Robust-to-outliers square-root LASSO, simultaneous inference with a MOM approach

Gianluca Finocchio[*][†], Alexis Derumigny[‡] and Katharina Proksch[*]

March 19, 2021

Abstract

We consider the least-squares regression problem with unknown noise variance, where the observed data points are allowed to be corrupted by outliers. Building on the median-of-means (MOM) method introduced by Lecue and Lerasle [15] in the case of known noise variance, we propose a general MOM approach for simultaneous inference of both the regression function and the noise variance, requiring only an upper bound on the noise level. Interestingly, this generalization requires care due to regularity issues that are intrinsic to the underlying convex-concave optimization problem. In the general case where the regression function belongs to a convex class, we show that our simultaneous estimator achieves with high probability the same convergence rates and a similar risk bound as if the noise level was unknown, as well as convergence rates for the estimated noise standard deviation.

In the high-dimensional sparse linear setting, our estimator yields a robust analog of the square-root LASSO. Under weak moment conditions, it jointly achieves with high probability the minimax rates of estimation $s^{1/p}\sqrt{(1/n)\log(p/s)}$ for the ℓ_p-norm of the coefficient vector, and the rate $\sqrt{(s/n)\log(p/s)}$ for the estimation of the noise standard deviation. Here n denotes the sample size, p the dimension and s the sparsity level. We finally propose an extension to the case of unknown sparsity level s, providing a jointly adaptive estimator $(\widetilde{\beta}, \widetilde{\sigma}, \widetilde{s})$. It simultaneously estimates the coefficient vector, the noise level and the sparsity level, with proven bounds on each of these three components that hold with high probability.

Keywords: Median-of-means, robustness, simultaneous adaptivity, unknown noise variance, minimax rates, sparse linear regression, high-dimensional statistics.

MSC 2020: Primary: 62G35, 62J07; Secondary: 62C20, 62F35.

[*]University of Twente, Enschede, the Netherlands.

[‡]Department of Applied Mathematics, Delft University of Technology, Delft, the Netherlands.

1 Introduction

We consider the statistical learning problem of predicting a real random variable Y by means of an explanatory variable $\mathbf{X}$ belonging to some measurable space $\mathcal{X}$. Given a dataset $\mathcal{D}$ of observations and a function class $\mathcal{F}$, the goal is to choose a function $\widehat{f} \in \mathcal{F}$ in such a way that $\widehat{f}(\mathbf{X})$ approximates Y as well as possible. In particular, we study the problem of predicting Y with the mean-squared loss, which corresponds to the estimation of an *oracle function* $f^* \in \arg\min_{f \in \mathcal{F}} \mathbb{E}[(Y - f(\mathbf{X}))^2]$. This setting has been formalized by [15] in the context of robust machine learning. In this framework, one observes a (possibly) contaminated dataset consisting of *informative observations* (sometimes called *inliers*), and *outliers*. The statistician does not know which data points are corrupted and nothing is usually assumed about the outliers, however one expects the informative observations to be sufficient to solve the problem at hand if the number of outliers is not too large. Even when the inliers are a sample of i.i.d. observations with finite second-moment, such a corrupted dataset can break naive estimators even in the simplest of problems: a single big outlier can push an empirical average towards infinity when estimating the mean of a real random variable. A much better choice of estimator in the presence of outliers is the so-called median-of-means, which is constructed as follows: given a partition of the dataset into some number K of blocks, one computes the empirical average relative to each block, and then takes the median of all these empirical averages. The resulting object is robust to $K/2$ outliers and has good performance even when the underlying distribution has no second moment, see [10, Section 4.1]. Some of the key ideas behind the median-of-means construction can be traced back to the work on stochastic optimization [21, 16], sampling from large discrete structures [12], and sketching algorithms [1].

Our work builds on the MOM method introduced in [15], which solves the least-squares problem by implementing a convex-concave optimization of a suitable functional. In the sparse linear case, this problem can be rewritten as the estimation of $\boldsymbol{\beta}^*$ in the model $Y = \mathbf{X}^T \boldsymbol{\beta}^* + \zeta$ for some noise ζ, where $\mathcal{F}_{s^*} = \{\mathbf{x} \mapsto \mathbf{x}^T \boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^d, \ |\boldsymbol{\beta}|_0 \leq s^*\}$ for some sparsity level $s^* > 0$ and $|\boldsymbol{\beta}|_0$ is the number of non-zero components of $\boldsymbol{\beta}$. The MOM-LASSO method [15] yields there a robust version of the LASSO estimator, which is known to be minimax optimal, see [2, 3, 4], but its optimal penalization parameter has to be proportional to the noise standard deviation σ^*. However, in practical applications this noise level σ^* is often unknown to the statistician, and, as a consequence, it may be difficult to apply the MOM-LASSO. We extend this MOM approach to the case of unknown noise variance and highlight the challenges that arise from this formulation of the problem. The main contribution of our paper is the choice of a new functional in the convex-concave procedure that yields, in the sparse linear case, a robust version of the square-root LASSO introduced in [5], which was shown to be minimax optimal by [8], while its penalization parameter does not require knowledge of σ^*. Interestingly, intuitive and seemingly innocuous choices of functional end up requiring too restrictive assumptions, such as a known (or estimated) lower bound $\sigma_- > 0$ on the noise standard deviation as in [9], whereas in this article, we only require a known (or estimated)

2

upper bound σ_+.

Our main results deal with the simultaneous estimation of the oracle function f^* and standard deviation σ^* of the residual $\zeta := Y - f^*(\mathbb{X})$. In the high-dimensional sparse linear regression setting with unknown σ^*, if the sparsity level $s^* \leq d$ is known and the number of outliers is no more than $O(s^* \log(ed/s^*))$, we prove that our MOM achieves the optimal rates of estimation of $\boldsymbol{\beta}^*$ using a number of blocks K of order $O(s^* \log(ed/s^*))$. We also prove that our estimator of the noise standard deviation satisfies $|\widehat{\sigma}_{K,\mu} - \sigma^*| \lesssim \sigma_+ \sqrt{\frac{s^*}{n} \log\left(\frac{ed}{s^*}\right)}$ with high probability, improving the rates compared to the previous best estimator $\hat{\sigma}$, see [6, Corollary 2], which satisfies $|\hat{\sigma}^2 - \sigma^2| \lesssim \sigma^{*2} \left(\frac{s^* \log(n \vee d \log n)}{n} + \sqrt{\frac{s^* \log(d \vee n)}{n}} + \frac{1}{\sqrt{n}} \right)$ whenever the noise has a finite fourth moment. Note that these rates for the estimation of σ^* derived in [6] correspond to a different penalty level than the one used in [8] that allows to derive optimal rates for the estimation of $\boldsymbol{\beta}^*$. A related paper is [7], which studies optimal noise level estimation for the sparse Gaussian sequence model.

Since the sparsity level may be unknown in practice, we provide an aggregated adaptive procedure based on Lepski's method, that is, we first infer an estimated sparsity $\widetilde{s}$ and then an estimated number of blocks $\widetilde{K}$ of order $O(\widetilde{s} \log(ed/\widetilde{s}))$. We show that the resulting adaptive estimator $(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}, \widetilde{s})$ attains the minimax rates for the estimation of $\boldsymbol{\beta}^*$ while still being adaptive to the unknown noise variance σ^2 and selecting a sparse model ($\widetilde{s} \leq s^*$) with high probability.

Estimator	Rate on $\boldsymbol{\beta}$	Adapt. to s	Rate and adapt. to σ^*	Robustness
Lasso	Optimal [3]	-	-	-
Aggreg. Lasso	Optimal [3]	Yes	-	-
Square-root Lasso	Optimal [8]	-	Yes, complicated rate [6]	-
Aggreg. Square-root Lasso	Optimal [8]	Yes	Yes, but no rate	-
MOM-Lasso	Optimal [15]	-	-	Yes
Aggreg. MOM-Lasso	Optimal [15]	Yes	-	Yes
Robust SR-Lasso	Optimal (Th. 4.4)	-	$\sqrt{\frac{s^*}{n} \log\left(\frac{ed}{s^*}\right)}$ (Th. 4.4)	Yes
Aggreg. Robust SR-Lasso	Optimal (Th. 4.7)	Yes	$\sqrt{\frac{s^*}{n} \log\left(\frac{ed}{s^*}\right)}$ (Th. 4.7)	Yes

Table 1: Comparison of estimators of sparse high-dimensional regressions and their main theoretical properties. Names in bold print refer to the new estimators that we propose in this article.

In Table 1, we detail a comparison of the Lasso-type estimators and their different theoretical properties in this sparse high-dimensional regression framework. The two new estimators that we propose solve the problem of minimax-optimal robust estimation of $\boldsymbol{\beta}$. Even in the setting where no outliers are present, our estimators still improve the best-known bounds on the estimation of the noise variance σ^{*2}. Moreover, the second estimator $(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}, \widetilde{s})$ attains the same rate of simultaneous estimation of $\boldsymbol{\beta}^*$ and σ^* adaptively to the sparsity level s^*. Finally, the estimator $(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma})$ is robust to the same number of outliers as the estimator which uses the knowledge of the true sparsity level s^*. For every $\sigma^* > 0$, let $\mathcal{P}(\sigma^*)$ be a class of distributions of $(\mathbf{X}, \zeta)$ such that the kurtosis of ζ is bounded, $\mathrm{Var}[\zeta] = \sigma^{*2}$ and $\mathbf{X}$ is isotropic, satisfies a weak moment condition and is such that the weighted norms $L^1(\mathbb{P}_{\mathbf{X}}), L^2(\mathbb{P}_{\mathbf{X}})$, and $L^4(\mathbb{P}_{\mathbf{X}})$

are equivalent on $\mathbb{R}^d$. We work with a dataset $\mathcal{D} = (\mathbf{X}_i, Y_i)_{i=1,\ldots,n}$ that might be contaminated by a set of outliers $(\mathbf{X}_i, Y_i)_{i \in \mathcal{O}}$ (for some $\mathcal{O} \subset \{1, \ldots, n\}$) in the sense that, for $i \in \mathcal{O}$, $(\mathbf{X}_i, Y_i)$ is an arbitrary outlier while for $i \notin \mathcal{O}$, $(\mathbf{X}_i, Y_i)$ is i.i.d. distributed as $(\mathbf{X}, Y)$. We denote by $\mathcal{D}(N)$ the set of all possible modifications of $\mathcal{D}$ by at most N observations. To sum up, our joint estimator $(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}, \widetilde{s})$ satisfies the following worst-case simultaneous deviation bound

$$\inf_{\substack{s^*=1,\ldots,s_+ \\ \sigma^* < \sigma_+}} \inf_{\boldsymbol{\beta}^* \in \mathcal{F}_{s^*}} \inf_{P_{\mathbf{X},\zeta} \in \mathcal{P}(\sigma^*)} P_{\boldsymbol{\beta}^*, P_{\mathbf{X},\zeta}}^{\otimes n} \left(\mathcal{A}_{\sigma^*, \boldsymbol{\beta}^*, s^*}(\mathcal{D}) \right) \geq 1 - \phi(s_+, d),$$

where the event $\mathcal{A}_{\sigma^*, \boldsymbol{\beta}^*, s^*}(\mathcal{D})$ describes the performance of the aggregated estimator over a class of contaminations of the dataset $\mathcal{D}$ by arbitrary outliers. Formally,

$$\mathcal{A}_{\sigma^*, \boldsymbol{\beta}^*, s^*}(\mathcal{D}) := \bigcap_{\mathcal{D}' \in \mathcal{D}\left(cs^* \log(ed/s^*)\right)} \mathcal{A}_{\sigma^*}(\mathcal{D}') \cap \mathcal{A}_{\boldsymbol{\beta}^*}(\mathcal{D}') \cap \mathcal{A}_{s^*}(\mathcal{D}'),$$

$$\mathcal{A}_{\sigma^*}(\mathcal{D}') := \left\{ \left| \widetilde{\sigma}(\mathcal{D}') - \sigma^* \right| \leq C\sigma_+ \sqrt{\frac{s^*}{n} \log\left(\frac{ed}{s^*}\right)} \right\},$$

$$\mathcal{A}_{\boldsymbol{\beta}^*}(\mathcal{D}') := \left\{ \left| \widetilde{\boldsymbol{\beta}}(\mathcal{D}') - \boldsymbol{\beta}^* \right|_p \leq C\sigma_+ s^{*1/p} \sqrt{\frac{1}{n} \log\left(\frac{ed}{s^*}\right)} \right\},$$

$$\mathcal{A}_{s^*}(\mathcal{D}') := \left\{ \widetilde{s}(\mathcal{D}') \leq s^* \right\},$$

where $(\widetilde{\boldsymbol{\beta}}(\mathcal{D}'), \widetilde{\sigma}(\mathcal{D}'), \widetilde{s}(\mathcal{D}'))$ is the joint estimator obtained from the perturbed dataset $\mathcal{D}'$. Our method only requires the knowledge of the upper bounds (σ_+, s_+), where $\mathcal{F}_s$ is the set of s-sparse vectors, $|\cdot|_p$ is the ℓ_p norm, $\phi(s, d) := 4(\log_2(s) + 1)^2 (2s/ed)^{C's}$ for a universal constant $C' > 0$, the constants $c, C > 0$ only depend on the class $\mathcal{P}(\sigma^*)$, $|\mathcal{O}|$ denotes the cardinality of the set $\mathcal{O}$ and $P_{\boldsymbol{\beta}^*, P_{\mathbf{X},\zeta}}$ is the distribution of $(\mathbf{X}, Y)$ when $(\mathbf{X}, \zeta) \sim P_{\mathbf{X},\zeta}$ and $Y = \mathbf{X}^\top \boldsymbol{\beta}^* + \zeta$.

The manuscript is organized as follows. In Section 2, we introduce the main framework and notation, as well as the step-by-step construction of the MOM estimator. In Section 3 we present our results in the general situation of a convex class $\mathcal{F}$ of regression functions. The results for the high-dimensional sparse linear regression framework are presented in Section 4. In Section 5 we discuss the contraction rates, the construction of the MOM estimator and some known results from the literature. The proofs are gathered in the appendix.

2 Notation and framework

2.1 General notation

Vectors are denoted by bold letters, e.g. $\mathbf{x} := (x_1, \ldots, x_d)^\top$. For $S \subseteq \{1, \ldots, d\}$, we write $|S|$ for the cardinality of S. As usual, we define $|\mathbf{x}|_p := (\sum_{i=1}^d |\mathbf{x}_i|^p)^{1/p}$, $|\mathbf{x}|_\infty := \max_i |\mathbf{x}_i|$, $|\mathbf{x}|_0 := \sum_{i=1}^d \mathbf{1}(\mathbf{x}_i \neq 0)$, where $\mathbf{1}$ is the indicator function and write $\|f\|_{L^p(D)}$ for the L^p norm of f on D. If there is no ambiguity concerning the domain D, we also write $\|\cdot\|_p$. We set

$|\mathbf{x}|_{2,n} := |\mathbf{x}|_2/\sqrt{n}$ and, for a measure μ on $\mathbb{R}^d$ and a function f in a class of functions $\mathcal{F}$, we define $\|f\|_{2,\nu} := \|f\|_{L^2(\nu)}$. The expected value of a random variable X with respect to a measure P is denoted PX. For two sequences $(a_n)_n$ and $(b_n)_n$ we write $a_n \lesssim b_n$ if there exists a constant C such that $a_n \leq Cb_n$ for all n. Moreover, $a_n \asymp b_n$ means that $(a_n)_n \lesssim (b_n)_n$ and $(b_n)_n \lesssim (a_n)_n$.

2.2 Mathematical framework

The goal is to predict a square-integrable random variable $Y \in \mathbb{R}$ by means of an explanatory random variable $\mathbf{X}$, on a measurable space $\mathcal{X}$, and a dataset $\mathcal{D} = \{(\mathbf{X}_i, Y_i) \in \mathcal{X} \times \mathbb{R} : i = 1, \ldots, n\}$. Let $\mathbb{P}_{\mathbf{X}}$ be the law of $\mathbf{X}$ and $L^2(\mathbb{P}_{\mathbf{X}})$ the corresponding weighted L^2-space. Let $\mathcal{F} \subseteq L^2(\mathbb{P}_{\mathbf{X}})$ be a convex class of functions from $\mathcal{X}$ to $\mathbb{R}$, so that, for any $f \in \mathcal{F}$, $\|f\|_{2,\mathbf{X}}^2 := \int_{\mathcal{X}} f(\mathbf{x})^2 d\mathbb{P}_{\mathbf{X}}(\mathbf{x})$ is finite. We consider the least-squares problem, which requires to minimize the *risk* $\mathrm{Risk}(f) := \mathbb{E}[(Y - f(\mathbf{X}))^2]$ among all possible predictions $f(\mathbf{X})$ for Y, which in turn minimizes the variance of the residuals $\zeta_f := Y - f(\mathbf{X})$. The best predictor on $L^2(\mathbb{P}_{\mathbf{X}})$ is the conditional mean $\overline{f}(\mathbf{X}) = E[Y|\mathbf{X}]$, which can only be computed when the joint distribution of $(\mathbf{X}, Y)$ is given. Therefore, one solves the least-squares problem by estimating any *oracle solution*

$$f^* \in \mathcal{F}^* := \underset{f \in \mathcal{F}}{\arg \min} \, \mathbb{E}[(Y - f(\mathbf{X}))^2], \tag{2.1}$$

which is unique, i.e. $\mathcal{F}^* = \{f^*\}$, if the class $\mathcal{F} \subseteq L^2(\mathbb{P}_{\mathbf{X}})$ is closed (on top of being convex). The resulting representation is

$$Y = f^*(\mathbf{X}) + \zeta, \quad \zeta := Y - f^*(\mathbf{X}), \tag{2.2}$$

where the residual ζ and $\mathbf{X}$ may not be independent.

Assumption 2.1. *We make the following assumptions on the residual ζ,*

$$\mathbb{E}[\zeta] = 0, \quad \sigma^* := \mathbb{E}[\zeta^2]^{\frac{1}{2}} \leq \sigma_+, \quad \mathfrak{m}^* := \mathbb{E}[\zeta^4]^{\frac{1}{4}} \leq \mathfrak{m}_+ := \sigma_+\kappa_+, \quad \kappa^* := \frac{\mathfrak{m}^{*4}}{\sigma^{*4}} \leq \kappa_+, \tag{2.3}$$

with possibly unknown $\sigma^*, \mathfrak{m}^*, \kappa^*$ and upper bounds σ_+, κ_+ either given or estimated from the data. We use the convention that $\kappa^* = 0$ if both σ^* and $\mathfrak{m}^*$ are zero.

Without loss of generality we have $\sigma_+ \leq \mathfrak{m}_+$, since any upper bound on $\mathfrak{m}^*$ is also an upper bound on the standard deviation σ^*. The requirement of a known upper bound on the fourth moment of the noise is natural when dealing with MOM procedures, this is in line with Assumption 3.1 in [18]. We aim at simultaneously estimating (f^*, σ^*) from the dataset $\mathcal{D}$, but the problem is made more difficult due to possible outliers in the observations.

Assumption 2.2. *We assume the dataset $\mathcal{D}$ can be partitioned into an informative set $\mathcal{D}_{\mathcal{I}}$ and an outlier set $\mathcal{D}_{\mathcal{O}}$ satisfying the following.*

- ***Informative data.*** *We assume that the pairs $(\mathbf{X}_i, Y_i)_{i \in \mathcal{I}} =: \mathcal{D}_{\mathcal{I}}$ with $\mathcal{I} \subseteq \{1, \ldots, n\}$ are independent and distributed as $(\mathbf{X}, Y)$ in the regression model (2.2).*

- **_Outliers._** _Nothing is assumed on the pairs_ $(\mathbf{X}_i, Y_i)_{i \in \mathcal{O}} =: \mathcal{D}_{\mathcal{O}}$ _with_ $\mathcal{O} \subseteq \{1, \ldots, n\}$. _They might be deterministic or even adversarial, in the sense that they might depend on the informative sample_ $(\mathbf{X}_i, Y_i)_{i \in \mathcal{I}}$ _defined above, or on the choice of estimator._

The i.i.d. requirement on the informative data can be weakened, as in [15], by assuming that the observations $(\mathbf{X}_i, Y_i)_{i \in \mathcal{I}}$ are independent and, for all $i \in \mathcal{I}$

$$\mathbb{E}[(Y_i - f^*(\mathbf{X}_i))(f - f^*)(\mathbf{X}_i)] = \mathbb{E}[(Y - f^*(\mathbf{X}))(f - f^*)(\mathbf{X})],$$
$$\mathbb{E}[(f - f^*)^2(\mathbf{X}_i)] = \mathbb{E}[(f - f^*)^2(\mathbf{X})].$$

In other words, the distributions of $(\mathbf{X}_i, Y_i)$ and $(\mathbf{X}, Y)$ induce the same L^2-metric on the function space $\mathcal{F} - f^* = \{f - f^* : f \in \mathcal{F}\}$.

By construction, $\mathcal{I} \cup \mathcal{O} = \{1, \ldots, n\}$ and $\mathcal{I} \cap \mathcal{O} = \emptyset$, but the statistician does not know whether any fixed index $i \in \{1, \ldots, n\}$ belongs to $\mathcal{I}$ or $\mathcal{O}$. Otherwise, one could just remove this group from the dataset and perform the inference of the informative part. In order to achieve robust inference, we implement a median-of-means approach.

The sparse linear case. We highlight the special case when $\mathcal{X} = \mathbb{R}^d$, with a fixed dimension $d > 0$. For $\boldsymbol{\beta} \in \mathbb{R}^d$, set $f_{\boldsymbol{\beta}} : \mathbb{R}^d \to \mathbb{R}$ the linear map $f_{\boldsymbol{\beta}}(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$. For any $1 \leq s \leq d$, we define

$$\mathcal{F} := \{f_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathbb{R}^d\}, \quad \mathcal{F}_s := \{f_{\boldsymbol{\beta}} \in \mathcal{F} : \boldsymbol{\beta} \in \mathbb{R}^d, \ |\boldsymbol{\beta}|_0 \leq s\},$$

here $|\boldsymbol{\beta}|_0$ is the number of non-zero entries of $\boldsymbol{\beta} \in \mathbb{R}^d$.

2.3 Convex-concave formulation

We follow the formalization made in [15]. For any function $f \in \mathcal{F}$, and any $(\mathbf{x}, y) \in \mathcal{X} \times \mathbb{R}$, set $\ell_f(\mathbf{x}, y) := (y - f(\mathbf{x}))^2$. In our setting we find

$$f^* \in \underset{f \in \mathcal{F}}{\arg\min} \, \mathbb{E}[\ell_f(\mathbf{X}, Y)], \quad \sigma^* = \mathbb{E}[\ell_{f^*}(\mathbf{X}, Y)]^{\frac{1}{2}},$$

since $\mathbb{E}[\ell_{f^*}(\mathbf{X}, Y)] = \mathbb{E}[\zeta^2]$ is the risk of the oracle function f^*. The oracle pair (f^*, σ^*) is a solution of the convex-concave problem

$$f^* \in \underset{f \in \mathcal{F}}{\arg\min} \sup_{g \in \mathcal{F}} \mathbb{E}[\ell_f(\mathbf{X}, Y) - \ell_g(\mathbf{X}, Y)], \quad \sigma^* = \mathbb{E}[\ell_{f^*}(\mathbf{X}, Y)]^{\frac{1}{2}}, \tag{2.4}$$

and the goal is to build an estimator $(\widehat{f}, \widehat{\sigma})$ such that, with probability as high as possible, the quantities

$$\mathrm{Risk}(\widehat{f}) - \mathrm{Risk}(f^*), \quad \|\widehat{f} - f^*\|_{2, \mathbf{X}}, \quad |\widehat{\sigma} - \sigma^*|,$$

are as small as possible. The quantity $\mathrm{Risk}(\widehat{f}) - \mathrm{Risk}(f^*)$ is the _excess risk_, whereas the quantity $\|\widehat{f} - f^*\|_{2, \mathbf{X}}$ is the convergence rate in $L^2(\mathbb{P}_{\mathbf{X}})-$norm of the random function $\widehat{f}$ to f^*. Since $\widehat{f}$ is a function of the dataset $\mathcal{D}$, we always mean that the expectation is conditional on $\mathcal{D}$, i.e. $\|\widehat{f} - f^*\|_{2, \mathbf{X}} = \mathbb{E}[(\widehat{f} - f^*)^2(\mathbf{X})|\mathcal{D}]$. Finally, the quantity $|\widehat{\sigma} - \sigma^*|$ is the convergence rate of $\widehat{\sigma}$ to σ^*.

2.4 Construction of the estimator

The starting point of our approach is the regularized median-of-means (MOM) tournament introduced in [17], which has been proposed as a procedure to outperform the *regularized empirical risk minimizer* (RERM)

$$\widehat{f}_\lambda^{RERM} := \underset{f \in \mathcal{F}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - f(\mathbf{X}_i))^2 + \lambda \|f\| \right\},$$

with $\|\cdot\|$ a *penalization* norm on the linear span of $\mathcal{F}$ and $\lambda > 0$ a *penalization parameter*. The penalization term reduces overfitting by assigning a higher cost to functions that are big with respect to $\|\cdot\|$. The RERM estimator above is susceptible to outliers since it involves all the pairs $(\mathbf{X}_i, Y_i)$ in the dataset $\mathcal{D}$, whereas replacing the empirical average by the corresponding median-of-means over a number of blocks leads to robustness. The MOM method in [15] builds directly on the theory of the MOM tournaments and it exploits the fact that $\widehat{f}_\lambda^{RERM}$ is computed by minimizing $n^{-1} \sum_{i=1}^n \ell_f(\mathbf{X}_i, Y_i) + \lambda \|f\|$. From this, the authors deal with the convex-concave equivalent

$$\widehat{f}_\lambda^{RERM} := \underset{f \in \mathcal{F}}{\arg\min} \sup_{g \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell_f(\mathbf{X}_i, Y_i) - \frac{1}{n} \sum_{i=1}^n \ell_g(\mathbf{X}_i, Y_i) + \lambda(\|f\| - \|g\|) \right\},$$

by replacing the empirical average $n^{-1} \sum_{i=1}^n \left(\ell_f(\mathbf{X}_i, Y_i) - \ell_g(\mathbf{X}_i, Y_i) \right)$ with the median-of-means over a chosen number of blocks. Our goal is to extend the scope of this procedure to the estimation of the unknown σ^*. To this end, we modify the convex-concave RERM by replacing the functional $R(\ell_g, \ell_f) = \ell_f - \ell_g$ with a new $R_c(\ell_g, \chi, \ell_f, \sigma)$ that incorporates $\chi, \sigma \in I_+ = (0, \sigma_+]$. This leads to a generalized empirical estimator

$$(\widehat{f}_\mu, \widehat{\sigma}_\mu) := \underset{(f,\sigma) \in \mathcal{F} \times I_+}{\arg\min} \sup_{(g,\chi) \in \mathcal{F} \times I_+} \left\{ \frac{1}{n} \sum_{i=1}^n R_c\left(\ell_g(\mathbf{X}_i, Y_i), \chi, \ell_f(\mathbf{X}_i, Y_i), \sigma \right) + \mu(\|f\| - \|g\|) \right\},$$

which we robustify using the MOM. The choice of the functional R_c is crucial for the performance of the procedure and a main contribution of our paper is providing a suitable $R_c(\ell_g, \chi, \ell_f, \sigma)$, we refer to Section 5 for a detailed discussion motivating our choice.

We give the step-by-step construction of a family of MOM estimators for (f^*, σ^*) from model (2.1)–(2.3). We start with a preliminary definition.

Quantiles. For any $K \in \mathbb{N}$, set $[K] = \{1, \ldots, K\}$. For all $\alpha \in (0,1)$ and $\mathbf{x} = (x_1, \ldots, x_K) \in \mathbb{R}^K$, we call $\alpha-quantile$ of $\mathbf{x}$ any element $Q_\alpha[\mathbf{x}]$ of the set

$$\mathcal{Q}_\alpha[\mathbf{x}] := \Big\{ u \in \mathbb{R} : \big|\{k = 1, \ldots, K : x_k \geq u\}\big| \geq (1 - \alpha)K,$$

$$\text{and } \big|\{k = 1, \ldots, K : x_k \leq u\}\big| \geq \alpha K \Big\}. \tag{2.5}$$

This means that $Q_\alpha[\mathbf{x}]$ is a $\alpha-quantile$ of $\mathbf{x}$ if at least $(1 - \alpha)K$ components of $\mathbf{x}$ are bigger than $Q_\alpha[\mathbf{x}]$ and at least αK components of $\mathbf{x}$ are smaller than $Q_\alpha[\mathbf{x}]$. For all $t \in \mathbb{R}$, we write $Q_\alpha[\mathbf{x}] \geq t$ when there exists $J \subset [K]$ such that $|J| \geq (1 - \alpha)K$ and, for all $k \in J$, $x_k \geq t$. We write $Q_\alpha[\mathbf{x}] \leq t$ if there exists $J \subset [K]$ such that $|J| \geq \alpha K$ and, for all $k \in J$, $x_k \leq t$.

STEP 1. Partition of the dataset.

Let $K \in \mathbb{N}$ be a fixed positive integer. Partition the dataset $\mathcal{D} = \{(\mathbf{X}_i, Y_i) : i = 1, \ldots, n\}$ into K blocks $\mathcal{D}_1, \ldots, \mathcal{D}_K$ of size n/K (assumed to be an integer). This corresponds to a partition of $\{1, \ldots, n\}$ into blocks $B_1, \ldots, B_K$.

STEP 2. Local criterion.

With $c > 1$ and $f, g \in \mathcal{F}$, $\sigma, \chi \in \mathbb{R}_+$, define the functional

$$R_c(\ell_g, \chi, \ell_f, \sigma) := (\sigma - \chi)\left(1 - 2\frac{\ell_f + \ell_g}{(\sigma + \chi)^2}\right) + 2c\frac{\ell_f - \ell_g}{\sigma + \chi}. \tag{2.6}$$

Since $\ell_f(\mathbf{x}, y) = (y - f(\mathbf{x}))^2$ for all $(\mathbf{x}, y) \in \mathcal{X} \times \mathbb{R}$, the latter definition induces the map $(\mathbf{x}, y) \mapsto R_c(\ell_g(\mathbf{x}, y), \chi, \ell_f(\mathbf{x}, y), \sigma)$ over $(\mathbf{x}, y) \in \mathcal{X} \times \mathbb{R}$. For each $k = [K]$, we define the *criterion of (f, σ) against (g, χ) on the block B_k* as the empirical mean of the functional $R_c(\ell_g, \chi, \ell_f, \sigma)$ on that block, that is,

$$\mathbb{P}_{B_k}\left(R_c(\ell_g, \chi, \ell_f, \sigma)\right) := \frac{1}{|B_k|} \sum_{i \in B_k} R_c\left(\ell_g(\mathbf{X}_i, Y_i), \chi, \ell_f(\mathbf{X}_i, Y_i), \sigma\right), \tag{2.7}$$

for all $(g, \chi, f, \sigma) \in \mathcal{F} \times \mathbb{R}_+ \times \mathcal{F} \times \mathbb{R}_+$. Here $|B_k| = n/K$ denotes the cardinality of B_k.

STEP 3. Global criterion.

For any $\alpha \in (0, 1)$ and number of blocks K, set

$$Q_{\alpha, K}\left[R_c(\ell_g, \chi, \ell_f, \sigma)\right] := Q_\alpha\left[\left(\mathbb{P}_{B_k}\left(R_c(\ell_g, \chi, \ell_f, \sigma)\right)\right)_{k \in [K]}\right],$$

the α-quantile of the vector of local criteria defined in the previous step. For $\alpha = 1/2$ we get the *median*. We define the *global criterion of (f, σ) against (g, χ)* as

$$MOM_K\left(R_c(\ell_g, \chi, \ell_f, \sigma)\right) := Q_{1/2, K}\left[R_c(\ell_g, \chi, \ell_f, \sigma)\right], \tag{2.8}$$

for all $(g, \chi, f, \sigma) \in \mathcal{F} \times \mathbb{R}_+ \times \mathcal{F} \times \mathbb{R}_+$. With some norm $\|\cdot\|$ on the span of $\mathcal{F}$, we denote

$$T_{K, \mu}(g, \chi, f, \sigma) := MOM_K\left(R_c(\ell_g, \chi, \ell_f, \sigma)\right) + \mu(\|f\| - \|g\|), \tag{2.9}$$

where $\mu > 0$ is a tuning parameter, the functional $T_{K, \mu}$ is the penalized version of the global criterion.

STEP 4. MOM estimator.

With σ_+ the known upper bound in (2.3), we define the *MOM$-K$ estimator* of (f^*, σ^*) as

$$(\widehat{f}_{K, \mu, \sigma_+}, \widehat{\sigma}_{K, \mu, \sigma_+}) := \underset{f \in \mathcal{F}, \ \sigma \leq \sigma_+}{\arg\min} \ \underset{g \in \mathcal{F}, \ \chi \leq \sigma_+}{\max} \ T_{K, \mu}(g, \chi, f, \sigma), \tag{2.10}$$

where $T_{K, \mu}$ is the penalized functional in (2.9). Furthermore, set

$$\mathcal{C}_{K, \mu}(f, \sigma) := \underset{g \in \mathcal{F}, \ \chi \leq \sigma_+}{\max} \ T_{K, \mu}(g, \chi, f, \sigma). \tag{2.11}$$

The estimator $(\widehat{f}_{K, \mu, \sigma_+}, \widehat{\sigma}_{K, \mu, \sigma_+})$ only depends on the upper bound σ_+, the number K of blocks and the tuning parameter μ.

3 Results for a general class $\mathcal{F}$

We assume the following regularity condition on the function class $\mathcal{F}$ and the inliers.

Assumption 3.1. *There exist constants $\theta_0, \theta_1 > 1$ such that, for all $i \in \mathcal{I}$ and $f \in \mathcal{F}$,*

1. $\|f - f^*\|_{2,\mathbf{X}}^2 = \mathbb{E}[(f - f^*)^2(\mathbf{X}_i)] \leq \theta_0^2 \mathbb{E}[|f - f^*|(\mathbf{X}_i)]^2 = \theta_0^2 \|f - f^*\|_{1,\mathbf{X}}^2.$

2. $\|f - f^*\|_{4,\mathbf{X}}^2 = \mathbb{E}[(f - f^*)^4(\mathbf{X}_i)]^{1/2} \leq \theta_1^2 \mathbb{E}[(f - f^*)^2(\mathbf{X}_i)] = \theta_1^2 \|f - f^*\|_{2,\mathbf{X}}^2.$

This assumption guarantees that the $L^1(\mathbb{P}_{\mathbf{X}}), L^2(\mathbb{P}_{\mathbf{X}}), L^4(\mathbb{P}_{\mathbf{X}})-$norms are equivalent on the function space $\mathcal{F} - f^*$. The equivalence between $\|\cdot\|_{1,\mathbf{X}}$ and $\|\cdot\|_{2,\mathbf{X}}$ in the first condition matches Assumption 3 in [15]. The equivalence between $\|\cdot\|_{2,\mathbf{X}}$ and $\|\cdot\|_{4,\mathbf{X}}$ in the second condition, together with the finiteness of fourth moment of the noise in Assumption 2.1, helps controlling the dependence between ζ and $\mathbf{X}$; this also matches Assumption 3.1 in [18]. We do not necessarily assume that ζ is independent of $\mathbf{X}$, but the Cauchy-Schwarz inequality gives

$$
\begin{aligned}
\|\zeta(f - f^*)\|_{2,\mathbf{X}}^2 &= \mathbb{E}[\zeta^2(f - f^*)^2(\mathbf{X})] \\
&\leq \mathbb{E}[\zeta^4]^{\frac{1}{2}} \mathbb{E}[(f - f^*)^4(\mathbf{X})]^{\frac{1}{2}} \\
&\leq \theta_1^2 \mathfrak{m}^{*2} \mathbb{E}[(f - f^*)^2(\mathbf{X})].
\end{aligned}
$$

The bound $\|\zeta(f - f^*)\|_{2,\mathbf{X}}^2 \leq \theta_1^2 \mathfrak{m}^{*2} \|f - f^*\|_{2,\mathbf{X}}^2$ is Assumption 2 in [15] with $\theta_m^2 = \theta_1^2 \mathfrak{m}^{*2}$, whereas in our setting this is a consequence of Assumption 2.1 and Assumption 3.1.

