TWO PERSPECTIVES ON HIGH-DIMENSIONAL ESTIMATION PROBLEMS

Posterior contraction and median-of-means

Gianluca Finocchio

TWO PERSPECTIVES ON HIGH-DIMENSIONAL ESTIMATION PROBLEMS: POSTERIOR CONTRACTION AND MEDIAN-OF-MEANS

Gianluca Finocchio

TWO PERSPECTIVES ON HIGH-DIMENSIONAL ESTIMATION PROBLEMS: POSTERIOR CONTRACTION AND MEDIAN-OF-MEANS

DISSERTATION

to obtain
the degree of doctor at the Universiteit Twente,
on the authority of the rector magnificus,
prof. dr. ir. A. Veldkamp,
on account of the decision of the Doctorate Board
to be publicly defended
on Wednesday 13 October 2021 at 16.45 hours

by

Gianluca Finocchio

born on the 1st of December, 1991
in Tychy, Poland

This dissertation has been approved by:

Supervisor
prof. dr. ir. A.J. Schmidt-Hieber

Co-supervisors
dr. K. Proksch
dr. A.F.F. Derumigny

Printed by:
Lay-out:
ISBN: 978-90-365-5235-6
DOI: 10.3990/1.9789036552356
https://doi.org/10.3990/1.9789036552356

Graduation Committee:

Chair / secretary:	prof.dr. J.N. Kok
Supervisor:	prof.dr.ir. A.J. Schmidt-Hieber
Co-supervisors:	dr. K. Proksch
	dr. A.F.F. Derumigny
Committee Members:	prof.dr. N.V. Litvak prof.dr. C. Brune prof. dr. J.H. Van Zanten prof. dr. I. Castillo

Acknowledgements

This thesis summarizes my research results spanning the four years as a PhD candidate in The Netherlands. It would have not existed without the contribution and support of many people and organizations.

My co-authors and I are grateful to an anonymous referee and Associate Editor of the *Electronic Journal of Statistics* for many helpful suggestions resulting in a major improvement of the work in Chapter 1. We are grateful to Joris Bierkens and Ismaël Castillo for helpful discussions on the work in Chapter 2. Part of the work in Chapter 3 was done while Alexis Derumigny was employed at the University of Twente. We are grateful to Alexandre Tsybakov, Guillaume Lecué and Matthieu Lerasle for useful discussions on a very early version of the manuscript, that appeared in [36, Chapter 3].

The thesis is part of the project *Nonparametric Bayes for high-dimensional models: contraction, credible sets, computations* (with project number 613.001.604) which is financed by the Dutch Research Council (NWO).

I would like to thank the committee members for accepting to read the manuscript and to assess it.

My sincere gratitude goes to my supervisors Johannes Schmidt-Hieber, Katharina Proksch and Alexis Derumigny. Their constructive feedback, high-quality standards and deep mathematical understanding have made me grow as a researcher and as a person. It has been fun to discuss with them many challenging problems and to take part in as many social activities.

I want to thank my colleagues at the University of Twente and Leiden University, who have made for a stimulating environment and have been inclusive and open-minded. A special mention to those with whom I have shared lunch breaks or evenings playing futsal (*Pi Hard* and *Mathletico*), climbing, running or playing board games. I am grateful to Lotte Weedage for helping me with the Dutch translation of the summary.

Ringrazio la mia famiglia per essermi stata vicina ed avermi sostenuto nonostante questi anni di lontananza.

Last but not least, I would like to thank Francesca, for her unconditional love and support.

Contents

Introduction

Statistical estimation problems require to reconstruct an unknown parameter of interest given data. The general setting can be expressed in terms of a parameter space Θ and a dataset $\mathcal{D}_n$ of $n \geq 1$ random variables whose distributions are parametrized by the model $\mathcal{D}_n|\theta \sim \mathbb{P}_\theta^{(n)}$, $(\mathbb{P}_\theta^{(n)} : \theta \in \Theta)$. This incorporates the idea that the information on the parameter can be (partially) inferred from the law of the observations.

This thesis consists of three chapters that investigate the theoretical properties of different approaches for benchmark estimation problems. They study many desirable features of estimation procedures, providing both positive and negative results. We introduce below the relevant terminology, starting from the different interpretations of randomness itself.

The frequentist approach assumes there exists some objective random process generating the observations, and that the experiment can be repeated (in principle) multiple times. That is, the dataset is distributed according to $\mathbb{P}_{\theta_0}^{(n)}$ for some 'true' parameter of interest $\theta_0 \in \Theta$. The parameter of interest θ_0 is viewed as some deterministic entity that can be inferred by a suitable estimator, which is a function of the random dataset. Maximum likelihood is a famous example, but frequentist methods are not restricted to be based on the likelihood.

The Bayesian interpretation assigns a prior probability distribution π to the whole parameter space Θ, in order to model *a priori* information. The prior distribution might be intended as an objective or a subjective belief, but the purpose is the same. Even though the observations might be the result of a random process, they are treated as given known quantities, and they are employed into the computation of the posterior distribution through Bayes' theorem. The posterior is the conditional distribution of the parameter θ, given the dataset $\mathcal{D}_n$ and the prior π, and is the sole tool that a Bayesian has (and needs) to make statements about estimation and uncertainty.

In this thesis, Bayesian methods are evaluated under the frequentist paradigm. This means that it is always assumed that the observations follow some unknown (but fixed) distribution $\mathbb{P}_{\theta_0}^{(n)}$ for some $\theta_0 \in \Theta$. The hope is that the posterior distribution contracts, that is, assigns most of its mass, to a small neighbourhood of θ_0, with high probability or in expectation (with respect to $\mathbb{P}_{\theta_0}^{(n)}$). In all these cases, the posterior density will be proportional to the product between the likelihood of the sample and the prior, and this makes posterior contraction a likelihood-based method.

A statistical model $(\mathbb{P}_\theta^{(n)} : \theta \in \Theta)$ can be classified depending on the size of the parameter space Θ, in the following sense. The model is parametric if Θ is embedded in some Euclidean space $\mathbb{R}^d$. The model is semiparametric if the indexing parameter is actually a pair (θ, η) of a Euclidean parameter θ and an infinite-dimensional nuisance parameter η. For our purposes, this means that the latter belongs to either an infinite-dimensional vector space, some subset of real-valued functions or some subset of probability measures on the real line. The model is nonparametric if Θ itself is infinite-dimensional. The main difference between semiparametric and nonparametric models is that, in the former, the parameter of interest $\theta_0 \in \Theta$ is always finite-dimensional. Lastly, we say that the model is high-dimensional if the number of parameters grows with the number of observations.

Among the many desirable properties that a good estimation method might satisfy, the bare minimum is consistency. This is an asymptotic property, thus only depends on the limiting behavior as the sample size n tends to infinity. A frequentist estimator $\widehat{\theta}_n$ is consistent if it converges, in a suitable sense, to the parameter of interest θ_0. Similarly, a Bayesian posterior distribution is consistent if it converges to the point mass at θ_0. Consistency can be quantified by convergence rates, which measure the speed of convergence to θ_0. They are usually given as decreasing sequences r_n depending of the sample size n, and possibly some other model-related parameters.

Fast rates make a procedure more appealing from a practical point of view, more so if it can be easily implemented. Even when this is not the case, the effort of recovering optimal convergence rates usually provides insights into the specific problem at hand. For example, it is known that the minimax rate of estimation (in supremum norm) in the nonparametric regression problem depends on the smoothness of the regression function. Another example is the sparse linear regression problem, where the minimax rates (in L^1 or L^2 norm) depend on the sparsity level, that is, the number of non-zero components of the regression vector.

In situations where the optimal rates depend on an underlying hyperparameter, it might be difficult to obtain fast rates when no prior information is available. To overcome this issue, the concept of adaptivity has been introduced. An estimator is adaptive if it can achieve optimal contraction rates without requiring knowledge of the hyperparameter. A well-established technique in the Bayes literature to deal with adaptivity involves hierarchical priors: one puts first a hyperprior on the hyperparameters and then, given a fixed hyperparameter, a prior on the corresponding parameter space. If optimal rates can be achieved when the true hyperparameter is known, and the prior is carefully selected, then the posterior will be adaptive. On the frequentist side, there is no general approach and each problem has to be tackled on its own. In the sparse linear regression setting, one can employ for instance a Lepski-type procedure as in [9], and show that the Lasso estimator in [79] achieves adaptivity with respect to the sparsity level.

A more refined property for an estimator $\widehat{\theta}_n$ with rates r_n is the asymptotic shape of the rescaled $r_n^{-1}(\widehat{\theta}_n - \theta_0)$. In parametric models, the frequentist statistician aims to find asymptotically efficient estimators, for which the sequence $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ converges in distribution to the normal $\mathcal{N}(0, I_{\theta_0}^{-1})$ where I_{θ_0} denotes the Fisher information matrix at θ_0. This is true for the maximum likelihood estimator (MLE) if the model is regular. The

Bayesian statistician aims to invoke the parametric Bernstein-von Mises (BvM) theorem, see Theorem 10.1 in [81], which states that, under weak conditions on the model and the prior $\theta_0 \sim \pi$, the posterior distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ also converges to $\mathcal{N}(0, I_{\theta_0}^{-1})$. A remarkable feature of this result is that the contribution of the prior washes out in the limit. The major implication is that the posterior distribution is a valid inference tool from a frequentist perspective.

The situation is dramatically different for high-dimensional models, where the prior choice becomes crucial in determining the asymptotic properties of the posterior. Semiparametric BvM theorems have been obtained in [12, 19, 23, 70]. It has been observed that there can be a large bias in the posterior limit, for example in [18, 23, 75], and it is unclear whether the bias is due to the specific choice of prior or whether this is a fundamental limitation of the Bayesian method. The nonparametric BvM phenomenon has been studied for Gaussian regression models in [21, 22], whereas the non-asymptotic accuracy of the normal approximation of the posterior has been the focus of a recent effort in [77].

Another valuable property of an estimation principle, such as Bayes or maximum likelihood, is the ability to deal with observations that are non-i.i.d. as a result of the combination of different datasets or the contamination by outliers. In these models, a fraction of the observations is informative, that is, it behaves as an i.i.d. sample of the true underlying distribution; the remaining fraction can instead be only slightly informative or even adversarial. A method that can lead to fast rates under contaminated datasets is said to be robust. A frequentist statistician has an advantage in this situation because it is sometimes possible to discard part of the data and reduce the fraction of contaminated observations. An example of a robust frequentist estimator for the mean of a heavy-tailed variable is the median-of-means in [37, Section 4.1]. On the other hand, a purely Bayesian approach does not allow to assign a prior after having looked at the data, and so the whole dataset should be used instead. In the Bayesian framework, the concept of robustness can also refer to the sensitivity of the posterior with respect to different choices of prior. If the posterior is well-behaved for large families of prior, then the procedure is robust.

With the terminology introduced so far, we can briefly synthesize the three chapters of the thesis in the next table. This is meant as a broad overview to compare their different facets.

	Chapter 1	**Chapter 2**	**Chapter 3**
Interpretation	Bayesian	Bayesian	frequentist
Dimension	semiparametric	nonparametric	nonparametric
Model	Gaussian sequence	regression	regression
Observations	indep. Gaussian	i.i.d. Gaussian noise	i.i.d. with outliers heavy-tailed noise
Param. of interest	model variance	regression function	regression function noise variance
Hyperparam.		underlying structure smoothness index	sparsity level
Nuisance param.	vector of means		noise distribution
Principle	posterior contraction	posterior contraction	median-of-means
Consistency	fails in general holds in special case	yes	unknown in general holds in linear case
Type of rates	asymptotic	asymptotic	non-asymptotic
Optimal rates	yes	yes	yes and maybe
Asymptotic shape	yes		
Adaptivity		yes	yes
Robustness (data)	mild		adversarial
Robustness (prior)	no	no	

Chapter 1: Bernstein-von Mises for a non-standard semiparametric model

In the first chapter, we consider the following semiparametric experiment. For given $0 \leq \alpha \leq 1$, one observes n independent and normally distributed random variables

$$X_i \sim \mathcal{N}\big(\mu_i^* \mathbf{1}(i > n\alpha), \sigma^{*2}\big), \quad i = 1, \dots, n.$$

The parameters in the model are $\{\mu_i^* : i > n\alpha\}$ and $\sigma^* > 0$. The goal is to estimate the variance σ^{*2} while treating the mean vector $\boldsymbol{\mu}^* := (\mu^*_{\lceil n\alpha \rceil}, \dots, \mu^*_n)$ as nuisance. We approach the problem from a Bayesian perspective, by studying the asymptotic properties of the posterior distribution arising from different priors on the parameters.

The observations can be divided into two sets, one with $\lfloor n\alpha \rfloor$ i.i.d. normal variables $\mathcal{N}(0, \sigma^{*2})$ and the other with $n - \lfloor n\alpha \rfloor$ independent normal variables $\mathcal{N}(\mu_i^*, \sigma^{*2})$. The statistician knows which fraction of the dataset has mean zero and which fraction is corrupted, but the contamination is not adversarial, that is, both fractions of the sample share the same variance σ^{*2}. This might be the result of combining datasets coming from different experiments measuring the same physical quantity.

The family of models taken into consideration generalizes the Neymann-Scott model in [70], which has been labelled ‘disturbing’ by L. Le Cam [54] since it naturally leads to inconsistent MLE. A frequentist approach would allow throwing away the fraction of the sample that is contaminated by the non-zero means, and this would yield consistent estimates. It is easy to show that the MLE obtained using all the data points converges to $\alpha\sigma^{*2}$, it therefore underestimates the true variance by the factor α. One could dismiss the issue by multiplying the MLE by the factor α^{-1}.

We investigate whether a Bayesian methods is robust enough to deal with the combined dataset by itself. This approach to the problem is more involved and requires a suitable choice of priors for the pair $(\sigma^{*2}, \boldsymbol{\mu}^*)$. Since the posterior assigns more mass to regions with large likelihood, it is not clear whether the Bayesian method is able to correct for the flaws of the MLE. Such a correction has been observed before in some irregular models, see for example [26, 40, 75].

In the first part of the chapter, we investigate whether the posterior is consistent when the nuisances are modelled as i.i.d. variables. Surprisingly, the answer is negative in a very general sense. Whenever the nuisances are independently drawn from a proper distribution, the posterior does not contract around the true variance. Thus the Bayesian method fails to correct the flaws of the MLE for a large class of natural priors. We can show this by means of lower bounds on the logarithm of the posterior density. Our arguments heavily rely on the specifics of the Gaussian sequence model, but they do not require any decay condition on the tails of the prior.

The lack of structure on the nuisance parameter would suggest that a correlated prior on the means should perform worse, but it turns out that this is not the case. In the second part of the chapter we construct a Gaussian mixture prior for which the posterior is consistent and contracts with parametric rate. For this prior, the limit distribution in the BvM sense is derived and it is shown that it is non-Gaussian in the case of small means. It remains open whether a similar behavior carries over to more general prior choices or whether this is a fortuitous feature of the Gaussian mixture.

Another counter-intuitive fact is the non-Gaussianity of the asymptotic shape in the case of small means. This is motivated by the fact that the posterior does not throw away the non-zero mean observations, and a simulation study shows that the maximum a posteriori (MAP) estimate based on the limit distribution has better frequentist properties than the adjusted MLE that only uses the observations with zero mean.

Chapter 2: deep Gaussian process priors

In the second chapter we consider the multivariate nonparametric regression model with random design supported on $[-1,1]^d$, where we observe n i.i.d. pairs $(\mathbf{X}_i, Y_i) \in [-1,1]^d \times \mathbb{R}$, $i = 1, \ldots, n$, with

$$Y_i = f^*(\mathbf{X}_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

and ε_i independent and standard normal random variables that are independent of the design vectors $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$. The problem is tackled from a Bayesian perspective, under an additional structural assumption that includes important cases such as (generalized) additive models.

We assume that the regression function f^* can be represented as the composition of q^*+1 functions, that is, $f^* = h^*_{q^*} \circ h^*_{q^*-1} \circ \cdots h^*_1 \circ h^*_0$. Each component h^*_i is β^*_i-Hölder and maps $\mathbb{R}^{d^*_i}$ to $\mathbb{R}^{d^*_{i+1}}$, thus taking in input $d^*_i \geq 1$ variables. We allow h^*_i to only depend on a number $t^*_i \leq d^*_i$ of variables, which we call effective dimension. This results in a hyperparameter $(\lambda^*, \boldsymbol{\beta}^*)$ consisting of a graph λ^* and a vector of smoothness indices $\boldsymbol{\beta}^*$. We are interested in recovering the regression function f^* only, while being adaptive with respect to the unknown composition graph and smoothness index.

