ almost surely, $l = 1, \ldots, d$, $j = k_0, \ldots, k$. Applying the maximal Burkholder inequality for $p > 1$ and the Davis inequality for $p = 1$ (cf. Shiryaev, 1996) yields

$$\mathbb{E}\max_{k_0\le i\le k}\|A_i\|_p^p = \mathbb{E}\max_{k_0\le i\le k}\sum_{l=1}^{d}\left|\sum_{j=k_0}^{i}\gamma_jT_{jl}\right|^p \le \sum_{l=1}^{d}\mathbb{E}\max_{k_0\le i\le k}\left|\sum_{j=k_0}^{i}\gamma_jT_{jl}\right|^p$$
$$\le B_p\sum_{l=1}^{d}\mathbb{E}\left[\sum_{j=k_0}^{k}\gamma_j^2T_{jl}^2\right]^{p/2} \le dB_p\left[\sum_{j=k_0}^{k}\gamma_j^2\right]^{p/2}, \tag{A.10}$$

for some constant B_p. One can take $B_p = ((18p^{5/2})/(p-1)^{3/2})^p$ for $p > 1$; cf. Shiryaev (1996). The statement of the theorem now follows by taking expectations on both sides of the bound on $\|\delta_{k+1}\|_p^p$ above and by using the last inequality. Note that for $p \ge 1$, $\|\delta_{k_0}\|_p \le d^{1/2}\|\delta_{k_0}\|_2 \le C$ almost surely by Lemma 3.1 so that $\mathbb{E}\|\delta_{k_0}\|_p \le d^{p/2}C^p$. □

Proof of Theorem 3.2. The proof of Theorem 3.2 is the same as that of Theorem 3.1 by making some particular choices for the quantities involved in the proof. We take $d = 1$, replace K_p with 1, replace everywhere the matrices $\boldsymbol{M}_k$ with the random variables M_k, replace the constant C from Lemma 3.1 with the constant c from Lemma 3.2, let $\theta_k = \theta_k(\boldsymbol{X}_{k-1}, \alpha_k)$ be the α_k-level quantile of $\mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1})$, invoke assumption (C) instead of assumption (A), replace $\|\cdot\|_p$ with $|\cdot|$, and use the triangle inequality instead of the Minkowski inequality.

Further, in (A.10) we have that $|T_{jl}| = |T_j| \le 2$, which leads to an extra multiplicative factor 2^p in the bound. □

APPENDIX B: TECHNICAL RESULTS

In this appendix we have some technical lemmas used in the proofs of the results from Appendix A.

Lemma B.1. *Let $x, y \in \mathbb{R}^d$. If $0 < \lambda_1' \|x\|^2 \le x^T y \le \lambda_2' \|x\|^2 < \infty$ and $\|y\| \le L\|x\|$ for some $\lambda_1', \lambda_2', L \in \mathbb{R}$ such that $0 < \lambda_1' \le \lambda_2' < \infty$ and $L > 0$, then there exists a symmetric positive definite matrix $\boldsymbol{M}$ such that $y = \boldsymbol{M}x$ and $0 < \lambda_1 \le \lambda_{(1)}(\boldsymbol{M}) \le \lambda_{(d)}(\boldsymbol{M}) \le \lambda_2 < \infty$ for some constants $\lambda_1, \lambda_2 \in \mathbb{R}$ depending only on λ_1', λ_2' and L.*

Proof. Suppose $0 < \lambda_1' \|x\|^2 \le x^T y \le \lambda_2' \|x\|^2 < \infty$ for some $\lambda_1', \lambda_2' \in \mathbb{R}$ such that $0 < \lambda_1' \le \lambda_2' < \infty$ and $\|y\| \le L\|x\|$. Let $V = \{v = ax + by : a, b \in \mathbb{R}\}$ be the linear space spanned by x and y. First consider the case $\dim(V) = 1$; that is, $y = \alpha x$ for some $\alpha \in \mathbb{R}$. Then $x^T y = \alpha \|x\|^2$ so that $0 < \lambda_1' \le \alpha \le \lambda_2' < \infty$. Thus, $y = \alpha x = \boldsymbol{M}x$ with symmetric and positive $\boldsymbol{M} = \alpha \boldsymbol{I}$ so that $0 < \lambda_1' \le \alpha = \lambda_{(1)}(\boldsymbol{M}) = \lambda_{(d)}(\boldsymbol{M}) \le \lambda_2' < \infty$.