3.1 Complexity parameters

With the introduction of MOM tournaments procedures, see [18] and references therein, the authors have characterized the underlying geometric features that drive the performance of a learning method. For any $\rho > 0, r > 0$, and $f \in \mathcal{F}$, we set

$$
\mathbb{B}(f, \rho) := \big\{g \in \mathcal{F} : \|g - f\| \leq \rho\big\}, \quad \mathbb{B}_2(f, r) := \big\{g \in \mathcal{F} : \|g - f\|_{2,\mathbf{X}} \leq r\big\},
$$

respectively the $\|\cdot\|-$ball of radius ρ and the $\|\cdot\|_{2,\mathbf{X}}-$ball of radius r, both centered around $f \in \mathcal{F}$. We denote by $\mathbb{B}(\rho)$ and $\mathbb{B}_2(r)$ the balls centered around zero. We define the *regular ball around f^** of radii $\rho > 0, r > 0$ as

$$
\mathbb{B}(f^*, \rho, r) := \{f \in \mathcal{F} : \|f - f^*\| \leq \rho, \ \|f - f^*\|_{2,\mathbf{X}} \leq r\}.
$$

For any subset of inlier indexes $J \subseteq \mathcal{I}$, we define the *standard empirical process on J* as

$$
f \mapsto \mathbb{P}_J(f - f^*) := \frac{1}{|J|} \sum_{i \in J} (f - f^*)(\mathbf{X}_i).
$$

Similarly, we define the *quadratic empirical process on J* and the *multiplier empirical process on J* as

$$f \mapsto \mathbb{P}_J\left((f - f^*)^2\right) := \frac{1}{|J|}\sum_{i \in J}(f - f^*)^2(\mathbf{X}_i),$$

$$f \mapsto \mathbb{P}_J\left(-2\zeta(f - f^*)\right) := -\frac{2}{|J|}\sum_{i \in J}\zeta_i(f - f^*)(\mathbf{X}_i),$$

where $\zeta_i = (Y_i - f^*(\mathbf{X}_i))$. These processes arise naturally when dealing with the *empirical excess risk on J*, which is

$$\begin{aligned}
\text{Risk}_J(f) - \text{Risk}_J(f^*) :&= \frac{1}{|J|}\sum_{i \in J}(Y_i - f(\mathbf{X}_i))^2 - \frac{1}{|J|}\sum_{i \in J}(Y_i - f^*(\mathbf{X}_i))^2 \\
&= \frac{1}{|J|}\sum_{i \in J}(f - f^*)^2(\mathbf{X}_i) - \frac{2}{|J|}\sum_{i \in J}\zeta_i(f - f^*)(\mathbf{X}_i) \\
&= \mathbb{P}_J\left((f - f^*)^2\right) + \mathbb{P}_J\left(-2\zeta(f - f^*)\right).
\end{aligned}$$

The empirical processes defined above only involve observations that are not contaminated by outliers and we are interested in controlling them when the indexing function class is a regular ball $\mathbb{B}(f^*, \rho, r)$.

Let ξ_i be *Rademacher variables*, that is, independent random variables uniformly distributed on $\{-1, 1\}$, and independent from the dataset $\mathcal{D}$. For any $r > 0$ and $\rho > 0$, consider the regular ball $\mathbb{B}(f^*, \rho, r)$ defined above. For every $\gamma_P, \gamma_Q, \gamma_M > 0$, we define the *complexity parameters*

$$r_P(\rho, \gamma_P) := \inf\left\{r > 0 : \sup_{J \subset \mathcal{I}, |J| \geq \frac{n}{2}} \mathbb{E}\left[\sup_{f \in \mathbb{B}(f^*, \rho, r)} \left|\frac{1}{|J|}\sum_{i \in J}\xi_i(f - f^*)(\mathbf{X}_i)\right|\right] \leq \gamma_P r\right\},$$

$$r_Q(\rho, \gamma_Q) := \inf\left\{r > 0 : \sup_{J \subset \mathcal{I}, |J| \geq \frac{n}{2}} \mathbb{E}\left[\sup_{f \in \mathbb{B}(f^*, \rho, r)} \left|\frac{1}{|J|}\sum_{i \in J}\xi_i(f - f^*)^2(\mathbf{X}_i)\right|\right] \leq \gamma_Q r^2\right\}, \quad (3.1)$$

$$r_M(\rho, \gamma_M) := \inf\left\{r > 0 : \sup_{J \subset \mathcal{I}, |J| \geq \frac{n}{2}} \mathbb{E}\left[\sup_{f \in \mathbb{B}(f^*, \rho, r)} \left|\frac{1}{|J|}\sum_{i \in J}\xi_i\zeta_i(f - f^*)(\mathbf{X}_i)\right|\right] \leq \gamma_M r^2\right\},$$

and let $r = r(\cdot, \gamma_P, \gamma_M)$ be a continuous non-decreasing function $r : \mathbb{R}_+ \to \mathbb{R}_+$ depending on γ_P, γ_M, such that

$$r(\rho) \geq \max\left\{r_P(\rho, \gamma_P), r_M(\rho, \gamma_M)\right\}, \tag{3.2}$$

for every $\rho > 0$. The definitions above depend on f^* and require that $|\mathcal{I}| \geq n/2$. The function $r(\cdot)$ matches the one defined in Definition 3 in [15]. We refer to Section 5 for a detailed discussion on the role of complexity parameters, here we only mention that in the sub-Gaussian setting of [13], for some choice of γ_P, γ_M, the quantity $r^*(\rho) = \max\{r_P(\rho, \gamma_P), r_M(\rho, \gamma_M)\}$ is the minimax convergence rate over the function class $\mathbb{B}(f^*, \rho)$.

3.2   Sparsity equation

We follow the setup of [15], that we restate here for convenience.

Subdifferential. Let $\mathcal{E}$ be the vector space generated by $\mathcal{F}$ and $\|\cdot\|$ a norm on $\mathcal{E}$. We denote by $(\mathcal{E}^*, \|\cdot\|_*)$ the dual normed space of $(\mathcal{E}, \|\cdot\|)$, that is, the space of all linear functionals z^* from $\mathcal{E}$ to $\mathbb{R}$. The *subdifferential* of $\|\cdot\|$ at any $f \in \mathcal{F}$ is denoted by

$$(\partial\|\cdot\|)_f := \{z^* \in \mathcal{E}^* : \|f + h\| \geq \|f\| + z^*(h), \ \forall h \in \mathcal{E}\}.$$

The penalization term of the functional $T_{K,\mu}$ in Section 2.4 is of the form $\mu(\|f\| - \|g\|)$, for $f, g \in \mathcal{F}$, and the subdifferential is useful in obtaining lower bounds for $\|f\| - \|f^*\|$. For any $\rho > 0$ and complexity parameter $r(\rho)$ as in (3.2), we denote $H_\rho = \{f \in \mathcal{F} : \|f - f^*\| = \rho, \ \|f - f^*\|_{2,\mathbf{X}} \leq r(\rho)\}$. Furthermore, we set

$$\Gamma_{f^*}(\rho) := \bigcup_{f \in \mathcal{F}: \|f - f^*\| \leq \rho/20} \left(\partial\|\cdot\|\right)_f,$$

$$\Delta(\rho) := \inf_{f \in H_\rho} \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(f - f^*). \tag{3.3}$$

The set $\Gamma_{f^*}(\rho)$ is the set of subdifferentials of all functions that are close to f^* (no more than $\rho/20$) in penalization norm $\|\cdot\|$. The quantity $\Delta(\rho)$ measures the smallest level $\Delta > 0$ for which the chain $\|f\| - \|f^*\| \geq \Delta - \rho/20$ holds. In fact, if $f^{**} \in \mathcal{F}$ is such that $\|f^* - f^{**}\| \leq \rho/20$, then $\|f\| - \|f^*\| \geq \|f\| - \|f^{**}\| - \|f^{**} - f^*\| \geq z^*(f - f^{**}) - \rho/20$, for any subdifferential $z^* \in (\partial\|\cdot\|)_{f^{**}}$.

Sparsity equation. The *sparsity equation* and its smallest solution are

$$\Delta(\rho) \geq \frac{4}{5}\rho, \quad \rho^* := \inf\left\{\rho > 0 : \Delta(\rho) \geq \frac{4\rho}{5}\right\}, \tag{3.4}$$

if ρ^* exists, the sparsity equation holds for any $\rho \geq \rho^*$.

3.3 Main result in the general case

We now present a result dealing with the simultaneous estimation of (f^*, σ^*) by means of a family of MOM estimators constructed as in Section 2.4. Fix any constant $c > 2$ in the definition on the functional R_c in (2.6) and, with $\sigma_+, \mathfrak{m}_+, \kappa_+$ the known bounds on the moments of the noise $\zeta = Y - f^*(\mathbf{X})$, set

$$c_\mu := 200(c+2)\kappa_+^{1/2},$$

$$\varepsilon := \frac{c - 2}{192\,\theta_0^2(c+2)\left(8 + 134\,\kappa_+^{1/2}((1 + \frac{\sigma_+}{\sigma^*}) \vee \frac{6}{5})\right)}, \tag{3.5}$$

$$c_\alpha^2 := \frac{3(c-2)}{5\theta_0^2},$$

and $\gamma_P = 1/(1488\,\theta_0^2)$, $\gamma_M = \varepsilon/744$ and $\gamma_Q = \varepsilon/372$. Let ρ^* be the smallest solution of the sparsity equation in (3.4) and $r(\cdot)$ any function such that $r(\rho) \geq \max\{r_P(\rho, \gamma_P), r_M(\rho, \gamma_M)\}$ as in (3.2). Define K^* as the smallest integer satisfying

$$K^* \geq \frac{n\varepsilon^2 r^2(\rho^*)}{384\,\theta_1^2\mathfrak{m}^{*2}}, \tag{3.6}$$

and, for any integer $K \geq K^*$, also define ρ_K as the implicit solution of

$$r^2(\rho_K) = \frac{384\,\theta_1^2 \mathfrak{m}^{*2} K}{n\varepsilon^2}. \tag{3.7}$$

Assumption 3.2. *We assume that there exists an absolute constant $c_r \geq 1$ such that, for all $\rho > 0$, we have $r(\rho) \leq r(2\rho) \leq c_r r(\rho)$.*

The role of the latter assumption is to simplify the statement of the main result. We are mainly interested in the sparse linear case, where this holds with $c_r = 2$ by construction of the function $r(\cdot)$, see Section 5.4.

Theorem 3.3. *With the notation above, let Assumptions 2.1–3.1 and Assumption 3.2 hold. With $C^2 := 384\,\theta_1^2 c_r^2 c_\alpha^2 \kappa_+^{1/2}$, suppose that $n\varepsilon^2 > 32C^2$ and $|\mathcal{O}| \leq n\varepsilon^2/(32C^2)$. Then, for any integer $K \in \left[K^* \vee 32|\mathcal{O}|,\ n\varepsilon^2/C^2\right]$, and for every $\iota_\mu \in [1/4, 4]$, the MOM$-K$ estimator $(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$ defined in (2.10) with K blocks and penalization parameter*

$$\mu := \iota_\mu c_\mu \varepsilon \frac{r^2(\rho_K)}{\mathfrak{m}^* \rho_K}, \tag{3.8}$$

satisfies, with probability at least $1 - 4\exp(-K/8920)$, for any possible $|\mathcal{O}|$ outliers,

$$\|\widehat{f}_{K,\mu,\sigma_+} - f^*\| \leq 2\,\rho_K, \quad \|\widehat{f}_{K,\mu,\sigma_+} - f^*\|_{2,\mathbf{X}} \leq r(2\rho_K), \quad |\widehat{\sigma}_{K,\mu,\sigma_+} - \sigma^*| \leq c_\alpha r(2\rho_K), \tag{3.9}$$

$$R(\widehat{f}_{K,\mu,\sigma_+}) \leq R(f^*) + \left(2 + 2c_\alpha + (44 + 5c_\mu)\,\varepsilon + \frac{25\kappa^{*1/2}}{8\theta_1^2}\varepsilon^2\right) r^2(2\rho_K) \tag{3.10}$$
$$+ 4\,\theta_1^2 \varepsilon \left(r^2(2\rho_K) \vee r_Q^2(2\rho_K, \gamma_Q)\right).$$

The proof of Theorem 3.3 is given in Appendix A. It provides theoretical guarantees for the MOM$-K$ estimator $(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$: this estimator recovers (f^*, σ^*), with high probability, whenever the number K of blocks is chosen to be at least $K^* \vee 32|\mathcal{O}|$ and at most $n\varepsilon^2/C^2$. Specifically, the random function $\widehat{f}_{K,\mu,\sigma_+}$ belongs to the regular ball $\mathbb{B}(f^*, 2\rho_K, r(2\rho_K))$, whereas the random standard deviation $\widehat{\sigma}_{K,\mu,\sigma_+}$ is at most $c_\alpha r(2\rho_K)$ away from σ^*. The best achievable rates are obtained for $K = K^*$ when $|\mathcal{O}| \leq K^*/32$. Any estimator $(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$ only depends on the penalization parameter μ, the number of blocks K and the upper bound σ_+, thus the result is mainly of interest when these quantities can be chosen without knowledge of (f^*, σ^*). Our Theorem 3.3 extends the scope of Theorem 1 in [15] to the case of unknown noise variance. In the latter reference, the authors obtain the same convergence rates for a MOM$-K$ estimator $\widehat{f}_{K,\lambda}$ defined by using a penalization parameter λ that we compare to our μ,

$$\lambda := 16\varepsilon \frac{r^2(\rho_K)}{\rho_K}, \quad \mu := c_\mu \varepsilon \frac{r^2(\rho_K)}{\mathfrak{m}^* \rho_K},$$

so that μ is proportional to $\lambda/\mathfrak{m}^*$. For the sparse linear case, [15] shows that the optimal choice is $\lambda \sim \mathfrak{m}^* \sqrt{\log(ed/s^*)/n}$, which is proportional to the noise level σ^*. This in turn guarantees that our penalization parameter can be chosen of the form $\mu \sim \sqrt{\log(ed/s^*)/n}$ to obtain the optimal rates, and that such a choice does not depend on the moments of the noise.

12

4 The high-dimensional sparse linear regression

4.1 Results for known sparsity

In this section, we will give non-asymptotic bounds that will hold adaptively and uniformly over a certain class of joint distributions for $(\mathbf{X}, \zeta)$. We now define the class of interest $\mathcal{P}_I$, parametrized by an interval I. This interval I represents the set of possible values for the standard deviation σ^* of the noise ζ.

Definition 4.1 (Class of distributions of interest). *For $I \subset \mathbb{R}_+$, $\theta_0, \theta_1, c_0, L, \kappa_+ > 1$, let us define $\mathcal{P}_I = \mathcal{P}_I(\theta_0, \theta_1, c_0, L, \kappa_+)$ to be the class of distributions $P_{\mathbf{X},\zeta}$ on $\mathbb{R}^{d+1}$ satisfying:*

1. *The standard deviation σ^* of ζ belongs to I and the kurtosis of ζ is smaller than κ_+.*

2. *For all $\boldsymbol{\beta} \in \mathbb{R}^d$, $\mathbb{E}\big[(\mathbf{X}^\top \boldsymbol{\beta})^2\big]^{\frac{1}{2}} \leq \theta_0 \mathbb{E}\big[|\mathbf{X}^\top \boldsymbol{\beta}|\big]$, and $\mathbb{E}\big[(\mathbf{X}^\top \boldsymbol{\beta})^4\big]^{\frac{1}{2}} \leq \theta_1^2 \mathbb{E}\big[(\mathbf{X}^\top \boldsymbol{\beta})^2\big]$.*

3. *$\mathbf{X}$ is isotropic: for all $\boldsymbol{\beta} \in \mathbb{R}^d$, $\|f_{\boldsymbol{\beta}}\|_{2,\mathbf{X}} := \mathbb{E}[(\mathbf{X}^\top \boldsymbol{\beta})^2] = |\boldsymbol{\beta}|_2$, where $f_{\boldsymbol{\beta}}(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$.*

4. *$\mathbf{X}$ satisfies the weak moment condition: for all $1 \leq p \leq c_0 \log(ed)$, $1 \leq j \leq d$, $\mathbb{E}\big[|\mathbf{X}^\top \mathbf{e}_j|^p\big]^{\frac{1}{p}} \leq L\sqrt{p}\,\mathbb{E}\big[|\mathbf{X}^\top \mathbf{e}_j|^2\big]^{\frac{1}{2}}$.*

The class $\mathcal{P}_I$ only requires a finite fourth moment on ζ, allowing it to follow heavy-tailed distributions. The weak moment condition only bounds moments of $\mathbf{X}$ up to the order $\log(d)$, which is weaker than the sub-Gaussian assumption, see [13] and the references therein for a discussion and a list of examples.

Definition 4.2 (Contaminated datasets). *For a dataset $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1,\ldots,n} \in \mathbb{R}^{(d+1)\times n}$ and for $N \in [n]$, we denote by $\mathcal{D}(N)$ the set of all datasets $\mathcal{D}' = (\mathbf{x}'_i, y'_i)_{i=1,\ldots,n} \in \mathbb{R}^{(d+1)\times n}$ that differ from $\mathcal{D}$ by at most N observations, i.e.*

$$\mathcal{D}(N) := \Big\{ \mathcal{D}' \in \mathbb{R}^{(d+1)\times n} : \big|\mathcal{D} \setminus \mathcal{D}'\big| \leq N \Big\},$$

where $\mathcal{D} \setminus \mathcal{D}'$ is defined as the difference between the (multi-)sets $\mathcal{D}$ and $\mathcal{D}'$, meaning that if there exists duplicated observations in $\mathcal{D}$ that appear also in $\mathcal{D}'$, they are removed from $\mathcal{D}$ up to their multiplicities in $\mathcal{D}'$. This encodes all the possible corrupted versions of $\mathcal{D}$ by means of up to N arbitrary outliers.

Definition 4.3. *Let $P_{\boldsymbol{\beta}^*, P_{\mathbf{X},\zeta}}$ be the distribution of $(\mathbf{X}, Y)$ when $(\mathbf{X}, \zeta) \sim P_{\mathbf{X},\zeta}$ and $Y := \mathbf{X}^\top \boldsymbol{\beta}^* + \zeta$.*

In the following, we will use the minimax optimal rates of convergence defined for $p \in [1, 2]$ by $\mathfrak{r}_p := s^{*1/p}\sqrt{(1/n)\log(ed/s^*)}$ and the allowed maximum number of outliers defined by $\mathfrak{r}_{\mathcal{O}} := s^* \log(ed/s^*) = n\mathfrak{r}_2^2$.

Theorem 4.4. *Assume that $\mathfrak{r}_2 < 1$. For every $\theta_0, \theta_1, c_0, L, \kappa_+ > 1$, there exists universal constants $\widetilde{c}_1, \ldots, \widetilde{c}_5 > 0$ such that for every σ_+ and for every $\iota_K, \iota_\mu \in [1/2, 2]^2$, setting*

$$K = \lceil \iota_K \widetilde{c}_1 s^* \log(ed/s^*) \rceil, \quad \mu = \iota_\mu \widetilde{c}_2 \sqrt{\frac{1}{n} \log\left(\frac{ed}{s^*}\right)},$$

the estimator $(\widehat{\boldsymbol{\beta}}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$ satisfies

$$
\inf_{\substack{P_{\mathbf{X},\zeta} \in \mathcal{P}_{[0,\sigma_+]} \\ \boldsymbol{\beta}^* \in \mathcal{F}_{s^*}}} \mathbb{P}_{\mathcal{D} \sim P_{\boldsymbol{\beta}^*, P_{\mathbf{X},\zeta}}^{\otimes n}} \left(\sup_{\mathcal{D}' \in \mathcal{D}(\widetilde{c}_3 \mathfrak{r}_{\mathcal{O}})} \left\{ \mathfrak{r}_2^{-1} |\widehat{\sigma}(\mathcal{D}') - \sigma^*| \right. \right.
$$

$$
\left. \left. \vee \sup_{p \in [1,2]} \mathfrak{r}_p^{-1} |\widehat{\boldsymbol{\beta}}(\mathcal{D}') - \boldsymbol{\beta}^*|_p \right\} \leq \widetilde{c}_4 \sigma_+ \right) \geq 1 - 4 \left(\frac{s^*}{ed} \right)^{\widetilde{c}_5 s^*},
$$

This theorem is proved in Section B.1. Theorem 4.4 ensures that, with high probability, the estimator $(\widehat{\boldsymbol{\beta}}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$ achieves the rates $|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_p \lesssim \sigma_+ s^{*1/p} \sqrt{(1/n) \log(ed/s^*)}$ and $|\widehat{\sigma} - \sigma^*| \lesssim \sigma_+ \sqrt{(s^*/n) \log(ed/s^*)}$, uniformly over the class of distributions $\mathcal{P}_{[0,\sigma_+]}$ with bounded variance while being robust to up to $\widetilde{c}_3 s^* \log(ed/s^*)$ arbitrary outliers. However, the uniform constants appearing in the statement might be difficult to compute in practice, to obtain precise values, one would need to quantify the constants in Theorem 1.6 in [20] and Lemma 5.3 in [14]. As usual for MOM estimators, the maximum number of outliers is of the same order as the number of blocks. Note that the estimator needs the knowledge of an upper bound on the noise level σ_+ and the sparsity level s.

In [3], it has been proved that the optimal minimax rate of estimation of $\boldsymbol{\beta}^*$ in the $|\cdot|_p$ norm is $\sigma^* \sqrt{(s^*/n) \log(ed/s^*)}$ when σ^* is fixed and the noise is sub-Gaussian. Our theorem shows that the rate of estimation of $\boldsymbol{\beta}$ over $\mathcal{P}_{[0,\sigma_+]}$ is the optimal minimax rate of estimation for the worst-case noise level σ_+. In particular, this means that in the noiseless case when $\sigma^* = 0$, the estimator $\widehat{\boldsymbol{\beta}}_{K,\mu,\sigma_+}$ does not achieve perfect reconstruction of the signal $\boldsymbol{\beta}^*$. This is worse than the square-root Lasso [8] which achieves the minimax optimal rate $|\widehat{\boldsymbol{\beta}}^{SR\text{-}Lasso} - \boldsymbol{\beta}^*|_p \lesssim \sigma^* s^{*1/p} \sqrt{(1/n) \log(ed/s^*)}$ adaptively over $\sigma^* \in \mathbb{R}_+$. However, the square-root Lasso is not robust to even one outlier in the dataset. Furthermore, this optimal rate for the square-root Lasso has only been proved for sub-Gaussian noise ζ whereas in Theorem 4.4, we allow for any distribution of ζ with finite fourth moment. The MOM-Lasso [15] achieves the optimal rate $|\widehat{\boldsymbol{\beta}}^{MOM-Lasso} - \boldsymbol{\beta}^*|_p \lesssim \sigma^* s^{*1/p} \sqrt{(1/n) \log(ed/s^*)}$, but needs the knowledge of σ^*. Therefore, this bound can uniformly hold only on a class of the form $\mathcal{P}_{[C_1 \sigma^*, C_2 \sigma^*]}$ for some fixed $0 < C_1 \leq C_2$.

To our knowledge, the estimator $\widehat{\sigma}$ is the first estimator of σ^* that achieves robustness. Its rate of estimation $\sqrt{(s^*/n) \log(ed/s^*)}$ is slower than the parametric rate $1/\sqrt{n}$ that one would get if β^* was known. Theorem 5 in [7] suggests that this rate $\mathfrak{r}_2$ might be minimax as well: the authors show that, albeit in a Gaussian sequence model, the factor $\sqrt{s^* \log(ed/s^*)}$ arises naturally in the estimation of σ^* by means of any adaptive procedure in a setting where the distribution of the noise ζ is unknown. Even in the case where no outliers are present, we improve on the best known bound on the estimation of σ^*, [6, Corollary 2] which was $\left| (\widehat{\sigma}^{SR\text{-}Lasso})^2 - \sigma^2 \right| \lesssim \sigma^2 \left(\frac{s^* \log(n \vee d \log n)}{n} + \sqrt{\frac{s^* \log(d \vee n)}{n}} + \frac{1}{\sqrt{n}} \right)$.

Remark 4.5. *When $\boldsymbol{\beta}^*$ is not sparse but very close to a sparse vector (i.e. $|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{**}|_1 \lesssim \sigma^* \sqrt{s^* \log(ed/s^*)/n}$ for a sparse vector $\boldsymbol{\beta}^{**} \in \mathcal{F}_{s^*}$, the complexity parameter $r(\rho)$ is in fact unchanged compared to the sparse case and the upper bounds on the rates of estimation $|\widehat{\boldsymbol{\beta}} -$*

$\boldsymbol{\beta}^*|_p \lesssim \sigma_+ s^* 1/p \sqrt{(1/n)\log(ed/s^*)}$ *and* $|\widehat{\sigma} - \sigma^*| \lesssim \sigma_+ \sqrt{(s^*/n)\log(ed/s^*)}$ *still hold, extending Theorem 4.4.*

In practice, it may not be obvious to choose what a good value for σ_+ could be. This means that the (unknown) distribution belongs in fact to the class $\mathcal{P}_{[0,+\infty]} = \bigcup_{\sigma_+ > 0} \mathcal{P}_{[0,\sigma_+]}$. A natural idea is to cut the data into two parts. On the first half of the data, we estimate the variance $\mathrm{Var}[Y]$ by the MOM estimator $\widehat{\sigma}_{K,+}^2 := Q_{1/2,K}\left[Y^2\right] - \left(Q_{1/2,K}\left[Y\right]\right)^2$. On the second half of the data, we use $\widehat{\sigma}_{K,+}$ as the "known" upper bound σ_+ and apply our algorithm as defined in Equation (2.10). The following corollary, proved in Section B.3, gives a bound on the performance of this estimator on the larger class $\mathcal{P}_{[0,+\infty]}$.

Corollary 4.6 (Performance of the estimator with estimated σ_+ on $\mathcal{P}_{[0,+\infty]}$). *Let $s^* > 0$. Then, for every $P_{\mathbf{X},\zeta} \in \mathcal{P}_{[0,+\infty]}$ and $\boldsymbol{\beta}^* \in \mathcal{F}_{s^*}$, there exists a constant $C > 0$ such that, for any $n > Cs^* \log(p/s^*)$ the estimator $(\widehat{\boldsymbol{\beta}}_{K,\mu,\widehat{\sigma}_{K,+}}, \widehat{\sigma}_{K,\mu,\widehat{\sigma}_{K,+}})$ satisfies*

$$\mathbb{P}_{\mathcal{D} \sim P_{\boldsymbol{\beta}^*,P_{\mathbf{X},\zeta}}^{\otimes n}} \left(\sup_{\mathcal{D}' \in \mathcal{D}(\widetilde{c}_3 \mathfrak{r}_{\mathcal{O}})} \left\{ \mathfrak{r}_2^{-1} |\widehat{\sigma}(\mathcal{D}') - \sigma^*| \vee \sup_{p \in [1,2]} \mathfrak{r}_p^{-1} |\widehat{\boldsymbol{\beta}}(\mathcal{D}') - \boldsymbol{\beta}^*|_p \right\} \right.$$
$$\left. \le 4\widetilde{c}_4 \sqrt{1 + SNR}\,\sigma^* \right) \ge 1 - 4\left(\frac{s^*}{ed}\right)^{\widetilde{c}_5 s^*} - 2\left(\frac{s^*}{ed}\right)^{\widetilde{c}_6 s^*},$$

where $\widetilde{c}_6$ is a universal constant and SNR denotes the signal-to-noise ratio, defined by $SNR := \mathrm{Var}[\mathbf{X}^\top \boldsymbol{\beta}^*]/\sigma^{*2} = \boldsymbol{\beta}^{*\top} \mathrm{Var}[X]\boldsymbol{\beta}^*/\sigma^{*2}$.

This corollary ensures that, with high probability, the estimator $(\widehat{\boldsymbol{\beta}}_{K,\mu,\widehat{\sigma}_{K,+}}, \widehat{\sigma}_{K,\mu,\widehat{\sigma}_{K,+}})$ achieves the rates of estimation $|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_p \lesssim \sqrt{1 + SNR}\,\sigma^* s^{*1/p} \sqrt{(1/n)\log(ed/s^*)}$ and $|\widehat{\sigma} - \sigma^*| \lesssim \sqrt{1 + SNR}\,\sigma^* \sqrt{(s^*/n)\log(ed/s^*)}$. The factor $\sqrt{1 + SNR}$ describes how the estimation rates of $\boldsymbol{\beta}^*$ and σ^* are degraded as a function of the signal-to-noise ratio. Indeed, when the noise level is of the same order or higher than the standard deviation of $f^*(\mathbf{X})$, the rates are optimal. On the contrary, when the noise level is very small ($SNR \ll 1$), the rates of estimation are dominated by $\sqrt{\mathrm{Var}\left[\mathbf{X}^\top \beta\right]}\mathfrak{r}_p$.

4.2 Adaptation to the unknown sparsity

We now provide an adaptive to s version of Theorem B.1 by introducing an estimator $(\widetilde{\beta}, \widetilde{\sigma}, \widetilde{s})$ that simultaneously estimates the vector of coefficients, the noise standard deviation and the sparsity level. This procedure is inspired by [8, Section 4] that proposes a general Lepski-type method for constructing an adaptive to s estimator from a sequence of estimators that attains the same rate for each value of s. This method is different from the one proposed in [15] for making the MOM-LASSO estimator adaptive to the sparsity level s, which seems difficult to adapt for the case of unknown noise level.

The main idea of this procedure is to compute different estimators for several possible sparsity levels. Starting from a sparsity of 2, we try different estimators by increasing each time the

sparsity by a factor of 2 unless the difference between an estimator and the next one is too small. We choose this stopping value as the estimated sparsity level, and it gives directly an estimated number of blocks to use, since there exists an optimal number of blocks for each sparsity level. More precisely, given a sparsity estimator $\widetilde{s}$, we take $\widetilde{K} = \lceil \widetilde{c}_2 \widetilde{s} \log(ed/\widetilde{s}) \rceil$.

Given a known upper bound $s_+ \leq d$ on the sparsity, we define the sequence of MOM$-K$ estimators $(\widehat{\boldsymbol{\beta}}_{(s),\sigma_+}, \widehat{\sigma}_{(s),\sigma_+})_{s=1,\dots,s_+}$ by $\widehat{\boldsymbol{\beta}}_{(s)} := \widehat{\boldsymbol{\beta}}_{K_s,\mu_s,\sigma_+}$, $\widehat{\sigma}_{(s),\sigma_+} := \widehat{\sigma}_{K_s,\mu_s,\sigma_+}$ and

$$K_s := \left\lceil \widetilde{c}_2 s \log\left(\frac{ed}{s}\right) \right\rceil, \quad \mu_s := \widetilde{c}_\mu \sqrt{\frac{1}{n} \log\left(\frac{ed}{s}\right)}. \tag{4.1}$$

The adaptive procedure yields an estimator of the form $\widetilde{s} = 2^{\widetilde{m}}$ for some integer $\widetilde{m} \in \{1, \dots, \lceil \log_2(s_+) \rceil + 1\}$, from which we get the simultaneous adaptive (to s and σ^*) MOM estimator $(\widetilde{\boldsymbol{\beta}}_{\sigma_+}, \widetilde{\sigma}_{\sigma_+}, \widetilde{s}_{\sigma_+}) = (\widehat{\boldsymbol{\beta}}_{(\widetilde{s}),\sigma_+}, \widehat{\sigma}_{(\widetilde{s}),\sigma_+}, \widetilde{s}_{\sigma_+})$.

Algorithm for adaptation to sparsity. The steps of the adaptive procedure are as follows.

- Set $M := \lceil \log_2(s_+) \rceil$.

- For every $m \in \{1, \dots, M+1\}$, compute $(\widehat{\boldsymbol{\beta}}_{(2^m),\sigma_+}, \widehat{\sigma}_{(2^m)}, \sigma_+) = \left(\widehat{\boldsymbol{\beta}}_{K_{2^m},\mu_{2^m},\sigma_+}, \widehat{\sigma}_{K_{2^m},\mu_{2^m},\sigma_+} \right)$, with K_{2^m} and μ_{2^m} as defined in Equation (4.1).

- For $u \in \{1, \dots, 2s_+\}$, let $\mathfrak{r}_p(u) = u^{1/p} \sqrt{(1/n) \log(ed/u)}$ and

$$\mathcal{M} := \Big\{ m \in \{1, \dots, M\} : \text{for all } k \geq m, \ |\widehat{\boldsymbol{\beta}}_{(2^{k-1})} - \widehat{\boldsymbol{\beta}}_{(2^k)}|_1 \leq C_1 \widehat{\sigma}_{(2^{M+1})} \mathfrak{r}_1(2^k),$$

$$|\widehat{\boldsymbol{\beta}}_{(2^{k-1})} - \widehat{\boldsymbol{\beta}}_{(2^k)}|_2 \leq C_2 \widehat{\sigma}_{(2^{M+1})} \mathfrak{r}_2(2^k) \text{ and } |\widehat{\sigma}_{(2^{k-1})} - \widehat{\sigma}_{(2^k)}| \leq C_3 \widehat{\sigma}_{(2^{M+1})} \mathfrak{r}_2(2^k) \Big\}.$$

- Set $\widetilde{m} := \min \mathcal{M}$, with the convention that $\widetilde{m} := M + 1$ if $\mathcal{M} = \emptyset$.

- Define $\widetilde{s}_{\sigma_+} := 2^{\widetilde{m}}$ and $(\widetilde{\boldsymbol{\beta}}_{\sigma_+}, \widetilde{\sigma}_{\sigma_+}) := (\widehat{\boldsymbol{\beta}}_{(\widetilde{s}),\sigma_+}, \widehat{\sigma}_{(\widetilde{s}),\sigma_+})$.

The following theorem is proved in Section C.2 and gives uniform bounds for the performance of the aggregated estimator $(\widetilde{\boldsymbol{\beta}}_{\sigma_+}, \widetilde{\sigma}_{\sigma_+}, \widetilde{s}_{\sigma_+})$.

Theorem 4.7. *Let $\theta_0, \theta_1, c_0, L, \kappa_+ > 1$. Let $s_+ \in \{1, \dots, d/(2e)\}$ and assume that $\mathfrak{r}_2(2s^+) < 1$. Then the aggregated estimator $(\widetilde{\boldsymbol{\beta}}_{\sigma_+}, \widetilde{\sigma}_{\sigma_+}, \widetilde{s}_{\sigma_+})$ satisfies*

$$\inf_{\substack{s^*=1,\dots,s_+ \\ P_{\mathbf{X},\zeta} \in \mathcal{P}_{[0,\sigma_+]} \\ \beta^* \in \mathcal{F}_{s^*}}} \inf P_{\beta^*,P_{\mathbf{X},\zeta}}^{\otimes n} \left(\sup_{\mathcal{D}' \in \mathcal{D}(\widetilde{c}_3 \mathfrak{r}_{\mathcal{O}})} \left\{ \mathfrak{r}_2(s^*)^{-1} |\widehat{\sigma}(\mathcal{D}') - \sigma^*| \vee \sup_{p \in [1,2]} \mathfrak{r}_p(s^*)^{-1} |\widehat{\boldsymbol{\beta}}(\mathcal{D}') - \beta^*|_p \right\} \right.$$

$$\left. \leq 4\widetilde{c}_4 \sigma_+ \right) \geq 1 - 4(\log_2(s_+) + 1)^2 \left(\frac{2s_+}{ed} \right)^{2\widetilde{c}_5 s_+}$$

and for all $\mathcal{D}' \in \mathcal{D}(\widetilde{c}_3 \mathfrak{r}_{\mathcal{O}})$, $\widetilde{s}_{\sigma_+}(\mathcal{D}') \leq s^*$ on the same event.

This theorem guarantees that for every $s^* \in \{1, \ldots, s_+\}$, both estimators $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\sigma}$ converge to their true values at the rate $\sigma_+ s^{*1/p}\sqrt{(1/n)\log(ed/s^*)}$ as if the true sparsity level s^* was known. However, the probability bounds are slightly deteriorated due to the knowledge of an upper bound s_+ only.

Note that the estimator presented above uses the knowledge of the upper bound on the standard deviation σ_+. If σ_+ is not available, the estimator presented in Corollary 4.6 can be aggregated in the same way. It will satisfy the same bounds up to some small degradation in the probability of the event.

5 From the choice of the functional R_c to empirical process bounds

Our construction in Section 2.4 produces a family of MOM estimators

$$(\widehat{f}_{K,\mu\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+}) = \arg\min_{f \in \mathcal{F}, \, \sigma \leq \sigma_+} \max_{g \in \mathcal{F}, \, \chi \leq \sigma_+} \left\{ MOM_K\Big(R_c(\ell_g, \chi, \ell_f, \sigma)\Big) + \mu\big(\|f\| - \|g\|\big) \right\},$$

where R_c is a carefully chosen functional in (2.6). As mentioned in Section 2.3, this extends the scope of the MOM estimator in [15]

$$\widehat{f}_{K,\lambda} = \arg\min_{f \in \mathcal{F}} \max_{g \in \mathcal{F}} \left\{ MOM_K\big(R(\ell_g, \ell_f)\big) + \lambda\big(\|f\| - \|g\|\big) \right\},$$

where $R(\ell_g, \ell_f) = \ell_f - \ell_g$, which was constructed in the setting of known σ^*. In this section we discuss in detail the role of the functional R_c. In Section 5.1 we motivate our choice by showing that, in the sparse linear setting, we recover a robust version of the square-root LASSO. In Section 5.2 we lay down our proving strategy and highlight the contribution of R_c in recovering convergence rates and excess risk bounds in terms of complexity parameters. In Section 5.3 and Section 5.4 we reproduce the main results on complexity parameters in the sub-Gaussian and sparse linear case respectively.

5.1 Adaptivity to σ^*: choice of the functional R_c and corresponding conditions

Since we implement the same proving strategy as in [15], we introduce the following properties as natural assumptions that the functional R_c should satisfy.

P1. Anti-symmetry. For all $f, g \in \mathcal{F}$, $\chi, \sigma \in R_+$ and $(\mathbf{x}, y) \in \mathcal{X} \times \mathbb{R}$, we have

$$R_c\big(\ell_g(\mathbf{x}, y), \chi, \ell_f(\mathbf{x}, y), \sigma\big) = -R_c\big(\ell_f(\mathbf{x}, y), \sigma, \ell_g(\mathbf{x}, y), \chi\big),$$

in short, we write $R_c(\ell_g, \chi, \ell_f, \sigma) = -R_c(\ell_f, \sigma, \ell_g, \chi)$.