This particular setting is inspired by deep learning methods [44], which are most successful when performing tasks that involve some underlying modular structure, that is, when complex objects have to be built using a small number of simpler features. A prototypical example is writing. The page of a book can be assembled using a small number sentences, each sentence using a small number of words, and each word using a small number of letters. In [76] it has been shown that sparsely connected deep neural networks are able to pick up the underlying composition structure and achieve near minimax estimation rates. On the contrary, wavelet thresholding methods are shown to be unable to adapt to the underlying structure resulting in potentially much slower convergence rates.

Gaussian process (GP) priors are a natural choice in the classical Bayesian nonparametric regression setting. Posterior contraction rates have been established in [41, Section 11] and are known to be optimal in certain cases. Gaussian processes priors are also widely used in machine learning [73]. This motivates the study of priors induced by the composition of GPs, which are known as deep Gaussian processes (DGPs) in the literature [33, 32] and are the Bayesian analogue of deep networks.

In this chapter we derive posterior contraction rates for DGPs, by extending the theory of GP priors. We implement a hierarchical procedure, where a hyperprior is assigned to the possible composition structures and then, given a composition structure, a suitable DGP prior is assigned to the corresponding function class. For such a DGP prior construction we show that the posterior contraction rate matches nearly the minimax estimation rate. In particular, if there is some low-dimensional structure in the composition, the posterior will not suffer from the curse of dimensionality.

The main tool of our analysis is an extension of the concentration function for Gaussian processes introduced in [82]. Furthermore, our proving strategy requires some regularization in the construction of the DGP prior. For a fully Bayesian approach, stability is enforced by conditioning each individual Gaussian process to be in a set of 'stable' paths. Specifically, these sets are obtained by inflating Hölder balls and, to achieve near optimal contraction rates, the size of the inflations has to be carefully selected and depends on the optimal contraction rate itself. It is not clear whether this regularization is indeed necessary, but it has the same flavor of other stabilization enhancing methods that improve the performance of deep learning, such as dropout and batch normalization [44].

Chapter 3: median-of-means for robust inference in least-squares regression

In the third chapter we consider the setting of nonparametric least-squares regression. Let Y be an unknown square-integrable real variable and let $\mathbf{X}$ be some explanatory variable on a measurable space $\mathcal{X}$ and law $\mathbb{P}_{\mathbf{X}}$. The statistician is given a (closed) convex function class $\mathcal{F} \subseteq L^2(\mathbb{P}_{\mathbf{X}})$ and some dataset of observations $\mathcal{D}_n = (\mathbf{X}_i, Y_i)_{i=1,\dots,n}$. The goal is to provide a frequentist estimator for the oracle pair (f^*, σ^*) given by

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

and this estimator should have as good as possible non-asymptotic guarantees.

The additional complication is that the dataset $\mathcal{D}_n$ may be contaminated by a subset $\mathcal{D}_{\mathcal{O}} = (\mathbf{X}_i, Y_i)_{i \in \mathcal{O}}$ of $|\mathcal{O}| \leq n/2$ arbitrary outliers. One expects to be able to solve the problem at hand as long as the number of outliers is not too large and the remaining informative observations in $\mathcal{D}_{\mathcal{I}} = \mathcal{D}_n \setminus \mathcal{D}_{\mathcal{O}}$ are i.i.d. as $(\mathbf{X}, Y)$, which satisfy $Y = f^*(\mathbf{X}) + \zeta$ with residual $\zeta := Y - f^*(\mathbf{X})$ that may be heavy-tailed and not independent of $\mathbf{X}$.

A frequentist nonparametric method to recover the unknown regression function f^* is the regularized empirical risk minimizer

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

for some tuning parameter $\lambda > 0$ and a penalty functional Ψ on $\mathcal{F}$. The penalty functional reduces overfitting by assigning a large value to functions that are too big, in some sense. This method has two drawbacks: on one hand, if the residuals are heavy-tailed this leads to suboptimal non-asymptotic properties; on the other, the empirical average involves all the observations in the dataset and can be strongly influenced by the presence of even one outlier.

The method introduced in [60, 55] makes the RERM robust by replacing the empirical averages by the median-of-means (MOM) over a number K of blocks: one partitions of the dataset into K blocks, 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 [37, Section 4.1]. This results in a robust MOM-K estimator $\widehat{f}_{\lambda,K}$ with penalization parameter $\lambda > 0$, for which non-asymptotic guarantees are obtained in high probability.

In the sparse linear case, this problem is equivalent to estimating $\boldsymbol{\beta}^*$ in the model $Y = \mathbf{X}^T \boldsymbol{\beta}^* + \zeta$ for the function space $\mathcal{F}_{s^*} = \{\mathbf{x} \mapsto \mathbf{x}^T \boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^d,\ |\boldsymbol{\beta}|_0 \leq s^*\}$ for some sparsity level $s^* > 0$ and $|\boldsymbol{\beta}|_0$ the number of non-zero components of $\boldsymbol{\beta}$. In this case, the MOM method outlined above yields a robust version of the Lasso estimator [7, 8, 9], which is minimax optimal but its optimal penalization parameter has to be proportional to σ^*. In a special instance, the Lasso has the following Bayesian interpretation: it is the maximum a posteriori (MAP) estimate arising from the model $Y|\mathbf{X}, \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{X}^\top \boldsymbol{\beta}, \sigma^2 I)$ and $\boldsymbol{\beta}$ drawn from a Laplace prior.

In the third chapter we extend the scope of the MOM approach to the case of unknown noise standard deviation σ^*. A new method is proposed that yields, in the sparse linear case, a robust version of the square-root Lasso [10, 35], which is minimax optimal and its penalization parameter does not depend on σ^*. The square-root Lasso is a scale-invariant method [43, Section 5] that modifies the Lasso in order to achieve adaptivity, but the modifications have no obvious Bayesian interpretation.

We also show that, in the high-dimensional sparse linear regression setting with unknown σ^* and known sparsity level $s^* \leq d$, our MOM estimator achieves the optimal rates of estimation of $\boldsymbol{\beta}^*$ using a number of blocks K of the order of the number of outliers. The convergence rate for σ^* improves on previously available estimators, but we do not show this being optimal. Since the sparsity level may be unknown in practice, an aggregated adaptive procedure based on Lepski's method is proposed. For that, one first infers an estimated sparsity level $\widetilde{s}$ and then a number of blocks $\widetilde{K}$. It is shown that the resulting adaptive estimator $(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}, \widetilde{s})$ attains similar frequentist properties as the estimator with known true sparsity level.

Chapter 1

Bayesian variance estimation in the Gaussian sequence model with partial information on the means

This chapter is based on:
G. Finocchio and J. Schmidt-Hieber. Bayesian variance estimation in the Gaussian sequence model with partial information on the means. *Electron. J. Statist. 14(1): 239-271 (2020).*

Abstract

Consider the Gaussian sequence model under the additional assumption that a fixed fraction of the means is known. We study the problem of variance estimation from a frequentist Bayesian perspective. The maximum likelihood estimator (MLE) for σ^2 is biased and inconsistent. This raises the question whether the posterior is able to correct the MLE in this case. By developing a new proving strategy that uses refined properties of the posterior distribution, we find that the marginal posterior is inconsistent for any i.i.d. prior on the mean parameters. In particular, no assumption on the decay of the prior needs to be imposed. Surprisingly, we also find that consistency can be retained for a hierarchical prior based on Gaussian mixtures. In this case we also establish a limiting shape result and determine the limit distribution. In contrast to the classical Bernstein-von Mises theorem, the limit is non-Gaussian. We show that the Bayesian analysis leads to new statistical estimators outperforming the correctly calibrated MLE in a numerical study.

1.1 Introduction

For given $0 \leq \alpha \leq 1$, suppose we observe n independent and normally distributed random variables

$$X_i \sim \mathcal{N}\big(\mu_i^* \mathbf{1}(i > n\alpha), \sigma^{*2}\big), \quad i = 1, \ldots, n. \tag{1.1.1}$$

The parameters in the model are μ_i^*, $i > n\alpha$ and $\sigma^* > 0$. The goal is to estimate the variance σ^{*2} while treating the mean vector $\boldsymbol{\mu}^* := (\mu^*_{\lceil n\alpha \rceil}, \ldots, \mu^*_n)$ as nuisance. For $\alpha = 0$, we recover the Gaussian sequence model. For $\alpha > 0$, this can be viewed as the Gaussian sequence model with additional knowledge that the means of the first $\lfloor n\alpha \rfloor$ observations are known (in which case we can subtract them from the data).

One can think of model (1.1.1) as a simple prototype of a combined dataset. Using for instance different measurement devices, one often faces merged datasets collected from multiple sources. The different sources might not be of the same quality concerning the underlying parameter, see [65] for an example. An alternative viewpoint is to interpret model (1.1.1) as a sparse sequence model with known support. Since a $(1-\alpha)$-fraction of the data is perturbed, we are in the dense regime. Knowledge of the support is then crucial as otherwise there is no consistent estimator for σ^{*2}.

If n is even and $\alpha = 1/2$, then (1.1.1) is equivalent to the Neyman-Scott model [70] up to a reparametrization. Model (1.1.1) is in this case equivalent to observing $U_i := (X_{n/2+i} + X_i)$ and $V_i := (X_{n/2+i} - X_i)$ for $i = 1, \ldots, n/2$. Since U_i and V_i are independent, this is thus equivalent to observing independent random variables $U_i, V_i \sim \mathcal{N}(\mu^*_{n/2+1}, 2\sigma^{*2})$. Estimation of σ^* in the latter model is known as Neyman-Scott problem.

Although σ^{*2} can be estimated with parametric rate based on the first $n\alpha$ observations, a striking feature of the model is that the MLE for σ^{*2} is inconsistent. In fact the MLE $\widehat{\sigma}^2_{\text{mle}}$ converges to $\alpha\sigma^{*2}$ therefore underestimating the true variance by the factor α. The reason is that the likelihood over the observations with non-zero mean significantly affects the total likelihood viewed as a function in σ^2.

We study what happens when a Bayesian approach is implemented for the estimation of the variance and whether a posterior distribution can correct for the bias of the MLE. The Bayesian method can be viewed as a weighted likelihood method: instead of taking the parameter with the largest likelihood the posterior puts mass on parameter sets with large likelihood. Because of this, the posterior can in some cases correct the flaws of the MLE. An example are irregular models, see [40, 26, 75].

In the first part of the paper, we prove that whenever the nuisances are independently generated from a proper distribution, the posterior does not contract around the true variance. This shows that, for a large class of natural priors, the Bayesian method is unable to correct the MLE. In frequentist Bayes, several lower bound techniques have been developed in order to describe when Bayesian methods do not work, [17, 25, 24, 80, 20, 49]. These results can be used for instance to show that a certain decay of the prior is necessary to ensure posterior contraction. Our lower bounds are of a different flavor and do not require a condition on the tail behavior.

Since for the non-zero means no additional structure is assumed, there is no way to get a better estimate of one mean from the knowledge of all other means. Therefore, one might be tempted to think that a correlated prior on the means cannot perform better than an i.i.d. prior and consequently must lead to an inconsistent posterior as well. Surprisingly, this is not true and we construct in the second part of the article a Gaussian mixture prior for which the posterior contracts with the parametric rate around the true variance. For this prior we derive the limit distribution in the Bernstein-von Mises sense. In contrast with the Bernstein-von Mises theorem, the posterior limit is non-Gaussian in the case of small means. In this case the posterior also incorporates information about the second part of the sample into the estimator and we show in a simulation study that the maximum a posteriori estimate based on the limit distribution outperforms the $\sqrt{n}$-consistent estimator that only uses the observations with zero mean.

Estimation of the variance in model 1.1.1 can also be interpreted as a semi-parametric problem. The results in this article therefore contribute to the recent efforts to understand frequentist Bayes in semiparametric models. Semiparametric Bernstein-von Mises theorems are derived under various conditions in [70, 19, 12, 23]. For specific priors, it has been observed that there can be a large bias in the posterior limit, see [18, 23, 75]. In all the cases studied so far, it is unclear whether the bias is due to the specific choice of prior or whether this is a fundamental limitation of the Bayesian method. To the best of our knowledge, our results show for the first time that the posterior can be inconsistent for all natural priors.

Related to model 1.1.1, [34] studies Bayes for variance estimation of the errors in the nonparametric regression model. It is shown that if the posterior contracts around the true regression function with rate $o(n^{-1/4})$, the marginal posterior for the variance contracts with parametric rate around the true error variance and Bernstein-von Mises result holds.

The article is organized as follows. In Section 1.2, we discuss aspects of the problem related to the likelihood and the posterior distribution. A crucial identity for the log-posterior is derived in Section 1.3. This leads then to the general negative result in Section 1.4. The Gaussian mixture prior with parametric posterior contraction is constructed in Section 1.5. This section also contains the limiting shape result and a numerical simulation study. All the proofs are deferred to the appendix.

Notation: Vectors are denoted in bold letters, that is, $\mathbf{u} = (u_1, \ldots, u_d) \in \mathbb{R}^d$. For a vector $\mathbf{u} = (u_1, \ldots, u_k)$, we write $|\mathbf{u}|_2^2 = \sum_{i=1}^k u_i^2$ and $\overline{\mathbf{u}^2} = |u|_2^2/k$ for the averages of the squares (not to be confused with the squared averages). We write $n_1 = \lfloor n\alpha \rfloor$ and $n_2 = n - n_1$. The probability and expectation induced by model (1.1.1) are denoted by $\mathbb{P}_0^n$ and $\mathbb{E}_0^n$.

1.2 Likelihood and posterior

The MLE. For the subsequent analysis, it is convenient to split the data vector $\mathbf{X} = (X_1, \ldots, X_n)$ in the part with zero means $\mathbf{Y} = (X_1, \ldots, X_{n_1})$ and the observations with non-zero means $\mathbf{Z} = (X_{n_1+1}, \ldots, X_n)$ such that $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$. The likelihood function of the model is

$$L\big(\sigma^2, \boldsymbol{\mu} \big| \mathbf{Y}, \mathbf{Z}\big) = \underbrace{\frac{1}{(2\pi\sigma^2)^{n_1/2}} e^{-\frac{|\mathbf{Y}|_2^2}{2\sigma^2}}}_{L(\sigma^2, \boldsymbol{\mu}|\mathbf{Y})} \underbrace{\frac{1}{(2\pi\sigma^2)^{n_2/2}} e^{-\frac{|\mathbf{Z}-\boldsymbol{\mu}|_2^2}{2\sigma^2}}}_{L(\sigma^2, \boldsymbol{\mu}|\mathbf{Z})} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{|\mathbf{Y}|_2^2 + |\mathbf{Z}-\boldsymbol{\mu}|_2^2}{2\sigma^2}}. \tag{1.2.1}$$

Maximizing over $(\sigma^2, \boldsymbol{\mu})$ yields the MLE

$$\big(\widehat{\sigma}^2_{\mathrm{mle}}, \widehat{\boldsymbol{\mu}}_{\mathrm{mle}}\big) = \bigg(\frac{|\mathbf{Y}|_2^2}{n}, \mathbf{Z}\bigg).$$

If only based on the subsample $\mathbf{Y}$, the MLE for σ^{*2} would be $|\mathbf{Y}|_2^2/n_1$ and this converges to σ^{*2} with the parametric rate $n^{-1/2}$. Hence $|\mathbf{Y}|_2^2/n$ converges to $\alpha\sigma^{*2}$. The MLE for σ^{*2} is therefore inconsistent and misses the true parameter σ^{*2} by a factor α. It is clear that there is very little extractable information about the parameter σ^{*2} in $\mathbf{Z}$. A frequentist estimator can simply discard $\mathbf{Z}$ and only use the subsample $\mathbf{Y}$. The MLE also does this but leads to an incorrect scaling of the estimator.

The incorrect scaling factor of the MLE can be explained in different ways. One interpretation is that the MLE can be written as

$$\widehat{\sigma}^2_{\mathrm{mle}} = \frac{n_1}{n}\widehat{\sigma}^2_{Y,\mathrm{mle}} + \frac{n_2}{n}\widehat{\sigma}^2_{Z,\mathrm{mle}},$$

with $\widehat{\sigma}^2_{Y,\mathrm{mle}} = |\mathbf{Y}|_2^2/n_1$ the MLE based on the subsample $\mathbf{Y}$ and $\widehat{\sigma}^2_{Z,\mathrm{mle}} = 0$ the MLE based on the subsample $\mathbf{Z}$. The fact that the overall MLE just forms a linear combination of the MLEs for the subsamples shows again that too much weight is given to $\mathbf{Z}$.

Another explanation for the incorrect scaling of the MLE is to observe that in (1.2.1) the likelihood based on the second subsample is $L(\sigma^2, \boldsymbol{\mu}|\mathbf{Z}) \propto \sigma^{-n_2}$ if $\boldsymbol{\mu} = \widehat{\boldsymbol{\mu}}_{\mathrm{mle}}$. If we would take the likelihood only over the first part of the sample $\mathbf{Y}$ we would obtain the optimal estimator $|\mathbf{Y}|_2^2/n_1$, but since the likelihhod over the full sample is the product of the likelihood functions for $\mathbf{Y}$ and $\mathbf{Z}$, an additional factor σ^{-n_2} occurs in the overall likelihood which leads to the incorrect scaling. We conjecture that likelihood methods do not perform well for combined datasets where one part of the data is informative about a parameter and the other part is affected by nuisance parameters.