Now consider the case $\dim(V) = 2$. Let $e_1 = x/\|x\|$ and $\{e_1, e_2\}$ be an orthonormal basis of V. Then

$$x = \|x\| e_1 \quad \text{and} \quad y = \alpha e_1 + \beta e_2.$$

The conditions $\lambda_1' \|x\|^2 \le x^T y = \alpha \|x\| \le \lambda_2' \|x\|^2$ and $\|y\| = \sqrt{\alpha^2 + \beta^2} \le L\|x\|$ imply that

$$\lambda_1' \|x\| \le \alpha \le \min\{\lambda_2', L\} \|x\|, \quad |\beta| \le L\|x\|.$$

Let e_2 be chosen in such a way that $\beta > 0$ (which is always possible). We change the basis of V as follows:

$$e_1' = \cos(\theta) e_1 - \sin(\theta) e_2,$$
$$e_2' = \sin(\theta) e_1 + \cos(\theta) e_2.$$

We thus rotate the basis $\{e_1, e_2\}$ by the angle θ. In these new basis we have

$$x = \|x\| \cos(\theta) e_1' + \|x\| \sin(\theta) e_2' = \alpha_x e_1' + \beta_x e_2',$$
$$y = (\alpha \cos(\theta) - \beta \sin(\theta)) e_1' + (\alpha \sin(\theta) + \beta \cos(\theta)) e_2' = \alpha_y e_1' + \beta_y e_2'.$$

Recall that $\alpha, \beta > 0$. Take $\theta \in (0, \pi/2)$ such that $\alpha \cos(\theta) - \beta \sin(\theta) = \frac{1}{2} \alpha \cos(\theta)$; that is, $\tan(\theta) = \frac{\alpha}{2\beta}$. Then we have that

$$\frac{\lambda_1'}{2} \le \frac{\alpha}{2\|x\|} = \frac{\alpha_y}{\alpha_x} \le \frac{\min\{\lambda_2', L\}}{2}, \quad \lambda_1' \le \frac{\alpha}{\|x\|} \le \frac{\beta_y}{\beta_x} \le \frac{\alpha}{\|x\|} + \frac{2\beta^2}{\alpha\|x\|} \le \min\{\lambda_2', L\} + \frac{2L^2}{\lambda_1'}.$$

Take then $\lambda_1 = \lambda_1'/2$ and $\lambda_2 = \min\{\lambda_2', L\} + 2L^2/\lambda_1'$.

Let $\{e_3', \ldots, e_d'\}$ be the orthonormal basis of $V^\perp$, so that $\boldsymbol{B} = \{e_1', e_2', e_3', \ldots, e_d'\}$ is an orthonormal basis of $\mathbb{R}^d$. Take

$$\tilde{\boldsymbol{M}} = \left[\begin{array}{c|c} \boldsymbol{D} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{I}_{d-2} \end{array}\right] \quad \text{with} \quad \boldsymbol{D} = \begin{bmatrix} \alpha_y/\alpha_x & 0 \\ 0 & \beta_y/\beta_x \end{bmatrix},$$

where the $\boldsymbol{O}$s indicate null matrices of the appropriate dimensions. We then have $y = \widetilde{\boldsymbol{M}}x$ in the basis $\boldsymbol{B}$ and $\lambda_1 \le \lambda_{(1)}(\widetilde{\boldsymbol{M}}) \le \lambda_{(d)}(\widetilde{\boldsymbol{M}}) \le \lambda_2$. We can finally obtain $\boldsymbol{M}$ by using the orthogonal matrix $\boldsymbol{T}$ to change the basis $\boldsymbol{B}$ to the canonical basis of $\mathbb{R}^d$ as $\boldsymbol{M} = \boldsymbol{T}^{-1} \widetilde{\boldsymbol{M}} \boldsymbol{T} = \boldsymbol{T}^T \widetilde{\boldsymbol{M}} \boldsymbol{T}$. Clearly, $\boldsymbol{M}$ has the same eigenvalues as $\widetilde{\boldsymbol{M}}$ and is symmetric. □