The latter is a crucial requirement for the whole convex-concave procedure to work, as we show in the next section. It is automatically satisfied when σ^* is known, since $R(\ell_g, \ell_f) = \ell_f - \ell_g = -R(\ell_f, \ell_g)$.

P2. Concavity in χ, given $f = g$. For any fixed $f = g \in \mathcal{F}$, $\sigma \in \mathbb{R}_+$ and $(\mathbf{x}, y) \in \mathcal{X} \times \mathbb{R}$, the function $\chi \mapsto R_c(\ell_f(\mathbf{x}, y), \chi, \ell_f(\mathbf{x}, y), \sigma)$ is concave and has a unique maximum for $\chi \in \mathbb{R}_+$.

This is an additional requirement that has no counterpart when σ^* is known. In fact, for $f = g$, we have $R(\ell_g, \ell_f) = \ell_f - \ell_g \equiv 0$.

P3. Maximization over g. For any fixed $f \in \mathcal{F}$ and $\chi, \sigma \in \mathbb{R}_+$, the problems of maximizing the functionals

$$g \mapsto MOM_K\Big(R_c(\ell_g, \chi, \ell_f, \sigma)\Big), \quad g \mapsto MOM_K\Big(\ell_f - \ell_g\Big),$$

over $g \in \mathcal{F}$ are equivalent.

The latter condition requires that our functional $R_c(\ell_g, \chi, \ell_f, \sigma)$ behaves similarly to $R(\ell_g, \ell_f) = \ell_f - \ell_g$ when viewed as a functional on $g \in \mathcal{F}$.

As a consequence of anti-symmetry, the following properties are equivalent to P1–P3 above:

P1'. Anti-symmetry. For all $f, g \in \mathcal{F}$ and $\chi, \sigma \in \mathbb{R}_+$, we have $R_c(\ell_g, \chi, \ell_f, \sigma) = -R_c(\ell_f, \sigma, \ell_g, \chi)$.

P2'. Convexity in σ, given $f = g$. For any fixed $f = g \in \mathcal{F}$, $\chi \in \mathbb{R}_+$ and $(\mathbf{x}, y) \in \mathcal{X} \times \mathbb{R}$, the function $\sigma \mapsto R_c(\ell_f(\mathbf{x}, y), \chi, \ell_f(\mathbf{x}, y), \sigma)$ is convex and has a unique minimum for $\sigma \in \mathbb{R}_+$.

P3'. Minimization over f. For any fixed $g \in \mathcal{F}$ and $\chi, \sigma \in \mathbb{R}_+$, the problems of minimizing the functionals

$$f \mapsto MOM_K\Big(R_c(\ell_g, \chi, \ell_f, \sigma)\Big), \quad f \mapsto MOM_K\Big(\ell_f - \ell_g\Big),$$

over $f \in \mathcal{F}$ are equivalent.

Consider the sparse linear setting, where we want to recover oracle solutions

$$\boldsymbol{\beta}^* \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^d} \mathbb{E}\Big[(Y - \mathbf{X}^\top \boldsymbol{\beta})^2\Big], \quad \sigma^* = \mathbb{E}\Big[(Y - \mathbf{X}^\top \boldsymbol{\beta}^*)^2\Big]^{\frac{1}{2}}.$$

Any linear function $f : \mathcal{X} \to \mathbb{R}$ can be identified with some $\boldsymbol{\beta}_f \in \mathbb{R}^d$ such that $f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}_f$ and $\ell_f(\mathbf{x}, y) = \ell_{\boldsymbol{\beta}_f}(\mathbf{x}, y) = (y - \mathbf{x}^\top \boldsymbol{\beta}_f)^2$. The MOM method in [15] yields a robust version of the LASSO estimator

$$\widehat{\boldsymbol{\beta}}^L \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})^2 + \lambda |\boldsymbol{\beta}|_1 \right\},$$

which has been shown to be minimax optimal in [4, 2, 3], but its optimal tuning parameter λ is proportional to σ^*. An adaptive version of the LASSO is the square-root LASSO introduced in [5], which is also minimax optimal, as shown in [8]. This adaptive method uses

$$\widehat{\boldsymbol{\beta}}^{SR\text{-}Lasso} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\arg\min} \left\{ \left(\frac{1}{n}\sum_{i=1}^n (Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})^2 \right)^{\frac{1}{2}} + \mu|\boldsymbol{\beta}|_1 \right\},$$

and its optimal tuning parameter μ does not require the knowledge of σ^*. The key insight behind the square-root LASSO, see for example Section 5 in [11], is that when $\boldsymbol{\beta}$ is close to $\boldsymbol{\beta}^*$ one can approximate σ^{*2} by $\mathbb{E}[(Y - \mathbf{X}^\top\boldsymbol{\beta})^2]$. Thus, with $\lambda = \sigma^*\mu$, one finds

$$\frac{\mathbb{E}[(Y - \mathbf{X}^\top\boldsymbol{\beta})^2]}{\sigma^*} + \frac{\lambda}{\sigma^*}|\boldsymbol{\beta}|_1 \simeq \mathbb{E}[(Y - \mathbf{X}^\top\boldsymbol{\beta})^2]^{\frac{1}{2}} + \mu|\boldsymbol{\beta}|_1,$$

and the minimization problem is independent of σ^*.

In view of the discussion above, a candidate natural implementation of the robust square-root LASSO is given by

$$\widetilde{R}_c(\ell_g, \chi, \ell_f, \sigma) = \frac{\ell_f}{\sigma} + \sigma - \frac{\ell_g}{\chi} - \chi,$$

$$= (\sigma - \chi)\left(1 - \frac{\ell_f}{\sigma\chi}\right) + \frac{\ell_f - \ell_g}{\chi},$$

$$\widetilde{T}_{K,\mu}(g, \chi, f, \sigma) = MOM_K\left(\widetilde{R}_c(\ell_g, \chi, \ell_f, \sigma)\right) + \mu\big(\|f\| - \|g\|\big),$$

since $\widetilde{R}_c$ implements the idea that, in the linear setting, dividing ℓ_f by σ should lead to the square-root of ℓ_f. Also, this choice satisfies the properties P1–P3:

- Anti-symmetry holds by construction.

- When $f = g$, replace $\ell_f(\mathbf{x}, y) = \ell_g(\mathbf{x}, y)$ by some positive real number $a^2 > 0$, then the function

$$\chi \mapsto \widetilde{R}_c(a^2, \chi, a^2, \sigma) = (\sigma - \chi)\left(1 - \frac{a^2}{\sigma\chi}\right),$$

  is concave and has a unique maximum for $\chi \in \mathbb{R}_+$.

- By definition, maximizing $g \mapsto MOM_K(\ell_f - \ell_g)$ with fixed $f \in \mathcal{F}$ is equivalent to maximizing the empirical average

$$g \mapsto -\frac{1}{|B_k|}\sum_{i\in B_k} \ell_g(\mathbf{X}_i, Y_i),$$

  where the block B_k realizes the median. For the same reason, maximising $g \mapsto MOM_K(\widetilde{R}_c(\ell_g, \chi, \ell_f, \sigma))$ is equivalent to maximizing the empirical average

$$g \mapsto \frac{1}{|B_k|}\sum_{i\in B_k} \left(\frac{\ell_f(\mathbf{X}_i, Y_i)}{\sigma} + \sigma - \frac{\ell_g(\mathbf{X}_i, Y_i)}{\chi} - \chi\right),$$

  where the block B_k realizes the median. Since the quantities f, σ, χ are fixed, this coincides with the above.

However, this choice comes with a drawback. The proof of our main result is based on the argument proposed in [15], which requires sharp bounds for the functional $\widetilde{T}_{K,\mu}(\ell_g, \chi, \ell_{f^*}, \sigma^*)$ over the possible values of (g, χ). This is done by carefully slicing the domain and assessing the contribution of each term appearing in $\widetilde{T}_{K,\mu}$. In particular, one finds a slice in which $\chi < \sigma^* - c_\alpha r(2\rho_K)$ and the leading term of $\widetilde{T}_{K,\mu}$ is of the form $2\varepsilon/\chi$, with some small fixed $\varepsilon > 0$. Since $2\varepsilon/\chi \to +\infty$, for $\chi \to 0$, we cannot control the supremum of $\widetilde{T}_{K,\mu}(\ell_g, \chi, \ell_{f^*}, \sigma^*)$ over this slice. The only way around it would be to assume from the start that $\sigma^* > \sigma_-$, for some known lower bound $\sigma_- > 0$, but this would be a stronger assumption than the upper bound σ_+ we use in (2.3). This issue is caused by the fact that the two terms of $\widetilde{R}_c(\ell_g, \chi, \ell_f, \sigma)$ are

$$(\ell_g, \chi, \ell_f, \sigma) \mapsto (\sigma - \chi)\left(1 - \frac{\ell_f}{\sigma\chi}\right), \quad (\ell_g, \chi, \ell_f, \sigma) \mapsto \frac{\ell_f - \ell_g}{\chi},$$

and the second one cannot be controlled if $\chi \to 0$. A way to introduce stability is to replace the denominator χ by the average $(\sigma + \chi)/2$, which is always bounded away from zero when σ is fixed. However, making this substitution alone breaks the anti-symmetry of the functional, so we have to take care of both terms simultaneously. To this end, we use

$$R_c(\ell_g, \chi, \ell_f, \sigma) = (\sigma - \chi)\left(1 - 2\frac{\ell_f + \ell_g}{(\sigma + \chi)^2}\right) + 2c\frac{\ell_f - \ell_g}{\sigma + \chi},$$

$$T_{K,\mu}(g, \chi, f, \sigma) = MOM_K\Big(R_c(\ell_g, \chi, \ell_f, \sigma)\Big) + \mu\big(\|f\| - \|g\|\big),$$

for all $(f, g) \in \mathcal{F} \times \mathcal{F}$ and $(\sigma, \chi) \in (0, \sigma_+] \times (0, \sigma_+]$, which guarantees that R_c satisfies properties P1–P3. In fact, anti-symmetry holds for both terms

$$(\ell_g, \chi, \ell_f, \sigma) \mapsto (\sigma - \chi)\left(1 - 2\frac{\ell_f + \ell_g}{(\sigma + \chi)^2}\right), \quad (\ell_g, \chi, \ell_f, \sigma) \mapsto 2c\frac{\ell_f - \ell_g}{\sigma + \chi},$$

separately. Also, for any fixed $f = g \in \mathcal{F}$, $\sigma \in \mathbb{R}_+$, we have

$$\chi \mapsto R_c(\ell_f, \chi, \ell_f, \sigma) = (\sigma - \chi)\left(1 - \frac{4\ell_f}{(\sigma + \chi)^2}\right),$$

which satisfies property P2. Finally, for any fixed $f \in \mathcal{F}$, $\sigma, \chi \in \mathbb{R}_+$, we can rewrite

$$g \mapsto MOM_K\left(R_c(\ell_g, \chi, \ell_f, \sigma)\right)$$
$$= MOM_K\left((\sigma - \chi) + \frac{2\ell_f}{\sigma + \chi}\left(c - \frac{\sigma - \chi}{\sigma + \chi}\right) - \frac{2\ell_g}{\sigma + \chi}\left(c + \frac{\sigma - \chi}{\sigma + \chi}\right)\right).$$

Since the quantity $c + (\sigma - \chi)/(\sigma + \chi)$ belongs to the interval $[c - 1, c + 1]$ and $c > 1$, property P3 is satisfied.

5.2 From R_c to convergence rates and excess risk bounds

The choice of R_c induces a penalized functional $T_{K,\mu}$ which characterizes the $MOM-K$ estimator

$$(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+}) = \underset{f \in \mathcal{F}, \ \sigma \in I_+}{\arg\min} \ \underset{g \in \mathcal{F}, \ \chi \in I_+}{\max} T_{K,\mu}(g, \chi, f, \sigma), \quad I_+ = (0, \sigma_+].$$

20

Our goal is to guarantee that, with as high probability as possible, the function estimator $\widehat{f}_{K,\mu,\sigma_+}$ recovers f^* with as small as possible rates in $\|\cdot\|$ and $\|\cdot\|_{2,\mathbf{X}}$, and that the standard deviation estimator $\widehat{\sigma}_{K,\mu,\sigma_+}$ recovers σ^* with as small as possible rates in absolute value. With the same high probability, we also want that the excess risk $\mathrm{Risk}(\widehat{f}_{K,\mu}) - \mathrm{Risk}(f^*)$ is as small as possible.

Starting with the convergence rates, they can be obtained by showing that the estimator $(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$ belongs to a bounded ball of the form

$$\mathbb{B}^*(2\rho) := \big\{(f,\sigma) \in \mathcal{F} \times I_+ : \|f - f^*\| \leq 2\rho,\ \|f - f^*\|_{2,\mathbf{X}} \leq r(2\rho),\ |\sigma - \sigma^*| \leq c_\alpha r(2\rho)\big\},$$

with appropriate radius ρ and complexity measure $r(2\rho)$. In the proof of Theorem 3.3, we show that this can be achieved with $\rho = \rho_K$ and any $r(\rho) \geq \max\{r_P(\rho, \gamma_P),\ r_M(\rho, \gamma_M)\}$, which only requires the complexities r_P, r_M. The convergence rates $2\rho_K, r(2\rho_K)$ are perfectly in line with those obtained with the MOM tournaments procedure in [18] and the robust MOM method in [15]. The key idea behind this result is to essentially show that the evaluation of $T_{K,\mu}$ at the point $(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+}, f^*, \sigma^*)$ is too big for $(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$ to be outside of the bounded ball $\mathbb{B}^*(2\rho_K)$. Precisely, we show that, for some $B_{1,1} > 0$,

$$T_{K,\mu}(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+}, f^*, \sigma^*) \geq -B_{1,1}, \qquad \sup_{(g,\chi) \notin \mathbb{B}^*(2\rho_K, r(2\rho_K))} T_{K,\mu}(g, \chi, f^*, \sigma^*) < -B_{1,1},$$

which guarantees that $(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+}, f^*, \sigma^*) \in \mathbb{B}^*(2\rho_K)$. The problem of finding a suitable bound $B_{1,1}$ is solved as follows.

- The problem is equivalent to $-T_{K,\mu}(\widehat{f}_{K,\mu}, \widehat{\sigma}_{K,\mu}, f^*, \sigma^*) \leq B_{1,1}$.

- By the anti-symmetry property P1 of R_c, together with the quantile properties in Lemma D.2, we have $-T_{K,\mu}(f, \sigma, f^*, \sigma^*) \leq T_{K,\mu}(f^*, \sigma^*, f, \sigma)$ and it is sufficient to find $T_{K,\mu}(f^*, \sigma^*, \widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+}) \leq B_{1,1}$.

- The evaluation at (f^*, σ^*) can be bounded with the supremum over the domain, that is, we look for $\sup_{(g,\chi) \in \mathcal{F} \times I_+} T_{K,\mu}(g, \chi, \widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+}) \leq B_{1,1}$.

- By definition, the $\mathrm{MOM}-K$ estimator $(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$ minimizes the latter supremum if we allow for other pairs (f, σ). In particular, with $(f, \sigma) = (f^*, \sigma^*)$, it is enough to find $\sup_{(g,\chi) \in \mathcal{F} \times I_+} T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq B_{1,1}$.

- Finally, in Lemma A.11 we show that the supremum is achieved on the bounded ball $\mathbb{B}^*(\rho_K)$, that is, the solution to the problem is the sharpest bound such that

$$\sup_{(g,\chi) \in \mathbb{B}^*(\rho_K)} T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq B_{1,1}.$$

The argument we just sketched can be found in the proof of the main result in [15], it is a clever exploitation of the convex-concave formulation of the problem. One key element of the argument is that the computations only require lower bounds on the quantiles of the

21

quadratic and multiplier empirical processes, which in turn can be obtained by means of the complexities r_P and r_M alone. These facts has been established in [14, 17] and we provide them in Lemma D.5, Lemma D.6.

The fact that the estimator $(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$ belongs to the ball $\mathbb{B}^*(2\rho_K)$ is instrumental in obtaining excess risk bounds. First, one writes

$$\mathrm{Risk}(\widehat{f}_{K,\mu,\sigma_+}) - \mathrm{Risk}(f^*) = \|\widehat{f}_{K,\mu,\sigma_+} - f^*\|_{2,\mathbf{X}}^2 + \mathbb{E}[-2\zeta(\widehat{f}_{K,\mu,\sigma_+} - f^*)(\mathbf{X})],$$

and then bounds $\|\widehat{f}_{K,\mu,\sigma_+} - f^*\|_{2,\mathbf{X}}^2 \leq r^2(2\rho_K)$. By applying a quantile inequality, see Lemma D.7, and adding the quadratic term $(\widehat{f}_{K,\mu,\sigma_+} - f^*)^2$, the expectation term becomes

$$\mathbb{E}[-2\zeta(\widehat{f}_{K,\mu,\sigma_+} - f^*)(\mathbf{X})] \leq Q_{1/4,K}\left[-2\zeta(\widehat{f}_{K,\mu,\sigma_+} - f^*)\right] + \alpha_M^2$$
$$\leq Q_{1/4,K}\left[\ell_{\widehat{f}_{K,\mu,\sigma_+}} - \ell_{f^*}\right] + \alpha_M^2,$$

since $\ell_f - \ell_{f^*} = (f - f^*)^2 - 2\zeta(f - f^*)$. Since the $1/4$–quantile is always smaller than the $1/2$–quantile, which is the median, some algebraic manipulations allow to rewrite the difference $\ell_{\widehat{f}_{K,\mu,\sigma_+}} - \ell_{f^*}$ in terms of our functional $R_c(\ell_{f^*}, \sigma^*, \ell_{\widehat{f}_{K,\mu,\sigma_+}}, \widehat{\sigma}_{K,\mu,\sigma_+})$ and to recover the penalized $T_{K,\mu}(f^*, \sigma^*, \widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$. Specifically, in Lemma D.9 we find

$$\mathbb{E}[-2\zeta(\widehat{f}_{K,\mu,\sigma_+} - f^*)(\mathbf{X})] \leq \frac{\widehat{\sigma}_{K,\mu,\sigma_+} + \sigma^*}{2c} T_{K,\mu}(f^*, \sigma^*, \widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+}) + \text{remainder},$$
$$\leq \frac{\widehat{\sigma}_{K,\mu,\sigma_+} + \sigma^*}{2c} B_{1,1} + \text{remainder},$$

where $B_{1,1}$ is the upper bound we found when dealing with the convergence rates. It is easy to show that $B_{1,1} \lesssim r^2(2\rho_K)$, the majority of the work is spent on bounding the remainder terms. In the same lemma, we show that they are: the quantity $\mu\rho_K \lesssim r^2(\rho_K)$ where $\mu \simeq r^2(\rho_K)/\rho_K$ is the penalization parameter, the quantity $\alpha_M^2 \lesssim r^2(2\rho_K)$ related to the quantiles of the multiplier process, the mixed terms

- $|\widehat{\sigma}_{K,\mu,\sigma_+} - \sigma^*| \cdot Q_{15/16,K}\left[(\widehat{f}_{K,\mu,\sigma_+} - f^*)^2\right],$

- $|\widehat{\sigma}_{K,\mu,\sigma_+} - \sigma^*| \cdot Q_{15/16,K}\left[-2\zeta(\widehat{f}_{K,\mu,\sigma_+} - f^*)\right],$

involving the quantiles of the quadratic and multiplier processes. The standard deviation estimator satisfies $|\widehat{\sigma}_{K,\mu,\sigma_+} - \sigma^*| \lesssim r(2\rho_K)$. In Lemma D.7 we show that $Q_{15/16,K}[-2\zeta(\widehat{f}_{K,\mu} - f^*)] \leq \mathbb{E}[-2\zeta(\widehat{f}_{K,\mu,\sigma_+} - f^*)] + \alpha_M^2$, so that the Cauchy-Schwarz inequality is sufficient for $\mathbb{E}[-2\zeta(\widehat{f}_{K,\mu,\sigma_+} - f^*)] \leq 4\sigma^*\|\widehat{f}_{K,\mu,\sigma_+} - f^*\|_{2,\mathbf{X}} \lesssim r(2\rho_K)$. Finally, in Lemma D.8 we find $Q_{15/16,K}[(\widehat{f}_{K,\mu,\sigma_+} - f^*)^2] \leq r^2(2\rho_K) + \alpha_Q^2 \lesssim r^2(2\rho_K) \vee r_Q^2(2\rho_K, \gamma_Q)$.

5.3 Complexity parameters in the sub-Gaussian setting

We follow the construction presented in [13]. Let $G = (G(f) : f \in L^2(\mathbb{P}_\mathbf{X}))$ the Gaussian process indexed on $L^2(\mathbb{P}_\mathbf{X})$ and such that $\mathbb{E}[G(f)] = 0$ and $\mathbb{E}[G(f)G(h)] = \mathbb{E}[f(\mathbf{X})h(\mathbf{X})]$. For

any $\mathcal{F}' \subseteq \mathcal{F}$, we set

$$\mathbb{E}\left[\|G\|_{\mathcal{F}'}\right] := \sup\left\{\mathbb{E}\left[\sup_{h \in \mathcal{H}} G(h)\right] : \mathcal{H} \subseteq \mathcal{F}' \text{ is finite}\right\}.$$

As an example, if $\mathcal{F}' = \{\mathbf{x} \mapsto \mathbf{x}^\top \boldsymbol{\beta} : \boldsymbol{\beta} \in T \subset \mathbb{R}^d\}$ and $\mathbf{X}$ is a random vector in $\mathbb{R}^d$ with covariance matrix Σ, then $G \sim \mathcal{N}(0, \Sigma)$ and

$$\mathbb{E}\left[\|G\|_{\mathcal{F}'}\right] = \mathbb{E}\left[\sup_{\boldsymbol{\beta} \in T} G^\top \boldsymbol{\beta}\right].$$

Sub-Gaussian class. We say that $\mathcal{F}$ is sub-Gaussian if there exists a constant L such that, for all $f, h \in \mathcal{F}$ and $p \geq 2$, one has $\|f - h\|_{p,\mathbf{X}} \leq L\sqrt{p}\|f - h\|_{2,\mathbf{X}}$.

Gaussian complexities. For any $r \geq 0$, set $\mathbb{B}_2(r) = \{f \in L^2(\mathbb{P}_{\mathbf{X}}) : \|f\|_{2,\mathbf{X}} \leq r\}$ and $\mathcal{F} - \mathcal{F} = \{f - h : f, h \in \mathcal{F}\}$. For any $\gamma, \gamma' > 0$, take

$$\begin{aligned}
s_n^*(\gamma) &:= \inf\{r > 0 : \mathbb{E}\left[\|G\|_{\mathbb{B}_2(r) \cap (\mathcal{F}-\mathcal{F})}\right] \leq \gamma r^2 \sqrt{n}\}, \\
r_n^*(\gamma') &:= \inf\{r > 0 : \mathbb{E}\left[\|G\|_{\mathbb{B}_2(r) \cap (\mathcal{F}-\mathcal{F})}\right] \leq \gamma' r \sqrt{n}\}.
\end{aligned} \tag{5.1}$$

The goal of this section is to provide the following bounds.

Lemma 5.1. *Under the sub-Gaussian assumption, there exist absolute constants c_2, c_3 such that the complexity parameters r_P, r_Q, r_M defined in (3.1) satisfy*

$$r_P(\rho, \gamma_P) \leq r_n^*\left(\frac{\gamma_P}{c_2 L^2}\right), \quad r_Q(\rho, \gamma_Q) \leq r_n^*\left(\frac{\gamma_Q}{c_2 L^2}\right), \quad r_M(\rho, \gamma_M) \leq s_n^*\left(\frac{\gamma_M}{c_3 L\mathfrak{m}^*}\right). \tag{5.2}$$

In particular, any continuous non-decreasing function $\rho \mapsto r(\rho)$ with

$$r(\rho) \geq \max\left\{r_n^*\left(\frac{\gamma_P}{c_2 L^2}\right), s_n^*\left(\frac{\gamma_M}{c_3 L\mathfrak{m}^*}\right)\right\},$$

is a valid choice in (3.2).

Proof of Lemma 5.1. We invoke Lemma 5.2, Lemma 5.3 and Lemma 5.4 below. They are all based on a symmetrization argument in [19], which controls the processes

$$\sup_{f \in \mathcal{F}: \|f - f^*\|_{2,\mathbf{X}} \leq r} \left|\frac{1}{n}\sum_{i=1}^n (f - f^*)(\mathbf{X}_i) - \mathbb{E}[(f - f^*)(\mathbf{X})]\right|,$$

$$\sup_{f \in \mathcal{F}: \|f - f^*\|_{2,\mathbf{X}} \leq r} \left|\frac{1}{n}\sum_{i=1}^n (f - f^*)^2(\mathbf{X}_i) - \mathbb{E}[(f - f^*)^2(\mathbf{X})]\right|,$$

$$\sup_{f \in \mathcal{F}: \|f - f^*\|_{2,\mathbf{X}} \leq r} \left|\frac{1}{n}\sum_{i=1}^n \zeta_i(f - f^*)(\mathbf{X}_i) - \mathbb{E}[\zeta(f - f^*)(\mathbf{X})]\right|,$$

23

in terms of the processes

$$\sup_{f\in\mathcal{F}:\|f-f^*\|_{2,\mathbf{X}}\leq r}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i(f-f^*)(\mathbf{X}_i)\right|,$$

$$\sup_{f\in\mathcal{F}:\|f-f^*\|_{2,\mathbf{X}}\leq r}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i(f-f^*)^2(\mathbf{X}_i)\right|,$$

$$\sup_{f\in\mathcal{F}:\|f-f^*\|_{2,\mathbf{X}}\leq r}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\zeta_i(f-f^*)(\mathbf{X}_i)\right|,$$

with Rademacher variables $(\xi_i)_{i=1,\ldots,n}$. These processes play a role in the definition of the complexities in (3.1).

Lemma 5.2 below shows that, for any $r > r_n^*(\gamma')$,

$$\sup_{f,h\in\mathcal{F}:\|f-h\|_{2,\mathbf{X}}\leq r}\left|\frac{1}{n}\sum_{i=1}^{n}(f-h)(\mathbf{X}_i)-\mathbb{E}[(f-h)(\mathbf{X})]\right|\leq c_2\gamma'Lr,$$

with probability bigger than $1-2\exp(-c_1\gamma'^2 n)$. Choosing $\gamma' = \gamma_P/(c_2L)$ and $h=f^*$ gives, for all $r > r_n^*(\gamma_Q/(c_2L))$,

$$\sup_{f\in\mathcal{F}:\|f-f^*\|_{2,\mathbf{X}}\leq r}\left|\frac{1}{n}\sum_{i=1}^{n}(f-f^*)(\mathbf{X}_i)-\mathbb{E}[(f-f^*)(\mathbf{X})]\right|\leq \gamma_Q r.$$

By definition, the complexity $r_P(\rho,\gamma_P)$ is the smallest level r at which the latter display holds for all functions f in the smaller set $\mathbb{B}(f^*,\rho,r)$. Thus $r_P(\rho,\gamma_P) \leq r_n^*(\gamma_P/(c_2L))$.

Lemma 5.3 below shows that, for any $r > r_n^*(\gamma')$,

$$\sup_{f,h\in\mathcal{F}:\|f-h\|_{2,\mathbf{X}}\leq r}\left|\frac{1}{n}\sum_{i=1}^{n}(f-h)^2(\mathbf{X}_i)-\mathbb{E}[(f-h)^2(\mathbf{X})]\right|\leq c_2\gamma'L^2r^2,$$

with probability bigger than $1-2\exp(-c_1\gamma'^2 n)$. Choosing $\gamma' = \gamma_Q/(c_2L^2)$ and $h=f^*$ gives, for all $r > r_n^*(\gamma_Q/(c_2L^2))$,

$$\sup_{f\in\mathcal{F}:\|f-f^*\|_{2,\mathbf{X}}\leq r}\left|\frac{1}{n}\sum_{i=1}^{n}(f-f^*)^2(\mathbf{X}_i)-\mathbb{E}[(f-f^*)^2(\mathbf{X})]\right|\leq \gamma_Q r^2.$$

By definition, the complexity $r_Q(\rho,\gamma_Q)$ is the smallest level r at which the latter display holds for all functions f in the smaller set $\mathbb{B}(f^*,\rho,r)$. Thus $r_Q(\rho,\gamma_Q) \leq r_n^*(\gamma_Q/(c_2L^2))$.

With $\mathbb{E}[\zeta^4]^{1/4} = \mathfrak{m}^*$, Lemma 5.4 below shows that, for any $r > s_n^*(\gamma)$,

$$\sup_{f,h\in\mathcal{F}:\|f-h\|_{2,\mathbf{X}}\leq r}\left|\frac{1}{n}\sum_{i=1}^{n}\zeta_i(f-h)(\mathbf{X}_i)-\mathbb{E}[\zeta(f-h)(\mathbf{X})]\right|\leq c_3\gamma\mathfrak{m}^*Lr^2,$$

with probability bigger than $1-4\exp(-c_1 n\min\{\gamma^2 r^2,1\})$. Choosing $\gamma = \gamma_M/(c_3L\mathfrak{m}^*)$ and $h=f^*$ gives, for all $r > s_n^*(\gamma_M/(c_3L\mathfrak{m}^*))$,

$$\sup_{f\in\mathcal{F}:\|f-f^*\|_{2,\mathbf{X}}\leq r}\left|\frac{1}{n}\sum_{i=1}^{n}\zeta_i(f-f^*)(\mathbf{X}_i)-\mathbb{E}[\zeta(f-f^*)(\mathbf{X})]\right|\leq \gamma_M r^2.$$

By definition, the complexity $r_M(\rho,\gamma_M)$ is the smallest display r at which the latter display holds for all functions f in the smaller set $\mathbb{B}(f^*,\rho,r)$. Thus $r_M(\rho,\gamma_M) \leq s_n^*(\gamma_M/(c_3L\mathfrak{m}^*))$. $\quad\square$

Lemma 5.2 (Corollary 1.8 in [19]). *There exist absolute constants c_1, c_2 for which the following holds. Let $\mathcal{F}$ be an $L-$sub-Gaussian class, assume that $\mathcal{F} - \mathcal{F}$ is star-shaped around 0. If $\gamma' \in (0,1)$ and $r > r_n^*(\gamma')$, then with probability at least $1 - 2\exp(-c_1\gamma'^2 n)$, we have*

$$\sup_{f,h \in \mathcal{F}: \|f-h\|_{2,\mathbf{X}} \leq r} \left| \frac{1}{n} \sum_{i=1}^{n} (f-h)(\mathbf{X}_i) - \mathbb{E}[(f-h)(\mathbf{X})] \right| \leq c_2 \gamma' L r.$$

Lemma 5.3 (Lemma 2.6 in [13]). *There exist absolute constants c_1, c_2 for which the following holds. Let $\mathcal{F}$ be an $L-$sub-Gaussian class, assume that $\mathcal{F} - \mathcal{F}$ is star-shaped around 0. If $\gamma' \in (0,1)$ and $r > r_n^*(\gamma')$, then with probability at least $1 - 2\exp(-c_1\gamma'^2 n)$, we have*

$$\sup_{f,h \in \mathcal{F}: \|f-h\|_{2,\mathbf{X}} \leq r} \left| \frac{1}{n} \sum_{i=1}^{n} (f-h)^2(\mathbf{X}_i) - \mathbb{E}[(f-h)^2(\mathbf{X})] \right| \leq c_2 \gamma' L^2 r^2.$$

Lemma 5.4 (Corollary of Theorem 2.7 in [13]). *Let $\mathcal{F}$ be an $L-$sub-Gaussian class, assume that $\mathcal{F} - \mathcal{F}$ is star-shaped around 0. Let $\mathbb{E}[|\zeta|^q]^{1/q} = \mathfrak{m}^*$ for some $q > 2$, there exists an absolute constant $c_3(q)$, depending on q only, for which the following holds. For some $\gamma > 0$ and $r > s_n^*(\gamma)$, with probability at least $1 - 4\exp(-c_1 n \min\{\gamma^2 r^2, 1\})$, we have*

$$\sup_{f,h \in \mathcal{F}: \|f-h\|_{2,\mathbf{X}} \leq r} \left| \frac{1}{n} \sum_{i=1}^{n} \zeta_i (f-h)(\mathbf{X}_i) - \mathbb{E}[\zeta(f-h)(\mathbf{X})] \right| \leq c_3(q) \gamma \mathfrak{m}^* L r^2.$$

5.4 Complexity parameters in the sparse linear setting

The next result shows that, in the linear setting, it is possible to weaken the sub-Gaussian assumption and still be able to control the complexity parameters r_P, r_M as in (5.2).

Theorem 5.5 (Theorem 1.6 in [20]). *There exists an absolute constant c_1 and for $K \geq 1$, $L \geq 1$ and $q_0 > 2$ there exists a constant c_2 that depends only on K, L, q_0 for which the following holds. Consider*

- *$V \subset \mathbb{R}^d$ for which the norm $\|\cdot\|_V = \sup_{\mathbf{v} \in V} |\langle \mathbf{v}, \cdot \rangle|$ is $K-$unconditional with respect to the basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$;*

- *$\mathfrak{m}^* = \mathbb{E}[|\zeta|^{q_0}]^{1/q_0} < +\infty$;*

- *an isotropic random vector $\mathbf{X} \in \mathbb{R}^d$ which satisfies the weak moment condition: for some constants $c_0, L > 1$, for all $\mathbf{y} \in \mathbb{R}^d$, $1 \leq p \leq c_0 \log(ed)$, $1 \leq j \leq d$,*

$$\mathbb{E}[|\mathbf{X}^\top \mathbf{e}_j|^p]^{\frac{1}{p}} \leq L\sqrt{p}\,\mathbb{E}[|\mathbf{X}^\top \mathbf{e}_j|^2]^{\frac{1}{2}}.$$

If $(\mathbf{X}_i, \zeta_i)_{i=1}^n$ are i.i.d. copies of $(\mathbf{X}, \zeta)$, then

$$\mathbb{E}\left[\sup_{v \in V} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\zeta_i \mathbf{X}_i^\top \mathbf{v} - \mathbb{E}[\zeta \mathbf{X}^\top \mathbf{v}] \right) \right| \right] \leq c_2 \mathfrak{m}^* \mathbb{E}[\|G\|_V].$$

Since this result deals with the multiplier empirical process and, when $\zeta \equiv 1$, with the standard empirical process, by arguing as in the proof of Lemma 5.1 we find that any function

$$\rho \mapsto r(\rho) \geq \max\left\{ r_n^*\left(\frac{\gamma_P}{c_2}\right), s_n^*\left(\frac{\gamma_M}{c_2 \mathfrak{m}^*}\right) \right\},$$

is a valid choice in (3.2). Our Definition 4.1 restricts our analysis to settings where the assumptions of the previous theorem are satisfied.

By following Section 4 in [13], we provide bounds for the complexity parameters r_n^*, s_n^* in (5.2). For any $\boldsymbol{\beta} \in \mathbb{R}^d$, set $f_{\boldsymbol{\beta}} : \mathbb{R}^d \to \mathbb{R}$ the linear map $f_{\boldsymbol{\beta}}(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$, consider $\mathcal{F} = \{ f_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathbb{R}^d \}$ and, for any $\rho > 0$,

$$\mathbb{B}_1(\rho) = \{ f_{\boldsymbol{\beta}} \in \mathcal{F} : |\boldsymbol{\beta}|_1 \leq \rho \}.$$

Assume that $\mathbf{X}$ is an isotropic random vector that satisfies the weak moment condition of Theorem 5.5, recall that $\mathfrak{m}^* = \mathbb{E}[\zeta^4]^{1/4}$. By symmetry, $\mathbb{B}_1(\rho) - \mathbb{B}_1(\rho) = \mathbb{B}_1(2\rho)$ and it is sufficient to control the function $r \mapsto \mathbb{E}\big[\|G\|_{\mathbb{B}_1(2\rho) \cap \mathbb{B}_2(r)}\big]$. One finds, for every $2\rho/\sqrt{d} \leq r$,

$$\mathbb{E}\big[\|G\|_{\mathbb{B}_1(2\rho) \cap \mathbb{B}_2(r)}\big] = \mathbb{E}\left[\sup_{\boldsymbol{\beta} \in \mathbb{R}^d : |\boldsymbol{\beta}|_1 \leq 2\rho, |\boldsymbol{\beta}|_2 \leq r} \left| \sum_{i=0}^d g_i \beta_i \right| \right] \sim \rho\sqrt{\log(ed \min\{r^2/\rho^2, 1\})},$$

and if $r \leq 2\rho/\sqrt{d}$, then

$$\mathbb{E}\big[\|G\|_{\mathbb{B}_1(2\rho) \cap \mathbb{B}_2(r)}\big] = \mathbb{E}\left[\sup_{\boldsymbol{\beta} \in \mathbb{R}^d : |\boldsymbol{\beta}|_1 \leq 2\rho, |\boldsymbol{\beta}|_2 \leq r} \left| \sum_{i=0}^d g_i \beta_i \right| \right] \sim \rho\sqrt{d}.$$

With C_{γ_P} some constants only depending on L and γ_P, one finds

$$r_n^{*2}\left(\frac{\gamma_P}{c_2}\right) \leq C_{\gamma_P}^2 \times \begin{cases} \frac{\rho^2}{n} \log\left(\frac{ed}{n}\right) & \text{if } n \leq c_3 d, \\ \frac{\rho^2}{d} & \text{if } c_3 d \leq n \leq c_4 d, \\ 0 & n > c_4 d, \end{cases}$$

the constants c_3, c_4 depend only on L. Similarly, with C_{γ_M} some constants only depending on L and γ_M,

$$s_n^{*2}\left(\frac{\gamma_M}{c_2 \mathfrak{m}^*}\right) \leq C_{\gamma_M}^2 \times \begin{cases} \rho \mathfrak{m}^* \sqrt{\frac{\log d}{n}} & \text{if } \rho^2 n \leq \mathfrak{m}^{*2} \log d, \\ \rho \mathfrak{m}^* \sqrt{\frac{1}{n} \log\left(\frac{ed^2 \mathfrak{m}^{*2}}{\rho^2 n}\right)} & \text{if } \mathfrak{m}^{*2} \log d \leq \rho^2 n \leq \mathfrak{m}^{*2} d^2, \\ \mathfrak{m}^{*2} \frac{d}{n} & \rho^2 n \geq \mathfrak{m}^{*2} d^2. \end{cases}$$

The bounds given above are valid for any regime of n and d, but we continue the discussion for the more interesting high-dimensional case, that is $d \gg n$. This simplifies the notation and allows to choose, for some constant C_{γ_P, γ_M} only depending on L, γ_P, γ_M,

$$r^2(\rho) = C_{\gamma_P, \gamma_M}^2 \begin{cases} \max\left\{ \rho \mathfrak{m}^* \sqrt{\frac{\log d}{n}}, \frac{\rho^2}{n} \log\left(\frac{ed}{n}\right) \right\}, & \text{if } \rho \leq \frac{\mathfrak{m}^* \sqrt{\log d}}{\sqrt{n}}, \\ \max\left\{ \rho \mathfrak{m}^* \sqrt{\frac{1}{n} \log\left(\frac{ed^2 \mathfrak{m}^{*2}}{\rho^2 n}\right)}, \frac{\rho^2}{n} \log\left(\frac{ed}{n}\right) \right\}, & \text{if } \frac{\mathfrak{m}^* \sqrt{\log d}}{\sqrt{n}} \leq \rho \leq \frac{\mathfrak{m}^* d}{\sqrt{n}}, \end{cases} \tag{5.3}$$

which coincides with the function obtained in Section 4.4 in [15].