Adjusted profile likelihood. For the profile likelihood, we first compute the maximum likelihood estimator of the nuisance parameter for fixed σ^2, denoted by, say $\widehat{\boldsymbol{\mu}}_{\sigma^2}$, and then maximize

$$\sigma^2 \mapsto L\big(\sigma^2, \widehat{\boldsymbol{\mu}}_{\sigma^2}\big|\mathbf{Y}, \mathbf{Z}\big).$$

Obviously $\widehat{\boldsymbol{\mu}}_{\sigma^2} = Z$ for any $\sigma^2 > 0$ and the profile likelihood estimator coincides with the MLE for σ^2 in the Neyman-Scott problem. If the parameter of interest and the nuisance parameters are orthogonal with respect to the expected Fisher information, that is,

$$E\Big[\frac{\partial^2}{\partial\sigma^2\partial\mu_j}\log L\big(\sigma^2, \boldsymbol{\mu}\big|\mathbf{Y}, \mathbf{Z}\big)\Big] = 0, \quad \text{for all } j \tag{1.2.2}$$

the adjusted profile likelihood estimator [29, 64, 30] is the maximizer of

$$\sigma^2 \mapsto \mathcal{L}(\sigma^2) := \det\big(M(\sigma^2, \widehat{\boldsymbol{\mu}}_{\sigma^2})\big)^{-1/2} L\big(\sigma^2, \widehat{\boldsymbol{\mu}}_{\sigma^2}\big|\mathbf{Y}, \mathbf{Z}\big) \tag{1.2.3}$$

for the matrix valued function

$$M(\sigma^2, \boldsymbol{\mu}) := \Big(-\frac{\partial^2}{\partial\mu_j\partial\mu_\ell}\log L\big(\sigma^2, \boldsymbol{\mu}\big|\mathbf{Y}, \mathbf{Z}\big)\Big)_{j,\ell}$$

and $\det(\cdot)$ the determinant. It is easy to check that (1.2.2) holds for model (1.1.1). Since $-\partial^2/(\partial\mu_j\partial\mu_\ell)\ \log L\big(\sigma^2,\boldsymbol{\mu}\big|\mathbf{Y},\mathbf{Z}\big) = \sigma^{-2}\mathbf{1}(j=\ell)$, the adjusted profile likelihood estimator for σ^2 coincides with the MLE for the subsample $\mathbf{Y}$,

$$\widehat{\sigma}^2 = \frac{|\mathbf{Y}|_2^2}{n_1}.$$

In particular, the adjusted profile likelihood results in an unbiased $\sqrt{n}$-consistent estimator for σ^2.

The posterior distribution. From a Bayesian perspective it is quite natural to draw σ^2 and the mean vector $\boldsymbol{\mu}$ from independent distributions. Due to the orthogonality with respect to the expected Fisher information (1.2.2), we also expect no strong interactions of σ^2 and the mean parameters in the likelihood that could be taken care of by a dependent prior. Suppose that $\boldsymbol{\mu}\sim\nu$ and that the prior for σ^2 has Lebesgue density π. The marginal posterior distribution is then given by Bayes formula

$$\pi\big(\sigma^2\big|\mathbf{Y},\mathbf{Z}\big) = \frac{L(\sigma^2|\mathbf{Y},\mathbf{Z})\pi(\sigma^2)}{\int_{\mathbb{R}_+} L(\sigma^2|\mathbf{Y},\mathbf{Z})\pi(\sigma^2)\,d\sigma^2}, \tag{1.2.4}$$

with

$$L(\sigma^2|\mathbf{Y},\mathbf{Z}) = \sigma^{-n}e^{-\frac{|\mathbf{Y}|_2^2}{2\sigma^2}}\Big(\int_{\mathbb{R}^n} e^{-\frac{|\mathbf{Z}-\boldsymbol{\mu}|_2^2}{2\sigma^2}}\,d\nu(\boldsymbol{\mu})\Big). \tag{1.2.5}$$

In [78] it has been argued that by using multivariate Laplace approximation,

$$L(\sigma^2|\mathbf{Y},\mathbf{Z}) = \mathcal{L}(\sigma^2)\nu\big(\widehat{\boldsymbol{\mu}}_{\sigma^2}\big)\big(1+O_{\mathbb{P}}(n^{-1})\big) = \mathcal{L}(\sigma^2)\nu\big(Z\big)\big(1+O_{\mathbb{P}}(n^{-1})\big), \tag{1.2.6}$$

with $\mathcal{L}(\sigma^2)$ the adjusted profile likelihood in (1.2.3). This suggests that the posterior distribution should be centered around the adjusted profile likelihood estimator $|\mathbf{Y}|_2^2/n_1$, therefore correcting the MLE.

Associated sequence model with random means. For the Gaussian sequence model with partial information (1.1.1) equipped with the product prior $\pi\otimes\nu$, define the *associated sequence model with random means,* where we observe independent random variables

$$Y_i\sim\mathcal{N}(0,\sigma^{*2}),\ i=1,\ldots,n_1 \ \text{ and } \ Z_i|\boldsymbol{\mu}\sim\mathcal{N}(\mu_i,\sigma^{*2}),\ i=n_1+1,\ldots,n, \tag{1.2.7}$$

with $\boldsymbol{\mu}\sim\nu$ and ν known. In this model, the nuisance parameters are replaced by additional randomness. The only parameter in this model is σ^{*2} and the model is therefore parametric.

Remark 1.2.1. *The likelihood function of model* (1.2.7) *is* $L(\sigma^2|\mathbf{Y},\mathbf{Z})$ *and models* (1.1.1) *and* (1.2.7) *lead to the same formula in terms of* $\mathbf{Y},\mathbf{Z}$ *for the posterior distribution of* σ^2.

Bayes with improper uniform prior. If the prior on the mean vector in the Bayes formula is chosen as the Lebesgue measure, the formula for the posterior simplifies to

$$\pi\big(\sigma^2\big|\mathbf{Y},\mathbf{Z}\big)\propto\sigma^{-n_1}e^{-\frac{|\mathbf{Y}|_2^2}{2\sigma^2}}\pi(\sigma^2).$$

This is the same posterior we would get if we discarded the subsample $\mathbf{Z}$. It follows from the parametric Bernstein-von Mises theorem that if π is positive and continuous in a neighbourhood of σ^{*2}, the posterior contracts around the true variance σ^{*2}. Notice that in the case of uniform prior, the Laplace approximation in (1.2.6) is exact and does not involve any remainder terms. Obviously the Lebesgue measure is not a probability measure and the prior is improper. This raises then the question whether there are also proper priors for which the marginal posterior is consistent on the whole parameter space. We will address this problem in the next sections.

1.3 On the derivative of the log-posterior

We first derive a differential equation for the posterior. Denote by $\boldsymbol{\mu}|(\mathbf{Z}, \sigma^2)$ the posterior distribution of $\boldsymbol{\mu}$ for the sample $\mathbf{Z}$, that is,

$$d\Pi(\boldsymbol{\mu}|\mathbf{Z}, \sigma^2) = \frac{e^{-\frac{|\mathbf{Z}-\boldsymbol{\mu}|_2^2}{2\sigma^2}} d\nu(\boldsymbol{\mu})}{\int_{\mathbb{R}^n} e^{-\frac{|\mathbf{Z}-\boldsymbol{\mu}|_2^2}{2\sigma^2}} d\nu(\boldsymbol{\mu})}. \tag{1.3.1}$$

In particular, we set

$$V\big(\boldsymbol{\mu}|(\mathbf{Z}, \sigma^2)\big) := \int_{\mathbb{R}^n} |\mathbf{Z} - \boldsymbol{\mu}|_2^2 d\Pi(\boldsymbol{\mu}|\mathbf{Z}, \sigma^2). \tag{1.3.2}$$

The quantity $V(\boldsymbol{\mu}|(\mathbf{Z}, \sigma^2))$ measures the spread of $\Pi(\boldsymbol{\mu}|\mathbf{Z}, \sigma^2)$ around the vector $\mathbf{Z}$. Recall moreover the definition of $L(\mathbf{Y}, \mathbf{Z}|\sigma^2)$ in (1.2.5).

Proposition 1.3.1. *The marginal posterior satisfies*

$$\partial_{\sigma^2} \log \frac{\pi(\sigma^2|\mathbf{Y}, \mathbf{Z})}{\pi(\sigma^2)} = \partial_{\sigma^2} \log L(\sigma^2|\mathbf{Y}, \mathbf{Z}) = \frac{|\mathbf{Y}|_2^2 + V(\boldsymbol{\mu}|(\mathbf{Z}, \sigma^2))}{2\sigma^4} - \frac{n}{2\sigma^2}. \tag{1.3.3}$$

By Remark 1.2.1, the right hand side is a closed-form expression of the score function for σ^2 in the random means model (1.2.7). If the MLE in (1.2.7) does not lie on the boundary, the score function vanishes at the MLE. From the Bernstein-van Mises phenomenon it is conceivable that the posterior will concentrate around this MLE. For the MLE to be close to the truth σ^{*2}, the score function evaluated at σ^{*2} must be $o_{\mathbb{P}}(1)$. Since $|\mathbf{Y}|_2^2 = n\alpha\sigma^{*2} + O_{\mathbb{P}}(\sqrt{n})$, this leads to the condition

$$\frac{V(\boldsymbol{\mu}|(\mathbf{Z}, \sigma^{*2}))}{n} = (1-\alpha)\sigma^{*2} + o_{\mathbb{P}}(1).$$

In the next section, we derive a very general negative result. The main part of the argument is to show that the previous equality does not hold in a neighborhood of σ^{*2}, see (1.A.12).

1.4 Posterior inconsistency for product priors

In this section we study posterior contraction under the following condition.

Prior. The prior on $\boldsymbol{\mu}$ is independent of the prior on σ^2. Under the prior, each component of the mean vector $\boldsymbol{\mu}$ is drawn independently from a distribution ν on $\mathbb{R}$. The prior on σ^2 has a positive and continuously differentiable Lebesgue density on $\mathbb{R}_+$.

So far ν denoted the prior on the mean vector. By a slight abuse of language we denote the prior on the individual components also by ν. The assumptions on the prior are mild enough to account for proper priors with heavy tails and possibly no moments.

The i.i.d. prior is the natural choice, if we believe that there is no structure in the nonzero means. From (1.2.7) it follows that the corresponding sequence model with random means is

$$Y_i \sim \mathcal{N}(0, \sigma^{*2}),\ i = 1, \ldots, n_1 \text{ and } Z_i|\mu_i \sim \mathcal{N}(\mu_i, \sigma^{*2}),\ i = n_1 + 1, \ldots, n, \tag{1.4.1}$$

with $\mu_i \sim \nu$. For $\alpha = 1/2$ and unknown ν, this model has been studied in [51]. It is shown that the MLE for σ^{*2} and the MLE for the distribution function of the means are consistent. Since the random means model leads to the same posterior distribution as explained in Remark 1.2.1, this suggests that the posterior might concentrate around the truth.

We now provide a second heuristic that leads to a different conclusion indicating that it makes a huge difference whether the distribution of the means ν is known or unknown. In the framework of (1.4.1), ν is known. If $\int u^2 d\nu(u) < \infty$, then $\overline{\boldsymbol{\mu}^2} = \int u^2 d\nu(u) + O_{\mathbb{P}}(n^{-1/2})$ and $\overline{\mathbf{Z}^2} = \overline{\boldsymbol{\mu}^2} + \sigma^{*2} + O_{\mathbb{P}}(n^{-1/2})$, so we have $\overline{\mathbf{Z}^2} - \int u^2 d\nu(u) = \sigma^{*2} + O_{\mathbb{P}}(n^{-1/2})$. This means that model (1.4.1) carries a lot of information about σ^{*2} in the sense that σ^{*2} can be estimated with parametric rate from the subsample $\mathbf{Z}$ only. Since the posterior only sees model (1.4.1) it is therefore natural to give a lot of weight to the subsample $\mathbf{Z}$ as well, which, from a frequentist perspective, is wrong.

This heuristic does not say anything about heavy-tailed priors with $\int u^2 d\nu(u) = \infty$. But even in this case, we will show that the posterior is inconsistent. The first result states that in a neighborhood of σ^{*2} the posterior is increasing extremely fast with high probability.

Proposition 1.4.1. *Given $\alpha < 1$ and the prior above, then, for all sufficiently large σ^{*2}, there exists a mean vector $\boldsymbol{\mu}^*$, such that*

$$\lim_{n\to\infty} \mathbb{P}_0^n \Big(\Big\{ \partial_{\sigma^2} \log \pi(\sigma^2|\mathbf{Y}, \mathbf{Z}) \geq \sigma^{*-2} n,\ \forall \sigma^2 \in \Big[\frac{\sigma^{*2}}{2}, 2\sigma^{*2}\Big] \Big\} \Big) = 1.$$

The proof of Proposition 1.4.1 constructs a lower bound on σ^{*2} that is independent of n and moreover guarantees that ν has sufficiently small mass outside $[-\sigma^{*2}, \sigma^{*2}]$. It therefore depends on the tail behavior of the prior mean distribution ν. The mean vector μ^* is subsequently chosen with all means being equal to an expression depending on σ^*. Thus the means in μ^* are uniformly bounded and independent of n as well.

Suppose that almost all posterior mass is close to σ^{*2}. By the previous proposition, the posterior is increasing at least up to $2\sigma^{*2}$. Hence, there must be even more mass around $2\sigma^{*2}$. This is a contradiction and shows that the posterior does not concentrate around σ^{*2}. The proof of the next theorem is based on this argument. For this result, the means in the vector $\boldsymbol{\mu}^*$ can again be chosen to be uniformly bounded.

Theorem 1.4.2. *Given $\alpha < 1$ and the prior above, then, for all sufficiently large σ^{*2}, there exists a mean vector $\boldsymbol{\mu}^*$ such that*

$$\lim_{n\to\infty} \mathbb{E}_0^n\Big[\Pi\Big(\Big|\frac{\sigma^2}{\sigma^{*2}} - 1\Big| \leq \frac{1}{2}\Big|\mathbf{Y}, \mathbf{Z}\Big)\Big] = 0.$$

Consequently, the posterior is inconsistent and assigns all its mass outside of a neighbourhood of the true variance.

The posterior is therefore inferior if compared to the frequentist variance estimator $\overline{\mathbf{Y}^2}$, which achieves the parametric rate $n^{-1/2}$ in the sense that

$$\sup_{\sigma^{*2}>0} \mathbb{E}_0^n\Big[\Big|\frac{\overline{\mathbf{Y}^2}}{\sigma^{*2}} - 1\Big|\Big] \lesssim n^{-1/2}.$$

It is remarkable that no conditions on the tail behavior of the prior distribution ν are required for Theorem 1.4.2. Recall that for the improper uniform prior the posterior contract around σ^{*2}. This shows that for distributions with heavy tailed densities, we need very sharp bounds.

To the best of our knowledge there are no negative results in the nonparametric Bayes literature that hold for such a large class of priors. The proof strategy to establish Proposition 1.4.1 is based on a highly non-standard shrinkage argument that will be sketched here. By expanding the square term in (1.3.2) we can lower bound (1.3.3) by

$$\partial_{\sigma^2} \log \pi(\sigma^2|\mathbf{Y}, \mathbf{Z}) \geq \frac{|\mathbf{Y}|_2^2}{2\sigma^4} + \frac{|\mathbf{Z}|_2^2}{2\sigma^4} - \frac{n}{2\sigma^2} - \frac{1}{\sigma^4}\sum_{i=1}^{n_2} V_i + O_{\mathbb{P}}(1),$$

where $V_i := |Z_i| \int |\mu_i| d\Pi(\boldsymbol{\mu}|Z_i, \sigma^2)$. For σ^2 close to σ^{*2}, we have

$$\partial_{\sigma^2} \log \pi(\sigma^2|\mathbf{Y}, \mathbf{Z}) \geq \frac{n_2\overline{\boldsymbol{\mu}^{*2}}}{2\sigma^{*4}} - \frac{1}{\sigma^{*4}}\sum_{i=1}^{n_2} V_i + O_{\mathbb{P}}(\sqrt{n}).$$

For improper uniform prior, one can check that $V_i \geq Z_i^2$, making the lower bound negative and useless. For proper prior, there is a shrinkage phenomenon in the sense that for all $c > 0$ there are parameters $(\mu_i^*)^2 \asymp \sigma^{*2}$ such that $V_i \leq cZ_i^2$, with high $\mathbb{P}_0^n$-probability. If this is the case then

$$\partial_{\sigma^2} \log \pi(\sigma^2|\mathbf{Y}, \mathbf{Z}) \geq \Big(\frac{1}{2} - 2c\Big)\frac{n_2}{2\sigma^{*2}} + O_{\mathbb{P}}(\sqrt{n}),$$

which yields the conclusion by choosing $c > 0$ small enough.