Lemma B.2. *Let $\boldsymbol{M}$ be a symmetric positive definite $(d \times d)$-matrix, $p \geq 1$ and a constant $\gamma > 0$ be such that $\gamma\lambda_{(d)}(\boldsymbol{M}) < 1$. Then $\|\boldsymbol{I} - \gamma\boldsymbol{M}\| = 1 - \gamma\lambda_{(1)}(\boldsymbol{M})$ and*

$$0 < 1 - \gamma\lambda_{(d)}(\boldsymbol{M}) = \lambda_{(1)}(\boldsymbol{I} - \gamma\boldsymbol{M}) \leq \lambda_{(d)}(\boldsymbol{I} - \gamma\boldsymbol{M}) = 1 - \gamma\lambda_{(1)}(\boldsymbol{M}) < 1.$$

Besides, $\|\boldsymbol{M}\|_p \leq K_p\|\boldsymbol{M}\| = K_p\lambda_{(d)}(\boldsymbol{M})$ for some constant $K_p = K_p(d) > 0$.

Proof. Let λ_is be the eigenvalues of $\boldsymbol{M}$, so that the matrix $\boldsymbol{I} - \gamma\boldsymbol{M}$ has eigenvalues $1 - \gamma\lambda_i$, $i = 1, \ldots, d$. Since $\gamma\lambda_{(d)}(\boldsymbol{M}) < 1$, then, for all $i = 1, \ldots, d$, $0 < \gamma\lambda_{(1)}(\boldsymbol{M}) \leq \gamma\lambda_i \leq \gamma\lambda_{(d)}(\boldsymbol{M}) < 1$, implying $1 > 1 - \gamma\lambda_{(1)}(\boldsymbol{M}) \geq 1 - \gamma\lambda_i \geq 1 - \gamma\lambda_{(d)}(\boldsymbol{M}) > 0$, so that $\|\boldsymbol{I} - \gamma\boldsymbol{M}\| = \max_i |1 - \gamma\lambda_i| = 1 - \gamma\lambda_{(1)}(\boldsymbol{M}) < 1$. The first two assertions follow.

It remains to prove the last assertion. For $x \in \mathbb{R}^d$, let $R_2^p = R_2^p(d) = \max_{x\neq 0} \|x\|_p/\|x\|_2$ and $R_p^2 = R_p^2(d) = \max_{x\neq 0} \|x\|_2/\|x\|_p$. According to Theorem 5.6.18 from Horn and Johnson (1988),

$$\max_{M\neq O} \frac{\|\boldsymbol{M}\|_p}{\|\boldsymbol{M}\|_2} = R_2^p R_p^2 = K_p.$$

Recall that $\|\boldsymbol{M}\|_2 = \|\boldsymbol{M}\| = \lambda_{(d)}(\boldsymbol{M})$ and $\|x\|_s \leq \|x\|_r \leq d^{1/r-1/s}\|x\|_s$ for any $x \in \mathbb{R}^d$ and $s \geq r \geq 1$. From the last relation it is easy to get the following bounds: $R_2^p \leq 1$ if $p \geq 2$, $R_2^p \leq d^{(2-p)/(2p)}$ if $1 \leq p < 2$; $R_p^2 \leq d^{(p-2)/(2p)}$ if $p \geq 2$, $R_p^2 \leq 1$ if $1 \leq p < 2$. These bounds imply that $K_p \leq d^{(p-2)/(2p)}$ if $p \geq 2$ and $K_p \leq d^{(2-p)/(2p)} \leq d^{1/2}$ if $1 \leq p < 2$. This completes the proof of the lemma. □

Lemma B.3 (Abel tranformation). *Suppose $d_1, d, k_0, k \in \mathbb{N}$ and $k_0 \leq k$. Let $\boldsymbol{B}_i$ be $(d_1 \times d)$-matrices, $\boldsymbol{A}_i \in \mathbb{R}^d$ and $\boldsymbol{A}_i = \sum_{j=k_0}^{i} a_j$, $i = k_0, \ldots, k$. Then*

$$\sum_{i=k_0}^{k} \boldsymbol{B}_i\boldsymbol{A}_i = \sum_{i=k_0}^{k-1}(\boldsymbol{B}_i - \boldsymbol{B}_{i+1})\boldsymbol{A}_i + \boldsymbol{B}_k\boldsymbol{A}_k.$$