Solution of the sparsity equation. We study the case $n \geq s \log(ed/s)$ and assume there exists a $s-$sparse vector in $\boldsymbol{\beta}^* + \mathbb{B}_1(\rho/20)$. In the proof of Theorem 1.4 in [14], it is shown that the smallest solution of the sparsity equation (3.4) is

$$\rho^* = C^*_{\gamma_P, \gamma_M} \mathfrak{m}^* s^* \sqrt{\frac{1}{n} \log\left(\frac{ed}{s^*}\right)},$$

for some constant $C^*_{\gamma_P, \gamma_M}$ only depending on L, γ_P, γ_M. We now compute $r^2(\rho^*)$. Up to multiplying ρ^* by a big constant, we have $\rho^* \gtrsim \mathfrak{m}^* \sqrt{\log d}/\sqrt{n}$, since $s^* \sqrt{\log(ed/s^*)} > \sqrt{\log d}$ for all $1 < s^* \leq d$. By definition, we have

$$
\begin{aligned}
r^2(\rho^*) &= C^2_{\gamma_P, \gamma_M} \max\left\{\rho^* \mathfrak{m}^* \sqrt{\frac{1}{n} \log\left(\frac{ed^2 \mathfrak{m}^{*2}}{\rho^{*2} n}\right)}, \ \frac{\rho^{*2}}{n} \log\left(\frac{ed}{n}\right)\right\} \\
&= C^2_{\gamma_P, \gamma_M} \rho^* \mathfrak{m}^* \sqrt{\frac{1}{n} \log\left(\frac{ed^2 \mathfrak{m}^{*2}}{\rho^{*2} n}\right)} \\
&= C^2_{\gamma_P, \gamma_M} C^*_{\gamma_P, \gamma_M} \frac{\mathfrak{m}^{*2} s^*}{n} \sqrt{\log\left(\frac{ed}{s^*}\right)} \sqrt{\log\left(\frac{ed^2}{C^{*2}_{\gamma_P, \gamma_M} s^{*2} \log\left(\frac{ed}{s^*}\right)}\right)} \\
&\leq \sqrt{2} C^2_{\gamma_P, \gamma_M} C^*_{\gamma_P, \gamma_M} \frac{\mathfrak{m}^{*2} s^*}{n} \log\left(\frac{ed}{s^*}\right),
\end{aligned}
$$

in the last inequality we have used that $\log(a^2) = 2\log(|a|)$ and $C^*_{\gamma_P, \gamma_M} > 1/\sqrt{\log(ed/s^*)}$. The latter is true without loss of generality in the high-dimensional setting $d \gg n \geq s^* \log(ed/s^*)$. The quantity $r(\rho^*)$ is the convergence rate of the LASSO estimator with penalization parameter $\lambda \sim r^2(\rho^*)/\rho^* \sim \mathfrak{m}^* \sqrt{\log(ed/s^*)/n}$. This choice of λ requires the knowledge of the true sparsity parameter s^*.

Acknowledgements

Appendix A   Proof of Theorem 3.3

The structure of the proof is as follows. First, we control the supremum of the functional $T_{K,\mu}(g, \chi, f^*, \sigma^*)$ over possible values of (g, χ) by partitioning the domain in slices. Each slice is treated separately by the results from Lemma A.2 to Lemma A.10. Then, we compare the bounds over different slices in Lemma A.11 and show that the leading contribution comes from a bounded ball of the form

$$\mathbb{B}^*(\rho_K) = \big\{(g, \chi) \in \mathcal{F} \times (0, \sigma_+] : \|g - f^*\| \leq \rho_K, \ \|g - f^*\|_{2,\mathbf{X}} \leq r(\rho_K), \ |\chi - \sigma^*| \leq c_\alpha r(\rho_K)\big\}.$$

In Lemma A.12 we translate the supremum bounds into convergence rates by showing that the MOM$-K$ estimator belongs to a bounded ball $\mathbb{B}^*(2\rho_K)$. We finalize the proof by computing the excess risk bound in Lemma A.13.

In the notation of Theorem 3.3, for any $c > 2$ we have

$$c_\mu := 200(c + 2)\kappa_+^{1/2},$$

$$\varepsilon := \frac{c - 2}{192\theta_0^2(c + 2)\big(8 + 134\kappa_+^{1/2}((1 + \frac{\sigma_+}{\sigma^*}) \vee \frac{6}{5})\big)},$$

$$c_\alpha^2 := \frac{3(c - 2)}{5\theta_0^2},$$

furthermore, we use the auxiliary parameters

$$\gamma_P = \frac{1}{1488\theta_0}, \quad \gamma_Q = \frac{\varepsilon}{360}, \quad \gamma_M = \frac{\varepsilon}{744}, \quad \eta = \frac{1}{16}, \quad \gamma = \frac{31}{32}, \quad \alpha = x = \frac{1}{93}. \tag{A.1}$$

We denote by $r(\cdot)$ a function such that $r(\rho) \geq \max\{r_P(\rho, \gamma_P), r_M(\rho, \gamma_M)\}$. By Assumption 3.2, there exists an absolute constant such that $r(\rho) \leq r(2\rho) < c_r r(\rho)$. With $C^2 = 384\theta_1^2 c_r^2 c_\alpha^2 \kappa_+^{1/2}$, we allow for $K \in \big[K^* \vee 32|\mathcal{O}|, \ n\varepsilon^2/C^2\big]$. We denote by $\Omega(K)$ the intersection of the event $\Omega_1(K)$ in Lemma D.4, the event $\Omega_2(K)$ in Lemma D.7 and the event $\Omega_3(K)$ in Lemma D.8. The probability of $\Omega(K) = \Omega_1(K) \cap \Omega_2(K) \cap \Omega_3(K)$ is at least $1 - \mathbb{P}(\Omega_1(K)) - \mathbb{P}(\Omega_2(K)) - \mathbb{P}(\Omega_3(K)) \geq 1 - 4\exp(-K/8920)$. For any $c_\rho \in \{1, 2\}$, we denote

$$\alpha_{K,c_\rho} := c_\alpha r(c_\rho \rho_K), \quad \delta_{K,n}^2 := \frac{25\mathfrak{m}^{*4}K}{n}, \quad r^2(\rho_K) = \frac{384\theta_1^2 \delta_{K,n}^2}{25\mathfrak{m}^{*2}\varepsilon^2}, \tag{A.2}$$

the last equation rewrites the implicit definition of ρ_K in (3.7).

The next lemma checks that the choices made in Theorem 3.3 satisfy a set of sufficient conditions that are required by our proving strategy. In principle, our main result is valid for different choices as long as the relevant quantities satisfy the conditions below.

Lemma A.1. *The assumptions of Theorem 3.3 imply, with $c_K^2 = 384$ and any $\iota_\mu \in [1/4, 4]$,*

$$n\varepsilon^2 > K c_K^2 \theta_1^2 c_r^2 c_\alpha^2 \kappa_+^{1/2}, \tag{A.3}$$

$$\iota_\mu c_\mu > \frac{1600 \kappa_+^{3/4} \varepsilon}{c_K^2 \theta_1^2} + 48 \kappa_+^{1/2}(c+2), \tag{A.4}$$

$$\frac{c-2}{24\theta_0^2} > \frac{800\kappa_+^{1/2}\varepsilon^2}{c_K^2 \theta_1^2} + 16(c+2)\varepsilon + \left(\frac{1 + \frac{\sigma_+}{\sigma^*}}{3} \vee \frac{36}{10}\right)\iota_\mu c_\mu \varepsilon, \tag{A.5}$$

$$c_\alpha^2 > \frac{1800\kappa_+^{1/2}\varepsilon^2}{c_K^2 \theta_1^2} + 108(c+2)\varepsilon + \frac{144\iota_\mu c_\mu \varepsilon}{10}. \tag{A.6}$$

Conditions (A.3) *and* (A.6) *imply* $4\delta_{K,n}/\sigma^* < \alpha_{K,c_\rho} < \sigma^*$. *Condition* (A.5) *implies both*

$$\frac{1}{16\theta_0^2} > 4\varepsilon + \frac{(\sigma^* + \sigma_+)\iota_\mu c_\mu \varepsilon}{2(c-2)\mathfrak{m}^*}, \tag{A.7}$$

$$\frac{c-2}{24\theta_0^2} > \frac{800\kappa^{*1/2}\varepsilon^2}{c_K^2 \theta_1^2} + 16(c+2)\varepsilon + \frac{36\iota_\mu c_\mu \varepsilon}{10}. \tag{A.8}$$

Proof of Lemma A.1. Condition (A.3) is equivalent to the upper bound $K \leq n\varepsilon^2/C^2$ on the number of blocks. We have $r^2(\rho_K) = c_K^2\theta_1^2\mathfrak{m}^{*2}K/(\varepsilon^2 n)$ and $r^2(2\rho_K) \leq c_r^2 r^2(\rho_K)$, by Assumption 3.2. Since $\alpha_{K,2} = c_\alpha r(2\rho_K)$, then also $\alpha_{K,2} \leq c_r\alpha_{K,1}$, therefore

$$\frac{\alpha_{K,1}^2}{\sigma^{*2}} \leq \frac{\alpha_{K,2}^2}{\sigma^{*2}} \leq \frac{c_r^2\alpha_{K,1}^2}{\sigma^{*2}} = c_r^2 c_\alpha^2 \frac{c_K^2\theta_1^2\mathfrak{m}^{*2}K}{\sigma^{*2}n\varepsilon^2} = c_r^2 c_\alpha^2 \frac{c_K^2\theta_1^2\kappa^{*1/2}K}{n\varepsilon^2} < 1,$$

where the last inequality is condition (A.3), then $\alpha_{K,c_\rho} < \sigma^*$. We show $4\delta_{K,n}/\sigma^* < \alpha_{K,c_\rho}$ using

$$\frac{16\delta_{K,n}^2}{\sigma^{*2}} = \frac{400\mathfrak{m}^{*4}K}{\sigma^{*2}n} = \kappa^{*1/2}\frac{400\mathfrak{m}^{*2}K}{n} < c_\alpha^2\frac{384\theta_1^2\mathfrak{m}^{*2}K}{n\varepsilon^2} = \alpha_{K,1}, \tag{A.9}$$

where the only inequality is implied by condition (A.6).

By definition of c_μ in (3.5), we have

$$\iota_\mu c_\mu \geq \frac{c_\mu}{4} = 50(c+2)\kappa_+^{1/2} = 2(c+2)\kappa_+^{1/2} + 48(c+2)\kappa_+^{1/2},$$

thus (A.4) is satisfied since, as we show below,

$$\varepsilon < \frac{c_K^2\theta_1^2(c+2)}{800\kappa_+^{1/4}} = \frac{12\theta_1^2(c+2)}{25\kappa_+^{1/4}}.$$

With $c_K^2 = 384$, we rewrite condition (A.5) as

$$\frac{50\kappa_+^{1/2}\theta_0^2\varepsilon^2}{(c-2)\theta_1^2} + \frac{384\theta_0^2(c+2)\varepsilon}{c-2} + \left(\frac{1 + \frac{\sigma_+}{\sigma^*}}{3} \vee \frac{36}{10}\right)\frac{24\theta_0^2\iota_\mu c_\mu\varepsilon}{c-2} < 1.$$

With the definition of c_μ in (3.5) and $\iota_\mu = 4$, this becomes

$$\frac{50\kappa_+^{1/2}\theta_0^2}{(c-2)\theta_1^2}\varepsilon^2 + \frac{48\theta_0^2(c+2)}{c-2}\left(8 + \frac{400\kappa_+^{1/2}}{3}\left(\left(1 + \frac{\sigma_+}{\sigma^*}\right) \vee \frac{12}{10}\right)\right)\varepsilon < 1.$$

29

The inequality above has the form $A\varepsilon^2 + B\varepsilon < 1$, which is satisfied by any ε smaller than $\min\{1/\sqrt{2A},\ 1/2B\}$. The definition of ε in (3.5) coincides with imposing $\varepsilon = c_\varepsilon \cdot \min\{1/\sqrt{2A},\ 1/2B\} = c_\varepsilon/2B$, with $c_\varepsilon = 1/2$ and

$$\frac{1}{\sqrt{2A}} = \sqrt{c-2}\,\frac{\theta_1}{10\theta_0\kappa_+^{1/4}},$$

$$\frac{1}{2B} = \frac{c-2}{96\theta_0^2(c+2)\big(8 + 134\kappa_+^{1/2}((1+\frac{\sigma_+}{\sigma^*}) \vee \frac{6}{5})\big)},$$

we have used that $400/3 < 134$. Thus, condition (A.5) is satisfied. It is immediate to verify that this implies both (A.7) and (A.8).

With $c_K^2 = 384$, the definition of c_μ in (3.5) and $\iota_\mu = 4$, we rewrite (A.6) as

$$c_\alpha^2 > \frac{75\kappa_+^{1/2}}{16\theta_1^2}\varepsilon^2 + 108(c+2)\left(1 + \frac{320}{3}\kappa_+^{1/2}\right)\varepsilon.$$

By the discussion on ε above, it is sufficient that, with $c_\varepsilon = 1/2$ and $320/3 < 107$,

$$c_\alpha^2 > \frac{75\kappa_+^{1/2}}{16\theta_1^2} \cdot \frac{c_\varepsilon^2(c-2)\theta_1^2}{100\theta_0^2\kappa_+^{1/2}} + 108(c+2)\left(1 + 107\kappa_+^{1/2}\right)\frac{c_\varepsilon(c-2)}{96\theta_0^2(c+2)\big(8 + 134\kappa_+^{1/2}((1+\frac{\sigma_+}{\sigma^*}) \vee \frac{6}{5})\big)}.$$

This is equivalent to

$$c_\alpha^2 > \frac{15(c-2)}{320\theta_0^2}c_\varepsilon^2 + \frac{27(c-2)(1+107\kappa_+^{1/2})}{24\theta_0^2\big(8 + 134\kappa_+^{1/2}((1+\frac{\sigma_+}{\sigma^*}) \vee \frac{6}{5})\big)}c_\varepsilon,$$

and, with

$$\frac{1+107\kappa_+^{1/2}}{8 + 134\kappa_+^{1/2}((1+\frac{\sigma_+}{\sigma^*}) \vee \frac{6}{5})} < 1,$$

and $c_\varepsilon = 1/2$, condition (A.6) holds if

$$c_\alpha^2 \geq \frac{15(c-2)}{320\theta_0^2}c_\varepsilon^2 + \frac{27(c-2)}{24\theta_0^2}c_\varepsilon = \frac{(c-2)}{16\theta_0^2}\left(\frac{15}{80} + \frac{27}{3}\right) = \frac{441(c-2)}{768\theta_0^2}.$$

This is exactly the case from the definition of c_α in (3.5), since $3/5 > 441/768$. The proof is complete. $\square$

A.1   Control of the supremum of $T_{K,\mu}(g,\chi,f^*,\sigma^*)$

With σ_+ the known upper bound on σ^*, set $I_+ = (0,\sigma_+]$ and, with $r(\cdot)$ any function such that $r(\rho) \geq \{r_P(\rho,\gamma_P), r_M(\rho,\gamma_M)\}$, any $c_\rho \in \{1,2\}$ and $\alpha_{K,c_\rho} = c_\alpha r(c_\rho \rho_K)$, let us define

$$\mathcal{F}_1^{(c_\rho)} := \{(g,\chi) \in \mathcal{F} \times I_+ : \|g - f^*\| \leq c_\rho \rho_K, \|g - f^*\|_{2,\mathbf{X}} \leq r(c_\rho \rho_K), \ |\sigma^* - \chi| \leq \alpha_{K,c_\rho}\}$$

$$\mathcal{F}_2^{(c_\rho)} := \{(g,\chi) \in \mathcal{F} \times I_+ : \|g - f^*\| \leq c_\rho \rho_K, \|g - f^*\|_{2,\mathbf{X}} > r(c_\rho \rho_K), \ |\sigma^* - \chi| \leq \alpha_{K,c_\rho}\}$$

$$\mathcal{F}_3^{(c_\rho)} := \{(g,\chi) \in \mathcal{F} \times I_+ : \|g - f^*\| > c_\rho \rho_K, \ |\sigma^* - \chi| \leq \alpha_{K,c_\rho}\}$$

$$\mathcal{F}_4^{(c_\rho)} := \{(g,\chi) \in \mathcal{F} \times I_+ : \|g - f^*\| \leq c_\rho \rho_K, \|g - f^*\|_{2,\mathbf{X}} \leq r(c_\rho \rho_K), \ \chi > \sigma^* + \alpha_{K,c_\rho}\}$$

$$\mathcal{F}_5^{(c_\rho)} := \{(g,\chi) \in \mathcal{F} \times I_+ : \|g - f^*\| \leq c_\rho \rho_K, \|g - f^*\|_{2,\mathbf{X}} > r(c_\rho \rho_K), \ \chi > \sigma^* + \alpha_{K,c_\rho}\}$$

$$\mathcal{F}_6^{(c_\rho)} := \{(g,\chi) \in \mathcal{F} \times I_+ : \|g - f^*\| > c_\rho \rho_K, \ \chi > \sigma^* + \alpha_{K,c_\rho}\}$$

$$\mathcal{F}_7^{(c_\rho)} := \{(g,\chi) \in \mathcal{F} \times I_+ : \|g - f^*\| \leq c_\rho \rho_K, \|g - f^*\|_{2,\mathbf{X}} \leq r(c_\rho \rho_K), \ \chi < \sigma^* - \alpha_{K,c_\rho}\}$$

$$\mathcal{F}_8^{(c_\rho)} := \{(g,\chi) \in \mathcal{F} \times I_+ : \|g - f^*\| \leq c_\rho \rho_K, \|g - f^*\|_{2,\mathbf{X}} > r(c_\rho \rho_K), \ \chi < \sigma^* - \alpha_{K,c_\rho}\}$$

$$\mathcal{F}_9^{(c_\rho)} := \{(g,\chi) \in \mathcal{F} \times I_+ : \|g - f^*\| > c_\rho \rho_K, \ \chi < \sigma^* - \alpha_{K,c_\rho}\}.$$

The sets above are a partition of the domain $\mathcal{F} \times I_+$ where the functional

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) = MOM_K\Big(R_c(\ell_g,\chi,\ell_{f^*},\sigma^*)\Big) + \mu(\|f^*\| - \|g\|)$$

takes inputs. For $c_\rho \in \{1,2\}$ and $i = 1,\ldots,9$, we set B_{i,c_ρ} some upper bound for the supremum of $T_{K,\mu}(g,\chi,f^*,\sigma^*)$ over $(g,\chi) \in \mathcal{F}_i^{(c_\rho)}$. That is,

$$\sup_{(g,\chi) \in \mathcal{F}_i^{(c_\rho)}} T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq B_{i,c_\rho}, \tag{A.10}$$

and the goal of this section is to give sharp bounds for each slice separately. Using the definition of $R_c(\ell_g,\chi,\ell_{f^*},\sigma^*)$ in (2.6), and $\ell_g = \ell_{f^*} + \ell_g - \ell_{f^*}$, we find

$$R_c(\ell_g,\chi,\ell_{f^*},\sigma^*) = (\sigma^* - \chi)\left(1 - 2\frac{\ell_{f^*} + \ell_g}{(\sigma^* + \chi)^2}\right) + 2c\frac{\ell_{f^*} - \ell_g}{\sigma^* + \chi}$$

$$= (\sigma^* - \chi)\left(1 - \frac{4\ell_{f^*}}{(\sigma^* + \chi)^2}\right) + 2\frac{\ell_{f^*} - \ell_g}{\sigma^* + \chi}\left(c + \frac{\sigma^* - \chi}{\sigma^* + \chi}\right)$$

$$= R_c(\ell_{f^*},\chi,\ell_{f^*},\sigma^*) + 2\Delta_c(\chi,\sigma^*)\frac{\ell_{f^*} - \ell_g}{\sigma^* + \chi}.$$

with

$$\Delta_c(\chi,\sigma) := \left(c + \frac{\sigma - \chi}{\sigma + \chi}\right) \in [c-1, c+1], \quad \forall \sigma, \chi \in (0, +\infty), \tag{A.11}$$

and $c > 2$ by construction. We plug this into the functional $T_{K,\mu}(g,\chi,f^*,\sigma^*)$, so that

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) = MOM_K\left(R_c(\ell_{f^*},\chi,\ell_{f^*},\sigma^*) + 2\Delta_c(\chi,\sigma^*)\frac{\ell_{f^*} - \ell_g}{\sigma^* + \chi}\right) + \mu(\|f^*\| - \|g\|).$$

For all $(\mathbf{x}, y) \in \mathcal{X} \times \mathbb{R}$, we have the decomposition

$$\ell_f(\mathbf{x},y) - \ell_g(\mathbf{x},y) = 2\big(y - f(\mathbf{x})\big)\big(g(\mathbf{x}) - f(\mathbf{x})\big) - \big(g(\mathbf{x}) - f(\mathbf{x})\big)^2,$$

and this gives $\ell_{f^*} - \ell_g = 2\zeta(g - f^*) - (g - f^*)^2$. By the triangular quantile property in Lemma D.2, we can write

$$
\begin{aligned}
T_{K,\mu}&(g, \chi, f^*, \sigma^*) \\
&= Q_{3/4,K}\left[R_c(\ell_{f^*}, \chi, \ell_{f^*}, \sigma^*) + 2\Delta_c(\chi, \sigma^*)\frac{\ell_{f^*} - \ell_g}{\sigma^* + \chi} \right] + \mu(\|f^*\| - \|g\|) \\
&\leq Q_{3/4,K}\left[R_c(\ell_{f^*}, \chi, \ell_{f^*}, \sigma^*) \right] + \frac{2\Delta_c(\chi, \sigma^*)}{(\sigma^* + \chi)} Q_{3/4,K}\left[2\zeta(g - f^*) - (g - f^*)^2 \right] \\
&\quad + \mu(\|f^*\| - \|g\|).
\end{aligned}
\tag{A.12}
$$

By arguing as in the proof of Lemma D.9, see bound for (D.2), the quantity

$$
Q_{3/4,K}\left[R_c(\ell_{f^*}, \chi, \ell_{f^*}, \sigma^*) \right] = Q_{3/4,K}\left[(\sigma^* - \chi)\left(1 - \frac{4\ell_{f^*}}{(\sigma^* + \chi)^2}\right) \right]
$$

is bounded above, when $\chi \geq \sigma^*$, by

$$
Q_{3/4,K}\left[R_c(\ell_{f^*}, \chi, \ell_{f^*}, \sigma^*) \right] \leq (\chi - \sigma^*)\left(\frac{4\sigma^{*2} + 4\delta_{K,n}}{(\sigma^* + \chi)^2} - 1\right),
\tag{A.13}
$$

or, when $\chi \leq \sigma^*$, by

$$
Q_{3/4,K}\left[R_c(\ell_{f^*}, \chi, \ell_{f^*}, \sigma^*) \right] \leq (\sigma^* - \chi)\left(1 - \frac{4\sigma^{*2} - 4\delta_{K,n}}{(\sigma^* + \chi)^2}\right).
\tag{A.14}
$$

The following lemmas show that, on the event $\Omega(K)$, one can choose bounds B_{i,c_ρ} in (A.10),

for $i = 1, \ldots, 9$ and $c_\rho \in \{1, 2\}$, as

$$B_{1,c_\rho} = \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,c_\rho}}r^2(c_\rho\rho_K) + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*}r^2(\rho_K),$$

$$B_{2,c_\rho} = \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* + \alpha_{K,c_\rho}}r^2(c_\rho\rho_K) + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*}r^2(\rho_K),$$

$$B_{3,c_\rho} = \max\left\{ \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + c_\rho\left(\frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,c_\rho}} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K), \right.$$
$$\left. \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + c_\rho\left(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* + \alpha_{K,c_\rho}} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K) \right\},$$

$$B_{4,c_\rho} = -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + \frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,c_\rho}}r^2(c_\rho\rho_K) + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*}r^2(\rho_K),$$

$$B_{5,c_\rho} = -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+}r^2(c_\rho\rho_K) + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*}r^2(\rho_K),$$

$$B_{6,c_\rho} = \max\left\{ -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + c_\rho\left(\frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,c_\rho}} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K), \right.$$
$$\left. -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + c_\rho\left(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K) \right\},$$

$$B_{7,c_\rho} = -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + \frac{8(c+2)\varepsilon}{\sigma^*}r^2(c_\rho\rho_K) + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*}r^2(\rho_K),$$

$$B_{8,c_\rho} = -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,c_\rho}}r^2(c_\rho\rho_K) + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*}r^2(\rho_K),$$

$$B_{9,c_\rho} = \max\left\{ -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + \left(\frac{8(c+2)\varepsilon c_\rho}{\sigma^*} - \frac{4c_\mu\varepsilon c_\rho}{5\mathfrak{m}^*} + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}\right)r^2(\rho_K), \right.$$
$$\left. -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + c_\rho\left(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,c_\rho}} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K) \right\}.$$

Lemma A.2. *On the event $\Omega(K)$, for all $c_\rho \in \{1, 2\}$, the supremum of $T_{K,\mu}(g, \chi, f^*, \sigma^*)$ over the set*

$$\mathcal{F}_1^{(c_\rho)} := \{(g, \chi) \in \mathcal{F} \times I_+ : \|g - f^*\| \leq c_\rho\rho_K, \|g - f^*\|_{2,\mathbf{x}} \leq r(c_\rho\rho_K), \ |\sigma^* - \chi| \leq \alpha_{K,c_\rho}\},$$

is bounded above by

$$B_{1,c_\rho} := \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,c_\rho}}r^2(c_\rho\rho_K) + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*}r^2(\rho_K).$$

Proof of Lemma A.2. Let $(g, \chi) \in \mathcal{F}_1^{(c_\rho)}$. Using the bound obtained in (A.12), the inequality $(g - f^*)^2 \geq 0$ and the triangular inequality, the quantity $T_{K,\mu}(g, \chi, f^*, \sigma^*)$ is bounded above by

$$Q_{3/4,K}\left[R_c(\ell_{f^*}, \chi, \ell_{f^*}, \sigma^*)\right] + \frac{2\Delta_c(\chi, \sigma^*)}{\sigma^* + \chi}Q_{3/4,K}\left[2\zeta(g - f^*) - (g - f^*)^2\right] + \mu(\|f^*\| - \|g\|)$$

$$\leq Q_{3/4,K}\left[R_c(\ell_{f^*}, \chi, \ell_{f^*}, \sigma^*)\right] + \frac{2\Delta_c(\chi, \sigma^*)}{\sigma^* + \chi}Q_{3/4,K}\left[2\zeta(g - f^*)\right] + \mu\|f^* - g\|.$$

By Lemma D.7, $Q_{3/4,K}[2\zeta(g-f^*)] \le \alpha_M^2 \le 4\varepsilon r^2(c_\rho\rho_K)$ and, with $\Delta_c(\chi,\sigma^*) \le c+2$, we find

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \le Q_{3/4,K}\big[R_c(\ell_{f^*},\chi,\ell_{f^*},\sigma^*)\big] + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,c_\rho}}r^2(c_\rho\rho_K) + \mu c_\rho\rho_K$$

$$= Q_{3/4,K}\big[R_c(\ell_{f^*},\chi,\ell_{f^*},\sigma^*)\big] + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,c_\rho}}r^2(c_\rho\rho_K) + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*}r^2(\rho_K),$$

where in the last step we put our choice $\mu = (c_\mu\varepsilon/\mathfrak{m}^*)r^2(\rho_K)/\rho_K$. We now bound the quantile term appearing in the latter display. Directly from (A.13) and (A.14), we get

$$Q_{3/4,K}\big[R_c(\ell_{f^*},\chi,\ell_{f^*},\sigma^*)\big]$$
$$\le \max\left\{ \sup_{\chi\in[\sigma^*,\sigma^*+\alpha_{K,c_\rho}]} |\sigma^* - \chi|\Big(\frac{4\sigma^{*2} + 4\delta_{K,n}}{(\sigma^* + \chi)^2} - 1\Big),\ \sup_{\chi\in[\sigma^*-\alpha_{K,c_\rho},\sigma^*]} |\sigma^* - \chi|\Big(1 - \frac{4\sigma^{*2} - 4\delta_{K,n}}{(\sigma^* + \chi)^2}\Big)\right\}.$$

By arguing as in the proof of Lemma D.9, see bounds on (D.2), with $\alpha_{K,c_\rho} > 2\delta_{K,n}/\sigma^*$ we obtain

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,c_\rho}}r^2(c_\rho\rho_K) + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*}r^2(\rho_K),$$

which is what we wanted. $\qquad\square$

Lemma A.3. *On the event $\Omega(K)$, for all $c_\rho \in \{1,2\}$, the supremum of $T_{K,\mu}(g,\chi,f^*,\sigma^*)$ over the set*

$$\mathcal{F}_2^{(c_\rho)} := \{(g,\chi) \in \mathcal{F} \times I_+ : \|g - f^*\| \le c_\rho\rho_K, \|g - f^*\|_{2,\mathbf{X}} > r(c_\rho\rho_K),\ |\sigma^* - \chi| \le \alpha_{K,c_\rho}\},$$

is bounded above by

$$B_{2,c_\rho} := \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* + \alpha_{K,c_\rho}}r^2(c_\rho\rho_K) + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*}r^2(\rho_K).$$

Proof of Lemma A.3. Let $(g,\chi) \in \mathcal{F}_2^{(c_\rho)}$. The space $\mathcal{F}_2^{(c_\rho)}$ shares with $\mathcal{F}_1^{(c_\rho)}$ the conditions $\|g - f^*\| \le c_\rho\rho_K$ and $|\chi - \sigma^*| \le \alpha_{K,c_\rho}$. By arguing as in the proof of Lemma A.2, we know already that

$$T_{K,\mu}(g,\chi,f^*,\sigma^*)$$
$$\le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + \frac{2\Delta_c(\chi,\sigma^*)}{(\sigma^* + \chi)}Q_{3/4,K}\big[2\zeta(g-f^*) - (g-f^*)^2\big] + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*}r^2(\rho_K).$$

An application of Lemma D.7 bounds from above the quantiles of $2\zeta(g-f^*)$ and from below the quantiles of $(g-f^*)^2$, together with $\Delta_c(\chi,\sigma^*) \ge c-2$ this leads to

$$T_{K,\mu}(g,\chi,f^*,\sigma^*)$$
$$\le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + \frac{2\Delta_c(\chi,\sigma^*)}{(\sigma^* + \chi)}Q_{3/4,K}\big[2\zeta(g-f^*) - (g-f^*)^2\big] + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*}r^2(\rho_K)$$
$$\le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + \frac{2\Delta_c(\chi,\sigma^*)}{(\sigma^* + \chi)}\big(\alpha_M^2 - (4\theta_0)^{-2}\|g - f^*\|_{2,\mathbf{X}}^2\big) + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*}r^2(\rho_K)$$
$$\le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* + \alpha_{K,c_\rho}}r^2(c_\rho\rho_K) + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*}r^2(\rho_K),$$

since $4\varepsilon < 1/(4\theta_0)^2$ by condition (A.7), so $\alpha_M^2 - \|g-f^*\|_{2,\mathbf{X}}^2(4\theta_0)^{-2} \le (4\varepsilon - (4\theta_0)^{-2})r^2(c_\rho\rho_K)$. $\qquad\square$

Lemma A.4. *On the event $\Omega(K)$, for all $c_\rho \in \{1,2\}$, the supremum of $T_{K,\mu}(g,\chi,f^*,\sigma^*)$ over the set*

$$\mathcal{F}_3^{(c_\rho)} := \{(g,\chi) \in \mathcal{F} \times I_+ : \|g - f^*\| > c_\rho \rho_K, \ |\sigma^* - \chi| \le \alpha_{K,c_\rho}\},$$

is bounded above by

$$B_{3,c_\rho} := \max\left\{ \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2} \delta_{K,n}^2 + c_\rho \left(\frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,c_\rho}} - \frac{4c_\mu \varepsilon}{5\mathfrak{m}^*} \right) r^2(\rho_K) + \frac{c_\mu \varepsilon}{10\mathfrak{m}^*} r^2(\rho_K), \right.$$

$$\left. \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2} \delta_{K,n}^2 + c_\rho \left(2(c-2) \frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* + \alpha_{K,c_\rho}} + \frac{c_\mu \varepsilon}{\mathfrak{m}^*} \right) r^2(\rho_K) + \frac{c_\mu \varepsilon}{10\mathfrak{m}^*} r^2(\rho_K) \right\}.$$

Proof of Lemma A.4. Let $(g,\chi) \in \mathcal{F}_3^{(c_\rho)}$. The space $\mathcal{F}_3^{(c_\rho)}$ shares with $\mathcal{F}_1^{(c_\rho)}, \mathcal{F}_2^{(c_\rho)}$ the constraint $|\chi - \sigma^*| \le \alpha_{K,c_\rho}$. By arguing as in the proofs of Lemma A.2 and Lemma A.3, the bound in (A.12) becomes

$$T_{K,\mu}(g,\chi,f^*,\sigma^*)$$

$$\le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2} \delta_{K,n}^2 + \frac{2\Delta_c(\chi,\sigma^*)}{(\sigma^* + \chi)} Q_{3/4,K}\left[2\zeta(g - f^*) - (g - f^*)^2\right] + \mu(\|f^*\| - \|g\|)$$

$$\le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2} \delta_{K,n}^2 + \frac{2\Delta_c(\chi,\sigma^*)}{(\sigma^* + \chi)} Q_{3/4,K}\left[2\zeta(g - f^*) - (g - f^*)^2\right]$$

$$- \mu \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g - f^*) + \frac{\mu \rho_K}{10},$$

where the last inequality follows from the application of Lemma D.1 with $\rho = \rho_K$. We follow now the proof of Lemma 5 in [15]. Let us define $f := f^* + \rho_K(g - f^*)/\|g - f^*\|$, this function belongs to the function class $\mathcal{F}$ by convexity. Let $\Upsilon := \|g - f^*\|/\rho_K$. By construction, $\|f - f^*\| = \rho_K$ and $g - f^* = \Upsilon(f - f^*)$. Then,

$$T_{K,\mu}(g,\chi,f^*,\sigma^*)$$

$$\le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2} \delta_{K,n}^2 + \frac{2\Upsilon \Delta_c(\chi,\sigma^*)}{(\sigma^* + \chi)} Q_{3/4,K}\left[2\zeta(f - f^*) - (f - f^*)^2\right]$$

$$- \mu \Upsilon \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) + \frac{\mu \rho_K}{10}.$$

From here, we separate the cases $\|f - f^*\|_{2,\mathbf{X}} \le r(\rho_K)$ and $\|f - f^*\|_{2,\mathbf{X}} > r(\rho_K)$.