In Proposition 1.4.1 we showed that the posterior overshoots the true variance σ^{*2} whenever the true means are large enough. By analyzing the Gaussian case in the next section, we see that for small means the posterior will in fact underestimate σ^{*2} and that only for a small range of means vectors, one can hope that the posterior will be able to concentrate around the true variance.

1.5 Gaussian mixture priors

1.5.1 Gaussian priors

To illustrate our approach, we first consider an i.i.d. Gaussian prior on the mean vector

$$\mu_i \sim \mathcal{N}(0, \theta^2), \text{ independently.}$$

From Theorem 1.4.2 we already know that the posterior will be inconsistent in this case. Nevertheless, the Gaussian assumptions yields more explicit formulas and this allows us to build a hierarchical prior that leads to good posterior contraction properties. By Remark 1.2.1, the marginal likelihood is the same as in the sequence model with random means (1.4.1). The marginal posterior is therefore

$$\pi\big(\sigma^2 \big| \mathbf{Y}, \mathbf{Z}\big) \propto \sigma^{-n_1} (\theta^2 + \sigma^2)^{-\frac{n_2}{2}} e^{-\frac{|\mathbf{Y}|_2^2}{2\sigma^2}} e^{-\frac{|\mathbf{Z}|_2^2}{2(\theta^2+\sigma^2)}} \pi(\sigma^2), \tag{1.5.1}$$

which can also be written as the product of two inverse Gamma densities. In view of the Bernstein-von Mises phenomenon, the posterior concentrates around the MLE for parametric problems. Similarly, we can argue here that the posterior will be concentrated around the value $\widehat{\sigma}^2$ maximizing the likelihood part of the posterior (1.5.1). By differentiation, we find $n_1 \widehat{\sigma}^2 + n_2 \widehat{\sigma}^4/(\widehat{\sigma}^2 + \theta^2) = |\mathbf{Y}|_2^2 + \widehat{\sigma}^4 |\mathbf{Z}|_2^2/(\theta^2 + \widehat{\sigma}^2)^2$ and rearranging yields

$$\widehat{\sigma}^2 - \overline{\mathbf{Y}^2} = \frac{n_2}{n_1} \left(\frac{\widehat{\sigma}^2}{\theta^2 + \widehat{\sigma}^2} \right)^2 \big[\overline{\mathbf{Z}^2} - \theta^2 - \widehat{\sigma}^2 \big].$$

This can be rewritten as

$$\widehat{\sigma}^2 - \sigma^{*2} + O_{\mathbb{P}}(n^{-1/2}) = \frac{1-\alpha}{\alpha} \big(1 + O(n^{-1})\big) \left(\frac{\widehat{\sigma}^2}{\theta^2 + \widehat{\sigma}^2} \right)^2 \left[\sigma^{*2} - \widehat{\sigma}^2 + \overline{\boldsymbol{\mu}^{*2}} + O_{\mathbb{P}}(n^{-1/2}) - \theta^2 \right], \tag{1.5.2}$$

where we set

$$\overline{\boldsymbol{\mu}^{*2}} = |\boldsymbol{\mu}^*|_2^2 / n_2$$

and suppress the dependence of the $O()$ term on σ^{*2} and $\boldsymbol{\mu}^*$. If θ is fixed, this shows that for $\widehat{\sigma}^2 = \sigma^{*2} + O_{\mathbb{P}}(n^{-1/2})$ we need

$$\overline{\boldsymbol{\mu}^{*2}} = \theta^2 + O_{\mathbb{P}}(n^{-1/2}). \tag{1.5.3}$$

Differently speaking, to force the maximum $\widehat{\sigma}^2$ to be close to σ^{*2}, the variance θ^2 of the prior has to match the empirical variance $\overline{\boldsymbol{\mu}^{*2}}$ of the nuisance parameter. We can also deduce from (1.5.2) that if $|\overline{\boldsymbol{\mu}^{*2}} - \theta^2| \gg n^{-1/2}$ and θ is fixed, then also $|\widehat{\sigma}^2 - \sigma^{*2}| \gg n^{-1/2}$. More precisely, we even have that $\overline{\boldsymbol{\mu}^{*2}} - \theta^2 \gg n^{-1/2}$ implies $\widehat{\sigma}^2 - \sigma^{*2} \gg n^{-1/2}$ and $\overline{\boldsymbol{\mu}^{*2}} - \theta^2 \ll -n^{-1/2}$ implies $\widehat{\sigma}^2 - \sigma^{*2} \ll -n^{-1/2}$. This shows that, depending on the size of $\overline{\boldsymbol{\mu}^{*2}}$ compared to θ^2, the posterior can either overestimate or underestimate the true variance.

If θ is allowed to vary with n, we can make the right hand side in (1.5.2) arbitrarily small by letting θ tend to infinity. As θ^2 is the variance of the prior, the behaviour resembles then that of the uniform improper prior, which, as we already know, leads to posterior consistency. If we think of a prior as a prior belief on parameters, then the prior should not change depending on the amount of available data and, in particular, it is unnatural that the prior becomes more vague if the sample size increases. In the next section we show that there are sample size independent mixture priors leading to a parametric posterior contraction rates.

1.5.2 Mixture priors

Section 1.4 explains the posterior inconsistency for i.i.d. prior on the nuisance. It seems not intuitive that adding dependency on the prior of the nuisance parameter can help avoiding posterior inconsistency for σ^{*2}. Surprisingly, this is not true. In this section, we first provide some intuition why mixture priors can resolve the issues of i.i.d. priors. Afterwards, we discuss and analyze a specific prior construction.

Analyzing Gaussian prior above, (1.5.3) suggests that for any nuisance parameter vector $\boldsymbol{\mu}^*$, there exists an i.i.d. prior which seems to work. This i.i.d. prior does, however, depend on the unknown μ^* and can therefore not be chosen without knowledge of the data. Intuitively, if the posterior had the chance to see all possible i.i.d. priors on $\boldsymbol{\mu}$, instead of just one, it is conceivable that it would automatically select one that is adapted to the unknown nuisance parameter and consequently leads to posterior consistency for the parameter of interest. De Finetti's theorem [47] states that an exchangeable prior ν over the infinite sequence $\boldsymbol{\mu} = (\mu^1, \mu^2, \dots)$ can be written as a mixture over i.i.d. priors in the sense that

$$\nu(A^1 \times \cdots \times A^k) := \int_{\mathcal{P}(\mathbb{R})} Q(A^1) \cdots Q(A^k) \lambda(dQ),$$

with λ a probability measure on the set of probability densities $\mathcal{P}(\mathbb{R})$ on $\mathbb{R}$. The posterior (1.2.4) then becomes

$$\begin{aligned}\pi\big(\sigma^2 \big| \mathbf{Y}, \mathbf{Z}\big) &\propto \pi(\sigma^2) \int_{\mathbb{R}^n} \frac{L(\sigma^2, \boldsymbol{\mu} | \mathbf{Y}, \mathbf{Z})}{L(\sigma^{*2}, \boldsymbol{\mu}^* | \mathbf{Y}, \mathbf{Z})} \nu(\boldsymbol{\mu}) d\boldsymbol{\mu}, \\ &= \pi(\sigma^2) \int_{\mathcal{P}(\mathbb{R})} \bigg(\int_{\mathbb{R}^n} \frac{L(\sigma^2, \boldsymbol{\mu} | \mathbf{Y}, \mathbf{Z})}{L(\sigma^{*2}, \boldsymbol{\mu}^* | \mathbf{Y}, \mathbf{Z})} \prod_{i=1}^n q(\mu^i) d\mu^i \bigg) \lambda(dq),\end{aligned}$$

where q denotes the probability density function of Q. Let q_0 be the i.i.d. prior maximizing the interior integral. Suppose that this is a unique maximum and that the outer integral is determined by the behavior of the integrand in a suitable neighborhood $\mathcal{S}$ of q_0. This means that

$$\begin{aligned}\pi\big(\sigma^2 \big| \mathbf{Y}, \mathbf{Z}\big) &\propto \pi(\sigma^2) \int_{\mathcal{P}(\mathbb{R})} \bigg(\int_{\mathbb{R}^n} \frac{L(\sigma^2, \boldsymbol{\mu} | \mathbf{Y}, \mathbf{Z})}{L(\sigma^{*2}, \boldsymbol{\mu}^* | \mathbf{Y}, \mathbf{Z})} \prod_{i=1}^n q(\mu^i) d\mu^i \bigg) \lambda(dq) \\ &\approx \pi(\sigma^2) \int_{\mathcal{S}} \bigg(\int_{\mathbb{R}^n} \frac{L(\sigma^2, \boldsymbol{\mu} | \mathbf{Y}, \mathbf{Z})}{L(\sigma^{*2}, \boldsymbol{\mu}^* | \mathbf{Y}, \mathbf{Z})} \prod_{i=1}^n q(\mu^i) d\mu^i \bigg) \lambda(dq) \\ &\approx \pi(\sigma^2) \bigg(\int_{\mathbb{R}^n} \frac{L(\sigma^2, \boldsymbol{\mu} | \mathbf{Y}, \mathbf{Z})}{L(\sigma^{*2}, \boldsymbol{\mu}^* | \mathbf{Y}, \mathbf{Z})} \prod_{i=1}^n q_0(\mu^i) d\mu^i \bigg) \int_{\mathcal{S}} \lambda(dq).\end{aligned}$$

The right hand side is the posterior density of σ^2 for i.i.d. prior $\prod_{i=1}^n q_0(\mu^i)$ on the components.

Although this argument is only a sketch, it suggests that something might be gained by mixing over i.i.d. priors instead of just choosing one. Maximizing the marginalized likelihood in (1.5.1) over θ^2 yields

$$\theta^2 = \overline{\mathbf{Z}^2} - \sigma^2, \tag{1.5.4}$$

if the r.h.s. is non-negative. For this choice of θ^2, (1.5.1) becomes $\pi(\sigma^2|\mathbf{Y},\mathbf{Z}) \propto \sigma^{-n_1}\exp(-|\mathbf{Y}|_2^2/(2\sigma^2))\pi(\sigma^2)$. The posterior therefore coincides with the posterior density based on the first part of the sample only which we know has good posterior contraction properties.

Prior. In a first step generate $\theta^2 \sim \gamma$, with γ a positive Lebesgue density on $\mathbb{R}_+$. Given θ^2, each non-zero mean is drawn independently from a centered normal distribution with variance θ^2, that is, $\mu_i|\theta^2 \sim \mathcal{N}(0,\theta^2)$, $i > n_1$.

Another heuristic about the posterior properties for this prior can again be derived by making the link to the associated sequence model with random means (1.2.7). For the prior considered here, the random means model has the form

$$Y_i \sim \mathcal{N}(0,\sigma^{*2}),\ i = 1,\ldots,n_1 \text{ and } Z_i|\theta^2 \sim \mathcal{N}(0,\theta^2+\sigma^{*2}),\ i = n_1+1,\ldots,n, \tag{1.5.5}$$

with $\theta^2 \sim \gamma$. If θ^2 were a second parameter and not generated from γ, the variance σ^{*2} would not be identifiable if only the Z_i's are observed. In model (1.5.5) we know the density γ, but this is not enough to consistently reconstruct σ^{*2} from the subsample $\mathbf{Z}$. By Remark 1.2.1, this model leads to the same posterior for σ^2. The posterior should therefore realize that there is little extractable information about σ^{*2} in $\mathbf{Z}$ and discard these observations. We will see in the limiting shape result below that this is roughly what happens.

We denote by $\ell(\sigma^2|\mathbf{Y})$ and $\ell(\sigma^2+\theta^2|\mathbf{Z})$ the log-likelihoods of the sub-samples $\mathbf{Y}$ and $\mathbf{Z}$ coming from model (1.5.5) with σ^2 replacing σ^{*2}, that is

$$\begin{aligned}\ell(\sigma^2|\mathbf{Y}) &= -\frac{n_1}{2}\log(2\pi\sigma^2) - \frac{n_1\overline{\mathbf{Y}^2}}{2\sigma^2},\\ \ell(\sigma^2+\theta^2|\mathbf{Z}) &= -\frac{n_2}{2}\log(2\pi(\sigma^2+\theta^2)) - \frac{n_2\overline{\mathbf{Z}^2}}{2(\sigma^2+\theta^2)}.\end{aligned} \tag{1.5.6}$$

The log-likelihoods appearing in (1.5.6) can be written in terms of inverse-gamma distributions. We denote by $\mathrm{IG}(\gamma,\beta)$ the inverse-gamma distribution with shape $\gamma > 0$ and scale $\beta > 0$. The corresponding p.d.f. is

$$f_{\mathrm{IG}(\gamma,\beta)}(x) = \frac{\beta^\gamma}{\Gamma(\gamma)}x^{-\gamma-1}e^{-\frac{\beta}{x}}, \tag{1.5.7}$$

where $\Gamma(\cdot)$ is the Gamma function. Rewriting the posterior, we have the following.

Lemma 1.5.1. *Under the Gaussian mixture prior, the marginal posterior density has the form*

$$\pi(\sigma^2|\mathbf{Y},\mathbf{Z}) \propto f_{\mathrm{IG}(\gamma_1,\beta_1)}(\sigma^2)\left(\int_0^{+\infty} f_{\mathrm{IG}(\gamma_2,\beta_2)}(\sigma^2+\theta^2)\gamma(\theta^2)d\theta^2\right)\pi(\sigma^2), \tag{1.5.8}$$

with $\gamma_1 = n_1/2 - 1$, $\beta_1 = n_1\overline{\mathbf{Y}^2}/2$ *and* $\gamma_2 = n_2/2 - 1$, $\beta_2 = n_2\overline{\mathbf{Z}^2}/2$. *The* $\mathrm{IG}(\gamma_1, \beta_1)$*-distribution has mode* $\beta_1/(\gamma_1+1) = \overline{\mathbf{Y}^2}$ *and variance* $\beta_1^2/(\gamma_1-1)^2(\gamma_1-2) = O(n^{-1})$*, whereas the* $\mathrm{IG}(\gamma_2, \beta_2)$*-distribution has mode* $\beta_2/(\gamma_2+1) = \overline{\mathbf{Z}^2}$ *and variance* $\beta_2^2/(\gamma_2-1)^2(\gamma_2-2) = O(n^{-1})$.

Starting from Lemma 1.5.1, we can develop a heuristic argument on how to recover the shape of the limit posterior distribution. We interpret the posterior $\Pi(\cdot|\mathbf{Y}, \mathbf{Z})$ with density (1.5.8) as the marginalized version, over the set $\theta^2 \in (0, +\infty)$, of the distribution $\widetilde{\Pi}(\cdot|\mathbf{Y}, \mathbf{Z})$ whose density is given by

$$\widetilde{\pi}(\sigma^2, \theta^2|\mathbf{Y}, \mathbf{Z}) \propto f_{\mathrm{IG}(\gamma_1,\beta_1)}(\sigma^2) f_{\mathrm{IG}(\gamma_2,\beta_2)}(\sigma^2 + \theta^2)\gamma(\theta^2)\pi(\sigma^2), \tag{1.5.9}$$

and refer to $\widetilde{\Pi}(\cdot|\mathbf{Y}, \mathbf{Z})$ as the joint posterior on $(\sigma^2, \theta^2) \in (0, +\infty)^2$. The first step is double localization. Thanks to the exponential tails of the inverse Gamma distribution, the joint posterior $\widetilde{\Pi}(\cdot|\mathbf{Y}, \mathbf{Z})$ asymptotically concentrates on the set $\{\sigma^2 \in B_1\} \cap \{\theta^2 \in B_2\}$, with B_1 a $O(\zeta_n)$-ball centered at $\overline{\mathbf{Y}^2}$ and B_2 a $O(\zeta_n)$-ball around $0 \vee (\overline{\mathbf{Z}^2} - \overline{\mathbf{Y}^2})$ for any sequence $\zeta_n \gg n^{-1/2}$. This also implies that the joint posterior (1.5.9) is arbitrarily close, in total variation distance, to the truncated posterior distribution with density $\widetilde{\pi}(\sigma^2, \theta^2|\mathbf{Y}, \mathbf{Z})\mathbf{1}(\{\sigma^2 \in B_1\} \cap \{\theta^2 \in B_2\})$. In particular, this means that the hyperparameter θ^2 concentrates on a neighborhood of the maximal value derived in (1.5.4).