Proof. We prove this by induction in k. For $k = k_0$ we simply have $\boldsymbol{B}_{k_0}\boldsymbol{A}_{k_0} = \boldsymbol{B}_{k_0}\boldsymbol{A}_{k_0} = \boldsymbol{B}_{k_0}\boldsymbol{A}_{k_0}$ and the assertion holds true. Assume that the equality holds for $k = n$ and let us prove the result for $k = n + 1$. We have

$$\begin{aligned}\sum_{i=k_0}^{n+1} \boldsymbol{B}_i\boldsymbol{A}_i &= \sum_{i=k_0}^{n} \boldsymbol{B}_i\boldsymbol{A}_i + \boldsymbol{B}_{n+1}\boldsymbol{A}_{n+1} = \sum_{i=k_0}^{n-1}(\boldsymbol{B}_i - \boldsymbol{B}_{i+1})\boldsymbol{A}_i + \boldsymbol{B}_n\boldsymbol{A}_n + \boldsymbol{B}_{n+1}\boldsymbol{A}_{n+1}\\ &= \sum_{i=k_0}^{n}(\boldsymbol{B}_i - \boldsymbol{B}_{i+1})\boldsymbol{A}_i - (\boldsymbol{B}_n - \boldsymbol{B}_{n+1})\boldsymbol{A}_n + \boldsymbol{B}_n\boldsymbol{A}_n + \boldsymbol{B}_{n+1}\boldsymbol{A}_{n+1}\\ &= \sum_{i=k_0}^{n}(\boldsymbol{B}_i - \boldsymbol{B}_{i+1})\boldsymbol{A}_i + \boldsymbol{B}_{n+1}\boldsymbol{A}_{n+1}.\end{aligned}$$

□

ACKNOWLEDGMENTS

We thank the Editor and the anonymous referees for suggesting improvements to this article.

REFERENCES

Bassett, G. and Koenker, R. (1978). Asymptotic Theory of Least Absolute Error Regression, *Journal of American Statistical Association* 73: 618–622.

Belitser, E. N. and Korostelev, A. P. (1992). Pseudovalues and Minimax Filtering Algorithms for the Nonparametric Median, *Advances in Soviet Mathematics* 12: 115–124.

Belitser, E. and Serra, P. (2013). On Properties of the Algorithm for Pursuing a Drifting Quantile, *Automation and Remote Control* 74: 613–627.

Belitser, E. and van de Geer, S. (2000). On Robust Recursive Nonparametric Curve Estimation, in *High Dimensional Probability II*, E. Giné, D. M. Mason, and J. A. Wellner, eds., pp. 391–403, Boston: Birkhäuser.

Cade, B. and Noon, B. (2003). A Gentle Introduction to Quantile Regression for Ecologists, *Frontiers in Ecology and the Environment* 1: 412–420.

Donoho, D. L. and Gasko, M. (1992). Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness, *Annals of Statistics* 20: 1803–1827.

Horn, R. and Johnson, C. (1988). *Matrix Analysis*, Cambridge: Cambridge University Press.

Koenker, R. (2005). *Quantile Regression*, Cambridge: Cambridge University Press.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method, *Annals of Mathematical Statistics* 22: 400–407.

Shiryaev, A. N. (1996). *Probability*, 2nd edition, New York: Springer.

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006). Nonparametric Quantile Estimation, *Journal of Machine Learning Research* 7: 1231–1264.

Tukey, J. W. (1974a). *Address to International Congress of Mathematicians*, Vancouver, Canada.

Tukey, J. W. (1974b). Order Statistics, in *Mimeographed Notes for Statistics 411*, Princeton: Princeton University.

Tukey, J. W. (1975). Mathematics and the Picturing of Data, in *Proceedings of International Congress of Mathematicians, vol. 2*, R. James, ed., pp. 523–531. Montreal: Canadian Mathematical Congress, August 21–29, 1974.

Zuo, Y. and Serfling, R. (2000). General Notions of Statistical Depth Function, *Annals of Statistics* 28: 461–482.