We start with $\|f - f^*\|_{2,\mathbf{X}} \le r(\rho_K)$. Since $\|f - f^*\| = \rho_K$, we have $f \in H_{\rho_K}$ with $H_{\rho_K} = \{f \in \mathcal{F} : \|f - f^*\| \le \rho_K, \ \|f - f^*\|_{2,\mathbf{X}} \le r(\rho_K)\}$ defined in Section 3.2. Recall that K^* is defined as the smallest integer satisfying $K^* \ge n\varepsilon r^2(\rho^*)/c_K^2 \theta_m^2$, with ρ^* the smallest value $\rho > 0$ satisfying the sparsity inequality

$$\inf_{f \in H_\rho} \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) \ge \frac{4}{5}\rho.$$

Since $K \ge K^*$, we get $\rho_K \ge \rho^*$ and ρ_K satisfies the sparsity inequality

$$\sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) \ge \frac{4}{5}\rho_K.$$

35

Using our choice of $\mu = (c_\mu \varepsilon / \mathfrak{m}^*) r^2(\rho_K)/\rho_K$, we get

$$-\mu \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) \le -\frac{4c_\mu \varepsilon}{5\mathfrak{m}^*} r^2(\rho_K).$$

The latter display, the fact that $(f - f^*)^2 \ge 0$, the bound $\Delta_c(\chi, \sigma^*) \le c + 2$, and the quantile bound $Q_{3/4,K}[2\zeta(f - f^*)] \le \alpha_M^2 \le 4\varepsilon r^2(\rho_K)$ in Lemma D.7, all together yield

$$T_{K,\mu}(g, \chi, f^*, \sigma^*) \le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2} \delta_{K,n}^2 + \Upsilon\left(\frac{8(c + 2)\varepsilon}{2\sigma^* - \alpha_{K,c_\rho}} - \frac{4c_\mu \varepsilon}{5\mathfrak{m}^*}\right) r^2(\rho_K) + \frac{\mu \rho_K}{10}.$$

By condition (A.4), the term multiplied by Υ is negative. This is true because $\kappa_+^{1/4} \ge \kappa^{*1/4} = \mathfrak{m}^*/\sigma^* > 1$ and

$$c_\mu > \frac{5\mathfrak{m}^*(c + 2)}{\sigma^*} \implies \frac{4c_\mu \varepsilon}{5\mathfrak{m}^*} > \frac{4(c + 2)\varepsilon}{\sigma^*} > \frac{4(c + 2)\varepsilon}{2\sigma^* - \alpha_{K,c_\rho}},$$

the last inequality follows from $\alpha_{K,c_\rho} < \sigma^*$, which is guaranteed by Lemma A.1. Since $\Upsilon > c_\rho$, we have

$$T_{K,\mu}(g, \chi, f^*, \sigma^*) \le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2} \delta_{K,n}^2 + c_\rho\left(\frac{8(c + 2)\varepsilon}{2\sigma^* - \alpha_{K,c_\rho}} - \frac{4c_\mu \varepsilon}{5\mathfrak{m}^*}\right) r^2(\rho_K) + \frac{c_\mu \varepsilon}{10\mathfrak{m}^*} r^2(\rho_K).$$

This concludes the first part of the proof.

We now consider the case $\|f - f^*\|_{2,\mathbf{X}} > r(\rho_K)$. Since $\|f - f^*\| = \rho_K$ and $\Delta_c(\chi, \sigma^*) \ge c - 2$, an application of Lemma D.7 bounds from above the quantiles of $2\zeta(g - f^*)$ and from below the quantiles of $(g - f^*)^2$, this gives

$$T_{K,\mu}(g, \chi, f^*, \sigma^*)$$
$$\le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2} \delta_{K,n}^2 + \Upsilon\left(2(c - 2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* + \alpha_{K,c_\rho}} r^2(\rho_K) + \mu \rho_K\right) + \frac{\mu \rho_K}{10}$$
$$\le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2} \delta_{K,n}^2 + c_\rho\left(2(c - 2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* + \alpha_{K,c_\rho}} + \frac{c_\mu \varepsilon}{\mathfrak{m}^*}\right) r^2(\rho_K) + \frac{c_\mu \varepsilon}{10\mathfrak{m}^*} r^2(\rho_K),$$

using that $\Upsilon > c_\rho$ and the term multiplied by Υ is negative, by condition (A.7). This can be seen by

$$\frac{1}{16\theta_0^2} > 4\varepsilon + \frac{(\sigma^* + \sigma_+)c_\mu \varepsilon}{2(c - 2)\mathfrak{m}^*}$$
$$\iff 0 > 2(c - 2)\frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+} + \frac{c_\mu \varepsilon}{\mathfrak{m}^*} > 2(c - 2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* + \alpha_{K,c_\rho}} + \frac{c_\mu \varepsilon}{\mathfrak{m}^*}.$$

This concludes the second part of the proof. $\qquad\square$

Lemma A.5. *On the event $\Omega(K)$, for all $c_\rho \in \{1, 2\}$, the supremum of $T_{K,\mu}(g, \chi, f^*, \sigma^*)$ over the set*

$$\mathcal{F}_4^{(c_\rho)} := \{(g, \chi) \in \mathcal{F} \times I_+ : \|g - f^*\| \le c_\rho \rho_K, \|g - f^*\|_{2,\mathbf{X}} \le r(c_\rho \rho_K), \chi > \sigma^* + \alpha_{K,c_\rho}\},$$

is bounded above by

$$B_{4,c_\rho} := -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,c_\rho})^2} \alpha_{K,c_\rho}^2 + \frac{8(c + 2)\varepsilon}{2\sigma^* + \alpha_{K,c_\rho}} r^2(c_\rho \rho_K) + \frac{c_\mu \varepsilon c_\rho}{\mathfrak{m}^*} r^2(\rho_K).$$

36

Proof of Lemma A.5. Let $(g, \chi) \in \mathcal{F}_4^{(c_\rho)}$. The space $\mathcal{F}_4^{(c_\rho)}$ shares with $\mathcal{F}_1^{(c_\rho)}$ the conditions $\|g - f^*\| \leq c_\rho \rho_K$ and $\|g - f^*\|_{2,\mathbf{X}} \leq r(c_\rho \rho_K)$. By arguing as in the proof of Lemma A.2 and using that $\chi > \sigma^* + \alpha_{K,c_\rho}$, from (A.13) we get

$$T_{K,\mu}(g, \chi, f^*, \sigma^*)$$

$$\leq \sup_{\chi > \sigma^* + \alpha_{K,c_\rho}} (\chi - \sigma^*) \left(\frac{4(\sigma^{*2} + \delta_{K,n})}{(\sigma^* + \chi)^2} - 1 \right) + \frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,c_\rho}} r^2(c_\rho \rho_K) + \frac{c_\mu \varepsilon c_\rho}{\mathfrak{m}^*} r^2(\rho_K)$$

$$= -\alpha_{K,c_\rho} \left(1 - \frac{8(\sigma^{*2} + \delta_{K,n})}{(2\sigma^* + \alpha_{K,c_\rho})^2} \right) + \frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,c_\rho}} r^2(c_\rho \rho_K) + \frac{c_\mu \varepsilon c_\rho}{\mathfrak{m}^*} r^2(\rho_K).$$

Since $\alpha_{K,c_\rho} > 2\delta_{K,n}/\sigma^*$, one has

$$1 - \frac{4(\sigma^{*2} + \delta_{K,n})}{(2\sigma^* + \alpha_{K,c_\rho})^2} = \frac{4(\sigma^* \alpha_{K,c_\rho} - \delta_{K,n})}{(2\sigma^* + \alpha_{K,c_\rho})^2} + \frac{\alpha_{K,c_\rho}^2}{(2\sigma^* + \alpha_{K,c_\rho})^2} > \frac{4(\sigma^* \alpha_{K,c_\rho} - \delta_{K,n})}{(2\sigma^* + \alpha_{K,c_\rho})^2} > \frac{2\sigma^* \alpha_{K,c_\rho}}{(2\sigma^* + \alpha_{K,c_\rho})^2},$$

and

$$T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,c_\rho})^2} \alpha_{K,c_\rho}^2 + \frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,c_\rho}} r^2(c_\rho \rho_K) + \frac{c_\mu \varepsilon c_\rho}{\mathfrak{m}^*} r^2(\rho_K).$$

This is enough to conclude. $\qquad\square$

Lemma A.6. *On the event $\Omega(K)$, for all $c_\rho \in \{1, 2\}$, the supremum of $T_{K,\mu}(g, \chi, f^*, \sigma^*)$ over the set*

$$\mathcal{F}_5^{(c_\rho)} := \{ (g, \chi) \in \mathcal{F} \times I_+ : \|g - f^*\| \leq c_\rho \rho_K, \|g - f^*\|_{2,\mathbf{X}} > r(c_\rho \rho_K), \ \chi > \sigma^* + \alpha_{K,c_\rho} \},$$

is bounded above by

$$B_{5,c_\rho} := -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,c_\rho})^2} \alpha_{K,c_\rho}^2 + 2(c-2) \frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+} r^2(c_\rho \rho_K) + \frac{c_\mu \varepsilon c_\rho}{\mathfrak{m}^*} r^2(\rho_K).$$

Proof of Lemma A.6. Let $(g, \chi) \in \mathcal{F}_5^{(c_\rho)}$. The space $\mathcal{F}_5^{(c_\rho)}$ shares with $\mathcal{F}_1^{(c_\rho)}$ the condition $\|g - f^*\| \leq c_\rho \rho_K$, with $\mathcal{F}_2^{(c_\rho)}$ the condition $\|g - f^*\|_{2,\mathbf{X}} > r(\rho_K)$, and with $\mathcal{F}_4^{(c_\rho)}$ the condition $\chi > \sigma^* + \alpha_{K,c_\rho}$. By arguing as in the proofs of Lemma A.2, Lemma A.3 and Lemma A.5, one gets

$$T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,c_\rho})^2} \alpha_{K,c_\rho}^2 + 2(c-2) \frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+} r^2(c_\rho \rho_K) + \frac{c_\mu \varepsilon c_\rho}{\mathfrak{m}^*} r^2(\rho_K),$$

where σ_+ is the upper bound on χ. $\qquad\square$

Lemma A.7. *On the event $\Omega(K)$, for all $c_\rho \in \{1, 2\}$, the supremum of $T_{K,\mu}(g, \chi, f^*, \sigma^*)$ over the set*

$$\mathcal{F}_6^{(c_\rho)} := \{ (g, \chi) \in \mathcal{F} \times I_+ : \|g - f^*\| > c_\rho \rho_K, \ \chi > \sigma^* + \alpha_{K,c_\rho} \},$$

is bounded above by

$$B_{6,c_\rho} := \max \left\{ -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,c_\rho})^2} \alpha_{K,c_\rho}^2 + c_\rho \left(\frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,c_\rho}} - \frac{4c_\mu \varepsilon}{5\mathfrak{m}^*} \right) r^2(\rho_K) + \frac{c_\mu \varepsilon}{10\mathfrak{m}^*} r^2(\rho_K), \right.$$

$$\left. -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,c_\rho})^2} \alpha_{K,c_\rho}^2 + c_\rho \left(2(c-2) \frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+} + \frac{c_\mu \varepsilon}{\mathfrak{m}^*} \right) r^2(\rho_K) + \frac{c_\mu \varepsilon}{10\mathfrak{m}^*} r^2(\rho_K) \right\}.$$

37

Proof of Lemma A.7. Let $(g, \chi) \in \mathcal{F}_6^{(c_\rho)}$. The space $\mathcal{F}_6^{(c_\rho)}$ shares with $\mathcal{F}_3^{(c_\rho)}$ the condition $\|g - f^*\| > c_\rho \rho_K$, and with $\mathcal{F}_5^{(c_\rho)}$ the condition $\chi > \sigma^* + \alpha_{K,c_\rho}$. By arguing as in the proofs of Lemma A.4 and Lemma A.6, we find

$$T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + \frac{2\Upsilon\Delta_c(\chi, \sigma^*)}{\sigma^* + \chi}Q_{3/4,K}\left[2\zeta(f - f^*) - (f - f^*)^2\right]$$
$$- \mu\Upsilon \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) + \frac{\mu\rho_K}{10},$$

with the function $f = f^* + \rho_K(g - f^*)/\|g - f^*\|$ and the quantity $\Upsilon = \|g - f^*\|/\rho_K$, as in the proof of Lemma A.4. By following the same argument, we split the cases $\|f - f^*\|_{2,\mathbf{X}} \leq r(\rho_K)$ and $\|f - f^*\|_{2,\mathbf{X}} > r(\rho_K)$.

We start with $\|f - f^*\|_{2,\mathbf{X}} \leq r(\rho_K)$. We find,

$$-\mu \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) \leq -\frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}r^2(\rho_K).$$

Combining this the fact that $(f - f^*)^2 \geq 0$, we get

$$T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + \Upsilon\left(\frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,c_\rho}} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{\mu\rho_K}{10}$$
$$\leq -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + c_\rho\left(\frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,c_\rho}} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K),$$

using that the quantity multiplied by Υ is negative by condition (A.4), and $\Upsilon > c_\rho$. This concludes the first part of the proof.

We now consider $\|f - f^*\|_{2,\mathbf{X}} > r(\rho_K)$. We have,

$$T_{K,\mu}(g, \chi, f^*, \sigma^*)$$
$$\leq -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + \Upsilon\left(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+}r^2(\rho_K) + \mu\rho_K\right) + \frac{\mu\rho_K}{10}$$
$$\leq -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + c_\rho\left(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K),$$

using that the quantity multiplied by Υ is negative by condition (A.7), and $\Upsilon > c_\rho$. This concludes the proof. $\qquad\square$

Lemma A.8. *On the event* $\Omega(K)$, *for all* $c_\rho \in \{1, 2\}$, *the supremum of* $T_{K,\mu}(g, \chi, f^*, \sigma^*)$ *over the set*

$$\mathcal{F}_7^{(c_\rho)} := \{(g, \chi) \in \mathcal{F} \times I_+ : \|g - f^*\| \leq c_\rho \rho_K, \|g - f^*\|_{2,\mathbf{X}} \leq r(c_\rho \rho_K), \ \chi < \sigma^* - \alpha_{K,c_\rho}\},$$

is bounded above by

$$B_{7,c_\rho} := -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + \frac{8(c+2)\varepsilon}{\sigma^*}r^2(c_\rho\rho_K) + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*}r^2(\rho_K).$$

38

Proof of Lemma A.8. Let $(g, \chi) \in \mathcal{F}_7^{(c_\rho)}$. The space $\mathcal{F}_7^{(c_\rho)}$ shares with $\mathcal{F}_1^{(c_\rho)}$ the conditions $\|g - f^*\| \le c_\rho \rho_K$ and $\|g - f^*\|_{2,\mathbf{X}} \le r(c_\rho \rho_K)$. By arguing as in the proof of Lemma A.2 and using $\chi < \sigma^* - \alpha_{K,c_\rho}$, from (A.14) we get

$$
\begin{aligned}
T_{K,\mu}&(g,\chi,f^*,\sigma^*) \\
&\le \sup_{\chi < \sigma^* - \alpha_{K,c_\rho}} (\sigma^* - \chi)\Big(1 - \frac{4(\sigma^{*2} - \delta_{K,n})}{(\sigma^* + \chi)^2}\Big) + \frac{8(c+2)\varepsilon}{\sigma^*} r^2(c_\rho \rho_K) + \frac{c_\mu \varepsilon c_\rho}{\mathfrak{m}^*} r^2(\rho_K) \\
&= -\alpha_{K,c_\rho}\Big(\frac{4(\sigma^{*2} - \delta_{K,n})}{(2\sigma^* - \alpha_{K,c_\rho})^2} - 1\Big) + \frac{8(c+2)\varepsilon}{\sigma^*} r^2(c_\rho \rho_K) + \frac{c_\mu \varepsilon c_\rho}{\mathfrak{m}^*} r^2(\rho_K).
\end{aligned}
$$

Since $4\delta_{K,n}/\sigma^* < \alpha_{K,c_\rho} < \sigma^*$ by Lemma A.1, we find

$$
\frac{4(\sigma^{*2} - \delta_{K,n})}{(2\sigma^* - \alpha_{K,c_\rho})^2} - 1 = \frac{4\sigma^* \alpha_{K,c_\rho} - 4\delta_{K,n} - \alpha_{K,c_\rho}^2}{(2\sigma^* - \alpha_{K,c_\rho})^2} > \frac{2\sigma^* \alpha_{K,c_\rho}}{(2\sigma^* - \alpha_{K,c_\rho})^2},
$$

and

$$
T_{K,\mu}(g,\chi,f^*,\sigma^*) \le -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + \frac{8(c+2)\varepsilon}{\sigma^*} r^2(c_\rho \rho_K) + \frac{c_\mu \varepsilon c_\rho}{\mathfrak{m}^*} r^2(\rho_K),
$$

which is sufficient to conclude. $\qquad\square$

Lemma A.9. *On the event $\Omega(K)$, for all $c_\rho \in \{1,2\}$, the supremum of $T_{K,\mu}(g,\chi,f^*,\sigma^*)$ over the set*

$$
\mathcal{F}_8^{(c_\rho)} := \{(g,\chi) \in \mathcal{F} \times I_+ : \|g - f^*\| \le c_\rho \rho_K, \|g - f^*\|_{2,\mathbf{X}} > r(c_\rho \rho_K), \ \chi < \sigma^* - \alpha_{K,c_\rho}\},
$$

is bounded above by

$$
B_{8,c_\rho} := -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,c_\rho}} r^2(c_\rho \rho_K) + \frac{c_\mu \varepsilon c_\rho}{\mathfrak{m}^*} r^2(\rho_K).
$$

Proof of Lemma A.9. Let $(g,\chi) \in \mathcal{F}_8^{(c_\rho)}$. The space $\mathcal{F}_8^{(c_\rho)}$ shares with $\mathcal{F}_1^{(c_\rho)}$ the condition $\|g - f^*\| \le c_\rho \rho_K$, with $\mathcal{F}_2^{(c_\rho)}$ the condition $\|g - f^*\| > r(c_\rho \rho_K)$, and with $\mathcal{F}_7^{(c_\rho)}$ the condition $\chi < \sigma^* - \alpha_{K,c_\rho}$. By arguing as in the proofs of Lemma A.2, Lemma A.3 and Lemma A.8, one finds

$$
T_{K,\mu}(g,\chi,f^*,\sigma^*) \le -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,c_\rho}} r^2(c_\rho \rho_K) + \frac{c_\mu \varepsilon c_\rho}{\mathfrak{m}^*} r^2(\rho_K),
$$

which concludes the proof. $\qquad\square$

Lemma A.10. *On the event $\Omega(K)$, for all $c_\rho \in \{1,2\}$, the supremum of $T_{K,\mu}(g,\chi,f^*,\sigma^*)$ over the set*

$$
\mathcal{F}_9^{(c_\rho)} := \{(g,\chi) \in \mathcal{F} \times I_+ : \|g - f^*\| > c_\rho \rho_K, \ \chi < \sigma^* - \alpha_{K,c_\rho}\},
$$

is bounded above by

$$
\begin{aligned}
B_{9,c_\rho} := \max\Big\{ &-\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + \Big(\frac{8(c+2)\varepsilon c_\rho}{\sigma^*} - \frac{4c_\mu \varepsilon c_\rho}{5\mathfrak{m}^*} + \frac{c_\mu \varepsilon}{10\mathfrak{m}^*}\Big) r^2(\rho_K), \\
&-\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + c_\rho\Big(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,c_\rho}} + \frac{c_\mu \varepsilon}{\mathfrak{m}^*}\Big) r^2(\rho_K) + \frac{c_\mu \varepsilon}{10\mathfrak{m}^*} r^2(\rho_K)\Big\}.
\end{aligned}
$$

Proof of Lemma A.10. Let $(g, \chi) \in \mathcal{F}_9^{(c_\rho)}$. The space $\mathcal{F}_9^{(c_\rho)}$ shares with $\mathcal{F}_6^{(c_\rho)}$ the condition $\|g - f^*\| > c_\rho \rho_K$, and with $\mathcal{F}_7^{(c_\rho)}$ the condition $\chi < \sigma^* - \alpha_{K,c_\rho}$. By arguing as in the proofs of Lemma A.7 and Lemma A.8, we get

$$T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + \frac{2\Upsilon \Delta_c(\chi, \sigma^*)}{\sigma^* + \chi}Q_{3/4,K}\left[2\zeta(f - f^*) - (f - f^*)^2\right]$$
$$- \mu\Upsilon \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) + \frac{\mu\rho_K}{10},$$

with the function $f = f^* + \rho_K(g - f^*)/\|g - f^*\|$ and the quantity $\Upsilon = \|g - f^*\|/\rho_K$. We now split the cases $\|f - f^*\|_{2,\mathbf{X}} \leq r(\rho_K)$ and $\|f - f^*\|_{2,\mathbf{X}} > r(\rho_K)$.

For $\|f - f^*\|_{2,\mathbf{X}} \leq r(\rho_K)$, we find

$$-\mu \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) \leq -\frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}r^2(\rho_K),$$

which we combine with the fact that $(f - f^*)^2 \geq 0$, this gives

$$T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho}^2)^2}\alpha_{K,c_\rho}^2 + \Upsilon\left(\frac{8(c+2)\varepsilon}{\sigma^*} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{\mu\rho_K}{10}$$
$$\leq -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho}^2)^2}\alpha_{K,c_\rho}^2 + c_\rho\left(\frac{8(c+2)\varepsilon}{\sigma^*} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K),$$

using that the quantity multiplied by Υ is negative by condition (A.4), and $\Upsilon > c_\rho$. This concludes the first part of the proof.

We now consider the case $\|f - f^*\|_{2,\mathbf{X}} > r(\rho_K)$. We find

$$T_{K,\mu}(g, \chi, f^*, \sigma^*)$$
$$\leq -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + \Upsilon\left(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,c_\rho}}r^2(\rho_K) + \mu\rho_K\right) + \frac{\mu\rho_K}{10}$$
$$\leq -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho})^2}\alpha_{K,c_\rho}^2 + c_\rho\left(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,c_\rho}} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K),$$

using that the quantity multiplied by Υ is negative by condition (A.7), and $\Upsilon > c_\rho$. This concludes the proof. □

A.2   Comparison between the bounds

This section compares the bounds $B_{1,c_\rho}, \ldots, B_{9,c_\rho}$ found above. We show that, for $c_\rho = 1$, the quantity $B_{1,1}$ dominates the bounds $B_{i,1}$ on the slices $i = 2, \ldots, 9$. Furthermore, for $c_\rho = 2$, the negative quantity $-B_{1,1}$ is also bigger than any other bound $B_{i,2}$ on the slices $i = 2, \ldots, 9$. This implicitly shows that the bounds $B_{i,2}$ are negative and bounded away from zero, if $i \neq 1$.

Lemma A.11. *We have $B_{1,1} = \max_{i=1,\ldots,9} B_{i,1}$ and $-B_{1,1} > \max_{i=2,\ldots,9} B_{i,2}$.*

Proof of Lemma A.11. We start by showing that $B_{1,1}$ is bigger than the other $B_{i,1}$, $i = 2, \ldots, 9$. By Lemma A.2, we have

$$B_{1,1} = \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\delta_{K,n}^2 + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}}r^2(\rho_K) + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}r^2(\rho_K).$$

Take $i = 2$. By Lemma A.3, we have

$$B_{2,1} = \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\delta_{K,n}^2 + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* + \alpha_{K,1}}r^2(c_\rho\rho_K) + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}r^2(\rho_K),$$

so that imposing $B_{2,1} \le B_{1,1}$ is equivalent to

$$2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* + \alpha_{K,1}} \le \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}},$$

which is always true since $4\varepsilon - (4\theta_0)^{-2} < 0$, by condition (A.7).

Take $i = 3$. By Lemma A.4, we have

$$B_{3,1} = \max\left\{ \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\delta_{K,n}^2 + \left(\frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K), \right.$$

$$\left. \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\delta_{K,n}^2 + \left(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* + \alpha_{K,1}} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K) \right\},$$

so that imposing $B_{3,1} \le B_{1,1}$ requires both

$$\frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} - \frac{17c_\mu\varepsilon}{10\mathfrak{m}^*} \le \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}},$$

$$2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* + \alpha_{K,1}} + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*} \le \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}}.$$

The first inequality is always true, whereas the second is equivalent to

$$\frac{8(c-2)\varepsilon}{2\sigma^* + \alpha_{K,1}} - \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*} \le 2(c-2)\frac{(4\theta_0)^{-2}}{2\sigma^* + \alpha_{K,1}}.$$

Since $2\sigma^* + \alpha_{K,1} > 2\sigma^* - \alpha_{K,1}$, the latter condition is implied by

$$\frac{c_\mu\varepsilon}{10\mathfrak{m}^*} - \frac{32\varepsilon}{2\sigma^* - \alpha_{K,1}} \le \frac{c-2}{8\theta_0^2(2\sigma^* + \alpha_{K,1})}.$$

By Lemma A.1, we have $0 < \alpha_{K,1} < \sigma^*$ and the above display is satisfied if

$$\frac{c_\mu\varepsilon}{10\mathfrak{m}^*} \le \frac{16\varepsilon}{\sigma^*} + \frac{c-2}{24\theta_0^2\sigma^*}.$$

We multiply by σ^* and use that $\kappa^{*1/4} = \mathfrak{m}^*/\sigma* \ge 1$, so it is sufficient that

$$\frac{c_\mu\varepsilon}{10} \le 16\varepsilon + \frac{c-2}{24\theta_0^2},$$

which holds by condition (A.8).

Take $i = 4$. By Lemma A.5, we have

$$B_{4,1} = -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,1})^2}\alpha_{K,1}^2 + \frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,1}}r^2(\rho_K) + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}r^2(\rho_K),$$

so that imposing $B_{4,1} \leq B_{1,1}$ is equivalent to

$$-\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,1})^2}\frac{\alpha_{K,1}^2}{r^2(\rho_K)} + \frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,1}} \leq \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\frac{\delta_{K,n}^2}{r^2(\rho_K)} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}},$$

which is always satisfied.

Take $i = 5$. By Lemma A.6, we have

$$B_{5,1} = -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,1})^2}\alpha_{K,1}^2 + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+}r^2(\rho_K) + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}r^2(\rho_K),$$

so that imposing $B_{5,1} \leq B_{1,1}$ is equivalent to

$$-\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,1})^2}\frac{\alpha_{K,1}^2}{r^2(\rho_K)} + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+} \leq \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\frac{\delta_{K,n}^2}{r^2(\rho_K)} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}},$$

which is always satisfied, since the term on the left is negative by condition (A.7).

Take $i = 6$. By Lemma A.7, we have

$$B_{6,1} = \max\left\{ -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,1})^2}\alpha_{K,1}^2 + \left(\frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,1}} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K), \right.$$

$$\left. -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,1})^2}\alpha_{K,1}^2 + \left(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K)\right\},$$

so that imposing $B_{6,1} \leq B_{1,1}$ is equivalent to both

$$\frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,1}} - \frac{7c_\mu\varepsilon}{10\mathfrak{m}^*} \leq \frac{2\sigma^*}{(2\sigma^* + \alpha_{K,1})^2}\frac{\alpha_{K,1}^2}{r^2(\rho_K)} + \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\frac{\delta_{K,n}^2}{r^2(\rho_K)} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*},$$

which is always true, and

$$2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+} + \frac{11c_\mu\varepsilon}{10\mathfrak{m}^*} \leq \frac{2\sigma^*}{(2\sigma^* + \alpha_{K,1})^2}\frac{\alpha_{K,1}^2}{r^2(\rho_K)} + \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\frac{\delta_{K,n}^2}{r^2(\rho_K)}$$
$$+ \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}.$$

The first term on the left side is negative, by condition (A.7). With the ratio $\delta_{K,n}^2/r^2(\rho_K)$ in (A.2), it is sufficient that

$$\frac{c_\mu\varepsilon}{10\mathfrak{m}^*} \leq \frac{2\sigma^*}{(2\sigma^* + \alpha_{K,1})^2}c_\alpha^2 + \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2} \cdot \frac{25\mathfrak{m}^{*2}\varepsilon^2}{c_K^2\theta_1^2} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}}.$$

By Lemma A.1, we have $0 < \alpha_{K,1} < \sigma^*$, so it is enough that

$$\frac{c_\mu\varepsilon}{10\mathfrak{m}^*} \leq \frac{2c_\alpha^2}{9\sigma^*} + \frac{400\mathfrak{m}^{*2}\varepsilon^2}{4\sigma^{*3}c_K^2\theta_1^2} + \frac{8(c+2)\varepsilon}{2\sigma^*}.$$

42

We now multiply by $\mathfrak{m}^*$ and use that $\kappa^{*1/4} = \mathfrak{m}^*/\sigma^* \geq 1$, this gives the sufficient condition

$$\frac{9c_\mu\varepsilon}{20} \leq c_\alpha^2 + \frac{450\varepsilon^2}{c_K^2\theta_1^2} + 18(c+2)\varepsilon,$$

which follows from condition (A.6).

Take $i = 7$. By Lemma A.8, we have

$$B_{7,1} = -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,1})^2}\alpha_{K,1}^2 + \frac{8(c+2)\varepsilon}{\sigma^*}r^2(\rho_K) + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}r^2(\rho_K),$$

so that imposing $B_{7,1} \leq B_{1,1}$ is equivalent to

$$\frac{8(c+2)\varepsilon}{\sigma^*} \leq \frac{2\sigma^*}{(2\sigma^* - \alpha_{K,1})^2}\frac{\alpha_{K,1}^2}{r^2(\rho_K)} + \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\frac{\delta_{K,n}^2}{r^2(\rho_K)} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}}.$$

We argue as for $i = 6$, we plug in the ratio $\delta_{K,n}^2/r^2(\rho_K)$ from (A.2) and use $0 < \alpha_{K,1} < \sigma^*$ and $2\sigma^* - \alpha_{K,1} < 2\sigma^* + \alpha_{K,1}$, it is enough that

$$\frac{8(c+2)\varepsilon}{\sigma^*} \leq \frac{2c_\alpha^2}{9\sigma^*} + \frac{400\mathfrak{m}^{*2}\varepsilon^2}{4\sigma^{*3}c_K^2\theta_1^2} + \frac{8(c+2)\varepsilon}{2\sigma^*}.$$

We now multiply by σ^* and use that $\kappa^{*1/4} = \mathfrak{m}^*/\sigma^* \geq 1$, this gives the sufficient condition

$$8(c+2)\varepsilon \leq \frac{2c_\alpha^2}{9} + 100\varepsilon^2 + 4(c+2)\varepsilon,$$

which is true if $18(c+2)\varepsilon \leq c_\alpha^2 + 450\varepsilon^2$, which holds thanks to condition (A.6).

Take $i = 8$. By Lemma A.9, we have

$$B_{8,1} = -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,1})^2}\alpha_{K,1}^2 + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,1}}r^2(\rho_K) + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}r^2(\rho_K),$$

so that imposing $B_{8,1} \leq B_{1,1}$ is equivalent to

$$-\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,1})^2}\frac{\alpha_{K,1}^2}{r^2(\rho_K)} + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,1}} \leq \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\frac{\delta_{K,n}^2}{r^2(\rho_K)} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}},$$

which holds since the left side is negative, thanks to condition (A.7).

Take $i = 9$. By Lemma A.10, we have

$$B_{9,1} = \max\left\{-\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,1})^2}\alpha_{K,1}^2 + \left(\frac{8(c+2)\varepsilon}{\sigma^*} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*} + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}\right)r^2(\rho_K),\right.$$

$$\left.-\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,1})^2}\alpha_{K,1}^2 + \left(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,1}} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K)\right\},$$

so that imposing $B_{9,1} \leq B_{1,1}$ is equivalent to both

$$\frac{8(c+2)\varepsilon}{\sigma^*} - \frac{7c_\mu\varepsilon}{10\mathfrak{m}^*} \leq \frac{2\sigma^*}{(2\sigma^* - \alpha_{K,1})^2}\frac{\alpha_{K,1}^2}{r^2(\rho_K)} + \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\frac{\delta_{K,n}^2}{r^2(\rho_K)} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*},$$

43

which is always true, and

$$2(c-2)\frac{4\varepsilon-(4\theta_0)^{-2}}{2\sigma^*-\alpha_{K,1}} + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*} \leq \frac{2\sigma^*}{(2\sigma^*-\alpha_{K,1})^2}\frac{\alpha_{K,1}^2}{r^2(\rho_K)} + \frac{16}{\sigma^*(2\sigma^*-\alpha_{K,1})^2}\frac{\delta_{K,n}^2}{r^2(\rho_K)} + \frac{8(c+2)\varepsilon}{2\sigma^*-\alpha_{K,1}}.$$

Arguing as in $i=6$, the first term on the left side is negative by condition (A.7), then it is sufficient that

$$\frac{c_\mu\varepsilon}{10\mathfrak{m}^*} \leq \frac{2\sigma^*}{(2\sigma^*-\alpha_{K,1})^2}\frac{\alpha_{K,1}^2}{r^2(\rho_K)} + \frac{16}{\sigma^*(2\sigma^*-\alpha_{K,1})^2}\frac{\delta_{K,n}^2}{r^2(\rho_K)} + \frac{8(c+2)\varepsilon}{2\sigma^*-\alpha_{K,1}},$$

which coincides with the bound obtained in $i=6$.

The first part of the proof is complete. We now show that $-B_{1,1}$ is bigger than $B_{i,2}$, for all $i=2,\ldots,9$. We recall that Lemma A.2 gives

$$B_{1,1} = \frac{16}{\sigma^*(2\sigma^*-\alpha_{K,1})^2}\delta_{K,n}^2 + \frac{8(c+2)\varepsilon}{2\sigma^*-\alpha_{K,1}}r^2(\rho_K) + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}r^2(\rho_K).$$

Take $i=2$. By Lemma A.3, we have

$$B_{2,2} = \frac{16}{\sigma^*(2\sigma^*-\alpha_{K,2})^2}\delta_{K,n}^2 + 2(c-2)\frac{4\varepsilon-(4\theta_0)^{-2}}{2\sigma^*+\alpha_{K,2}}r^2(2\rho_K) + \frac{2c_\mu\varepsilon}{\mathfrak{m}^*}r^2(\rho_K),$$

so that imposing $B_{2,2} + B_{1,1} < 0$ gives

$$\frac{16}{\sigma^*(2\sigma^*-\alpha_{K,2})^2}\delta_{K,n}^2 + \frac{8(c-2)\varepsilon}{2\sigma^*+\alpha_{K,2}}r^2(2\rho_K) + \frac{2c_\mu\varepsilon}{\mathfrak{m}^*}r^2(\rho_K)$$

$$+ \frac{16}{\sigma^*(2\sigma^*-\alpha_{K,1})^2}\delta_{K,n}^2 + \frac{8(c+2)\varepsilon}{2\sigma^*-\alpha_{K,1}}r^2(\rho_K) + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}r^2(\rho_K) < 2(c-2)\frac{(4\theta_0)^{-2}}{2\sigma^*+\alpha_{K,2}}r^2(2\rho_K),$$

Since $r^2(2\rho_K) \geq r^2(\rho_K)$, $\alpha_{K,2} \geq \alpha_{K,1}$, it is sufficient to show

$$\frac{32}{\sigma^*(2\sigma^*-\alpha_{K,2})^2}\frac{\delta_{K,n}^2}{r^2(2\rho_K)} + \frac{8(c-2)\varepsilon}{2\sigma^*-\alpha_{K,2}} + \frac{8(c+2)\varepsilon}{2\sigma^*-\alpha_{K,2}} + \frac{3c_\mu\varepsilon}{\mathfrak{m}^*} < \frac{c-2}{8\theta_0^2(2\sigma^*+\alpha_{K,2})}.$$

By Lemma A.1, we have $0 < \alpha_{K,2} < \sigma^*$ and, with the ratio $\delta_{K,n^2}/r^2(\rho_K)$ in (A.2), it is enough that

$$\frac{800\mathfrak{m}^{*2}\varepsilon^2}{\sigma^{*2}c_K^2\theta_1^2} + 8(c-2)\varepsilon + 8(c+2)\varepsilon + \frac{3c_\mu\varepsilon\sigma^*}{\mathfrak{m}^*} < \frac{c-2}{24\theta_0^2}.$$

Since $\kappa^{*1/4} = \mathfrak{m}^*/\sigma^* \geq 1$, we find the sufficient condition

$$\frac{800\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 16(c+2)\varepsilon + 3c_\mu\varepsilon < \frac{c-2}{24\theta_0^2},$$

which is true by condition (A.8).

Take $i=3$. By Lemma A.4, we have

$$B_{3,2} = \max\left\{\frac{16}{\sigma^*(2\sigma^*-\alpha_{K,2})^2}\delta_{K,n}^2 + 2\left(\frac{8(c+2)\varepsilon}{2\sigma^*-\alpha_{K,2}} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K),\right.$$

$$\left.\frac{16}{\sigma^*(2\sigma^*-\alpha_{K,2})^2}\delta_{K,n}^2 + 2\left(2(c-2)\frac{4\varepsilon-(4\theta_0)^{-2}}{2\sigma^*+\alpha_{K,2}} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K)\right\},$$

so that imposing $B_{3,2} + B_{1,1} < 0$ requires both

$$\frac{32}{\sigma^*(2\sigma^* - \alpha_{K,2})^2} \frac{\delta_{K,n}^2}{r^2(2\rho_K)} + \frac{16(c+2)\varepsilon}{2\sigma^* - \alpha_{K,2}} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} < \frac{c_\mu \varepsilon}{2\mathfrak{m}^*},$$

$$\frac{32}{\sigma^*(2\sigma^* - \alpha_{K,2})^2} \frac{\delta_{K,n}^2}{r^2(2\rho_K)} + \frac{16(c-2)\varepsilon}{2\sigma^* + \alpha_{K,2}} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} + \frac{31 c_\mu \varepsilon}{10\mathfrak{m}^*} < \frac{c-2}{4\theta_0^2(2\sigma^* + \alpha_{K,2})}.$$

By arguing as for $i = 2$, it is sufficient that both

$$\frac{800 \kappa^{*1/2} \varepsilon^2}{c_K^2 \theta_1^2} + 16(c+2)\varepsilon + 8(c+2)\varepsilon < \frac{c_\mu \varepsilon}{2\kappa^{*1/4}},$$

$$\frac{800 \kappa^{*1/2} \varepsilon^2}{c_K^2 \theta_1^2} + 8(c-2)\varepsilon + 8(c+2)\varepsilon + \frac{31 c_\mu \varepsilon}{10 \kappa^{*1/4}} < \frac{c-2}{12\theta_0^2}.$$

The first bound holds by condition (A.4), so we plug it into the second line using $\kappa^* \geq 1$, we obtain the sufficient condition $36 c_\mu \varepsilon / 10 < (c-2)/(12\theta_0^2)$, which follows from condition (A.8).