Arguing as in the classical proof of the Bernstein-von Mises theorem, we can then show that the truncated posterior distribution will asymptotically not depend on the prior and prove that the posterior given by (1.5.8) behaves asymptotically like

$$\pi_1(\sigma^2|\mathbf{Y}, \mathbf{Z}) = \mathbf{1}(\sigma^2 \in B_1) f_{\mathrm{IG}(\gamma_1,\beta_1)}(\sigma^2) \int_{B_2} f_{\mathrm{IG}(\gamma_2,\beta_2)}(\sigma^2 + \theta^2) d\theta^2. \tag{1.5.10}$$

Using essentially Laplace approximation, we show that the log-likelihoods $\ell(\sigma^2|\mathbf{Y})$ and $\ell(\sigma^2 + \theta^2|\mathbf{Z})$ in (1.5.6) can be always uniformly approximated by a second-order Taylor expansion around their maxima $\overline{\mathbf{Y}^2}$ and $\overline{\mathbf{Z}^2} - \sigma^2$, and thus the localized posterior converges in total variation distance to a distribution with density

$$\pi_2(\sigma^2|\mathbf{Y}, \mathbf{Z}) \propto \mathbf{1}(\sigma^2 \in B_1) e^{-\frac{n_1}{4\sigma^{*4}}(\sigma^2 - \overline{\mathbf{Y}^2})^2} \int_{B_2} e^{-\frac{n_2}{4(\sigma^{*2} + \overline{\boldsymbol{\mu}^{*2}})^2}(\theta^2 + \sigma^2 - \overline{\mathbf{Z}^2})^2} d\theta^2, \tag{1.5.11}$$

whose factors are a truncated Gaussian density with mode $\overline{\mathbf{Y}^2}$ and variance $2\sigma^{*4}/n_1 = O(n^{-1})$ and the integral of a truncated Gaussian density with mode $\overline{\mathbf{Z}^2} - \sigma^2$ and variance $2(\sigma^{*2} + \overline{\boldsymbol{\mu}^{*2}})^2/n_2 = O(n^{-1})$. By undoing the localization argument, we can show that the restriction to the sets B_1 and B_2 can be removed from (1.5.11) and the posterior given by (1.5.8) converges in total variation distance to the posterior limit distribution

$$\pi_\infty(\sigma^2|\mathbf{Y}, \mathbf{Z}) \propto \mathbf{1}(\sigma^2 \geq 0) e^{-\frac{n_1}{4\sigma^{*4}}(\sigma^2 - \overline{\mathbf{Y}^2})^2} \left[1 - \Phi\left(\frac{\sqrt{n_2}(\sigma^2 - \overline{\mathbf{Z}^2})}{\sqrt{2}(\sigma^{*2} + \overline{\boldsymbol{\mu}^{*2}})}\right)\right], \tag{1.5.12}$$

with Φ the c.d.f. of the standard normal distribution. Recall that $\overline{\mathbf{Z}^2} \approx \sigma^{*2} + \overline{\boldsymbol{\mu}^{*2}}$. This suggests that the term involving Φ in the posterior limit distribution should asymptotically disappear if $\overline{\boldsymbol{\mu}^{*2}} \gg n^{-1/2}$. The limit of the posterior should then be the truncated Gaussian

$$\widetilde{\pi}_\infty(\sigma^2|\mathbf{Y}) \propto \mathbf{1}(\sigma^2 \geq 0) \exp\Big(- \frac{n_1}{4\sigma^{*4}}(\sigma^2 - \overline{\mathbf{Y}^2})^2\Big), \tag{1.5.13}$$

with mode $\overline{\mathbf{Y}^2}$ and variance $2\sigma^{*4}/n_1 = O(n^{-1})$.

The next result is a formal statement of the arguments mentioned above. To pass to (1.5.13) involves an additional $\log n$-factor in the signal strength of $\overline{\boldsymbol{\mu}^{*2}}$. Denote by $\|\cdot\|_{\mathrm{TV}}$ the total variation distance and recall that the expectation $\mathbb{E}_0^n$ is taken with respect to model (1.1.1).

Theorem 1.5.2. *Let* $\Pi_\infty(\cdot|\mathbf{Y},\mathbf{Z})$ *and* $\widetilde{\Pi}_\infty(\cdot|\mathbf{Y})$ *be the distributions corresponding to the densities* (1.5.12) *and* (1.5.13)*, respectively. If the prior densities* $\gamma, \pi : [0,\infty) \to (0,\infty)$ *are positive and uniformly continuous, then, for any compact sets* $K \subset (0,\infty), K' \subset (-\infty,\infty)$*, and* $n \to \infty$*,*

$$\sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} \mathbb{E}_0^n \Big[\big\| \Pi(\cdot|\mathbf{Y},\mathbf{Z}) - \Pi_\infty(\cdot|\mathbf{Y},\mathbf{Z}) \big\|_{\mathrm{TV}} \Big] \to 0.$$

Moreover, if $\inf_{\mu_i^*\in K', \forall i} |\mu_i^*| \gg (\log n/n)^{1/4}$*, then*

$$\sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} \mathbb{E}_0^n \Big[\big\| \Pi(\cdot|\mathbf{Y},\mathbf{Z}) - \widetilde{\Pi}_\infty(\cdot|\mathbf{Y}) \big\|_{\mathrm{TV}} \Big] \to 0.$$

As a corollary of the proof, posterior contraction around the true variance σ^{*2} with contraction rate $O(\sqrt{\log n/n})$ can be established. In the case of large means this is an immediate consequence of the posterior limit $\widetilde{\Pi}_\infty(\cdot|\mathbf{Y})$ and the parametric Bernstein-von Mises theorem. For small means it is less obvious because of the non-standard limit of the posterior.

Corollary 1.5.3. *There exists a constant* $M = M(\alpha)$*, such that*

$$\sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} E_0^n \Big[\Pi\Big(\Big| \frac{\sigma^2}{\sigma^{*2}} - 1 \Big| \geq M\sqrt{\frac{\log n}{n}} \Big| \mathbf{Y},\mathbf{Z}\Big)\Big] \to 0.$$

The posterior limit distribution is closely related to the class of skew normal distributions, see [5, 6]. We now derive an alternative characterization of the limit distribution. From the argumentation above, the p.d.f.

$$\propto \mathbf{1}(\sigma^2, \theta^2 \geq 0) e^{-\frac{n_1}{4\sigma^{*4}}(\sigma^2 - \overline{\mathbf{Y}^2})^2} e^{-\frac{n_2}{4(\sigma^{*2}+\overline{\boldsymbol{\mu}^{*2}})^2}(\theta^2+\sigma^2-\overline{\mathbf{Z}^2})^2} \tag{1.5.14}$$

can be viewed as the joint posterior limit of (σ^2, θ^2). In particular, the posterior limit distribution is the marginal distribution with respect to σ^2. As this is clear from the context, we do not write explicitly that the following distributions are conditional on $\mathbf{Y}, \mathbf{Z}$, that is, $\mathbf{Y}, \mathbf{Z}$ are assumed to be fixed.

Lemma 1.5.4. *Let*

$$\xi \sim \mathcal{N}\Big(\overline{\mathbf{Y}^2}, \frac{2\sigma^{*4}}{n_1}\Big), \quad \eta \sim \mathcal{N}\Big(\overline{\mathbf{Z}^2}, \frac{2(\sigma^{*2} + \overline{\boldsymbol{\mu}^{*2}})^2}{n_2}\Big).$$

be independent. The distribution with p.d.f. (1.5.14) *coincides with the distribution of*

$$(\xi, \eta - \xi)\big|(0 \leq \xi \leq \eta).$$

In particular, the posterior limit distribution $\Pi_\infty(\cdot|\mathbf{Y}, \mathbf{Z})$ *coincides with the distribution of*

$$\xi\big|(0 \leq \xi \leq \eta).$$

If the standard deviations of η, ξ are small compared to the means, the posterior limit distribution essentially compares the means $\overline{\mathbf{Y}^2}$ and $\overline{\mathbf{Z}^2}$. This behavior is very reasonable because if $\overline{\boldsymbol{\mu}^{*2}}$ is small, $\overline{\mathbf{Y}^2} \approx \overline{\mathbf{Z}^2}$ and the subsample $\mathbf{Z}$ becomes informative about σ^2.

The posterior limit depends on unknown quantities. A frequentist estimator mimicking the posterior would be to estimate σ^2 from the MLE for zero means $\overline{\mathbf{X}^2}$ in the case that the means are small. To detect whether small means are present, we can check whether $\overline{\mathbf{Y}^2} \geq \overline{\mathbf{Z}^2}$, which leads then to the estimator

$$\widetilde{\sigma}^2 = \begin{cases} \overline{\mathbf{Y}^2}, & \text{if } \overline{\mathbf{Y}^2} < \overline{\mathbf{Z}^2}, \\ \overline{\mathbf{X}^2}, & \text{otherwise.} \end{cases}$$

1.5.3 Finite sample analysis

We compare the estimators $\widehat{\sigma}_Y^2 = \overline{\mathbf{Y}^2}$ and $\widetilde{\sigma}^2$ to the maximum $\widehat{\sigma}^2_{\text{map},\infty}$ and the mean $\widehat{\sigma}^2_{\text{mean},\infty}$ of the limit density $\sigma^2 \mapsto \pi_\infty(\sigma^2|\mathbf{Y}, \mathbf{Z})$ for sample sizes $n \in \{10, 100, 1000\}$. As discussed above, we expect to see some differences for small means. We study the performances for $\sigma^{*2} = 1$ and $\boldsymbol{\mu}$ the vector with all entries equal to $t/n^{1/4}$ for the values $t \in \{0, 1, 2, 5\}$. Since $\widehat{\sigma}_Y^2$ does not depend on the means, the estimator performs equally well in all setups. Table 1.1 reports the average of the squared errors and the corresponding standard errors based on 10.000 repetitions. The rescaled MLE $\widehat{\sigma}_Y^2$ performs worse than any of the other estimators for small signals. Among the other estimators there is no clear 'winner'. For $t = 5$, the risk of all estimators is nearly the same. For larger values of t, our simulation experiments did not show any changes compared to $t = 5$ and the results are therefore omitted from the table.

There has been a long-standing debate whether Bayesian methods perform well if interpreted as frequentist methods. Results like the complete class theorem and the Bernstein-von Mises theorem have been foundational in this regard, see [53, 41]. Our theory highlights another instance where Bayes leads to new estimators with good finite sample properties. The analysis moreover shows that the construction of a prior resulting in a posterior with good frequentist properties can be highly non-intuitive.

Table 1.1: Comparison of the estimators for $(\sigma^{*2}, \boldsymbol{\mu}^*) = (1, (t/n^{1/4}, \dots, t/n^{1/4}))$ and $t \in \{0, 1, 2, 5\}$.

Estim.	n	0	1	2	5
$\widehat{\sigma}^2_Y$	10	0.414 (± 8.7e-03)	0.411 (± 8.6e-03)	0.386 (± 8.2e-03)	0.399 (± 8.4e-03)
	100	0.040 (± 5.9e-04)	0.040 (± 5.9e-04)	0.390 (± 5.7e-04)	**0.041** (± 6.4e-04)
	1000	0.004 (± 5.7e-05)	0.004 (± 5.6e-05)	0.004 (± 5.8e-05)	**0.004** (± 5.8e-05)
$\widetilde{\sigma}^2$	10	0.235 (± 3.1e-03)	0.268 (± 4.2e-03)	0.336 (± 6.2e-03)	0.399 (± 8.4e-03)
	100	**0.028** (± 3.8e-04)	**0.031** (± 4.2e-04)	0.037 (± 5.2e-04)	**0.041** (± 6.4e-04)
	1000	**0.003** (± 4.3e-05)	**0.003** (± 4.4e-05)	0.004 (± 5.4e-05)	**0.004** (± 5.8e-05)
$\widehat{\sigma}^2_{\text{map},\infty}$	10	0.337 (± 3.3e-03)	0.330 (± 4.6e-03)	0.359 (± 6.9e-03)	0.398 (± 8.3e-03)
	100	0.036 (± 4.3e-04)	0.032 (± 4.2e-04)	**0.034** (± 4.7e-04)	**0.041** (± 6.3e-04)
	1000	**0.003** (± 4.9e-05)	**0.003** (± 4.5e-05)	**0.003** (± 4.9e-05)	**0.004** (± 5.8e-05)
$\widehat{\sigma}^2_{\text{mean},\infty}$	10	**0.167** (± 2.1e-03)	**0.182** (± 3.8e-03)	**0.232** (± 5.9e-03)	**0.283** (± 7.0e-03)
	100	0.040 (± 4.5e-04)	0.034 (± 4.3e-04)	**0.034** (± 4.7e-04)	**0.041** (± 6.2e-04)
	1000	0.004 (± 5.1e-05)	**0.003** (± 4.6e-05)	**0.003** (± 4.9e-05)	**0.004** (± 5.8e-05)

Appendix 1.A Proofs

1.A.1 Proofs for Section 1.3

Proof of Proposition 1.3.1. By direct computation,

$$\partial_{\sigma^2} \log L(\sigma^2|\mathbf{Y}, \mathbf{Z}) = -\frac{n}{2\sigma^2} + \frac{|\mathbf{Y}|_2^2}{2\sigma^4} + \frac{\partial_{\sigma^2}\big(\int e^{-\frac{|\mathbf{Z}-\boldsymbol{\mu}|_2^2}{2\sigma^2}} d\nu(\boldsymbol{\mu})\big)}{\int e^{-\frac{|\mathbf{Z}-\boldsymbol{\mu}|_2^2}{2\sigma^2}} d\nu(\boldsymbol{\mu})}.$$

Since

$$\partial_{\sigma^2}\Big(\int e^{-\frac{|\mathbf{Z}-\boldsymbol{\mu}|_2^2}{2\sigma^2}} d\nu(\boldsymbol{\mu})\Big) = \int \frac{|\mathbf{Z}-\boldsymbol{\mu}|_2^2}{2\sigma^4} e^{-\frac{|\mathbf{Z}-\boldsymbol{\mu}|_2^2}{2\sigma^2}} d\nu(\boldsymbol{\mu}),$$

we recover (1.3.3). □

1.A.2 Proofs for Section 1.4

Proof of Proposition 1.4.1. It is enough to show that the following statements hold for sufficiently large sample size n. Let $Q(u) = \nu([-u, u]^c)/\nu([-u, u])$. Since ν is a distribution function $Q(u) \to 0$ for $u \to \infty$. We work on $I = [\sigma^{*2}/2, 2\sigma^{*2}]$, where σ^{*2} is chosen such that

$$Q(\sigma^*) \leq \exp\Big(-48\Big(17 + 2e^2 + \frac{24}{1-\alpha}\Big)\Big), \tag{1.A.1}$$

and α denotes the fraction of known zero means in the model. Notice that

$$\frac{\sigma^2}{2} \leq \sigma^{*2} \leq 2\sigma^2 \quad \text{for all} \quad \sigma^2 \in I. \tag{1.A.2}$$

Let

$$R := \frac{\sigma^*}{\sqrt{6}}\sqrt{\log\Big(\frac{1}{Q(\sigma^*)}\Big)}. \tag{1.A.3}$$

We choose the non-zero means to be

$$\mu_i^* := \frac{R}{2}. \tag{1.A.4}$$

The interval I is compact and the prior π is continuous and positive on $\mathbb{R}_+$, $\inf_{\sigma^2\in I}\pi(\sigma^2) > 0$. Since we also assumed that π' is continuous, we find that

$$\sup_{\sigma^2\in I}\frac{\sigma^{*2}|\pi'(\sigma^2)|}{n\pi(\sigma^2)} \le 1$$

for all sufficiently large n. With (1.3.3) and (1.A.2),

$$\begin{aligned}\inf_{\sigma^2\in I}\partial_{\sigma^2}\log\pi(\sigma^2|\mathbf{Y},\mathbf{Z}) &\ge \frac{n}{\sigma^{*2}}\inf_{\sigma^2\in I}\Big(\frac{\sigma^{*2}V(\boldsymbol{\mu}|(\mathbf{Z},\sigma^2))}{2n\sigma^4} - \frac{\sigma^{*2}}{2\sigma^2} - 1\Big)\\ &\ge \frac{n}{\sigma^{*2}}\Big(\frac{\inf_{\sigma^2\in I}V(\boldsymbol{\mu}|(\mathbf{Z},\sigma^2))}{8\sigma^{*2}n} - 2\Big).\end{aligned} \tag{1.A.5}$$

Using (1.3.1) and (1.3.2), we expand $V(\boldsymbol{\mu}|(\mathbf{Z},\sigma^2)$,

$$\begin{aligned}\frac{V(\boldsymbol{\mu}|(\mathbf{Z},\sigma^2))}{n} &= \frac{|\mathbf{Z}|_2^2}{n} + \frac{1}{n}\int_{\mathbb{R}^n}(|\boldsymbol{\mu}|_2^2 - 2\mathbf{Z}^\top\boldsymbol{\mu})\pi(\mu|\mathbf{Z},\sigma^2)d\mu\\ &= \frac{|\mathbf{Z}|_2^2}{n} + \frac{1}{n}\sum_{i=1}^n\int_{\mathbb{R}}(\mu_i^2 - 2Z_i\mu_i)\pi(\mu_i|Z_i,\sigma^2)d\mu_i.\end{aligned}$$