Take $i = 4$. By Lemma A.5, we have

$$B_{4,2} = -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,2})^2} \alpha_{K,2}^2 + \frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,2}} r^2(2\rho_K) + \frac{2 c_\mu \varepsilon}{\mathfrak{m}^*} r^2(\rho_K),$$

so that imposing $B_{4,2} + B_{1,1} < 0$ gives

$$\frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2} \frac{\delta_{K,n}^2}{r^2(2\rho_K)} + \frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,2}} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} + \frac{3 c_\mu \varepsilon}{\mathfrak{m}^*} < \frac{2\sigma^*}{(2\sigma^* + \alpha_{K,2})^2} \frac{\alpha_{K,2}^2}{r^2(2\rho_K)}.$$

By arguing as for $i = 3$, it is sufficient that

$$\frac{400 \kappa^{*1/2} \varepsilon^2}{c_K^2 \theta_1^2} + 4(c+2)\varepsilon + 8(c+2)\varepsilon + \frac{3 c_\mu \varepsilon}{\kappa^{*1/4}} < \frac{2 c_\alpha^2}{9}.$$

With $\kappa^* \geq 1$, it is enough that

$$\frac{1800 \kappa^{*1/2} \varepsilon^2}{c_K^2 \theta_1^2} + 54(c+2)\varepsilon + \frac{27 c_\mu \varepsilon}{2} < c_\alpha^2,$$

which follows from condition (A.6).

Take $i = 5$. By Lemma A.6, we have

$$B_{5,2} = -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,2})^2} \alpha_{K,2}^2 + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+} r^2(2\rho_K) + \frac{2 c_\mu \varepsilon}{\mathfrak{m}^*} r^2(\rho_K),$$

so that imposing $B_{5,2} + B_{1,1} < 0$ gives

$$\frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2} \frac{\delta_{K,n}^2}{r^2(2\rho_K)} + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+} + \frac{3 c_\mu \varepsilon}{\mathfrak{m}^*} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} < \frac{2\sigma^*}{(2\sigma^* + \alpha_{K,2})^2} \frac{\alpha_{K,2}^2}{r^2(2\rho_K)}.$$

The second term in the latter display is negative by condition (A.7). By arguing as for $i = 4$, it is sufficient that

$$\frac{400 \kappa^{*1/2} \varepsilon^2}{c_K^2 \theta_1^2} + \frac{3 c_\mu \varepsilon}{\kappa^{*1/4}} + 8(c+2)\varepsilon < \frac{2 c_\alpha^2}{9}.$$

With $\kappa^* \geq 1$, it is enough that

$$\frac{1800\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + \frac{27c_\mu\varepsilon}{2} + 36(c+2)\varepsilon < c_\alpha^2,$$

which is true thanks to condition (A.6).

Take $i = 6$. By Lemma A.7, we have

$$B_{6,2} = \max\left\{ -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,2})^2}\alpha_{K,2}^2 + 2\left(\frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,2}} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K), \right.$$
$$\left. -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,2})^2}\alpha_{K,2}^2 + 2\left(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K) \right\},$$

so that imposing $B_{6,2} + B_{1,1} < 0$ requires both

$$\frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\frac{\delta_{K,n}^2}{r^2(\rho_K)} + \frac{16(c+2)\varepsilon}{2\sigma^* + \alpha_{K,2}} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} - \frac{c_\mu\varepsilon}{2\mathfrak{m}^*} < \frac{2\sigma^*}{(2\sigma^* + \alpha_{K,2})^2}\frac{\alpha_{K,2}^2}{r^2(\rho_K)},$$
$$\frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\frac{\delta_{K,n}^2}{r^2(\rho_K)} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} + \frac{31c_\mu\varepsilon}{10\mathfrak{m}^*} + 4(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+} < \frac{2\sigma^*}{(2\sigma^* + \alpha_{K,2})^2}\frac{\alpha_{K,2}^2}{r^2(\rho_K)}.$$

By condition (A.7), the last terms on the left side of both equations are negative. By arguing as in $i = 5$, we find the sufficient conditions

$$\frac{400\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 24(c+2)\varepsilon < \frac{2c_\alpha^2}{9},$$
$$\frac{400\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 8(c+2)\varepsilon + \frac{31c_\mu\varepsilon}{10\kappa^{*1/4}} < \frac{2c_\alpha^2}{9}.$$

With $\kappa^* \geq 1$, it is enough that

$$\frac{1800\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 108(c+2)\varepsilon < c_\alpha^2,$$
$$\frac{1800\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 36(c+2)\varepsilon + 14c_\mu\varepsilon < c_\alpha^2,$$

which follow from condition (A.6).

Take $i = 7$. By Lemma A.8, we have

$$B_{7,2} = -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,2})^2}\alpha_{K,2}^2 + \frac{8(c+2)\varepsilon}{\sigma^*}r^2(2\rho_K) + \frac{2c_\mu\varepsilon}{\mathfrak{m}^*}r^2(\rho_K),$$

so that imposing $B_{7,2} + B_{1,1} < 0$ gives

$$\frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\frac{\delta_{K,n}^2}{r^2(2\rho_K)} + \frac{8(c+2)\varepsilon}{\sigma^*} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} + \frac{3c_\mu\varepsilon}{\mathfrak{m}^*} < \frac{2\sigma^*}{(2\sigma^* - \alpha_{K,2})^2}\frac{\alpha_{K,2}^2}{r^2(2\rho_K)}.$$

By arguing as in $i = 6$, it is sufficient that

$$\frac{400\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 8(c+2)\varepsilon + 8(c+2)\varepsilon + \frac{3c_\mu\varepsilon}{\kappa^{*1/4}} < \frac{2c_\alpha^2}{9}.$$

With $\kappa^* \geq 1$, it is enough that

$$\frac{1800\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 72(c+2)\varepsilon + \frac{27c_\mu\varepsilon}{2} < c_\alpha^2,$$

which follows from condition (A.6).

Take $i = 8$. By Lemma A.9, we have

$$B_{8,2} = -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,2})^2}\alpha_{K,2}^2 + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,2}}r^2(2\rho_K) + \frac{2c_\mu\varepsilon}{\mathfrak{m}^*}r^2(\rho_K),$$

so that imposing $B_{8,2} + B_{1,1} < 0$ gives

$$\frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\frac{\delta_{K,n}^2}{r^2(2\rho_K)} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} + \frac{3c_\mu\varepsilon}{\mathfrak{m}^*} + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,2}} < \frac{2\sigma^*}{(2\sigma^* - \alpha_{K,2})^2}\frac{\alpha_{K,2}^2}{r^2(2\rho_K)}.$$

By condition (A.7), the last term on the left side is negative. By arguing as in $i = 7$, it is sufficient that

$$\frac{400\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 8(c+2)\varepsilon + \frac{3c_\mu\varepsilon}{\kappa^{*1/4}} < \frac{2c_\alpha^2}{9}.$$

With $\kappa^* \geq 1$, it is enough that

$$\frac{1800\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 36(c+2)\varepsilon + \frac{27c_\mu\varepsilon}{2} < c_\alpha^2,$$

which holds thanks to condition (A.6).

Take $i = 9$. By Lemma A.10, we have

$$B_{9,2} = \max\left\{ -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,2})^2}\alpha_{K,2}^2 + \left(\frac{16(c+2)\varepsilon}{\sigma^*} - \frac{8c_\mu\varepsilon}{5\mathfrak{m}^*} + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}\right)r^2(\rho_K),\right.$$

$$\left.-\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,2})^2}\alpha_{K,2}^2 + 2\left(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,2}} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\right)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K)\right\},$$

so that imposing $B_{9,2} + B_{1,1} < 0$ gives both

$$\frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\frac{\delta_{K,n}^2}{r^2(\rho_K)} + \frac{16(c+2)\varepsilon}{\sigma^*} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} - \frac{5c_\mu\varepsilon}{10\mathfrak{m}^*} < \frac{2\sigma^*}{(2\sigma^* - \alpha_{K,2})^2}\frac{\alpha_{K,2}^2}{r^2(\rho_K)},$$

$$\frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\frac{\delta_{K,n}^2}{r^2(\rho_K)} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} + \frac{32c_\mu\varepsilon}{10\mathfrak{m}^*} + 4(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,2}} < \frac{2\sigma^*}{(2\sigma^* - \alpha_{K,2})^2}\frac{\alpha_{K,2}^2}{r^2(\rho_K)}.$$

By condition (A.7), the last terms on the left side in the latter display are negative. By arguing as in $i = 8$, it is sufficient that

$$\frac{400\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 16(c+2)\varepsilon + 8(c+2)\varepsilon < \frac{2c_\alpha^2}{9},$$

$$\frac{400\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 8(c+2)\varepsilon + \frac{32c_\mu\varepsilon}{10\kappa^{*1/4}} < \frac{2c_\alpha^2}{9}.$$

47

With $\kappa^* \geq 1$, it is enough that

$$\frac{1800\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 108(c+2)\varepsilon < c_\alpha^2,$$

$$\frac{1800\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 36(c+2)\varepsilon + \frac{144c_\mu\varepsilon}{10} < c_\alpha^2,$$

which both follow from condition (A.6). $\qquad\square$

A.3 Contraction rates and risk bound

In this section we obtain convergence rates and risk bounds by exploiting the results of the previous section. We recall that we are using a function $r(\cdot)$ such that $r(\rho) \geq \max\{r_P(\rho, \gamma_P), r_M(\rho, \gamma_M)\}$. By Assumption 3.2, there exists an absolute constant c_r such that $r(\rho) \leq r(2\rho) < c_r r(\rho)$. With $C^2 = 384\theta_1^2 c_r^2 c_\alpha^2 \kappa_+^{1/2}$, we allow for $K \in [K^* \vee 32|\mathcal{O}|, n\varepsilon^2/C^2]$. We denote by $\Omega(K)$ the intersection of the event $\Omega_1(K)$ in Lemma D.4, the event $\Omega_2(K)$ in Lemma D.7 and the event $\Omega_3(K)$ in Lemma D.8. The probability of $\Omega(K) = \Omega_1(K) \cap \Omega_2(K) \cap \Omega_3(K)$ is at least $1 - \mathbb{P}(\Omega_1(K)) - \mathbb{P}(\Omega_2(K)) - \mathbb{P}(\Omega_3(K)) \geq 1 - 4\exp(-K/8920)$.

Lemma A.12. *On the event $\Omega(K)$ defined above, the $MOM-K$ estimator $(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$ belongs to the slice*

$$\mathcal{F}_1^{(2)} := \{(g, \chi) \in \mathcal{F} \times I_+ : \|g - f^*\| \leq 2\rho_K, \|g - f^*\|_{2,\mathbf{X}} \leq r(2\rho_K), \ |\sigma^* - \chi| \leq c_\alpha r(2\rho_K)\},$$

thus recovering the convergence rates in (3.9).

Proof of Lemma A.12. By definition (2.11), we have

$$\mathcal{C}_{K,\mu}(\widehat{f}_{K,\mu}, \widehat{\sigma}_{K,\mu}) \leq \mathcal{C}_{K,\mu}(f^*, \sigma^*) = \sup_{g \in \mathcal{F}, \ \chi < \sigma_+} T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq B_{1,1},$$

where the last inequality follows from Lemma A.11. Then,

$$B_{1,1} \geq \mathcal{C}_{K,\mu}(\widehat{f}_{K,\mu}, \widehat{\sigma}_{K,\mu}) = \sup_{g \in \mathcal{F}, \chi < \sigma_+} T_{K,\mu}(g, \chi, \widehat{f}_{K,\mu}, \widehat{\sigma}_{K,\mu})$$

$$\geq T_{K,\mu}(f^*, \sigma^*, \widehat{f}_{K,\mu}, \widehat{\sigma}_{K,\mu}) \geq -T_{K,\mu}(\widehat{f}_{K,\mu}, \widehat{\sigma}_{K,\mu}, f^*, \sigma^*),$$

in the last step we have used $Q_{1/2}[\mathbf{x}] \geq -Q_{1/2}[-\mathbf{x}]$ from Lemma D.2. We deduce that, on the event $\Omega(K)$, $T_{K,\mu}(\widehat{f}_{K,\mu}, \widehat{\sigma}_{K,\mu}, f^*, \sigma^*) \geq -B_{1,1}$. Applying Lemma A.11 again, we have $-B_{1,1} > \sup_{i=2,\dots 9} B_{i,2}$ and

$$\max_{i=2,\dots,9} \sup_{(g,\chi) \in \mathcal{F}_i^{(2)}} T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq \max_{i=2,\dots,9} B_{i,2} < -B_{1,1}.$$

Thus, the estimator $(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$ is outside $\cup_{i=2}^9 \mathcal{F}_i^{(2)}$, which means that $(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$ belongs to $\mathcal{F}_1^{(2)}$. By definition of $\mathcal{F}_1^{(2)}$, we have $\|\widehat{f}_{K,\mu,\sigma_+} - f^*\| \leq 2\rho_K$, $\|\widehat{f}_{K,\mu,\sigma_+} - f^*\|_{2,\mathbf{X}} \leq r(2\rho_K)$, and $|\widehat{\sigma}_{K,\mu,\sigma_+} - \sigma^*| \leq \alpha_{K,2} = c_\alpha r(2\rho_K)$. The proof is complete. $\qquad\square$

48

Lemma A.13. *On the event $\Omega(K)$ defined above, the MOM$-K$ estimator $(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$ satisfies*

$$R(\widehat{f}_{K,\mu,\sigma_+}) - R(f^*) \leq \left(2 + 2c_\alpha + (44 + 5c_\mu)\varepsilon + \frac{25\kappa^{*1/2}}{8\theta_1^2}\varepsilon^2 \right) r^2(2\rho_K)$$

$$+ 4\theta_1^2\varepsilon \left(r^2(2\rho_K) \vee r_Q^2(2\rho_K, \gamma_Q) \right),$$

thus recovering the excess risk bound in (3.10).

Proof of Lemma A.13. We apply Lemma D.9 with $\rho = 2\rho_K$ and $\alpha_{K,c_\rho} = \alpha_{K,2}$, which gives

$$R(\widehat{f}_{K,\mu}) - R(f^*) = \|\widehat{f}_{K,\mu} - f^*\|_{2,\mathbf{X}}^2 + \mathbb{E}[-2\zeta(\widehat{f}_{K,\mu} - f^*)(\mathbf{X})]$$

$$\leq r^2(2\rho_K) + \frac{2\sigma^* + \alpha_{K,2}}{2c}T_{K,\mu}(f^*, \sigma^*, \widehat{f}_{K,\mu}, \widehat{\sigma}_{K,\mu}) + \frac{2\sigma^* + \alpha_{K,2}}{c}\mu\rho_K + \alpha_M^2$$

$$+ \frac{8(2\sigma^* + \alpha_{K,2})}{c\sigma^*(2\sigma^* - \alpha_{K,2})^2}\delta_{K,n}^2 + \frac{\alpha_{K,2}}{c(2\sigma^* - \alpha_{K,2})}\left(2\sigma^* r(2\rho_K) + r^2(2\rho_K) + \alpha_Q^2 + \alpha_M^2 \right).$$

In the proof of Lemma A.12 we have shown that $T_{K,\lambda}(f^*, \sigma^*, \widehat{f}_{K,\mu}, \widehat{\sigma}_{K,\mu}) \leq \mathcal{C}_{K,\lambda}(\widehat{f}_{K,\mu}, \widehat{\sigma}_{K,\mu}) \leq B_{1,1}$. By Lemma A.2 and the ratio $\delta_{K,n}^2/r^2(2\rho_K)$ in (A.2), we have

$$B_{1,1} = \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\delta_{K,n}^2 + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}}r^2(\rho_K) + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}r^2(\rho_K)$$

$$= \left(\frac{25\mathfrak{m}^{*2}\varepsilon^2}{24\theta_1^2\sigma^*(2\sigma^* - \alpha_{K,1})^2} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*} \right) r^2(\rho_K)$$

$$\leq \left(\frac{25\kappa^{*1/2}\varepsilon^2}{24\theta_1^2\sigma^*} + \frac{8(c+2)\varepsilon}{\sigma^*} + \frac{c_\mu\varepsilon}{\sigma^*} \right) r^2(\rho_K),$$

in the last inequality we have used $\mathfrak{m}^* > \sigma^*$ and $\alpha_{K,1} < \sigma^*$, which holds by Lemma A.1. This gives

$$\frac{2\sigma^* + \alpha_{K,2}}{2c}B_{1,1} \leq \frac{3\sigma^*}{2c}\left(\frac{25\kappa^{*1/2}\varepsilon^2}{24\theta_1^2\sigma^*} + \frac{8(c+2)\varepsilon}{\sigma^*} + \frac{c_\mu\varepsilon}{\sigma^*} \right) r^2(2\rho_K)$$

$$= \left(\frac{25\kappa^{*1/2}\varepsilon^2}{16\theta_1^2 c} + \frac{12(c+2)\varepsilon}{c} + \frac{3c_\mu\varepsilon}{2c} \right) r^2(2\rho_K).$$

By construction, we have $\mu = (c_\mu\varepsilon/\mathfrak{m}^*)r^2(\rho_K)/\rho_K$, so that

$$\frac{2\sigma^* + \alpha_{K,2}}{c}\mu\rho_K \leq \frac{3\sigma^*}{c}\frac{c_\mu\varepsilon}{\mathfrak{m}^*}r^2(\rho_K) \leq \frac{3c_\mu\varepsilon}{c}r^2(\rho_K),$$

since $\mathfrak{m}^* > \sigma^*$ and $\alpha_{K,2} < \sigma^*$.

By Lemma D.7 we have $\alpha_M^2 \leq 4\varepsilon r^2(2\rho_K)$, whereas by Lemma D.8 we bound

$$\alpha_Q^2 \leq \varepsilon \max\left(\|f - f^*\|_{2,\mathbf{X}}^2\frac{1488\theta_1^4}{\varepsilon^2}\frac{K}{n}, \ r_Q^2(\rho, \gamma_Q), \ \|f - f^*\|_{2,\mathbf{X}}^2 \right)$$

$$\leq \varepsilon \left(r^2(2\rho_K) \vee r_Q^2(2\rho_K, \gamma_Q) \right) \max\left(\frac{1488\theta_1^4 K}{n\varepsilon^2}, \ 1 \right) \leq 4\theta_1^2\varepsilon \left(r^2(2\rho_K) \vee r_Q^2(2\rho_K, \gamma_Q) \right)$$

since $K \leq n\varepsilon^2/C^2$, $C^2 = 384\theta_1^2 c_r^2 c_\alpha^2 k_+^{1/2}$ and $1488/384 < 4$.

With $\alpha_{K,2} < \sigma^*$ and the ratio $\delta_{K,n}^2/r^2(\rho_K)$ in (A.2), we find

$$\frac{8(2\sigma^* + \alpha_{K,2})}{c\sigma^*(2\sigma^* - \alpha_{K,2})^2}\delta_{K,n}^2 \leq \frac{24}{c\sigma^{*2}}\delta_{K,n}^2 \leq \frac{25\kappa^{*1/2}\varepsilon^2}{16\theta_1^2 c}r^2(\rho_K).$$

By putting together all the previous bounds we have

$$R(\widehat{f}_{K,\mu,\sigma_+}) - R(f^*) \leq r^2(2\rho_K) + \left(\frac{25\kappa^{*1/2}\varepsilon^2}{16\theta_1^2 c} + \frac{12(c+2)\varepsilon}{c} + \frac{3c_\mu\varepsilon}{2c} + \frac{3c_\mu\varepsilon}{c} + 4\varepsilon\right)r^2(2\rho_K)$$

$$+ \frac{25\kappa^{*1/2}\varepsilon^2}{16\theta_1^2 c}r^2(2\rho_K) + \frac{c_\alpha}{c\sigma^*}\left(2\sigma^* r^2(2\rho_K) + (1+4\varepsilon)r^3(2\rho_K)\right)$$

$$+ \frac{4\theta_1^2 c_\alpha\varepsilon}{c\sigma^*}r(2\rho_K)\left(r^2(2\rho_K) \vee r_Q^2(2\rho_K, \gamma_Q)\right).$$

Using $c_\alpha r(2\rho_K) = \alpha_{K,2} < \sigma^*$ in the second and third lines of the latter display, we find

$$R(\widehat{f}_{K,\mu}) - R(f^*) \leq r^2(2\rho_K) + \left(\frac{25\kappa^{*1/2}\varepsilon^2}{16\theta_1^2 c} + \frac{12(c+2)\varepsilon}{c} + \frac{3c_\mu\varepsilon}{2c} + \frac{3c_\mu\varepsilon}{c} + 4\varepsilon\right)r^2(2\rho_K)$$

$$+ \frac{25\kappa^{*1/2}\varepsilon^2}{16\theta_1^2 c}r^2(2\rho_K) + \left(\frac{c_\alpha}{c}2r^2(2\rho_K) + \frac{1}{c}(1+4\varepsilon)r^2(2\rho_K)\right)$$

$$+ \frac{4\theta_1^2\varepsilon}{c}\left(r^2(2\rho_K) \vee r_Q^2(2\rho_K, \gamma_Q)\right).$$

With $c > 1$ and $(c+2)/c < 3$, this recovers

$$R(\widehat{f}_{K,\mu}) - R(f^*) \leq \left(2 + 2c_\alpha + (44 + 5c_\mu)\varepsilon + \frac{25\kappa^{*1/2}}{8\theta_1^2}\varepsilon^2\right)r^2(2\rho_K)$$

$$+ 4\theta_1^2\varepsilon\left(r^2(2\rho_K) \vee r_Q^2(2\rho_K, \gamma_Q)\right),$$

which completes the proof. $\qquad\square$

Appendix B   Proofs for the high-dimensional sparse linear regression

B.1   Proof of Theorem 4.4

In Section B.2, we prove the following Theorem B.1. We show now how this theorem can be used to derive our Theorem 4.4.

Theorem B.1. *Assume that $P_{\mathbf{X},\xi} \in \mathcal{P}_{[0,\sigma_+]}$. There exists universal constants $\widetilde{c}_\mu$, $(\widetilde{c}_i)_{i=0,\dots,5}$ that only depend on $\theta_0, \theta_1, \gamma_Q, \gamma_M$ such that the following holds. Assume that $|\mathcal{I}| \geq n/2$, $|\mathcal{O}| \leq \widetilde{c}_0 s^* \log(ed/s^*)$, $n \geq s^* \log(ed/s^*)$ and $\boldsymbol{\beta}^* \in \mathcal{F}_{s^*}$.*

For every $\iota_K, \iota_\mu \in [1/2, 2]^2$, let $K = \lceil \iota_K \widetilde{c}_2 s^* \log(ed/s^*) \rceil$ and let $(\widehat{\boldsymbol{\beta}}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$ be the MOM$-K$ estimator defined in (2.10) with penalization parameter

$$\mu := \iota_\mu \widetilde{c}_\mu \sqrt{\frac{1}{n} \log\left(\frac{ed}{s^*}\right)}.$$

Then, for all $p \in [1, 2]$, we have

$$|\widehat{\boldsymbol{\beta}}_{K,\mu,\sigma_+} - \boldsymbol{\beta}^*|_p \leq \widetilde{c}_3 \varepsilon^{-1} \kappa^* \sigma^* s^{*\frac{1}{p}} \sqrt{\frac{1}{n} \log\left(\frac{ed}{s^*}\right)},$$

$$|\widehat{\sigma}_{K,\mu,\sigma_+} - \sigma^*| \leq c_\alpha \widetilde{c}_3 \varepsilon^{-1} \kappa^* \sigma^* s^{*\frac{1}{2}} \sqrt{\frac{1}{n} \log\left(\frac{ed}{s^*}\right)}.$$

(B.1)

with probability at least $1 - 4\exp(-K/8920)$.

With high probability, we have

$$|\widehat{\boldsymbol{\beta}}_{K,\mu} - \boldsymbol{\beta}^*|_p \leq \widetilde{c}_3 \varepsilon^{-1} \kappa^* \sigma^* s^{*\frac{1}{p}} \sqrt{\frac{1}{n} \log\left(\frac{ed}{s^*}\right)}.$$

We can explicit the value of ε^{-1} as

$$\varepsilon^{-1} = \frac{192\theta_0^2(c+2)\left(8 + 134\kappa_+^{1/2}((1 + \frac{\sigma_+}{\sigma^*}) \vee \frac{6}{5})\right)}{c - 2} = C\left((1 + \frac{\sigma_+}{\sigma^*}) \vee \frac{6}{5}\right).$$

for a constant $C > 0$, and therefore

$$|\widehat{\boldsymbol{\beta}}_{K,\mu} - \boldsymbol{\beta}^*|_p \lesssim \left((1 + \frac{\sigma_+}{\sigma^*}) \vee \frac{6}{5}\right)\sigma^* s^{*\frac{1}{p}} \sqrt{\frac{1}{n} \log\left(\frac{ed}{s}\right)}.$$

Since by assumption $\sigma^* < \sigma_+$, we deduce

$$|\widehat{\boldsymbol{\beta}}_{K,\mu} - \boldsymbol{\beta}^*|_p \lesssim \sigma_+ s^{*\frac{1}{p}} \sqrt{\frac{1}{n} \log\left(\frac{ed}{s}\right)}.$$

The proof for the bound on $\widehat{\sigma}_{K,\mu,\sigma_+}$ follows the same computations as it involves a factor of ε^{-1}.

B.2   Proof of Theorem B.1

In this section we use the results in Theorem 3.3 and the computations in Section 5.4 for the sparse linear setting. For any fixed $\varepsilon \in (0, 1)$, the function

$$r_\varepsilon^2(\rho) = C_{\gamma_P,\gamma_M}^2 \begin{cases} \max\left\{\rho\mathfrak{m}^* \sqrt{\frac{\log d}{n\varepsilon^2}}, \ \frac{\rho^2}{n\varepsilon^2} \log\left(\frac{ed}{n\varepsilon^2}\right)\right\}, & \text{if } \rho \leq \frac{\mathfrak{m}^*\sqrt{\log d}}{\sqrt{n\varepsilon^2}}, \\ \max\left\{\rho\mathfrak{m}^* \sqrt{\frac{1}{n\varepsilon^2} \log\left(\frac{ed^2\mathfrak{m}^{*2}}{\rho^2 n\varepsilon^2}\right)}, \ \frac{\rho^2}{n\varepsilon^2} \log\left(\frac{ed}{n\varepsilon^2}\right)\right\}, & \text{if } \frac{\mathfrak{m}^*\sqrt{\log d}}{\sqrt{n\varepsilon^2}} \leq \rho \leq \frac{\mathfrak{m}^*d}{\sqrt{n\varepsilon^2}}, \end{cases}$$

(B.2)

is a strict upper bound on $r^2(\rho)$ defined in (5.3). By arguing as in the discussion above, the smallest solution of the sparsity equation is of the form

$$\rho^* = C^*_{\gamma_P,\gamma_M} \mathfrak{m}^* s^* \sqrt{\frac{1}{n\varepsilon^2} \log\left(\frac{ed}{s^*}\right)}, \quad r^2_\varepsilon(\rho^*) = C^{*2}_{\gamma_P,\gamma_M} \frac{\mathfrak{m}^{*2} s^*}{n\varepsilon^2} \log\left(\frac{ed}{s^*}\right).$$

For any fixed constant $C > 0$, let K^* be the smallest integer such that

$$K^* \geq \frac{n\varepsilon^2}{C^2 \mathfrak{m}^{*2}} r^2_\varepsilon(\rho^*),$$

this matches definition (3.6) in Theorem 3.3 with $C^2 = 384\theta_1^2$ and $r = r_\varepsilon$. By definition, this is equivalent to

$$K^* \geq \frac{C^{*2}_{\gamma_P,\gamma_M}}{C^2} s \log\left(\frac{ed}{s}\right),$$

which gives the heuristic that the minimum number of blocks is of order $K^* \sim s \log(ed/s)$. For any integer $K \geq K^*$, we compute the radii ρ_K solving

$$K = \frac{n\varepsilon^2}{C^2 \mathfrak{m}^{*2}} r^2_\varepsilon(\rho_K),$$

which is a rearrangement of definition (3.7) in Theorem 3.3. For all $\rho^* \leq \rho_K \lesssim \mathfrak{m}^* \sqrt{n\varepsilon^2}$, we have

$$r^2_\varepsilon(\rho_K) = C^2_{\gamma_P,\gamma_M} \rho_K \mathfrak{m}^* \sqrt{\frac{1}{n\varepsilon^2} \log\left(\frac{ed^2 \mathfrak{m}^{*2}}{\rho_K^2 n\varepsilon^2}\right)},$$

and the implicit solutions ρ_K are of the form

$$\rho_K = C_K K \mathfrak{m}^* \sqrt{\frac{1}{n\varepsilon^2} \left[\log\left(\frac{ed^2}{K^2}\right)\right]^{-1}},$$

with C_K some absolute constant, for all $K \lesssim n\varepsilon^2$. In fact, let us compute

$$\frac{n\varepsilon^2}{K \mathfrak{m}^{*2}} r^2_\varepsilon(\rho_K) = C^2_{\gamma_P,\gamma_M} C_K \sqrt{\left[\log\left(\frac{ed^2}{K^2}\right)\right]^{-1} \log\left(\frac{ed^2}{C_K^2 K^2} \log\left(\frac{ed^2}{K^2}\right)\right)}$$

$$= C^2_{\gamma_P,\gamma_M} C_K \sqrt{\frac{\log\left(\frac{ed^2}{K^2}\right) + \log\log\left(\frac{ed^2}{K^2}\right) - \log\left(C_K^2\right)}{\log\left(\frac{ed^2}{K^2}\right)}},$$

which we want to be equal to the given C^2. Since $d \gg n$ and $K \lesssim n\varepsilon^2$, without loss of generality $C_K^2 \ll d/n$, thus

$$\frac{1}{2} < 1 - \frac{\log\left(C_K^2\right)}{\log\left(\frac{ed^2}{K^2}\right)} < \frac{\log\left(\frac{ed^2}{K^2}\right) + \log\log\left(\frac{ed^2}{K^2}\right) - \log\left(C_K^2\right)}{\log\left(\frac{ed^2}{K^2}\right)} < 2 - \frac{\log\left(C_K^2\right)}{\log\left(\frac{ed^2}{K^2}\right)} < 2,$$

which allows for an absolute constant $C_K \in [C^2_{\gamma_P,\gamma_M}/(\sqrt{2}C^2), \sqrt{2}C^2_{\gamma_P,\gamma_M}/C^2]$ recovering the solution.

As mentioned earlier, we can write $K^* = \lceil \widetilde{c}s^* \log(ed/s^*) \rceil$ with $\widetilde{c} = C^{*2}_{\gamma_P, \gamma_M}/(384\theta_1^2)$ and, without loss of generality, $\widetilde{c} \geq 1$. Assume that the number of outliers is smaller than $\widetilde{c}_0 s^* \log(ed/s^*)$ with $\widetilde{c}_0 = \widetilde{c}/32$, this results in $32|\mathcal{O}| \leq K^*$ and the choice $K = K^*$ is valid in Theorem 3.3. Then set $\widetilde{c}_2 = 2\widetilde{c}$ and apply Theorem 3.3 separately for any choice $K = \lceil \iota_K \widetilde{c}_2 s^* \log(ed/s^*) \rceil$ for all $\iota_K \in [1/2, 2]$. Then, for any $\iota_\mu \in [1/4, 4]$, any penalization parameter of the form

$$\mu = \iota_\mu c_\mu \varepsilon \frac{r_\varepsilon^2(\rho_K)}{\mathfrak{m}^* \rho_K} = \iota_\mu c_\mu C^2_{\gamma_P, \gamma_M} \varepsilon \sqrt{\frac{1}{n\varepsilon^2} \log\left(\frac{ed^2 \mathfrak{m}^{*2}}{\rho_K^2 n\varepsilon^2}\right)} = \iota_\mu \widetilde{c}_\mu \sqrt{\frac{1}{n} \log\left(\frac{ed^2}{K^2}\right)},$$

with universal constant $\widetilde{c}_\mu = c_\mu C^2_{\gamma_P, \gamma_M}$, is a compatible choice. Furthermore, one finds

$$\mu = \iota_\mu c_\mu C^2_{\gamma_P, \gamma_M} \sqrt{\frac{1}{n} \left(\log\left(\frac{ed^2}{s^{*2}}\right) - 2\log\log\left(\frac{ed}{s^*}\right) - 2\log(\iota_K \widetilde{c}_2) \right)}.$$

We observe that, since $\iota_K \widetilde{c}_2 \geq 1$,

$$\log\left(\frac{ed^2}{s^2}\right) - 2\log\log\left(\frac{ed}{s^*}\right) - 2\log(\iota_K \widetilde{c}_2) \leq \log\left(\frac{ed^2}{s^{*2}}\right),$$

and, with $\log(ed/s^*) \leq (\sqrt{e}d/s^*)^{1/2}$ and $\iota_K \widetilde{c}_2 \leq (ed/s^*)^{1/4}$,

$$\log\left(\frac{ed^2}{s^{*2}}\right) - 2\log\log\left(\frac{ed}{*}\right) - 2\log(\iota_K \widetilde{c}_2) \geq \frac{1}{2}\log\left(\frac{ed^2}{s^{*2}}\right) - 2\log(\iota_K \widetilde{c}_2) \geq \frac{1}{4}\log\left(\frac{ed^2}{s^{*2}}\right).$$

Therefore, any penalization parameter in the smaller interval

$$\mu \in \left[\frac{1}{2}\widetilde{c}_\mu \sqrt{\frac{1}{n} \log\left(\frac{ed^2}{s^{*2}}\right)}, 2\widetilde{c}_\mu \sqrt{\frac{1}{n} \log\left(\frac{ed^2}{s^{*2}}\right)}\right],$$

with absolute constant $\widetilde{c}_\mu = c_\mu C^2_{\gamma_P, \gamma_M}$, is valid. This matches the construction required by Theorem B.1 for any $\iota_K, \iota_\mu \in [1/2, 2]^2$ and shows that the penalization parameter μ can be chosen without knowledge of the moments of the noise.

The convergence rates in Theorem 3.3 become

$$|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1 \leq 2\rho_K = 2C_K \varepsilon^{-1} \mathfrak{m}^* K \sqrt{\frac{1}{n}\left[\log\left(\frac{ed^2}{K^2}\right)\right]^{-1}},$$

$$|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2 \leq r_\varepsilon(2\rho_K) \leq 2C\varepsilon^{-1} \mathfrak{m}^* \sqrt{\frac{K}{n}},$$

$$|\widehat{\sigma}_{K,\mu} - \sigma^*| \leq c_\alpha r_\varepsilon(2\rho_K) \leq 2c_\alpha C \varepsilon^{-1} \mathfrak{m}^* \sqrt{\frac{K}{n}}.$$

Finally, for $K \simeq K^*$, one gets

$$|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1 \leq 2\rho_{K^*} \lesssim 2C^*_{\gamma_P, \gamma_M} \varepsilon^{-1} \mathfrak{m}^* s^* \sqrt{\frac{1}{n} \log\left(\frac{ed}{s^*}\right)},$$

$$|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2 \leq r_\varepsilon(2\rho_{K^*}) \lesssim 2C^*_{\gamma_P, \gamma_M} \varepsilon^{-1} \mathfrak{m}^* \sqrt{\frac{s^*}{n} \log\left(\frac{ed}{s^*}\right)},$$

$$|\widehat{\sigma}_{K,\mu} - \sigma^*| \leq c_\alpha r(2\rho_{K^*}) \lesssim 2c_\alpha C^*_{\gamma_P, \gamma_M} \varepsilon^{-1} \mathfrak{m}^* \sqrt{\frac{s^*}{n} \log\left(\frac{ed}{s^*}\right)}.$$

The bounds in (B.1) for $p \in [1, 2]$ are obtained by applying the interpolation inequality $|\boldsymbol{\beta}|_p \leq |\boldsymbol{\beta}|_1^{-1+2/p} |\boldsymbol{\beta}|_2^{2-2/p}$. This concludes the proof.

B.3 Proof of Corollary 4.6

Recall the definition of signal-to-noise ratio

$$SNR := \frac{\mathrm{Var}(f^*)}{\mathrm{Var}(\zeta)} = \frac{\mathrm{Var}(f^*)}{\sigma^{*2}},$$

and denote

$$A_Y^2 := \frac{\mathrm{Var}(Y^2)}{\mathrm{Var}(Y)^2}, \quad B_Y^2 := \frac{\mathbb{E}[Y]^2}{\mathrm{Var}(Y)}.$$

The following proposition allows us to bound above and below the estimator $\widehat{\sigma}_{K,+}$ on an event with high probability.