Since the integrands in the latter display are positive for $|\mu_i| \ge 2|Z_i|$, we can set $V_i := |Z_i|\int_{|\mu|\le 2|Z_i|}|\mu|\pi(\mu|Z_i,\sigma^2)d\mu$ and bound

$$\begin{aligned}\frac{V(\boldsymbol{\mu}|(\mathbf{Z},\sigma^2))}{n} &\ge \frac{|\mathbf{Z}|_2^2}{n} - \frac{2}{n}\sum_{i=1}^{n_2}Z_i\int_{|\mu_i|\le 2|Z_i|}\mu_i\pi(\mu_i|Z_i,\sigma^2)d\mu_i\\ &\ge \frac{|\mathbf{Z}|_2^2}{n} - \frac{2}{n}\sum_{i=1}^{n_2}V_i.\end{aligned}$$

As a next step in the proof, we show

$$\inf_{\sigma^2\in I}\frac{V(\boldsymbol{\mu}|(\mathbf{Z},\sigma^2))}{n} \ge \frac{|\mathbf{Z}|_2^2}{2n} - \frac{16}{n}\Big|\mathbf{Z} - \frac{R}{2}\Big|_2^2 - \frac{2n_2}{n}\sigma^{*2}e^2. \tag{1.A.6}$$

To prove this inequality, we distinguish the cases $|Z_i| > R$ and $|Z_i| \le R$, decomposing

$$V_i =: |Z_i|(A_i + B_i) \tag{1.A.7}$$

with

$$\begin{aligned}A_i &:= \mathbf{1}(|Z_i| > R)\int_{|\mu|\le 2|Z_i|}|\mu|\pi(\mu|Z_i,\sigma^2)d\mu\\ B_i &:= \mathbf{1}(|Z_i| \le R)\int_{|\mu|\le 2|Z_i|}|\mu|\pi(\mu|Z_i,\sigma^2)d\mu.\end{aligned} \tag{1.A.8}$$

For the term A_i of (1.A.8), observe that $A_i \leq 2|Z_i|\mathbf{1}(|Z_i| > R)$. If $|Z_i| > R$, $|Z_i| \leq 2|Z_i| - R \leq 2|Z_i - R/2|$ and therefore,

$$|Z_i|A_i \leq 8\Big(Z_i - \frac{R}{2}\Big)^2. \tag{1.A.9}$$

Next, we bound the term B_i in (1.A.8). In the sequel, we frequently make use of the fact that $\sigma^2 \in I$. The idea is to split the domain of integration $0 \leq |\mu| \leq 2|Z_i|$ into sets $|\mu| \leq \sigma^*$ and $\sigma^* < |\mu| \leq 2|Z_i|$. The contribution of the first part can be bounded by σ^*. More work is needed for the second part. By expanding the square $(\mu - Z_i)^2$ in the exponent, the Z_i^2-terms in the numerator and denominator cancel against each other, as they do not depend on μ, and we have

$$\begin{aligned} B_i &= \mathbf{1}(|Z_i| \leq R)\frac{\int_{|\mu|\leq 2|Z_i|} |\mu| e^{-\frac{(\mu - Z_i)^2}{2\sigma^2}} d\nu(\mu)}{\int e^{-\frac{(\mu-Z_i)^2}{2\sigma^2}} d\nu(\mu)} \\ &\leq \sigma^* + \mathbf{1}(|Z_i| \leq R)\frac{\int_{\sigma^*<|\mu|\leq 2R} |\mu| e^{-\frac{\mu^2}{2\sigma^2}} e^{\frac{\mu Z_i}{\sigma^2}} d\nu(\mu)}{\int e^{-\frac{\mu^2}{2\sigma^2}} e^{\frac{\mu Z_i}{\sigma^2}} d\nu(\mu)}. \end{aligned}$$

We now treat numerator and denominator separately. For the numerator, the function $y \mapsto ye^{-y^2/2}$ attains its maximum at $y = 1$ and is bounded by $e^{-1/2}$. This means that $|\mu|e^{-\frac{\mu^2}{2\sigma^2}} \leq \sigma e^{-1/2} \leq \sigma^*$, where the last step follows from (1.A.2). Together with (1.A.2), we obtain

$$\mathbf{1}(|Z_i| \leq R)\int_{\sigma^*<|\mu|\leq 2R} |\mu| e^{-\frac{\mu^2}{2\sigma^2}} e^{\frac{\mu Z_i}{\sigma^2}} \nu(\mu) d\mu \leq \sigma^* e^{\frac{4R^2}{\sigma^{*2}}} \nu\big([-\sigma^*, \sigma^*]^c\big),$$

using $\mu Z_i/\sigma^2 \leq 4R^2/\sigma^{*2}$ to bound the exponent in the integral. To derive a lower bound of the denominator, we replace the integral over $\mathbb{R}$ by an integral over $[-\sigma^*, \sigma^*]$. On this interval, $e^{-\mu^2/(2\sigma^2)} \geq e^{-1}$ and $\mathbf{1}(|Z_i| \leq R)e^{\frac{\mu Z_i}{\sigma^2}} \geq e^{-R^2/\sigma^2} \geq e^{-2R^2/\sigma^{*2}}$, since $\sigma^* \leq R$. We obtain

$$\mathbf{1}(|Z_i| \leq R)\int_{\mathbb{R}} e^{-\frac{\mu^2}{2\sigma^2}} e^{\frac{\mu Z_i}{\sigma^2}} d\nu(\mu) \geq e^{-1} e^{-\frac{2R^2}{\sigma^{*2}}} \nu\big([-\sigma^*, \sigma^*]\big).$$

Combining this with the upper bound for the numerator yields, with (1.A.1), (1.A.3) and the definition of the function $Q(u)$,

$$B_i \leq e^{1+\frac{6R^2}{\sigma^{*2}}} Q(\sigma^*)\sigma^* = e^{1-\log Q(\sigma^*)} Q(\sigma^*)\sigma^* = e\sigma^* \quad \text{for all } \sigma^2 \in I. \tag{1.A.10}$$

Together with (1.A.9) and (1.A.7),

$$V_i \leq 8\Big(Z_i - \frac{R}{2}\Big)^2 + |Z_i|\sigma^* e, \quad \text{for all } \ \sigma^2 \in I.$$

With $|Z_i|\sigma^* e \leq Z_i^2/4 + \sigma^{*2}e^2$, we finally obtain (1.A.6).

In a final step of the proof, we derive, on an event with large probability, a deterministic lower bound for the right hand side in (1.A.6). Let $U_1, \ldots, U_{n_2}$ be independent random variables. Rewriting Chebyshev's inequality yields $P(n^{-1}\sum_{i=1}^{n_2} U_i > n^{-1}\sum_{i=1}^{n_2}(E[U_i] - \sigma^{*2})) \geq$

$1-\sum_{i=1}^{n_2}\operatorname{Var}(U_i)/(n_2\sigma^{*2})^2$. We aply this with $U_i = Z_i^2/2 - 16(Z_i - R/2)^2$. Recall that $Z_i \sim \mathcal{N}(R/2, \sigma^{*2})$. Therefore, $E_0[Z_i^2] = R^2/4 + \sigma^{*2}$ and $E[(Z_i - R/2)^2] = \sigma^{*2}$. For the variance, $\operatorname{Var}_0(Z_i^2) = R^2\sigma^{*2} + \sigma^{*4}$ and $\operatorname{Var}((Z_i - R/2)^2) = \sigma^{*4}$. Since by assumption $\alpha < 1$, Chebyshev's inequality yields then $\mathbb{P}_0^n(\mathcal{A}_n) \to 1$ when $n \to \infty$ for the set

$$\mathcal{A}_n := \Big\{ \frac{|\mathbf{Z}|_2^2}{2n} - \frac{16}{n}\Big|\mathbf{Z} - \frac{R}{2}\Big|_2^2 \geq \frac{n_2}{n}\Big(\frac{R^2 + 4\sigma^{*2}}{8} - 17\sigma^{*2}\Big)\Big\}. \tag{1.A.11}$$

On $\mathcal{A}_n$, we have using (1.A.3), (1.A.6) and $Q(\sigma^*) \leq \exp(-48(17 + 2e^2 + 24/(1-\alpha)))$,

$$\inf_{\sigma^2 \in I} \frac{V(\boldsymbol{\mu}|(\mathbf{Z}, \sigma^2))}{8\sigma^{*2}n} \geq \frac{n_2}{8\sigma^{*2}n}\Big(\frac{R^2}{8} - \sigma^{*2}(17 + 2e^2)\Big) \geq 3. \tag{1.A.12}$$

The assertion follows with (1.A.5). □

Proof of Theorem 1.4.2. Proposition 1.4.1 shows that

$$\inf_{\sigma^2 \in [\sigma^{*2}/2, 2\sigma^{*2}]} \partial_{\sigma^2} \log \pi(\sigma^2|\mathbf{Y}, \mathbf{Z}) \geq \frac{n}{\sigma^{*2}}$$

has $\mathbb{P}_0^n$-probability tending to one. This means that for $\sigma^2, \widetilde{\sigma}^2 \in [\sigma^{*2}/2, 2\sigma^{*2}]$, with $\sigma^2 \leq \widetilde{\sigma}^2$, we must have $\log \pi(\sigma^2|\mathbf{Y}, \mathbf{Z}) \leq \log \pi(\widetilde{\sigma}^2|\mathbf{Y}, \mathbf{Z}) - n(\widetilde{\sigma}^2 - \sigma^2)/\sigma^{*2}$. Exponentiating this inequality for $\widetilde{\sigma}^2 = \sigma^2 + \sigma^{*2}/2$, yields

$$\begin{aligned}\Pi\Big(\sigma^2 \in \Big[\frac{\sigma^{*2}}{2}, 3\frac{\sigma^{*2}}{2}\Big]\Big|\mathbf{Y}, \mathbf{Z}\Big) &= \int_{\sigma^{*2}/2}^{3\sigma^{*2}/2} \pi_n(\sigma^2|\mathbf{Y}, \mathbf{Z}) d\sigma^2 \\ &\leq e^{-n/2} \int_{\sigma^{*2}}^{2\sigma^{*2}} \pi_n(\sigma^2|\mathbf{Y}, \mathbf{Z}) d\sigma^2 \leq e^{-n/2}\end{aligned}$$

and this completes the proof since $|\sigma^2/\sigma^{*2} - 1| \leq 1/2$ is equivalent to $\sigma^2 \in [\sigma^{*2}/2, 3\sigma^{*2}/2]$. □

1.A.3 Proofs for Section 1.5

Proof of Lemma 1.5.1. We can write the posterior as

$$\pi(\sigma^2|\mathbf{Y}, \mathbf{Z}) \propto \mathbf{1}(\sigma^2 \geq 0) e^{\ell(\sigma^2|\mathbf{Y})} \int_0^\infty e^{\ell(\sigma^2 + \theta^2|\mathbf{Z})} \gamma(\theta^2) d\theta^2 \pi(\sigma^2). \tag{1.A.13}$$

By using (1.5.6) and (1.5.7) we obtain (1.5.8). □

We now prepare for the proof of the limiting shape result. From (1.5.8), the density (1.5.9) of the joint posterior is

$$\widetilde{\pi}(\sigma^2, \theta^2|\mathbf{Y}, \mathbf{Z}) \propto \mathbf{1}(\sigma^2 \geq 0, \theta^2 \geq 0) e^{\ell(\sigma^2|\mathbf{Y})} e^{\ell(\sigma^2 + \theta^2|\mathbf{Z})} \gamma(\theta^2) \pi(\sigma^2).$$

With

$$\zeta_n := 4\sqrt{\Big(1 + \Big(\frac{\alpha}{1-\alpha} \vee \frac{1-\alpha}{\alpha}\Big)\Big)\frac{\log n}{n_1 \wedge n_2}} \wedge 1, \tag{1.A.14}$$

define

$$
\begin{aligned}
B_1 &:= \Big[\frac{\overline{\mathbf{Y}^2}}{1+\zeta_n}, \frac{\overline{\mathbf{Y}^2}}{1-\zeta_n}\Big],\\
B_2 &:= \Big[0 \vee \Big(\frac{\overline{\mathbf{Z}^2}}{1+\zeta_n} - \frac{\overline{\mathbf{Y}^2}}{1-\zeta_n}\Big), \frac{\overline{\mathbf{Z}^2}}{1-\zeta_n} - \frac{\overline{\mathbf{Y}^2}}{1+\zeta_n}\Big].
\end{aligned}
\tag{1.A.15}
$$

It is shown below that the posterior concentrates on $\{\sigma^2 \in B_1\}$ and $\{\theta^2 \in B_2\}$. The posterior can consequently be approximated by the distribution $\Pi_1(\cdot|\mathbf{Y},\mathbf{Z})$ defined through its density (1.5.10). On the localized set $(\sigma^2, \theta^2) \in B_1 \times B_2$, we are able to replace the log-likelihoods by a quadratic expansion. This then allows us to approximate the posterior by $\Pi_2(\cdot|\mathbf{Y},\mathbf{Z})$ which is defined as the distribution with density (1.5.11). We now state the single steps formally and provide the proofs.

Proposition 1.A.1. *If the prior densities $\gamma, \pi : [0,\infty) \to (0,\infty)$ are positive and uniformly continuous, then there exists a sequence of sets $(A_n)_n$ such that for any compact sets $K \subset (0,\infty), K' \subset (-\infty,\infty)$,*

(i) $\lim_{n\to\infty} \sup_{\sigma^{*2}\in K, \mu_i^*\in K', \forall i} P_0^n(A_n^c) = 0.$

(ii) With B_1, B_2 as defined in (1.A.15)*, we have for $n \to \infty$,*

$$\sup_{\sigma^{*2}\in K, \mu_i^*\in K', \forall i} \widetilde{\Pi}\big(\{\sigma^2 \notin B_1\} \cup \{\theta^2 \notin B_2\} \ |\mathbf{Y},\mathbf{Z}\big)\mathbf{1}\big((\mathbf{Y},\mathbf{Z}) \in A_n\big) \to 0.$$

(iii) For $n \to \infty$,

$$\sup_{\sigma^{*2}\in K, \mu_i^*\in K', \forall i} \Big\|\widetilde{\Pi}\big(\sigma^2 \in \cdot|\mathbf{Y},\mathbf{Z}\big) - \Pi_1(\cdot|\mathbf{Y},\mathbf{Z})\Big\|_{\mathrm{TV}} \mathbf{1}\big((\mathbf{Y},\mathbf{Z}) \in A_n\big) \to 0.$$

(iv) For $n \to \infty$,

$$\sup_{\sigma^{*2}\in K, \mu_i^*\in K', \forall i} \Big\|\Pi_1(\cdot|\mathbf{Y},\mathbf{Z}) - \Pi_2(\cdot|\mathbf{Y},\mathbf{Z})\Big\|_{\mathrm{TV}} \mathbf{1}\big((\mathbf{Y},\mathbf{Z}) \in A_n\big) \to 0.$$

(v) For $n \to \infty$,

$$\sup_{\sigma^{*2}\in K, \mu_i^*\in K', \forall i} \Big\|\Pi_2(\cdot|\mathbf{Y},\mathbf{Z}) - \Pi_\infty(\cdot|\mathbf{Y},\mathbf{Z})\Big\|_{\mathrm{TV}} \mathbf{1}\big((\mathbf{Y},\mathbf{Z}) \in A_n\big) \to 0.$$

(vi) For $n \to \infty$, and $\inf_{\mu_i^*\in K'} |\mu_i^*| \gg (\log n/n)^{1/4}$,

$$\sup_{\sigma^{*2}\in K, \mu_i^*\in K', \forall i} \Big\|\Pi_\infty(\cdot|\mathbf{Y},\mathbf{Z}) - \widetilde{\Pi}_\infty(\cdot|\mathbf{Y})\Big\|_{\mathrm{TV}} \mathbf{1}\big((\mathbf{Y},\mathbf{Z}) \in A_n\big) \to 0.$$

Proof of Proposition 1.A.1. Recall the definition of ζ_n in (1.A.14) and set

$$\delta_n := C^{-1}\zeta_n = \sqrt{2\frac{\log n}{n_1 \wedge n_2}} \wedge C^{-1}, \text{ with } C^2 := 16 + 16\Big(\frac{\alpha}{1-\alpha} \vee \frac{1-\alpha}{\alpha}\Big). \tag{1.A.16}$$

Let $\underline{\sigma}^{*2} = \inf\{\sigma^{*2} \in K\} > 0$. Define the event

$$A_n := \Big\{\overline{\mathbf{Z}^2} > \frac{\overline{\mathbf{Y}^2}}{1+\delta_n/2}\Big\} \cap \Big\{\Big|\frac{\overline{\mathbf{Z}^2} - \overline{\boldsymbol{\mu}^{*2}}}{\sigma^{*2}} - 1\Big| + \Big|\frac{\overline{\mathbf{Y}^2}}{\sigma^{*2}} - 1\Big| \leq \delta_n\Big\}. \tag{1.A.17}$$

Since $\delta_n \leq 1/2$, this implies in particular that on A_n, $\overline{\mathbf{Y}^2} \wedge \overline{\mathbf{Z}^2} \geq \underline{\sigma}^{*2}/2$.

Proof of (i): We simplify the notation by introducing the events

$$B_n := \Big\{\overline{\mathbf{Z}^2} > \frac{\overline{\mathbf{Y}^2}}{1+\delta_n/2}\Big\}, \quad D_n := \Big\{\Big|\frac{\overline{\mathbf{Z}^2} - \overline{\boldsymbol{\mu}^{*2}}}{\sigma^{*2}} - 1\Big| + \Big|\frac{\overline{\mathbf{Y}^2}}{\sigma^{*2}} - 1\Big| \leq \delta_n\Big\},$$

so that $A_n = B_n \cap D_n$. Thus $\mathbb{P}_0^n(A_n^c) \leq \mathbb{P}_0^n(B_n^c) + \mathbb{P}_0^n(D_n^c)$. We show that both $\mathbb{P}_0^n(B_n^c)$ and $\mathbb{P}_0^n(D_n^c)$ tend to zero uniformly over compact sets of parameters. By Chebyshev's inequality,

$$\begin{aligned}\mathbb{P}_0^n(D_n^c) &\leq \mathbb{P}_0^n\Big(\Big|\frac{\overline{\mathbf{Z}^2} - \overline{\boldsymbol{\mu}^{*2}}}{\sigma^{*2}} - 1\Big| > \frac{\delta_n}{2}\Big) + \mathbb{P}_0^n\Big(\Big|\frac{\overline{\mathbf{Y}^2}}{\sigma^{*2}} - 1\Big| > \frac{\delta_n}{2}\Big) \\ &\leq 4\frac{\mathrm{Var}_0\Big(\frac{\overline{\mathbf{Z}^2} - \overline{\boldsymbol{\mu}^{*2}}}{\sigma^{*2}}\Big) + \mathrm{Var}_0\Big(\frac{\overline{\mathbf{Y}^2}}{\sigma^{*2}}\Big)}{\delta_n^2}.\end{aligned}$$

Since

$$\mathrm{Var}_0\left(\frac{\overline{\mathbf{Z}^2} - \overline{\boldsymbol{\mu}^{*2}}}{\sigma^{*2}}\right) = \frac{2}{n_2} + \frac{4\overline{\boldsymbol{\mu}^{*2}}}{n_2\sigma^{*2}}, \quad \mathrm{Var}_0\left(\frac{\overline{\mathbf{Y}^2}}{\sigma^{*2}}\right) = \frac{2}{n_1},$$

we find

$$\sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} \mathbb{P}_0^n(D_n^c) \leq \frac{8}{n_1\delta_n^2} + \frac{8}{n_2\delta_n^2} + \frac{16H}{n_2\delta_n^2}$$

with $H := \sup_{\sigma^{*2}\in K, \mu_i^*\in K', \forall i} (\mu_i^*)^2/\sigma^{*2}$. Notice that H is a finite constant since $K \subset (0,\infty)$ and K' are compact sets. Because $\delta_n = O(\sqrt{\log n/n})$, the previous probability tends to zero as n increases. We now bound $\mathbb{P}_0^n(B_n^c)$. Rewriting B_n^c, we obtain

$$B_n^c = \left\{\left(1+\frac{\delta_n}{2}\right)\left(\frac{\overline{\mathbf{Z}^2} - \overline{\boldsymbol{\mu}^{*2}}}{\sigma^{*2}} - 1\right) + 1 - \frac{\overline{\mathbf{Y}^2}}{\sigma^{*2}} \leq -\frac{\delta_n}{2} - \left(1+\frac{\delta_n}{2}\right)\frac{\overline{\boldsymbol{\mu}^{*2}}}{\sigma^{*2}}\right\},$$

and again by Chebyshev's inequality

$$\begin{aligned}P_0^n(B_n^c) &\leq \frac{\big(1+\frac{\delta_n}{2}\big)^2 \mathrm{Var}_0\Big(\frac{\overline{\mathbf{Z}^2} - \overline{\boldsymbol{\mu}^{*2}}}{\sigma^{*2}} - 1\Big) + \mathrm{Var}_0\Big(1 - \frac{\overline{\mathbf{Y}^2}}{\sigma^{*2}}\Big)}{\Big(\frac{\delta_n}{2} + \big(1+\frac{\delta_n}{2}\big)\frac{\overline{\boldsymbol{\mu}^{*2}}}{\sigma^{*2}}\Big)^2} \\ &\leq \left(1+\frac{\delta_n}{2}\right)^2\left(\frac{8}{n_2\delta_n^2} + \frac{16H}{n_2\delta_n^2}\right) + \frac{8}{n_1\delta_n^2},\end{aligned}$$

which again tends to zero for $n \to \infty$ uniformly over $\sigma^{*2} \in K, \mu_i^* \in K', \forall i$.

Proof of (ii): We work on the event A_n defined in (1.A.17) deriving deterministic lower and upper bounds for the denominator and numerator in the Bayes formula. We start with

$$\widetilde{\Pi}(B_1^c \times \mathbb{R}_+ | \mathbf{Y}, \mathbf{Z}) = \frac{\int_{B_1^c} e^{\ell(\sigma^2|\mathbf{Y})} \int_0^\infty e^{\ell(\sigma^2+\theta^2|\mathbf{Z})}\gamma(\theta^2)d\theta^2\pi(\sigma^2)d\sigma^2}{\int_0^\infty e^{\ell(\sigma^2|\mathbf{Y})} \int_0^\infty e^{\ell(\sigma^2+\theta^2|\mathbf{Z})}\gamma(\theta^2)d\theta^2\pi(\sigma^2)d\sigma^2}, \tag{1.A.18}$$

and show that on the event A_n this quantity tends to 0 when n tends to infinity. The first part of the proof provides a lower bound for the denominator. For that, we restrict $\sigma^2 \in \Sigma := [\overline{\mathbf{Y}^2}/(1+\delta_n), \overline{\mathbf{Y}^2}/(1+\delta_n/2)]$ and $\theta^2 \in \Theta(\sigma^2) := [\overline{\mathbf{Z}^2} - \sigma^2, \overline{\mathbf{Z}^2}(1+\delta_n) - \sigma^2] \subset (0, \infty)$, where the last inclusion follows since by definition of the event A_n in (1.A.17), $Z^2 - \sigma^2 \geq Z^2 - \overline{\mathbf{Y}^2}/(1+\delta_n/2) \geq 0$. The inner integral in the denominator of (1.A.18) can be lower bounded by

$$\int_0^\infty e^{\ell(\sigma^2+\theta^2|\mathbf{Z})}\gamma(\theta^2)d\theta^2 \geq \int_{\Theta(\sigma^2)} e^{\ell(\sigma^2+\theta^2|\mathbf{Z})}d\theta^2 \inf_{\theta^2 \leq \overline{\mathbf{Z}^2}(1+\delta_n)} \gamma(\theta^2).$$

Thanks to the definition of A_n in (1.A.17) and $\delta_n \leq 1$, we have $\overline{\mathbf{Z}^2} \leq \overline{\boldsymbol{\mu}^{*2}} + \sigma^{*2}(1+\delta_n)$, so that $\overline{\mathbf{Z}^2}(1+\delta_n) \leq 2\overline{\boldsymbol{\mu}^{*2}} + 4\sigma^{*2}$. We then set

$$\underline{\gamma} := \inf_{\theta^2 \leq \sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} 2\overline{\boldsymbol{\mu}^{*2}}+4\sigma^{*2}} \gamma(\theta^2) \leq \inf_{\theta^2 \leq \overline{\mathbf{Z}^2}(1+\delta_n)} \gamma(\theta^2).$$

Since K, K' are compact sets and γ is continuous and positive, we must have $\underline{\gamma} > 0$. Differentiating (1.5.6) gives $\partial_{\theta^2}\ell(\sigma^2+\theta^2|\mathbf{Y}) = \frac{1}{2}n_2(\overline{\mathbf{Z}^2} - \sigma^2 - \theta^2)/(\sigma^2+\theta^2)^2$, so the function $\theta^2 \mapsto \ell(\sigma^2+\theta^2|\mathbf{Y})$ is decreasing on $\Theta(\sigma^2)$ for any σ^2. As a direct consequence of (1.5.6), we obtain

$$\ell\big(\overline{\mathbf{Z}^2}(1+\delta_n)|\mathbf{Z}\big) = \ell\big(\overline{\mathbf{Z}^2}|\mathbf{Z}\big) + \frac{n_2}{2}\big(\delta_n/(1+\delta_n) - \log(1+\delta_n)\big). \tag{1.A.19}$$

Consequently, for any $\sigma^2 \in \Sigma$,

$$\begin{aligned}\int_0^\infty e^{\ell(\sigma^2+\theta^2|\mathbf{Z})}\gamma(\theta^2)d\theta^2 &\geq \underline{\gamma}\overline{\mathbf{Z}^2}\delta_n e^{\ell(\overline{\mathbf{Z}^2}|\mathbf{Z})+\frac{n_2}{2}(\delta_n/(1+\delta_n)-\log(1+\delta_n))} \\ &\geq \frac{1}{2}\underline{\gamma}\underline{\sigma}^{*2}\delta_n e^{\ell(\overline{\mathbf{Z}^2}|\mathbf{Z})-\frac{n_2}{4}\delta_n^2},\end{aligned} \tag{1.A.20}$$

where the last inequality follows since $\overline{\mathbf{Z}^2} \geq \sigma^{*2}/2$ on A_n, $\delta_n \leq 1$, and $-\log(1+\delta_n) \geq -\delta_n$ for $\delta_n \leq 1$. The right hand side does not depend on σ^2 anymore. To lower bound the first integral in the denominator of (1.A.18) we apply a similar argument. By (1.5.6), $\partial_{\sigma^2}\ell(\sigma^2|\mathbf{Y}) = n_1(\overline{\mathbf{Y}^2} - \sigma^2)/(2\sigma^4)$. This means that the function $\sigma^2 \mapsto \ell(\sigma^2|\mathbf{Y})$ is increasing on Σ and (1.5.6) yields

$$\ell\big(\overline{\mathbf{Y}^2}/(1+\delta_n)\big) = \ell\big(\overline{\mathbf{Y}^2}|\mathbf{Y}\big) + \frac{n_1}{2}\big(\log(1+\delta_n) - \delta_n\big).$$

On A_n, $\overline{\mathbf{Y}^2} \leq \sigma^{*2}(1+\delta_n)$ and therefore $\overline{\mathbf{Y}^2}/(1+\delta_n/2) \leq 2\sigma^{*2}$. Set

$$\underline{\pi} := \inf_{\sigma^2 \leq \sup_{\sigma^{*2}\in K} 2\sigma^{*2}} \pi(\sigma^2) \leq \inf_{\sigma^2 \leq \overline{\mathbf{Y}^2}/(1+\delta_n/2)} \pi(\sigma^2),$$

so that $\underline{\pi} > 0$ because K is a compact set and π is continuous and positive. We bound

$$\begin{aligned}\int_0^\infty e^{\ell(\sigma^2|\mathbf{Y})}\pi(\sigma^2)d\sigma^2 &\geq \inf_{\sigma^2\in\Sigma} \pi(\sigma^2)\frac{\delta_n}{2}\overline{\mathbf{Y}^2}e^{\ell(\overline{\mathbf{Y}^2}/(1+\delta_n)|\mathbf{Y})} \\ &\geq \underline{\pi}\frac{\delta_n}{2}\overline{\mathbf{Y}^2}e^{\ell(\overline{\mathbf{Y}^2}|\mathbf{Y})+\frac{n_1}{2}(\log(1+\delta_n)-\delta_n)} \\ &\geq \frac{1}{4}\underline{\pi}\delta_n\underline{\sigma}^{*2}e^{\ell(\overline{\mathbf{Y}^2}|\mathbf{Y})-\frac{n_1}{16}\delta_n^2},\end{aligned} \tag{1.A.21}$$

using that on A_n, $\overline{\mathbf{Y}^2} \geq \sigma^{*2}/2$ and $\log(1+\delta_n) \geq \delta_n - \delta_n^2/8$ for $0 \leq \delta_n \leq 1$. The product of the lower bounds obtained in (1.A.20) and (1.A.21) is then a lower bound for the denominator of (1.A.18).

In the next step we upper bound the numerator of (1.A.18). Firstly, observe that $\ell(\sigma^2 + \theta^2|\mathbf{Z}) \leq \ell(\overline{\mathbf{Z}^2}|\mathbf{Z})$ and

$$\int_0^\infty e^{\ell(\sigma^2+\theta^2|\mathbf{Z})}\gamma(\theta^2)d\theta^2 \leq e^{\ell(\overline{\mathbf{Z}^2}|\mathbf{Z})}. \tag{1.A.22}$$

Secondly, since $\sigma^2 \mapsto \ell(\sigma^2|\mathbf{Y})$ is increasing on $(0, \overline{\mathbf{Y}^2}]$ and decreasing on $[\overline{\mathbf{Y}^2}, \infty)$,

$$\begin{aligned}
\int_0^{\overline{\mathbf{Y}^2}/(1+\zeta_n)} e^{\ell(\sigma^2|\mathbf{Y})}\pi(\sigma^2)d\sigma^2 &\leq e^{\ell(\overline{\mathbf{Y}^2}/(1+\zeta_n)|\mathbf{Y})} \\
&= e^{\ell(\overline{\mathbf{Y}^2}|\mathbf{Y})+\frac{n_1}{2}(\log(1+\zeta_n)-\zeta_n)} \\
&\leq e^{\ell(\overline{\mathbf{Y}^2}|\mathbf{Y})-\frac{n_1}{16}\zeta_n^2}, \qquad (1.A.23)\\
\int_{\overline{\mathbf{Y}^2}/(1-\zeta_n)}^\infty e^{\ell(\sigma^2|\mathbf{Y})}\pi(\sigma^2)d\sigma^2 &\leq e^{\ell(\overline{\mathbf{Y}^2}/(1-\zeta_n)|\mathbf{Y})} = e^{\ell(\overline{\mathbf{Y}^2}|\mathbf{Y})+\frac{n_1}{2}(\log(1-\zeta_n)+\zeta_n)} \\
&\leq e^{\ell(\overline{\mathbf{Y}^2}|\mathbf{Y})-\frac{n_1}{16}\zeta_n^2}.
\end{aligned}$$

The numerator of (1.A.18) is upper bounded by the product of the bounds obtained in (1.A.22) and (1.A.23). Together with the bounds on the denominator in (1.A.20) and (1.A.21), and $\zeta_n = C\delta_n$, we derive, on the event A_n, the following bound for (1.A.18):

$$\sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} \widetilde{\Pi}\big(\sigma^2 \notin B_1\big|\mathbf{Y}, \mathbf{Z}\big) \leq \frac{16}{\underline{\pi}\underline{\gamma}\underline{\sigma}^{*4}\delta_n^2} e^{-(C^2 n_1 - 4n_2 - n_1)\delta_n^2/16} \to 0. \tag{1.A.24}$$

The convergence to zero follows since by definition of the constant C in (1.A.16), $n_1 C^2 - 4n_2 - n_1 > 4n_1$ and because of $\delta_n = O(\sqrt{\log n/n})$.