Proposition B.2. *Assume that* $\mathrm{Var}(Y) > 0$ *and consider the quantities* A_Y, B_Y *defined above. For any integer*

$$K \in \left[8|\mathcal{O}|, \ \frac{n\varepsilon^2}{C^2} \wedge \frac{n}{177A_Y^2} \wedge \frac{n}{706B_Y^2} \right],$$

there exists an event $\Omega(K)$ *with probability at least* $1 - 2\exp(-7K/3600)$ *such that, on this event, the estimator*

$$\widehat{\sigma}_{K,+}^2 := Q_{1/2,K}\left[Y^2\right] - \left(Q_{1/2,K}\left[Y\right]\right)^2,$$

satisfies $\sigma^{*2} \leq 8\widehat{\sigma}_{K,+}^2 \leq 16\sigma^{*2}(SNR + 1)$.

Combining Proposition B.2 and Theorem 4.4 by replacing σ_+ by $\widehat{\sigma}_{K,+}$ and reasoning on the intersection of both events yields the conclusion.

We now prove Proposition B.2.

Proof. We start with

$$\mathrm{Var}(Y) = \mathrm{Var}(f^*(\mathbf{X}) + \zeta) = \mathrm{Var}(f^*(\mathbf{X})) + \sigma^{*2} + 2\,\mathrm{Cov}(f^*(\mathbf{X}), \zeta) = \mathrm{Var}(f^*(\mathbf{X})) + \sigma^{*2},$$

where in the last step we have used that $f^*(\mathbf{X}) = \mathbf{X}^\top \boldsymbol{\beta}^*$ is the orthogonal projection of the square-integrable random variable $Y = \mathbf{X}^\top \boldsymbol{\beta}^* + \zeta$ onto the closed and convex set of square-integrable random variables $\mathcal{A} := \{\mathbf{X}^\top \boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^d\}$. Thus, $\mathrm{Var}(Y) = \sigma^{*2}(SNR + 1)$.

We apply Lemma D.3 to the variable $Z = Y^2$. We choose $\eta = 1/2$ and $\gamma = 7/8$, $x = 1/15$, $\delta_{K,n}^2 = a_{K,n}^2 := 15(K/n)\,\mathrm{Var}(Y^2)$, so that $\gamma(1 - 1/15 - x) \geq 1/2$, in fact

$$\gamma\left(1 - \frac{1}{15} - x\right) = \frac{7}{8}\left(1 - \frac{1}{15} - \frac{1}{15}\right) = \frac{91}{120} > \frac{1}{2}.$$

Therefore, on an event $\Omega_1(K)$ with probability at least $1 - \exp(-7K/3600)$, we have $Q_{1/2,K}\left[Y^2\right] \in [\mathbb{E}[Y^2] - a_{K,n}, \mathbb{E}[Y^2] + a_{K,n}]$.

We now repeat the argument for $Z = Y$. We choose again $\eta = 1/2$ and $\gamma = 7/8$, $x = 1/15$, $\delta_{K,n}^2 = b_{k,n}^2 := 15(K/n)\,\mathrm{Var}(Y)$, so that $\gamma(1 - 1/15 - x) \geq 1/2$. Therefore, on an event $\Omega_2(K)$ with probability at least $1 - \exp(-7K/3600)$, we have $(Q_{1/2,K}\,[Y])^2 \in [(\mathbb{E}[Y] - b_{K,n})^2, (\mathbb{E}[Y] + b_{K,n})^2]$.

We now work on the event $\Omega(K) = \Omega_1(K) \cap \Omega_2(K)$ which has probability at least $1 - 2\exp(-7K/3600)$. We have

$$\widehat{\sigma}_{K,+}^2 \in \left[\, \mathrm{Var}(Y) - a_{K,n} - 2\mathbb{E}[Y]b_{K,n} - b_{K,n}^2, \ \mathrm{Var}(Y) + a_{K,n} + 2\mathbb{E}[Y]b_{K,n} - b_{K,n}^2 \right],$$

with $a_{K,n}^2 = 15(K/n)\,\mathrm{Var}(Y^2)$, $b_{K,n}^2 = 15(K/n)\,\mathrm{Var}(Y)$. We now show that

$$\frac{\sigma^{*2}}{4} \leq 2\widehat{\sigma}_{K,+}^2 \leq 4\,\mathrm{Var}(Y),$$

which would give the claim. We start with the lower bound, we want

$$1 \leq \frac{2\,\mathrm{Var}(Y) - 2a_{K,n} - 4\mathbb{E}[Y]b_{K,n} - 2b_{K,n}^2}{\sigma^{*2}/4},$$

and we show the stronger

$$\max\left\{ \frac{2a_{K,n}}{\sigma^{*2}/4}, \ \frac{4\mathbb{E}[Y]b_{K,n}}{\sigma^{*2}/4}, \ \frac{2b_{K,n}^2}{\sigma^{*2}/4} \right\} \leq \frac{1}{3}\left(\frac{2\,\mathrm{Var}(Y)}{\sigma^{*2}/4} - 1 \right).$$

By construction, we have

$$\frac{8a_{K,n}}{\sigma^{*2}} = \frac{\sqrt{\mathrm{Var}(Y^2)}}{\sigma^{*2}}\sqrt{\frac{960K}{n}},$$

$$\frac{16\mathbb{E}[Y]b_{K,n}}{\sigma^{*2}} = \frac{\mathbb{E}[Y]\sqrt{\mathrm{Var}(Y)}}{\sigma^{*2}}\sqrt{\frac{3840K}{n}},$$

$$\frac{8b_{K,n}^2}{\sigma^{*2}} = \frac{\mathrm{Var}(Y)}{\sigma^{*2}}\frac{120K}{n},$$

and the quantities A_Y, B_Y are defined in such a way that $\sqrt{\mathrm{Var}(Y^2)} = A_Y\,\mathrm{Var}(Y)$ and $\mathbb{E}[Y] = B_Y\sqrt{\mathrm{Var}(Y)}$. Therefore, it is enough that

$$A_Y(SNR + 1)\sqrt{\frac{8640K}{n}} \leq 8(SNR + 1) - 1,$$

$$B_Y(SNR + 1)\sqrt{\frac{34560K}{n}} \leq 8(SNR + 1) - 1,$$

$$(SNR + 1)\frac{360K}{n} \leq 8(SNR + 1) - 1.$$

We now divide by $(SNR + 1)$ and use $1/(SNR + 1) \leq 1$, the stronger condition

$$A_Y\sqrt{\frac{8640K}{n}} \leq 7,$$

$$B_Y\sqrt{\frac{34560K}{n}} \leq 7,$$

$$\frac{360K}{n} \leq 7,$$

is then satisfied if $K \le n/\max\{177A_Y^2,\ 706B_Y^2,\ 52\}$, which is true by assumption on the upper bound on the number of blocks. This completes the proof of $\sigma^{*2} \le 8\widehat{\sigma}_{K,+}^2$ on the event $\Omega(K)$.

We now deal with $2\widehat{\sigma}_{K,+}^2 \le 4\operatorname{Var}(Y)$. Since the quantity $-b_{K,n}^2$ is negative, it is sufficient that $2\operatorname{Var}(Y) + 2a_{K,n} + 2\mathbb{E}[Y]b_{K,n} \le 2\operatorname{Var}(Y)$ and, dividing by σ^{*2},

$$\frac{2a_{K,n}}{\sigma^{*2}} + \frac{2\mathbb{E}[Y]b_{K,n}}{\sigma^{*2}} \le \frac{2\operatorname{Var}(Y)}{\sigma^{*2}}.$$

We show the stronger inequalities

$$\frac{2a_{K,n}}{\sigma^{*2}} \le \frac{\operatorname{Var}(Y)}{\sigma^{*2}},$$
$$\frac{2\mathbb{E}[Y]b_{K,n}}{\sigma^{*2}} \le \frac{\operatorname{Var}(Y)}{\sigma^{*2}},$$

by arguing as for the previous step. It is sufficient that

$$A_Y(SNR+1)\sqrt{\frac{60K}{n}} \le (SNR+1),$$

$$B_Y(SNR+1)\sqrt{\frac{60K}{n}} \le (SNR+1),$$

which holds if $K \le n/\max\{60A_Y^2,\ 60B_Y^2\}$, and the latter is true by assumption on the upper bound on the number of blocks. This completes the proof of $2\widehat{\sigma}_{K,+}^2 \le 4\operatorname{Var}(Y)$ on the event $\Omega(K)$. $\qquad\square$

Appendix C Proofs for adaptivity to the sparsity level s

C.1 A general algorithm for simultaneous adaptivity

In this section, we prove a more general theorem, that will yield Theorem 4.7 as a particular case.

Algorithm for adaptation to sparsity. The steps of the adaptive procedure are as follows.

- Let w_1, w_2, w_3 be three functions $[1, d/e] \to \mathbb{R}_+$ and set $M := \lfloor \log_2(s_+) \rfloor$.

- For every $m \in \{1, \ldots, M+1\}$, compute $(\widehat{\boldsymbol{\beta}}_{(2^m)}, \widehat{\sigma}_{(2^m)})$.

- Set

$$\mathcal{M} := \left\{ m \in \{1, \ldots, M\} : \text{for all } k \ge m,\ |\widehat{\boldsymbol{\beta}}_{(2^{k-1})} - \widehat{\boldsymbol{\beta}}_{(2^k)}|_1 \le C_1\widehat{\sigma}w_1(2^k), \right.$$

$$\left. |\widehat{\boldsymbol{\beta}}_{(2^{k-1})} - \widehat{\boldsymbol{\beta}}_{(2^k)}|_2 \le C_2\widehat{\sigma}w_2(2^k) \text{ and } |\widehat{\sigma}_{(2^{k-1})} - \widehat{\sigma}_{(2^k)}| \le C_3\widehat{\sigma}w_3(2^k) \right\}.$$

- Set $\widetilde{m} := \min \mathcal{M}$, with the convention that $\widetilde{m} := M + 1$ if $\mathcal{M} = \emptyset$.

- Define $\widetilde{s} := 2^{\widetilde{m}}$ and $(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}) := (\widehat{\boldsymbol{\beta}}_{(\widetilde{s})}, \widehat{\sigma}_{(\widetilde{s})})$.

Definition C.1. *Let Θ be a subset of $\mathbb{R}^d \times \mathbb{R}_+$ and $\| \cdot \|$ a norm on Θ. For a given $s \in \{2, \ldots, d/(2e)\}$, we say that an estimator $\widehat{\theta}_{(s)} \in \Theta$ robustly converges to $\theta^* \in \Theta$ in norm $\| \cdot \|$ with bound $C_1 \sigma^* w(s)$ if*

$$\inf_{\boldsymbol{\beta}^* \in \mathcal{F}_s, \sigma^* > 0} P_{\boldsymbol{\beta}^*, P_{\mathbf{X}, \zeta}}^{\otimes n} \left(\forall \mathcal{D}' \in \mathcal{D}(N), \|\widehat{\theta}_{(s)}(\mathcal{D}') - \theta^*\| \leq C_1 \sigma^* w(s) \right) \geq 1 - \widetilde{c}_6 C_2 \left(\frac{s}{ed} \right)^{\widetilde{c}_5 s} - u_n,$$

(C.1)

$$\inf_{\boldsymbol{\beta}^* \in \widetilde{\mathcal{F}}_{2s}, \sigma^* > 0} P_{\boldsymbol{\beta}^*, P_{\mathbf{X}, \zeta}}^{\otimes n} \left(\forall \mathcal{D}' \in \mathcal{D}(N), \|\widehat{\theta}_{(s)}(\mathcal{D}') - \theta^*\| \leq C_1 \sigma^* w(s) \right) \geq 1 - \widetilde{c}_6 C_2 \left(\frac{2s}{ed} \right)^{2\widetilde{c}_5 s} - u_n.$$

(C.2)

and if the function $w(\cdot) : [1, d/e] \to \mathbb{R}_+$ satisfies the following conditions:

1. *$w(\cdot)$ is increasing on $[1, d/e]$;*

2. *There exists a constant $C' > 0$ such that, for all $m = 1, \ldots, \lfloor \log_2(s_+) \rfloor$, we have*

$$\sum_{k=1}^{m} w(2^k) \leq C' \cdot w(2^m) ;$$

3. *There exists a constant $C'' > 0$ such that, for all $b = 1, \ldots, s_+$,*

$$w(2b) \leq C'' w(b).$$

Theorem C.2 (Joint adaptation of $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})$ to s). *Let $s_+ \in \{2, \ldots, d/(2e)\}$ and for $s = 1, \ldots, 2s_+$, let $(\widehat{\boldsymbol{\beta}}_{(s)}, \widehat{\sigma}_{(s)})$ be a joint estimator of $(\boldsymbol{\beta}^*, \sigma^*)$ such that*

1. *$\widehat{\boldsymbol{\beta}}_{(s)}$ robustly converges to $\boldsymbol{\beta}^*$ in $| \cdot |_1$-norm with bound $C_1 \sigma^* w_1(s)$;*

2. *$\widehat{\boldsymbol{\beta}}_{(s)}$ robustly converges to $\boldsymbol{\beta}^*$ in $| \cdot |_2$-norm with bound $C_2 \sigma^* w_2(s)$;*

3. *$\widehat{\sigma}_{(s)}$ robustly converges to σ^* in $| \cdot |$-norm with bound $C_3 \sigma^* w_3(s)$;*

for some constants $N > 0$, $\widetilde{c}_6 > 0$ $C_1 > 0$, $u_n > 0$ and for some functions w_1, w_2, w_3 such that $C_3 w_3(2s_+) \leq 1/2$. Then, there exists constants $\widetilde{C}_1, \widetilde{C}_2, \widetilde{C}_3$ such that, for all $s^* \in \{1, \ldots, s_+\}$ and $\boldsymbol{\beta}^* \in \widetilde{\mathcal{F}}_{s^*}$, the aggregated estimator $(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}, \widetilde{s})$ satisfies

$$P_{\boldsymbol{\beta}^*, P_{\mathbf{X}, \zeta}}^{\otimes n} \left(\forall \mathcal{D}' \in \mathcal{D}(N), |\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1 \leq \widetilde{C}_1 \sigma^* w_1(s^*), |\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2 \leq \widetilde{C}_2 \sigma^* w_2(s^*), |\widetilde{\sigma} - \sigma^*| \leq \widetilde{C}_3 \sigma^* w_3(s^*) \right)$$

$$\geq 1 - 21(\log_2(s_+) + 1)^2 \left(\widetilde{c}_5 \left(\frac{2s^*}{d} \right)^{2\widetilde{c}_6 s^*} + u_n \right) - 21\widetilde{c}_6 \left(\frac{2^{M+1}}{d} \right)^{\widetilde{c}_5 2^{M+1}} - 21 u_n$$

and

$$\mathbb{P}_{\boldsymbol{\beta}^*} \left(\forall \mathcal{D}' \in \mathcal{D}(N), \widetilde{s} \leq s^* \right) \geq 1 - 6(\log_2(s_+) + 1)^2 \left(\widetilde{c}_6 \left(\frac{2s^*}{d} \right)^{2\widetilde{c}_5 s^*} + u_n \right) - 6\widetilde{c}_6 \left(\frac{2^{M+1}}{d} \right)^{\widetilde{c}_5 2^{M+1}} - 6 u_n.$$

We adapt the proof given in [8, Section 7.3.1] to this new setting where the adaptation is done on both estimators simultaneously. Proof of Theorem C.2 is given in Section C.3.

C.2   Proof of Theorem 4.7

To prove Theorem 4.7, we will apply Theorem C.2. We first check that its assumption are satisfied. We choose the functions $w_1(s) = s\sqrt{(1/n)\log(ed/s)}$, $w_2(s) = w_3(s) = w_1(s) = s^{1/2}\sqrt{(1/n)\log(ed/s)}$. By Lemma 4.4 in [8], w_1, w_2 and w_3 satisfy the 3 conditions in Definition C.1.

It remains to check that the following bounds in probability (C.1) and (C.2) hold for all $s^* = 1, \ldots, s_+$. Applying Theorem 4.4 gives

$$\inf_{\boldsymbol{\beta}^* \in \mathcal{F}_{s^*}, \sigma^* > 0} P^{\otimes n}_{\boldsymbol{\beta}^*, P_{\mathbf{X},\zeta}} \left(\sup_{\mathcal{D}' \in \mathcal{D}(\widetilde{c}_3 \mathfrak{r}_{\mathcal{O}})} \left\{ \mathfrak{r}_2^{-1} |\widehat{\sigma}(\mathcal{D}') - \sigma^*| \vee \sup_{p \in [1,2]} \mathfrak{r}_p^{-1} |\widehat{\boldsymbol{\beta}}(\mathcal{D}') - \boldsymbol{\beta}^*|_p \right\} \le \widetilde{c}_4 \sigma_+ \right) \ge 1 - 4\left(\frac{s^*}{ed}\right)^{\widetilde{c}_5 s^*},$$

proving that the bound (C.1) is satisfied.

Furthermore, we have

$$K_{2s} = \left\lceil \widetilde{c}_2 2s^* \log\left(\frac{ed}{2s^*}\right) \right\rceil = \left\lceil \widetilde{c}_2 2s^* \left(\log\left(\frac{ed}{s^*}\right) + \log(2) \right) \right\rceil = \gamma(2s^*) K_{s^*},$$

$$\mu_{2s^*} = \widetilde{c}_\mu \sqrt{\frac{1}{n} \log\left(\frac{ed}{2s^*}\right)} = \widetilde{c}_\mu \sqrt{\frac{1}{n} \log\left(\frac{ed}{s^*}\right) - \frac{\log(2)}{n}} = \widetilde{\gamma}(2s^*) \mu_s,$$

with some $\gamma(2s^*), \widetilde{\gamma}(2s^*) \in [1/2, 2]^2$. This gives $\widehat{\boldsymbol{\beta}}_{K_{2s^*}/\gamma(2s^*), \mu_{2s^*}/\widetilde{\gamma}(2s^*)} = \widehat{\boldsymbol{\beta}}_{K_{s^*}, \mu_{s^*}}$ and, applying Theorem 4.4 with $2s^*$ instead of s^*, yields

$$\inf_{\boldsymbol{\beta}^* \in \mathcal{F}_{2s^*}, \sigma^* > 0} P^{\otimes n}_{\boldsymbol{\beta}^*, P_{\mathbf{X},\zeta}} \left(\forall \mathcal{D}' \in \mathcal{D}(\widetilde{c}_3 \mathfrak{r}_{\mathcal{O}}), \left\{ |\widehat{\sigma}(\mathcal{D}') - \sigma^*| \le \widetilde{c}_4 \sigma_+ \sqrt{\frac{2s^*}{n} \log\left(\frac{ed}{2s^*}\right)} \right. \right.$$

$$\left. \left. \text{and } \forall p \in [1,2], |\widehat{\boldsymbol{\beta}}(\mathcal{D}') - \boldsymbol{\beta}^*|_1 \le \widetilde{c}_4 \sigma_+ (2s^*)^{1/p} \sqrt{\frac{1}{n} \log\left(\frac{ed}{2s^*}\right)} \right) \ge 1 - 4\left(\frac{2s^*}{ed}\right)^{\widetilde{c}_5 2s^*}, \right.$$

proving that the bound (C.2) is satisfied with $\widetilde{c}_4$ multiplied by 4.

C.3   Proof of Theorem C.2

We choose $s \in [1, s_+]$ and assume that $\boldsymbol{\beta}^* \in \mathcal{F}_s$. Define $\mathbb{P} := \mathbb{P}_{\boldsymbol{\beta}^*, \sigma^*}$ and $m_0 := \lfloor \log_2(s) \rfloor + 1$. For $p = 1, 2$, define $\widehat{\theta}^{(p)}_{(s)} := \widehat{\boldsymbol{\beta}}_{(s)}$, $\widetilde{\theta}^{(p)} := \widetilde{\boldsymbol{\beta}}$, $\theta^{(p),*} := \boldsymbol{\beta}^*$ and d_p be the distance on $\mathbb{R}$ induced by the norm $|\cdot|_p$. Define $\widehat{\theta}^{(3)}_{(s)} = \widehat{\sigma}_{(s)}$, $\widetilde{\theta}^{(3)} := \widetilde{\sigma}$, $\theta^{(3),*} := \sigma^*$ and d_3 be the distance on $\mathbb{R}$ induced by the absolute value.

Bound on $\widehat{\sigma}$ with high probability. Combining the definition $\widehat{\sigma} = \widehat{\sigma}_{2s_+}$ with the assumptions that $C_3 w_3(2s_+) \le 1/2$ and that $\widehat{\sigma}_{(s)}$ robustly converges to σ^* in $|\cdot|$-norm with bound $C_3 \sigma^* w_3(s)$, we get

$$\mathbb{P}\left(\forall \mathcal{D}' \in \mathcal{D}(N), \sigma^*/2 \le \widehat{\sigma} \le (3/2)\sigma^*\right) \ge 1 - \widetilde{c}_6 \left(\frac{2^{M+1}}{d}\right)^{\widetilde{c}_5 2^{M+1}} - u_n \qquad \text{(C.3)}$$

Bound on the probability $\mathbb{P}(\exists \mathcal{D}' \in \mathcal{D}(N), \tilde{m} \geq m_0 + 1)$. We have

$$\mathbb{P}(\exists \mathcal{D}' \in \mathcal{D}(N), \tilde{m} \geq m_0 + 1) \leq \sum_{m=m_0+1}^{M} \mathbb{P}(\exists \mathcal{D}' \in \mathcal{D}(N), \tilde{m} = m_0 + 1)$$

$$\leq \sum_{m=m_0+1}^{M} \sum_{k=m}^{M} \mathbb{P}\left(\exists \mathcal{D}' \in \mathcal{D}(N), |\widehat{\boldsymbol{\beta}}_{(2^{k-1})} - \widehat{\boldsymbol{\beta}}_{(2^k)}|_1 > 4C_1\widehat{\sigma}w_1(2^k)\right.$$

$$\left. \text{or } |\widehat{\boldsymbol{\beta}}_{(2^{k-1})} - \widehat{\boldsymbol{\beta}}_{(2^k)}|_2 > 4C_2\widehat{\sigma}w_2(2^k) \text{ or } |\widehat{\sigma}_{(2^{k-1})} - \widehat{\sigma}_{(2^k)}| > 4C_3\widehat{\sigma}w_3(2^k)\right)$$

$$\leq \sum_{m=m_0+1}^{M} \sum_{k=m}^{M} \mathbb{P}\left(\exists \mathcal{D}' \in \mathcal{D}(N), \exists p \in [3], d_p(\widehat{\theta}^{(p)}_{(2^{k-1})}, \widehat{\theta}^{(p)}_{(2^k)}) > 4C_p\widehat{\sigma}w_p(2^k)\right)$$

$$\leq \sum_{p=1}^{3} \sum_{m=m_0+1}^{M} \sum_{k=m}^{M} \mathbb{P}\left(\exists \mathcal{D}' \in \mathcal{D}(N), d_p(\widehat{\theta}^{(p)}_{(2^{k-1})}, \widehat{\theta}^{(p)}_{(2^k)}) > 4C_p\widehat{\sigma}w_p(2^k)\right)$$

$$\leq \sum_{p=1}^{3} \sum_{m=m_0+1}^{M} \sum_{k=m}^{M} \mathbb{P}\left(\exists \mathcal{D}' \in \mathcal{D}(N), d_p(\widehat{\theta}^{(p)}_{(2^{k-1})}, \theta^{(p),*}) > 4C_p\widehat{\sigma}w_p(2^k)\right)$$

$$+ \mathbb{P}\left(\exists \mathcal{D}' \in \mathcal{D}(N), d_p(\widehat{\theta}^{(p)}_{(2^k)}, \theta^{(p),*}) > 4C_p\widehat{\sigma}w_p(2^k)\right)$$

$$\leq 2\sum_{p=1}^{3} \sum_{m=m_0+1}^{M} \sum_{k=m-1}^{M} \mathbb{P}\left(\exists \mathcal{D}' \in \mathcal{D}(N), d_p(\widehat{\theta}^{(p)}_{(2^{k-1})}, \theta^{(p),*}) > 4C_p\widehat{\sigma}w_p(2^k)\right)$$

$$\leq 2\sum_{p=1}^{3} \sum_{m=m_0+1}^{M} \sum_{k=m-1}^{M} \mathbb{P}\left(\exists \mathcal{D}' \in \mathcal{D}(N), d_p(\widehat{\theta}^{(p)}_{(2^{k-1})}, \theta^{(p),*}) > 4C_p\widehat{\sigma}w_p(2^k), \widehat{\sigma} \geq \frac{\sigma}{2}\right)$$

$$+ 6\,\mathbb{P}\left(\exists \mathcal{D}' \in \mathcal{D}(N), \hat{\sigma} < \frac{\sigma}{2}\right).$$

Combining the previous equation with Equation (C.3), and then with the assumption on the bound on the estimator $\widehat{\theta}^{(p)}_{(2^{k-1})}$ for the distance d_p, we get

$$\mathbb{P}(\exists \mathcal{D}' \in \mathcal{D}(N), \tilde{m} \geq m_0 + 1)$$

$$\leq 2\sum_{p=1}^{3} \sum_{m=m_0+1}^{M} \sum_{k=m-1}^{M} \mathbb{P}\left(\exists \mathcal{D}' \in \mathcal{D}(N), d_p(\widehat{\theta}^{(p)}_{(2^{k-1})}, \theta^{(p),*}) > 2C_p\widehat{\sigma}w_p(2^k)\right)$$

$$- 6\widetilde{c}_6\left(\frac{2^{M+1}}{d}\right)^{\widetilde{c}_5 2^{M+1}} - 6u_n$$

$$\leq 6M^2\widetilde{c}_6\left(\left(\frac{2s}{p}\right)^{2\widetilde{c}_5 s} + u_n\right) - 6\widetilde{c}_6\left(\frac{2^{M+1}}{d}\right)^{2^{M+1}\widetilde{c}_5} - 6u_n$$

$$\leq 6(\log_2(s_+) + 1)^2\widetilde{c}_6\left(\left(\frac{2s}{p}\right)^{2\widetilde{c}_6 s} + u_n\right) - 6\widetilde{c}_6\left(\frac{2^{M+1}}{d}\right)^{\widetilde{c}_5 2^{M+1}} - 6u_n. \qquad (C.4)$$

This gives the bound on $\tilde{s}$ as claimed.

Bound on the deviation probability of $\widetilde{\theta}^{(p)}$. For any $a > 0$, we have

$$\mathbb{P}\big(\exists \mathcal{D}' \in \mathcal{D}(N), d_p(\widetilde{\theta}^{(p)}, \theta^{(p),*}) \geq a\big) \leq \mathbb{P}\big(\exists \mathcal{D}' \in \mathcal{D}(N), d_p(\widetilde{\theta}^{(p)}, \theta^{(p),*}) \geq a, \widetilde{m} \leq m_0\big)$$
$$+ \mathbb{P}(\exists \mathcal{D}' \in \mathcal{D}(N), \widetilde{m} \geq m_0 + 1). \tag{C.5}$$

On the event $\{\widetilde{m} \leq m_0\}$, we have the decomposition

$$d_p(\widetilde{\theta}^{(p)}, \theta^{(p),*}) \leq \sum_{k=\widetilde{m}+1}^{m_0} d_p\left(\widehat{\theta}^{(p)}_{(2^{k-1})}, \widehat{\theta}^{(p)}_{(2^k)}\right) + d_p(\widehat{\theta}^{(p)}_{(2^{m_0})}, \theta^{(p),*}). \tag{C.6}$$

Using the assumption on the function w_p, we get that,

$$\sum_{k=\widetilde{m}+1}^{m_0} d_p\left(\widehat{\theta}^{(p)}_{(2^{k-1})}, \widehat{\theta}^{(p)}_{(2^k)}\right) \leq \sum_{k=\widetilde{m}+1}^{m_0} 4\hat{\sigma} C_0 w(2^k)$$
$$\leq 4\hat{\sigma} C_p C' w_p(2^{m_0}) \leq 4\hat{\sigma} C_p C' C'' w_p(s). \tag{C.7}$$

We have $2^{m_0} \leq 2s$, therefore applying Assumption (C.2), we have with $\mathbb{P}_{\beta^*, \sigma^*}$-probability at least $1 - \widetilde{c}_5 (2s/p)^{2\widetilde{c}_6 s} - u_n$, for all $\mathcal{D}' \in \mathcal{D}(N)$,

$$d_p(\widehat{\theta}^{(p)}_{(2^{m_0})}, \theta^{(p),*}) \leq C_p\hat{\sigma} w(2s) \leq C_p C'' \hat{\sigma} w(s). \tag{C.8}$$

Combining Equations (C.6), (C.7), (C.8) and (C.3), we get with $\mathbb{P}_{\beta^*}$-probability at least $1 - \widetilde{c}_5(2s/p)^{2\widetilde{c}_6 s} - \widetilde{c}_5(2^{M+1}/p)^{\widetilde{c}_6 2^{M+1}} - 2u_n$, for all $\mathcal{D}' \in \mathcal{D}(N)$,

$$d_p(\widetilde{\theta}^{(p)}, \theta^{(p),*}) \leq \left(4C_p C' C'' + (3/2)C_p C''\right) \sigma w(s). \tag{C.9}$$

Combining Equation (C.4) with Equations (C.5) and (C.9), we finally get that

$$\mathbb{P}\left(\exists \mathcal{D}' \in \mathcal{D}(N), d_p(\widetilde{\theta}^{(p)}, \theta^{(p),*}) \geq \left(4C_p C' C'' + (3/2)C_p C''\right) \sigma w_p(s)\right)$$
$$\leq 7(\log_2(s_+) + 1)^2 \left(\widetilde{c}_6 \left(\frac{2s}{p}\right)^{2\widetilde{c}_5 s} + u_n\right) - 7\widetilde{c}_6 \left(\frac{2^{M+1}}{d}\right)^{\widetilde{c}_5 2^{M+1}} - 7u_n.$$

By a union bound, we then obtain

$$\mathbb{P}_{\beta^*, \sigma^*}\left(\forall \mathcal{D}' \in \mathcal{D}(N), \forall p = 1, 2, 3, d_p(\widetilde{\theta}^{(p)}, \theta^{(p),*}) \geq \left(4C'C'' + (3/2)C''\right) C_p \sigma w_p(s)\right)$$
$$\geq 1 - 21(\log_2(s_+) + 1)^2 \left(\widetilde{c}_6 \left(\frac{2s}{d}\right)^{2\widetilde{c}_5 s} + u_n\right) - 21\widetilde{c}_6 \left(\frac{2^{M+1}}{d}\right)^{\widetilde{c}_5 2^{M+1}} - 21u_n.$$

as claimed.

Appendix D Auxiliary results

In this section we give auxiliary results that are used in the proofs of the main results.

Lemma D.1 (Lemma 6 in [15]). *Let $\rho \geq 0$, $\Gamma_{f^*}(\rho) := \bigcup_{f \in \mathcal{F}: \|f-f^*\| \leq \rho/20} (\partial \|\cdot\|)_f$. For all $g \in \mathcal{F}$, we have*

$$\|f^*\| - \|g\| \leq \frac{\rho}{10} - \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(g - f^*).$$

We recall here the definition of quantiles we used in Section 2.4. For any $K \in \mathbb{N}$, set $[K] = \{1, \ldots, K\}$. For all $\alpha \in (0,1)$ the $\alpha-$quantile of a vector $\mathbf{x} = (x_1, \ldots, x_K) \in \mathbb{R}^K$ is any element $Q_\alpha[\mathbf{x}]$ of the set

$$\mathcal{Q}_\alpha[\mathbf{x}] := \Big\{ u \in \mathbb{R}: \ \big|\{k \in [K] : x_k \geq u\}\big| \geq (1-\alpha)K, \ \big|\{k \in [K] : x_k \leq u\}\big| \geq \alpha K \Big\}.$$

For all $t \in \mathbb{R}$, we write $Q_\alpha[\mathbf{x}] \geq t$ when there exists $J \subset [K]$ such that $|J| \geq (1-\alpha)K$ and, for all $j \in J$, $x_j \geq t$. We write $Q_\alpha[\mathbf{x}] \leq t$ if there exists $J \subset [K]$ such that $|J| \geq \alpha K$ and, for all $j \in J$, $x_j \leq t$.

Lemma D.2. *We have the following properties.*

1. **Monotonicity**
   For all $\alpha \in (0,1)$, $\beta \in (0,\alpha]$ and $\mathbf{x} \in \mathbb{R}^K$, $Q_\beta[\mathbf{x}] \leq Q_\alpha[\mathbf{x}]$.

2. **Opposite**
   For all $\alpha \in (0,1)$ and $\mathbf{x} \in \mathbb{R}^K$, $Q_\alpha[\mathbf{x}] \geq -Q_{1-\alpha}[-\mathbf{x}]$.

3. **Linearity**
   For all $\alpha \in (0,1)$, $\mathbf{x} \in \mathbb{R}^K$ and $a,b \in \mathbb{R}$, $Q_\alpha[a\mathbf{x} + b] = |a|Q_\alpha[\mathrm{sgn}(a)\mathbf{x}] + b$.

4. **Difference**
   For all $\alpha, \beta \in (0,1)$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^K$, $Q_\alpha[\mathbf{x} - \mathbf{y}] \leq Q_{\alpha+\beta}[\mathbf{x}] - Q_\beta[\mathbf{y}]$.

5. **Triangular**
   For all $\alpha, \beta \in (0,1)$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^K$, $Q_\alpha[\mathbf{x} + \mathbf{y}] \leq Q_{\alpha+\beta}[\mathbf{x}] + Q_{1-\beta}[\mathbf{y}]$.

Proof of Lemma D.2. We prove property 1. Write $\mathbf{x} = (x_j)_{j \in [K]}$. The property $Q_\beta[\mathbf{x}] \leq Q_\alpha[\mathbf{x}]$ is true by construction, because $Q_\alpha[\mathbf{x}] \leq u$ implies that there are at least $\alpha K \geq \beta K$ components such that $x_j \leq u$.

We prove property 2. Write $\mathbf{x} = (x_j)_{j \in [K]}$ and $Q_\alpha[\mathbf{x}] = u$, then there are at least $(1-\alpha)K$ components such that $x_j \geq u$ and at least αK components such that $x_j \leq u$. We now show that $u \geq -Q_{1-\alpha}[-\mathbf{x}]$. This is equivalent to $Q_{1-\alpha}[-\mathbf{x}] \geq -u$, which requires at least αK components such that $-x_j \geq -u$, that is, $x_j \leq u$. The latter is true by construction.

We prove property 3. Write $\mathbf{x} = (x_j)_{j \in [K]}$. The property $Q_\alpha[a\mathbf{x} + b] = Q_\alpha[a\mathbf{x}] + b$ follows from the definition, that is, if $Q_\alpha[a\mathbf{x}] = u$ then there are at least $(1-\alpha)K$ components such that $ax_j \geq u$ and at least αK components such that $ax_j \leq u$. Thus, the same components also satisfy $ax_j + b \geq u + b$ or $ax_j + b \leq u + b$. It remains to show that $Q_\alpha[a\mathbf{x}] = |a|Q_\alpha[\mathrm{sgn}(a)\mathbf{x}]$. Let $Q_\alpha[a\mathbf{x}] = u$. We show that we have at least $(1-\alpha)K$ components $\mathrm{sgn}(a)x_j \geq u/|a|$ and at least

61

αK components $\mathrm{sgn}(a)x_j \leq u/|a|$. The latter conditions are equivalent to $|a|\,\mathrm{sgn}(a)x_j \geq u$ and $|a|\,\mathrm{sgn}(a)x_j \leq u$. This is enough to conclude since $a = \mathrm{sgn}(a)|a|$ and $Q_\alpha[a\mathbf{x}] = u$.

We prove property 4. Write $\mathbf{x} = (x_j)_{j\in[K]}$, $\mathbf{y} = (y_i)_{i\in[K]}$ and $Q_{\alpha+\beta}[\mathbf{x}] = u$, $Q_\beta[\mathbf{y}] = l$. By construction:

- there are at least $(1-\alpha-\beta)K$ components $x_j \geq u$;

- there are at least $(\alpha+\beta)K$ components $x_j \leq u$;

- there are at least $(1-\beta)K$ components $y_i \geq l$;

- there are at least βK components $y_i \leq l$.

With $(\mathbf{x}-\mathbf{y}) = (x_k - y_k)_{k\in[K]}$, we want to show that $Q_\alpha[\mathbf{x}-\mathbf{y}] \leq u-l$, which means there are αK components $x_k - y_k \leq u-l$. We now count how many times this inequality fails. In order for a component to be $x_k - y_k \geq u-l$, it is necessary that either $x_k \geq u$, which can happen at most $(1-\alpha-\beta)K$ times, or $y_k \leq l$, which can happen at most βK times. Therefore, the inequality $x_k - y_k \geq u-l$ is satisfied by at most $(1-\alpha-\beta)K + \beta K = (1-\alpha)K$ components, leaving at least αK components where $x_k - y_k \leq u-l$. This is enough to conclude.

We prove property 5 as a consequence of property 4 and property 2. $\qquad\square$

In the following, we use the notation $[K] = \{1, \ldots, K\}$ and $[K]_I := \{k \in [K] : B_k \subset \mathcal{I}\}$. We denote by K_I the cardinality of $[K]_I$.

Lemma D.3. *Let $Z = Z(\mathbf{X}, Y)$ be a real-valued random variable. Let $\eta \in (0,1)$ and $\gamma, \delta_{K,n}, x > 0$ such that $\gamma(1 - KVar(Z)/(n\delta_{K,n}^2) - x) \geq \max\{\eta, 1-\eta\}$. Let $K \in [|\mathcal{O}|/(1-\gamma), n]$. There exists an event $\Omega = \Omega(Z, K)$ with $\mathbb{P}(\Omega) \geq 1 - \exp(-K\gamma x^2/2)$ such that, on this event*

$$\left|\{k \in [K] : |\mathbb{P}_{B_k}(Z) - \mathbb{E}[Z]| \leq \delta_{K,n}\}\right| \geq \max\{\eta, 1-\eta\}K,$$

thus the quantiles $Q_\eta[Z], Q_{1-\eta}[Z]$ belong to the interval $[\mathbb{E}[Z] - \delta_{K,n}, \mathbb{E}[Z] + \delta_{K,n}]$.

Proof of Lemma D.3. We have

$$\begin{aligned}
|\{k \in [K] : |\mathbb{P}_{B_k}(Z) - \mathbb{E}[Z]| \leq \delta_{K,n}\}| &\geq \sum_{k\in[K]_I} \mathbf{1}\{|\mathbb{P}_{B_k}(Z) - \mathbb{E}[Z]| \leq \delta_{K,n}\} \\
&= K_I - \sum_{k\in[K]_I} \mathbb{P}_{\mathbf{X}}\{|\mathbb{P}_{B_k}(Z) - \mathbb{E}[Z]| \geq \delta_{K,n}\} \\
&\quad - \sum_{k\in[K]_I} \Big(\mathbf{1}\{|\mathbb{P}_{B_k}(Z) - \mathbb{E}[Z]| \geq \delta_{K,n}\} - \mathbb{P}_{\mathbf{X}}\{|\mathbb{P}_{B_k}(Z) - \mathbb{E}[Z]| \geq \delta_{K,n}\}\Big).
\end{aligned}$$

We bound the second term using Chebychev's inequality

$$\sum_{k\in[K]_I} \mathbb{P}_{\mathbf{X}}\{|\mathbb{P}_{B_k}(Z) - \mathbb{E}[Z]| \geq \delta_{K,n}\} \leq K_I \frac{Var[P_{B_k}(Z) - \mathbb{E}[Z]]}{\delta_{K,n}^2} = K_I \frac{Var[Z]}{|B_k|\delta_{K,n}^2} = K_I \frac{KVar[Z]}{n\delta_{K,n}^2}.$$

We bound the last term using Hoeffding's inequality

$$\sum_{k \in [K]_I} \left(\mathbf{1}\{|\mathbb{P}_{B_k}(Z) - \mathbb{E}[Z]| \geq \delta_{K,n}\} - \mathbb{P}_{\mathbf{X}}\{|\mathbb{P}_{B_k}(Z) - \mathbb{E}[Z]| \geq \delta_{K,n}\} \right) \leq xK_I,$$

on an event $\Omega(Z, K)$ of probability greater than $1 - \exp(-x^2 K_I / 2)$. Combining the previous inequalities, we get that on $\Omega(Z, K)$,

$$|\{k \in [K]_I : |\mathbb{P}_{B_k}(Z) - \mathbb{E}[Z]| \leq \delta_{K,n}\}| \geq K_I \left(1 - \frac{K Var[Z]}{n\delta_{K,n}^2} - x\right) \geq K\gamma \left(1 - \frac{K Var[Z]}{n\delta_{K,n}^2} - x\right),$$

and the last term is bigger than $\max\{\eta, 1-\eta\}K$ by assumption. By definition, this also means that the quantiles $Q_\eta[Z], Q_{1-\eta}[Z]$ belong to the interval $[\mathbb{E}[Z] - \delta_{K,n}, \mathbb{E}[Z] + \delta_{K,n}]$. $\qquad\square$

Lemma D.4. *Let $K \in [16|\mathcal{O}|, n]$. On an event $\Omega(K)$ with probability $\mathbb{P}(\Omega(K)) \geq 1 - \exp(-K/4320)$, the quantiles $Q_{1/8,K}[\zeta^2], Q_{7/8,K}[\zeta^2]$ belong to the interval $[\sigma^{*2} - \delta_{K,n}, \sigma^{*2} + \delta_{K,n}]$, with $\delta_{K,n}$ defined in (A.2).*

Proof of Lemma D.4. We use Lemma D.3 with $\eta = 1/8$, $Z = \zeta^2$, $Var(Z) = \mathbb{E}[\zeta^4] - \mathbb{E}[\zeta^2]^2 = \sigma^{*4}(\kappa^* - 1)$, $\eta = 1/8$, $\gamma = 15/16$, $x = 1/45$, and $\delta_{K,n}^2 \geq 25(K/n) Var(Z)$. Then,

$$\gamma \left(1 - x - \frac{K Var(Z)}{n\delta_{K,n}^2}\right) \geq \frac{15}{16}\left(1 - \frac{1}{45} - \frac{1}{25}\right) = \frac{15}{16} - \frac{7}{120} > \frac{7}{8} = 1 - \eta.$$

The probability of the corresponding event is $\mathbb{P}(\Omega(K)) \geq 1 - \exp(-K\gamma x^2/2) = 1 - \exp(-K/4320)$. $\qquad\square$

Lemma D.5 (Lemma 3 in [15]). *Grant Assumption 3.1. Fix $\eta \in (0,1)$ and $\rho \in (0, +\infty)$. Let $\alpha, \gamma, \gamma_P, x$ be positive real numbers such that $\gamma(1 - \alpha - x - 16\gamma_P \theta_0) \geq 1 - \eta$. Assume that K is an integer in $[|\mathcal{O}|/(1 - \gamma), n\alpha/4\theta_0^2]$. Then, there exists an event $\Omega_Q(K)$ with probability $\mathbb{P}(\Omega_Q(K)) \geq 1 - 4\exp(-K\gamma x^2/2)$ and, on this event: for all $f \in \mathcal{F}$ with $\|f - f^*\| \leq \rho$, if $\|f - f^*\|_{2,\mathbf{X}} \geq r_P(\rho, \gamma_P)$ then*

$$\left|\{k \in [K] : \mathbb{P}_{B_k}(f - f^*)^2 \geq (4\theta_0)^{-2}\|f - f^*\|_{2,\mathbf{X}}^2\}\right| \geq (1 - \eta)K$$

In particular, $Q_{\eta,K}[(f - f^*)^2] \geq (4\theta_0)^{-2}\|f - f^*\|_{2,\mathbf{X}}^2$.

Lemma D.6 (Lemma 4 in [15]). *Grant Assumption 3.1. Fix $\eta \in (0,1)$ and $\rho \in (0, +\infty]$. Let $\alpha, \gamma, \gamma_M, x$ be positive real numbers such that $\gamma(1 - \alpha - x - 8\gamma_M/\varepsilon) \geq 1 - \eta$. Assume that K is an integer in $[|\mathcal{O}|/(1 - \gamma), n]$. Then, there exists an event $\Omega_M(K)$ with probability $\mathbb{P}(\Omega_M(K)) \geq 1 - \exp(-K\gamma x^2/2)$ and, on this event: for all $f \in \mathcal{F}$ with $\|f - f^*\| \leq \rho$,*

$$\left|\{k \in [K] : |(\mathbb{P}_{B_k} - \mathbb{E})(2\zeta(f - f^*)| \leq \alpha_M^2\}\right| \geq (1 - \eta)K,$$

with

$$\alpha_M^2 := \varepsilon \max\left(\frac{16\theta_m^2}{\varepsilon^2 \alpha} \frac{K}{n}, \ r_M^2(\rho, \gamma_M), \ \|f - f^*\|_{2,\mathbf{X}}^2\right).$$

Lemma D.7. *Let $K \in \left[32|\mathcal{O}|, \ n/(372\theta_0^2)\right]$. There exists an event $\Omega(K)$ of probability bigger than $1 - 2\exp(-K/8928)$ such that, for all $\rho \in \{\rho_K, 2\rho_K\}$, and all $f \in \mathcal{F}$ such that $\|f - f^*\| \leq \rho$, we have*

1. *if $\|f - f^*\|_{2,\mathbf{X}} \geq r_P(\rho, \gamma_P)$, then $Q_{1/16,K}\big((f - f^*)^2\big) \geq (4\theta_0)^{-2}\|f - f^*\|_{2,\mathbf{X}}^2$;*

2. $Q_{15/16,K}\big[-2\zeta(f - f^*)\big] \leq \mathbb{E}[-2\zeta(f - f^*)(\mathbf{X})] + \alpha_M^2,$

3. $Q_{1/16,K}[-2\zeta(f - f^*)] \geq \mathbb{E}[-2\zeta(f - f^*)(\mathbf{X})] - \alpha_M^2.$

4. $Q_{15/16,K}\big[2\zeta(f - f^*)\big] \leq \alpha_M^2,$

with

$$\alpha_M^2 := \varepsilon \max\left(\frac{1488\theta_m^2}{\varepsilon^2}\frac{K}{n}, \ r_M^2(\rho, \gamma_M), \ \|f - f^*\|_{2,\mathbf{X}}^2\right), \quad \theta_m = \theta_1 \mathfrak{m}^*.$$

Furthermore, for $r(\cdot)$ as in Theorem 3.3 and $\|f - f^*\|_{2,\mathbf{X}} \leq r(\rho)$, we find $\alpha_M^2 \leq 4\varepsilon r^2(\rho)$.

Proof of Lemma D.7. The first property follows from applying Lemma D.5 with $\eta = 1/16$, $\rho \in \{\rho_K, 2\rho_K\}$, $\alpha = x = 1/93$, $\gamma = 31/32$, $\gamma_P = 1/(1488\theta_0)$ and checking that $\gamma(1 - \alpha - x - 16\gamma_P\theta_0) \geq 1 - \eta$. With our choices, we find

$$\frac{31}{32}\left(1 - \frac{1}{93} - \frac{1}{93} - \frac{16}{1488}\right) = \frac{31}{32}\left(1 - \frac{1}{31}\right) = \frac{30}{32} = \frac{15}{16}.$$

The corresponding event Ω_1 has probability at least $1 - \exp(-K\gamma x^2/2) = 1 - \exp(-K/8928)$.

The second and third properties follow from applying Lemma D.6 with $\eta = 1/16$, $\rho \in \rho_K, 2\rho_K$, $\alpha = x = 1/93$, $\gamma = 31/32$, $\gamma_M = \varepsilon/744$ and checking that $\gamma(1 - \alpha - x - 8\gamma_M/\varepsilon) \geq 1 - \eta$. With our choices, we find

$$\frac{31}{32}\left(1 - \frac{1}{93} - \frac{1}{93} - \frac{8}{744}\right) = \frac{31}{32}\left(1 - \frac{1}{31}\right) = \frac{30}{32} = \frac{15}{16}.$$

The corresponding event Ω_2 has probability at least $1 - \exp(-K\gamma x^2/2) = 1 - \exp(-K/8928)$.

The fourth property holds on the same event Ω_2 given above, and is a consequence of the nearest point theorem and the convexity of the function class $\mathcal{F}$, which guarantee that $\mathbb{E}[2\zeta(f - f^*)(\mathbf{X})] \leq 0$.

Given all the above, the probability of the event $\Omega(K) = \Omega_1 \cap \Omega_2$ is at least $1 - \mathbb{P}(\Omega_1) - \mathbb{P}(\Omega_1) = 1 - 2\exp(-K/8928)$.

We finally bound, with $r^2(\rho_K) = 384\theta_m^2 K/(n\varepsilon^2)$,

$$\frac{\alpha_M^2}{r^2(2\rho_K)} \leq \frac{\alpha_M^2}{r^2(\rho_K)} = \varepsilon \max\left(\frac{1488\theta_m^2}{\varepsilon^2}\frac{K}{n}\frac{1}{r^2(\rho_K)}, \ 1\right) = \varepsilon\frac{1488}{384} < 4\varepsilon.$$

$\square$

64

Lemma D.8. *Let $K \in [32|\mathcal{O}|, n/(372\theta_0^2)]$. There exists an event $\Omega_Q(K)$ of probability bigger than $1 - \exp(-K/8928)$ such that, for all $\rho \in \{\rho_K, 2\rho_K\}$, and all $f \in \mathcal{F}$ such that $\|f - f^*\| \leq \rho$, we have*

$$Q_{15/16,K}\left[(f - f^*)^2\right] \leq \|f - f^*\|_{2,\mathbf{X}}^2 + \alpha_Q^2,$$

with

$$\alpha_Q^2 := \varepsilon \max\left(\|f - f^*\|_{2,\mathbf{X}}^2 \frac{1488\theta_1^4}{\varepsilon^2} \frac{K}{n}, \ r_Q^2(\rho, \gamma_Q), \ \|f - f^*\|_{2,\mathbf{X}}^2\right).$$

Proof of Lemma D.8. Take $\eta = 1/16$, $\gamma = 31/32$, $\alpha = x = 1/93$ and $\gamma_Q = \varepsilon/372$. We follow the steps of the proof of Lemma 4 in [15]. For all $f \in \mathcal{F}$ and $\rho > 0$, set $\mathbb{B}(f, \rho) = \{g \in \mathcal{F} : \|g - f\| \leq \rho\}$. For all $k \in [K]$, set $\mathcal{D}_k = (\mathbf{X}_i, Y_i)_{i \in B_k}$ and

$$g_f(\mathcal{D}_k) := (\mathbb{P}_{B_k} - \mathbb{E})[(f - f^*)^2],$$

$$\alpha_Q^2(f) := \varepsilon \max\left(\|f - f^*\|_{2,\mathbf{X}}^2 \frac{4\theta_1^4}{\varepsilon^2 \alpha} \cdot \frac{K}{n}, r_Q^2(\rho, \gamma_Q), \|f - f^*\|_{2,\mathbf{X}}^2\right).$$

Let $[K]_I = \{k \in [K] : B_k \subset \mathcal{I}\}$ and consider any $k \in [K]_I$. An application of Markov inequality gives

$$\mathbb{P}\left(2|g_f(\mathcal{D}_k)| \geq \alpha_Q^2(f)\right) \leq \frac{4\mathbb{E}\left[|g_f(\mathcal{D}_k)|^2\right]}{\alpha_Q^2(f) \cdot \alpha_Q^2(f)}.$$

The denominator of the last term in the previous display can be bounded below using both $\alpha_Q^2(f) \geq \varepsilon\|f - f^*\|_{2,\mathbf{X}}^2$ and $\alpha_Q^2(f) \geq \|f - f^*\|_{2,\mathbf{X}}^2 4\theta_1^4 K/(\varepsilon\alpha n)$. This gives

$$\begin{aligned}
\mathbb{P}\left(2|g_f(\mathcal{D}_k)| \geq \alpha_Q^2(f)\right) &\leq \frac{4\mathbb{E}\left[\left((\mathbb{P}_{B_k} - \mathbb{P}_{\mathbf{X}})(f - f^*)^2\right)^2\right]}{\|f - f^*\|_{2,\mathbf{X}}^2 \frac{4\theta_1^4}{\alpha} \frac{K}{n}\|f - f^*\|_{2,\mathbf{X}}^2} \\
&\leq \frac{\sum_{i \in B_k} \mathrm{Var}\left((f - f^*)^2(\mathbf{X}_i)\right)}{|B_k|^2 \frac{\theta_1^4}{\alpha} \frac{K}{n}\|f - f^*\|_{2,\mathbf{X}}^4} \\
&\leq \frac{\mathbb{E}[(f - f^*)^4(\mathbf{X})]}{|B_k|\frac{\theta_1^4}{\alpha} \frac{K}{n}\|f - f^*\|_{2,\mathbf{X}}^4} \\
&\leq \frac{\alpha\|f - f^*\|_{4,\mathbf{X}}^4}{\theta_1^4\|f - f^*\|_{2,\mathbf{X}}^4} \\
&\leq \alpha,
\end{aligned}$$

since $\|f - f^*\|_{4,\mathbf{X}} \leq \theta_1\|f - f^*\|_{2,\mathbf{X}}$ by Assumption 3.1. The following bound follows exactly from the proof of Lemma 4 in [15]. Take $J = \cup_{k \in [K]_I} B_k$ and write $r_Q(\rho) = r_Q(\rho, \gamma_Q)$. Take $\mathbb{B}(f^*, \rho, r_Q(\rho))$ the set of functions $f \in \mathbb{B}(f^*, \rho)$ such that $\|f - f^*\|_{2,\mathbf{X}} \leq r_Q(\rho)$. We have

$$\mathbb{E}\left[\sup_{f \in \mathbb{B}(f^*,\rho)} \sum_{k \in [K]_I} \xi_k \frac{g_f(\mathcal{D}_k)}{\alpha_Q^2(f)}\right] \leq \frac{2}{\varepsilon r_Q^2(\rho)} \mathbb{E}\left[\sup_{f \in \mathbb{B}(f^*,\rho,r_Q(\rho))} \left|\sum_{k \in [K]_I} \xi_k(\mathbb{P}_{B_k} - \mathbb{E})(f - f^*)^2\right|\right].$$

Furthermore, we can apply the symmetrization argument in the proof of Lemma 4 in [15]. Together with the definition of $r_Q(\cdot)$, we find

$$\mathbb{E}\left[\sup_{f\in\mathbb{B}(f^*,\rho)}\sum_{k\in[K]_I}\xi_k\frac{g_f(\mathcal{D}_k)}{\alpha_Q^2(f)}\right]\leq\frac{4K}{\varepsilon n}\gamma_Q|[K]_I|\frac{n}{K}=\frac{4\gamma_Q}{\varepsilon}|[K]_I|.$$

Now we utilize the function ψ found in the proof of Lemma 4 in [15]. On an event $\Omega(K)$ with probability at least $1-\exp(-K\gamma x^2/2)=1-\exp(-K/8928)$,

$$\sum_{k\in[K]_I}\mathbf{1}\left(|g_f(\mathcal{D}_k)|<\alpha_Q^2(f)\right)$$

$$\geq(1-\alpha)|[K]_I|-2\mathbb{E}\left[\sup_{f\in\mathbb{B}(f^*,\rho)}\sum_{k\in[K]_I}\psi\left(\frac{|g_f(\mathcal{D}_k)|}{\alpha_Q^2(f)}\right)\right]+|[K]_I|x$$

$$\geq(1-\alpha)|[K]_I|-2\mathbb{E}\left[\sup_{f\in\mathbb{B}(f^*,\rho)}\sum_{k\in[K]_I}\xi_k\frac{|g_f(\mathcal{D}_k)|}{\alpha_Q^2(f)}\right]-|[K]_I|x$$

$$\geq|[K]_I|\left(1-\alpha-x-\frac{4\gamma_Q}{\varepsilon}\right)$$

$$\geq\gamma K\left(1-\alpha-x-\frac{4\gamma_Q}{\varepsilon}\right).$$

We now check that the latter is bigger than $(1-\eta)K$. With our choices, this gives

$$\frac{31}{32}\left(1-\frac{1}{93}-\frac{1}{93}-\frac{4}{372}\right)=\frac{31}{32}\left(1-\frac{1}{31}\right)=\frac{30}{32}=\frac{15}{16},$$

which is what we want. As a consequence, $Q_{15/16,K}[(f-f^*)^2]\leq\|f-f^*\|_{2,\mathbf{X}}^2+\alpha_Q^2(f)$. $\quad\square$

In the next result we use the event $\Omega(K):=\Omega_1(K)\cap\Omega_2(K)\cap\Omega_3(K)$ with $\Omega_1(K),\Omega_2(K)$ and $\Omega_3(K)$ respectively defined as the events in Lemma D.4, Lemma D.7 and Lemma D.8. The event $\Omega(K)$ has probability at least $1-4\exp(-K/8920)$. We also denote by $r(\cdot)$ any function satisfying $r(\rho)\geq\max\{r_P(\rho,\gamma_P),r_M(\rho,\gamma_M)\}$. For any integer K and $c_\rho\in\{1,2\}$, we will use the notation $\alpha_{K,c_\rho}:=c_\alpha r(c_\rho\rho)$ and $\delta_{K,n}^2:=25\mathfrak{m}^{*4}K/n$.

Lemma D.9. *Let $C^2=384\theta_1^2c_r^2c_\alpha^2\kappa_+^{1/2}$ and*

$$K\in\left[32|\mathcal{O}|,\frac{n}{372\theta_0^2}\wedge\frac{n}{25\kappa_+}\wedge\frac{n\varepsilon^2}{C^2}\right].$$

On the event $\Omega(K)$ defined above, for all $f\in\mathcal{F}$ such that $\|f-f^*\|\leq c_\rho\rho_K$, $\|f-f^*\|_{2,\mathbf{X}}\leq r(c_\rho\rho_K)$ and $|\sigma-\sigma^*|\leq\alpha_{K,c_\rho}$,

$$\mathbb{E}[-2\zeta(f-f^*)(\mathbf{X})]\leq\frac{2\sigma^*+\alpha_{K,c_\rho}}{2c}T_{K,\mu}(f^*,\sigma^*,f,\sigma)+\frac{2\sigma^*+\alpha_{K,c_\rho}}{2c}\mu\rho+\alpha_M^2$$

$$+\frac{8(2\sigma^*+\alpha_{K,c_\rho})}{c\sigma^*(2\sigma^*-\alpha_{K,c_\rho})^2}\delta_{K,n}^2+\frac{\alpha_{K,c_\rho}}{c(2\sigma^*-\alpha_{K,c_\rho})}\left(2\sigma^*r(c_\rho\rho_K)+r^2(c_\rho\rho_K)+\alpha_Q^2+\alpha_M^2\right),$$

where α_M^2,α_Q^2 are given in Lemma D.7 and Lemma D.8.

66

Proof of Lemma D.9. We start by applying Lemma D.7, which gives

$$\mathbb{E}[-2\zeta(f-f^*)(\mathbf{X})] \leq Q_{1/4,K}[-2\zeta(f-f^*)] + \alpha_M^2 \leq Q_{1/4,K}[(f-f^*)^2 - 2\zeta(f-f^*)] + \alpha_M^2,$$

the second inequality follows from the fact that $(f-f^*)^2$ is positive. Using the definition of $T_{K,\mu}(f^*, \sigma^*, f, \sigma)$ in (2.9) and the quantile properties in Lemma D.2, we can rewrite

$$\mathbb{E}[-2\zeta(f-f^*)(\mathbf{X})]$$

$$\leq Q_{1/4,K}[(f-f^*)^2 - 2\zeta(f-f^*)] + \alpha_M^2$$

$$= \frac{\sigma + \sigma^*}{2c} Q_{1/4,K}\left[2c\frac{\ell_f - \ell_{f^*}}{\sigma + \sigma^*}\right] + \alpha_M^2$$

$$= \frac{\sigma + \sigma^*}{2c} Q_{1/4,K}\left[R_c(\ell_{f^*}, \sigma^*, \ell_f, \sigma) - (\sigma - \sigma^*)\left(1 - 2\frac{\ell_f + \ell_{f^*}}{(\sigma + \sigma^*)^2}\right)\right] + \alpha_M^2$$

$$\leq \frac{\sigma + \sigma^*}{2c}\left(Q_{1/2,K}\left[R_c(\ell_{f^*}, \sigma^*, \ell_f, \sigma)\right] - Q_{1/4,K}\left[(\sigma - \sigma^*)\left(1 - 2\frac{\ell_f + \ell_{f^*}}{(\sigma + \sigma^*)^2}\right)\right]\right) + \alpha_M^2$$

$$\leq \frac{\sigma + \sigma^*}{2c}\left(Q_{1/2,K}\left[R_c(\ell_{f^*}, \sigma^*, \ell_f, \sigma)\right] + \mu(\|f\| - \|f^*\|)\right) + \frac{\sigma + \sigma^*}{2c}\mu\rho + \alpha_M^2$$

$$\quad - \frac{\sigma + \sigma^*}{2c} Q_{1/4,K}\left[(\sigma - \sigma^*)\left(1 - 2\frac{\ell_f + \ell_{f^*}}{(\sigma + \sigma^*)^2}\right)\right]$$

$$= \frac{\sigma + \sigma^*}{2c} T_{K,\mu}(f^*, \sigma^*, f, \sigma) + \frac{\sigma + \sigma^*}{2c}\left(\mu\rho - Q_{1/4,K}\left[(\sigma - \sigma^*)\left(1 - 2\frac{\ell_f + \ell_{f^*}}{(\sigma + \sigma^*)^2}\right)\right]\right) + \alpha_M^2.$$

Since $\sigma + \sigma^* \leq 2\sigma^* + \alpha_{K,c_\rho}$, it remains to show that

$$-\frac{\sigma + \sigma^*}{2c} Q_{1/4,K}\left[(\sigma - \sigma^*)\left(1 - 2\frac{\ell_f + \ell_{f^*}}{(\sigma + \sigma^*)^2}\right)\right] \tag{D.1}$$

$$\leq \frac{8(2\sigma^* + \alpha_{K,c_\rho})}{c\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + \frac{\alpha_{K,c_\rho}}{c(2\sigma^* - \alpha_{K,c_\rho})}\left(2\sigma^* r(c_\rho\rho_K) + r^2(c_\rho\rho_K) + \alpha_Q^2 + \alpha_M^2\right).$$

First, by the quantile properties in Lemma D.2, we have

$$-\frac{\sigma + \sigma^*}{2c} Q_{1/4,K}\left[(\sigma - \sigma^*)\left(1 - 2\frac{\ell_f + \ell_{f^*}}{(\sigma + \sigma^*)^2}\right)\right] \leq \frac{\sigma + \sigma^*}{2c} Q_{3/4,K}\left[(\sigma - \sigma^*)\left(2\frac{\ell_f + \ell_{f^*}}{(\sigma + \sigma^*)^2} - 1\right)\right].$$

By expanding $\ell_f = \ell_{f^*} + \ell_f - \ell_{f^*}$, we get

$$\frac{\sigma + \sigma^*}{2c} Q_{3/4,K}\left[(\sigma - \sigma^*)\left(2\frac{\ell_f + \ell_{f^*}}{(\sigma + \sigma^*)^2} - 1\right)\right]$$

$$= \frac{\sigma + \sigma^*}{2c} Q_{3/4,K}\left[(\sigma - \sigma^*)\left(\frac{4\ell_{f^*}}{(\sigma + \sigma^*)^2} - 1\right) + (\sigma - \sigma^*)\frac{2(\ell_f - \ell_{f^*})}{(\sigma + \sigma^*)^2}\right]$$

$$\leq \frac{\sigma + \sigma^*}{2c} Q_{7/8,K}\left[(\sigma - \sigma^*)\left(\frac{4\ell_{f^*}}{(\sigma + \sigma^*)^2} - 1\right)\right] + \frac{Q_{7/8,K}\left[(\sigma - \sigma^*)(\ell_f - \ell_{f^*})\right]}{c(\sigma + \sigma^*)}.$$

Since the term $(\sigma - \sigma^*)$ has different signs for $\sigma < \sigma^*$ and $\sigma > \sigma^*$, we need to account for this in the bounds. We focus first on the term

$$Q_{7/8,K}\left[(\sigma - \sigma^*)\left(\frac{4\ell_{f^*}}{(\sigma + \sigma^*)^2} - 1\right)\right]$$

$$\leq \max\left\{\sup_{\sigma\in(\sigma^*, \sigma^* + \alpha_{K,c_\rho}]}(\sigma - \sigma^*)\left(\frac{4Q_{7/8,K}[\ell_{f^*}]}{(\sigma + \sigma^*)^2} - 1\right), \sup_{\sigma\in[\sigma^* - \alpha_{K,c_\rho}, \sigma^*)}(\sigma^* - \sigma)\left(1 - \frac{4Q_{7/8,K}[\ell_{f^*}]}{(\sigma + \sigma^*)^2}\right)\right\}.$$

Thanks to Lemma D.4, the quantile $Q_{7/8,K}[\ell_{f^*}] = Q_{7/8,K}[\zeta^2]$ is in the interval $[\sigma^{*2} - \delta_{K,n}, \sigma^{*2} + \delta_{K,n}]$, therefore

$$Q_{7/8,K}\left[(\sigma - \sigma^*)\left(\frac{4\ell_{f^*}}{(\sigma + \sigma^*)^2} - 1\right)\right]$$

$$\leq \max\left\{ \sup_{\sigma \in (\sigma^*, \sigma^* + \alpha_{K,c_\rho}]} (\sigma - \sigma^*)\left(\frac{4(\sigma^{*2} + \delta_{K,n})}{(\sigma + \sigma^*)^2} - 1\right), \sup_{\sigma \in [\sigma^* - \alpha_{K,c_\rho}, \sigma^*)} (\sigma^* - \sigma)\left(1 - \frac{4(\sigma^{*2} - \delta_{K,n})}{(\sigma + \sigma^*)^2}\right)\right\}.$$
$$(D.2)$$

We denote $a_+^2 = \sigma^{*2} + \delta_{K,n}$ and $a_-^2 = \sigma^{*2} - \delta_{K,n}$. The first function in the latter display is positive (or zero) for $\sigma \in [\sigma^*, 2a_+ - \sigma^*]$. Let σ_{a_+} be the point achieving the maximum, then σ_{a_+} belongs to the same interval and $|\sigma_{a_+} - \sigma^*| \leq 2a_+ - 2\sigma^* = 2\sigma^*(\sqrt{1 + \delta_{K,n}/\sigma^{*2}} - 1)$. By construction, the quantity $\delta_{K,n}/\sigma^{*2}$ is smaller than one, since

$$\frac{\delta_{K,n}^2}{\sigma^{*4}} = \frac{25\mu^{*4}K}{\sigma^{*4}n} = \frac{25\kappa^* K}{n} \leq \frac{25\kappa_+ K}{n} \leq 1$$

and $K \leq n/(25\kappa_+)$. For all $x \in (0,1)$, the inequality $\sqrt{1+x} \leq 1 + x$ holds, so that

$$|\sigma_{a_+} - \sigma^*| \leq 2\sigma^*\left(\sqrt{1 + \frac{\delta_{K,n}}{\sigma^{*2}}} - 1\right) \leq 2\sigma^*\left(1 + \frac{\delta_{K,n}}{\sigma^{*2}} - 1\right) = \frac{2\delta_{K,n}}{\sigma^*}.$$

Now we repeat the same argument for the second function in (D.2), using $\sqrt{1-x} \geq 1-x$ for all $x \in (0,1)$, thus getting a point σ_{a_-} achieving the maximum such that $|\sigma_{a_-} - \sigma^*| \leq 2\delta_{K,n}/\sigma^*$. By Lemma A.1, we have $2\delta_{K,n}/\sigma^* < \alpha_{K,c_\rho} < \sigma^*$. With $\delta_a = 2\delta_{K,n}/\sigma^*$, this yields

$$Q_{7/8,K}\left[(\sigma - \sigma^*)\left(\frac{4\ell_{f^*}}{(\sigma + \sigma^*)^2} - 1\right)\right]$$

$$\leq \max\left\{(\sigma^* - \sigma_{a_-})\left(1 - \frac{4a_-^2}{(\sigma_{a_-} + \sigma^*)^2}\right), (\sigma_{a_+} - \sigma^*)\left(\frac{4a_+^2}{(\sigma_{a_+} + \sigma^*)^2} - 1\right)\right\}$$

$$\leq \frac{2\delta_{K,n}}{\sigma^*} \max\left\{1 - \frac{4\sigma^{*2} - 4\delta_{K,n}}{(2\sigma^* - \delta_a)^2}, \frac{4\sigma^{*2} + 4\delta_{K,n}}{(2\sigma^* + \delta_a)^2} - 1\right\}$$

$$= \frac{2\delta_{K,n}}{\sigma^*} \max\left\{\frac{4\sigma^*\delta_a + \delta_a^2 + 4\delta_{K,n}}{(2\sigma^* - \delta_a)^2}, \frac{4\delta_{K,n} - 4\sigma^*\delta_a - \delta_a^2}{(2\sigma^* + \delta_a)^2}\right\}$$

$$\leq \frac{16\delta_{K,n}^2}{\sigma^*(2\sigma^* - \delta_a)^2}$$

$$\leq \frac{16\delta_{K,n}^2}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}.$$

One last term needs to be bounded in order to obtain (D.1). We only consider the case when $\sigma \in [\sigma^*, \sigma^* + \alpha_{K,c_\rho}]$, the case $\sigma \in [\sigma^* - \alpha_{K,c_\rho}, \sigma^*]$ follows the same steps. With $\ell_{f^*} - \ell_f = 2\zeta(f - f^*) - (f - f^*)^2$, we get

$$\frac{1}{c(\sigma + \sigma^*)} Q_{7/8,K}\left[(\sigma - \sigma^*)(\ell_f - \ell_{f^*})\right] = \frac{(\sigma - \sigma^*)}{c(\sigma + \sigma^*)} Q_{7/8,K}\left[(f - f^*)^2 - 2\zeta(f - f^*)\right]$$

$$\leq \frac{\alpha_{K,c_\rho}}{c(2\sigma^* - \alpha_{K,c_\rho})}\left(Q_{15/16,K}\left[(f - f^*)^2\right] + Q_{15/16,K}\left[-2\zeta(f - f^*)\right]\right)$$

$$\leq \frac{\alpha_{K,c_\rho}}{c(2\sigma^* - \alpha_{K,c_\rho})}\left(\|f - f^*\|_{2,\mathbf{X}}^2 + \alpha_Q^2 + \mathbb{E}[-2\zeta(f - f^*)(\mathbf{X})] + \alpha_M^2\right),$$

the last inequality follows from Lemma D.7 and Lemma D.8. By the Cauchy-Schwarz inequality, $\mathbb{E}[-2\zeta(f - f^*)(\mathbf{X})] \leq 2\sigma^*\|f - f^*\|_{2,\mathbf{X}} \leq 2\sigma^* r(c_\rho \rho_K)$. By putting everything together, we conclude

$$\mathbb{E}[-2\zeta(f - f^*)(\mathbf{X})] \leq \frac{2\sigma^* + \alpha_{K,c_\rho}}{2c} T_{K,\mu}(f^*, \sigma^*, f, \sigma) + \frac{2\sigma^* + \alpha_{K,c_\rho}}{2c}\mu\rho + \alpha_M^2$$
$$+ \frac{8(2\sigma^* + \alpha_{K,c_\rho})}{c\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + \frac{\alpha_{K,c_\rho}}{c(2\sigma^* - \alpha_{K,c_\rho})}\left(2\sigma^* r(c_\rho\rho_K) + r^2(c_\rho\rho_K) + \alpha_Q^2 + \alpha_M^2\right),$$

which gives the claim. $\qquad\square$

References

[1] ALON, N., MATIAS, Y., AND SZEGEDY, M. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences 58*, 1 (1999), 137 – 147.

[2] BELLEC, P. C., LECUÉ, G., AND TSYBAKOV, A. B. Towards the study of least squares estimators with convex penalty. *arXiv e-prints* (Jan. 2017), arXiv:1701.09120.

[3] BELLEC, P. C., LECUÉ, G., AND TSYBAKOV, A. B. Slope meets lasso: Improved oracle bounds and optimality. *Ann. Statist. 46*, 6B (12 2018), 3603–3642.

[4] BELLEC, P. C., AND TSYBAKOV, A. B. Bounds on the prediction error of penalized least squares estimators with convex penalty. *arXiv e-prints* (Sept. 2016), arXiv:1609.06675.

[5] BELLONI, A., CHERNOZHUKOV, V., AND WANG, L. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika 98*, 4 (2011), 791–806.

[6] BELLONI, A., CHERNOZHUKOV, V., AND WANG, L. Pivotal estimation via square-root lasso in nonparametric regression. *Annals of Statistics 42*, 2 (2014), 757–788.

[7] COMMINGES, L., COLLIER, O., NDAOUD, M., AND TSYBAKOV, A. B. Adaptive robust estimation in sparse vector model. *arXiv preprint arXiv:1802.04230* (2018).

[8] DERUMIGNY, A. Improved bounds for square-root Lasso and square-root Slope. *Electronic Journal of Statistics 12*, 1 (2018), 741–766.

[9] DERUMIGNY, A. *Some statistical results in high-dimensional dependence modeling.* PhD thesis, Université Paris-Saclay (ComUE), 2019.

[10] DEVROYE, L., LERASLE, M., LUGOSI, G., AND OLIVEIRA, R. I. Sub-gaussian mean estimators. *The Annals of Statistics 44*, 6 (2016), 2695–2725.

[11] GIRAUD, C. *Introduction to high-dimensional statistics*, vol. 138. CRC Press, 2014.

[12] JERRUM, M. R., VALIANT, L. G., AND VAZIRANI, V. V. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science 43* (1986), 169 – 188.

[13] LECUÉ, G., AND MENDELSON, S. Learning subgaussian classes: upper and mini-max bounds (2013). *Topics in Learning Theory-Societe Mathematique de France,(S. Boucheron and N. Vayatis Eds.)* (2013).

[14] LECUÉ, G., AND MENDELSON, S. Regularization and the small-ball method i: sparse recovery. *The Annals of Statistics 46*, 2 (2018), 611–641.

[15] LECUÉ, G., AND LERASLE, M. Robust machine learning by median-of-means: Theory and practice. *Ann. Statist. 48*, 2 (04 2020), 906–931.

[16] LEVIN, L. A. Notes for Miscellaneous Lectures. *arXiv e-prints* (Mar. 2005), cs/0503039.

[17] LUGOSI, G., AND MENDELSON, S. Regularization, sparse recovery, and median-of-means tournaments. *Bernoulli 25*, 3 (2019), 2075–2106.

[18] LUGOSI, G., AND MENDELSON, S. Risk minimization by median-of-means tournaments. *J. Eur. Math. Soc. 22* (2020), 925–965.

[19] MENDELSON, S. Upper bounds on product and multiplier empirical processes. *Stochastic Processes and their Applications 126*, 12 (2016), 3652–3680.

[20] MENDELSON, S. On multiplier processes under weak moment assumptions. In *Geometric Aspects of Functional Analysis*. Springer, 2017, pp. 301–318.

[21] NEMIROVSKIJ, A. S., AND YUDIN, D. B. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.