Along similar lines, we show now that, on the event A_n, $\widetilde{\Pi}(\theta^2 \notin B_2|\mathbf{Y}, \mathbf{Z}) \to 0$ as n tends to infinity. Since $\{\theta^2 \notin B_2\} \subset \{\sigma^2 \notin B_1\} \cup (\{\sigma^2 \in B_1\} \cap \{\theta^2 \notin B_2\})$, and $\widetilde{\Pi}(\sigma^2 \notin B_1|\mathbf{Y}, \mathbf{Z})$ tends to zero by (1.A.24), it is sufficient to establish convergence of

$$\widetilde{\Pi}(B_1 \times B_2^c|\mathbf{Y}, \mathbf{Z}) = \frac{\int_{B_1} e^{\ell(\sigma^2|\mathbf{Y})} \int_{B_2^c} e^{\ell(\sigma^2+\theta^2|\mathbf{Z})}\gamma(\theta^2)d\theta^2\pi(\sigma^2)d\sigma^2}{\int_0^\infty e^{\ell(\sigma^2|\mathbf{Y})} \int_0^\infty e^{\ell(\sigma^2+\theta^2|\mathbf{Z})}\gamma(\theta^2)d\theta^2\pi(\sigma^2)d\sigma^2} \tag{1.A.25}$$

to zero. We can argue similarly as for the upper bound above using that $\ell(\sigma^2|\mathbf{Y}) \leq \ell(\overline{\mathbf{Y}^2}|\mathbf{Y})$. By following the same steps as for (1.A.22) and (1.A.23) and using that $a \mapsto \ell(a|\mathbf{Z})$ is increasing on $(0, \overline{\mathbf{Z}^2}]$ and decreasing on $[\overline{\mathbf{Z}^2}, \infty)$, the numerator in (1.5.9) integrated over the set $\{\sigma^2 \in B_1\} \cap \{\theta^2 \notin B_2\}$ is upper bounded by

$$\leq e^{\ell(\overline{\mathbf{Y}^2}|\mathbf{Y})} \sup_{\sigma^2 \in B_1} \int_{B_2^c} e^{\ell(\sigma^2+\theta^2|\mathbf{Z})}\gamma(\theta^2)d\theta^2 \leq 2e^{\ell(\overline{\mathbf{Y}^2}|\mathbf{Y})+\ell(\overline{\mathbf{Z}^2}|\mathbf{Z})-\frac{n_2}{16}\zeta_n^2}.$$

Together with the lower bounds for the denominator in (1.A.20) and (1.A.21), we upper bound (1.A.25), on the event A_n, by

$$\sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} \widetilde{\Pi}(B_1 \times B_2^c|\mathbf{Y}, \mathbf{Z}) \leq \frac{32}{\underline{\pi}\underline{\gamma}\underline{\sigma}^{*4}\delta_n^2} e^{-(C^2 n_2 - 4n_2 - n_1)\delta_n^2/16}. \tag{1.A.26}$$

By definition (see (1.A.16)), the constant $C^2 > 0$ satisfies $n_2 C^2 - 4n_2 - n_1 > 4n_2$. Since $\delta_n = O(\sqrt{\log n/n})$, this implies that the right hand side of (1.A.26) is bounded above by $\lesssim n \exp(-n_2\delta_n^2/4) \to 0$, as $n \to \infty$. Together with (1.A.24), this completes the proof for part (ii).

Proof of (iii): It is well-known that for probability measures P, Q defined on the same measurable space $\mathcal{X}$,

$$\|P - P(\cdot|A)\|_{\mathrm{TV}} \le 2P(A^c), \tag{1.A.27}$$

see Lemma E.1 in [75]. With $A = B_1 \cap B_2$, $P = \widetilde{\Pi}(\cdot|\mathbf{Y},\mathbf{Z})$ and $\Pi_0(\cdot|\mathbf{Y},\mathbf{Z})$ the distribution with density

$$\pi_0(\sigma^2,\theta^2|\mathbf{Y},\mathbf{Z}) = \frac{e^{\ell(\sigma^2|\mathbf{Y})}e^{\ell(\sigma^2+\theta^2|\mathbf{Z})}\mathbf{1}(\sigma^2 \in B_1, \theta^2 \in B_2)}{\int_{B_1} e^{\ell(\sigma^2|\mathbf{Y})}(\int_{B_2} e^{\ell(\sigma^2+\theta^2|\mathbf{Z})}d\theta^2)d\sigma^2},$$

we have that

$$\begin{aligned}&\sup_{\sigma^{*2}\in K, \mu_i^*\in K', \forall i} \Big\|\widetilde{\Pi}\big(\sigma^2 \in \cdot\,\big|\mathbf{Y},\mathbf{Z}\big) - \Pi_0\big(\sigma^2 \in \cdot\,\big|\mathbf{Y},\mathbf{Z}\big)\Big\|_{\mathrm{TV}}\\ &\le \sup_{\sigma^{*2}\in K, \mu_i^*\in K', \forall i} \Big\|\widetilde{\Pi}\big(\sigma^2 \in \cdot, \theta^2 \in \cdot\,\big|\mathbf{Y},\mathbf{Z}\big) - \Pi_0\big(\sigma^2 \in \cdot, \theta^2 \in \cdot\,\big|\mathbf{Y},\mathbf{Z}\big)\Big\|_{\mathrm{TV}} \to 0.\end{aligned}$$

By bounding the L^1-distance between the densities, we now show that $\Pi_0(\sigma^2 \in \cdot|\mathbf{Y},\mathbf{Z})$ and $\Pi_1(\sigma^2 \in \cdot|\mathbf{Y},\mathbf{Z})$ are close in total variation using the following lemma.

Lemma 1.A.2 (Lemma E.3 in [75]). *If $h(\sigma^2) \propto d\Pi_0(\sigma^2 \in \cdot|\mathbf{Y},\mathbf{Z})/d\Pi_1(\sigma^2 \in \cdot|\mathbf{Y},\mathbf{Z})$ exists and $\int |h(\sigma^2) - 1| d\Pi_1(\sigma^2|\mathbf{Y},\mathbf{Z}) \le \delta$ for some $\delta \in (0,1)$, then also*

$$\big\|\Pi_0\big(\sigma^2 \in \cdot\,\big|\mathbf{Y},\mathbf{Z}\big) - \Pi_1\big(\sigma^2 \in \cdot\,\big|\mathbf{Y},\mathbf{Z}\big)\big\|_{\mathrm{TV}} \le \frac{\delta}{1-\delta}.$$

As h is the Radon-Nikodym derivative up to a multiplicative factor, we can choose

$$h(\sigma^2) = \frac{\pi(\sigma^2)\int_{B_2} e^{\ell(\sigma^2+\theta^2|\mathbf{Z})}\gamma(\theta^2)d\theta^2}{\inf_{\widetilde{\sigma}^2\in B_1, \widetilde{\theta}^2\in B_2} \pi(\widetilde{\sigma}^2)\gamma(\widetilde{\theta}^2)\int_{B_2} e^{\ell(\sigma^2+\theta^2|\mathbf{Z})}d\theta^2}\mathbf{1}(\sigma^2 \in B_1).$$

Then,

$$1 \le h(\sigma^2) \le \frac{\sup_{\sigma^2\in B_1, \theta^2\in B_2} \pi(\sigma^2)\gamma(\theta^2)}{\inf_{\widetilde{\sigma}^2\in B_1, \widetilde{\theta}^2\in B_2} \pi(\widetilde{\sigma}^2)\gamma(\widetilde{\theta}^2)}. \tag{1.A.28}$$

Using the argument above, it remains to prove that $\sup_{\sigma^2\in B_1} |h(\sigma^2) - 1| \to 0$ for $n \to \infty$. By the definition of A_n and due to $\delta_n \le \zeta_n$,

$$B_1 \subseteq B_1' := [\kappa_n\sigma^{*2}, \kappa_n^{-1}\sigma^{*2}] \quad \text{with } \kappa_n := \frac{1-\zeta_n}{1+\zeta_n} = 1 - 2\zeta_n + O(\zeta_n^2). \tag{1.A.29}$$

Recall that K is a compact set. Since π is positive and uniformly continuous,

$$\sup_{\sigma^{*2}\in K} \sup_{\sigma^2, \widetilde{\sigma}^2 \in [\kappa_n\sigma^{*2}, \kappa_n^{-1}\sigma^{*2}]} \Big|\frac{\pi(\sigma^2)}{\pi(\widetilde{\sigma}^2)} - 1\Big| \to 0. \tag{1.A.30}$$

Similarly, we have on the event A_n,

$$B_2 \subseteq B_2' := \Big[\frac{\overline{\boldsymbol{\mu}^{*2}}}{1+\zeta_n} + \Big(\kappa_n - \frac{1}{\kappa_n}\Big)\sigma^{*2}, \frac{\overline{\boldsymbol{\mu}^{*2}}}{1-\zeta_n} + \Big(\frac{1}{\kappa_n} - \kappa_n\Big)\sigma^{*2}\Big]. \tag{1.A.31}$$

Since $\mu_i^* \in K'$ for all i, the average of the squares $\overline{\boldsymbol{\mu}^{*2}}$ lies in the convex hull of K' and

$$\sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} \ \sup_{\theta^2, \widetilde{\theta}^2 \in B_2'} \Big|\frac{\gamma(\theta^2)}{\gamma(\widetilde{\theta}^2)} - 1\Big| \to 0.$$

For real numbers u, v, $uv = (u-1)(v-1) + (u-1) + (v-1) + 1$. We therefore obtain with (1.A.28) and (1.A.30), $\sup_{\sigma^2 \in B_1} |h(\sigma^2) - 1| \to 0$ for $n \to \infty$. This completes the proof of (iii).

Proof of (iv): We use the same strategy as in the proof of part (iii), applying Lemma 1.A.2 to

$$h(\sigma^2) = \mathbf{1}(\sigma^2 \in B_1) e^{\ell(\sigma^2|\mathbf{Y}) - \ell(\overline{\mathbf{Y}^2}|\mathbf{Y}) + \frac{n_1}{4\sigma^{*4}}(\sigma^2 - \overline{\mathbf{Y}^2})^2} \frac{\int_{B_2} e^{\ell(\sigma^2+\theta^2|\mathbf{Z}) - \ell(\overline{\mathbf{Z}^2}|\mathbf{Z})} d\theta^2}{\int_{B_2} e^{-\frac{n_2}{4(\sigma^{*2}+\overline{\boldsymbol{\mu}^{*2}})^2}(\theta^2+\sigma^2-\overline{\mathbf{Z}^2})^2} d\theta^2},$$

which is a constant multiple of the likelihood ratio of $\Pi_1(\sigma^2 \in \cdot|\mathbf{Y}, \mathbf{Z})$ and $\Pi_2(\sigma^2 \in \cdot|\mathbf{Y}, \mathbf{Z})$. To verify the assumptions of Lemma 1.A.2, we have to show that $\sup_{\sigma^{*2}\in K} \ |h(\sigma^2) - 1| \to 0$ for $n \to \infty$. Using again the identity $uv = (u-1)(v-1) + (u-1) + (v-1) + 1$ and the fact that $|\int f / \int g - 1| \leq \sup |f/g - 1|$, we find that it is enough to prove that on the event A_n,

$$\sup_{\sigma^{*2}\in K} \ \sup_{\sigma^2 \in B_1} \Big|\ell(\sigma^2|\mathbf{Y}) - \ell(\overline{\mathbf{Y}^2}|\mathbf{Y}) + \frac{n_1}{4\sigma^{*4}}(\sigma^2 - \overline{\mathbf{Y}^2})^2\Big| \to 0. \tag{1.A.32}$$

$$\sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} \ \sup_{\sigma^2 \in B_1, \theta^2 \in B_2} \Big|\ell(\sigma^2 + \theta^2|\mathbf{Z}) - \ell(\overline{\mathbf{Z}^2}|\mathbf{Z}) + \frac{n_2(\theta^2 + \sigma^2 - \overline{\mathbf{Z}^2})^2}{4(\sigma^{*2} + \overline{\boldsymbol{\mu}^{*2}})^2}\Big| \to 0. \tag{1.A.33}$$

To verify (1.A.32), differentiating (1.5.6) gives

$$\begin{aligned}
&\partial_{\sigma^2}\ell(\sigma^2|\mathbf{Y}) = \frac{n_1}{2\sigma^4}(\overline{\mathbf{Y}^2} - \sigma^2), \quad \partial_{\sigma^2}\ell(\overline{\mathbf{Y}^2}|\mathbf{Y}) = 0,\\
&\partial^2_{\sigma^2}\ell(\sigma^2|\mathbf{Y}) = \frac{n_1}{2\sigma^6}(\sigma^2 - 2\overline{\mathbf{Y}^2}), \quad \partial^2_{\sigma^2}\ell(\overline{\mathbf{Y}^2}|\mathbf{Y}) = -\frac{n_1}{2\overline{\mathbf{Y}^2}^2} < 0,\\
&\partial^3_{\sigma^2}\ell(\sigma^2|\mathbf{Y}) = \frac{n_1}{\sigma^8}(3\overline{\mathbf{Y}^2} - \sigma^2),
\end{aligned}$$

and by a third-order Taylor expansion around the maximum $\overline{\mathbf{Y}^2}$,

$$\begin{aligned}
&\ell(\sigma^2|\mathbf{Y}) - \ell(\overline{\mathbf{Y}^2}|\mathbf{Y})\\
&\quad = \frac{1}{2}\partial^2_{\sigma^2}\ell(\overline{\mathbf{Y}^2}|\mathbf{Y})(\sigma^2 - \overline{\mathbf{Y}^2})^2 + \frac{1}{6}\partial^3_{\sigma^2}\ell(s^2|\mathbf{Y})(\sigma^2 - \overline{\mathbf{Y}^2})^3\\
&\quad = -\frac{n_1}{4\overline{\mathbf{Y}^2}^2}(\sigma^2 - \overline{\mathbf{Y}^2})^2 + \frac{n_1}{6s^8}(3\overline{\mathbf{Y}^2} - s^2)(\sigma^2 - \overline{\mathbf{Y}^2})^3\\
&\quad = -\frac{n_1}{4\sigma^{*4}}(\sigma^2 - \overline{\mathbf{Y}^2})^2 + \frac{n_1(\overline{\mathbf{Y}^2} + \sigma^{*2})}{4\sigma^{*4}\overline{\mathbf{Y}^2}^2}(\overline{\mathbf{Y}^2} - \sigma^{*2})(\sigma^2 - \overline{\mathbf{Y}^2})^2\\
&\qquad + \frac{n_1}{6s^8}(3\overline{\mathbf{Y}^2} - s^2)(\sigma^2 - \overline{\mathbf{Y}^2})^3,
\end{aligned}$$

for some s^2 between σ^2 and $\overline{\mathbf{Y}^2}$. We now control the smaller order terms uniformly over $\sigma^2 \in B_1$. Observe that also $\overline{\mathbf{Y}^2}, s^2 \in B_1$. With (1.A.29), $\sup_{\sigma^2,\widetilde{\sigma}^2 \in B_1} |\sigma^2 - \widetilde{\sigma}^2| = O(\zeta_n)$ and $\sigma^{*2}/2 \leq \sigma^2 \leq 2\sigma^{*2}$ for all $\sigma^2 \in B_1$. Moreover, since $K \subset (0,\infty)$ is compact, $\inf \sigma^{*2} \in K > 0$. Together this shows that

$$\sup_{\sigma^{*2}\in K} \sup_{\sigma^2 \in B_1} \left| \ell(\sigma^2|\mathbf{Y}) - \ell(\overline{\mathbf{Y}^2}|\mathbf{Y}) + \frac{n_1}{4\sigma^{*4}}(\sigma^2 - \overline{\mathbf{Y}^2})^2 \right| = O(n_1\zeta_n^3) \to 0,$$

establishing (1.A.32). To prove (1.A.33) we argue similarly. Differentiating (1.5.6) gives

$$\begin{aligned}
\partial_{\theta^2}\ell(\sigma^2+\theta^2|\mathbf{Z}) &= \frac{n_2}{2(\sigma^2+\theta^2)^2}(\overline{\mathbf{Z}^2} - \sigma^2 - \theta^2), \quad \partial_{\theta^2}\ell(\overline{\mathbf{Z}^2}|\mathbf{Z}) = 0,\\
\partial_{\theta^2}^2\ell(\sigma^2+\theta^2|\mathbf{Z}) &= \frac{n_2}{2(\sigma^2+\theta^2)^3}(\theta^2 + \sigma^2 - 2\overline{\mathbf{Z}^2}), \quad \partial_{\theta^2}^2\ell(\overline{\mathbf{Z}^2}|\mathbf{Z}) = -\frac{n_2}{2\overline{\mathbf{Z}^2}^2} < 0,\\
\partial_{\theta^2}^3\ell(\sigma^2+\theta^2|\mathbf{Z}) &= \frac{n_2}{(\sigma^2+\theta^2)^4}(3\overline{\mathbf{Z}^2} - \sigma^2 - \theta^2),
\end{aligned}$$

and by a third-order Taylor expansion around the maximum $\theta_*^2 = \overline{\mathbf{Z}^2} - \sigma^2$,

$$\begin{aligned}
&\ell(\sigma^2+\theta^2|\mathbf{Z}) - \ell(\overline{\mathbf{Z}^2}|\mathbf{Z})\\
&= \frac{1}{2}\partial_{\theta^2}^2\ell(\overline{\mathbf{Z}^2}|\mathbf{Z})(\theta^2+\sigma^2-\overline{\mathbf{Z}^2})^2 + \frac{1}{6}\partial_{\theta^2}^3\ell(\sigma^2+s^2|\mathbf{Z})(\theta^2+\sigma^2-\overline{\mathbf{Z}^2})^3\\
&= -\frac{n_2}{4\overline{\mathbf{Z}^2}^2}(\theta^2+\sigma^2-\overline{\mathbf{Z}^2})^2 + \frac{n_2}{6(\sigma^2+s^2)^4}(3\overline{\mathbf{Z}^2}-\sigma^2-s^2)(\theta^2+\sigma^2-\overline{\mathbf{Z}^2})^3\\
&= -\frac{n_2}{4(\sigma^{*2}+\overline{\boldsymbol{\mu}^{*2}})^2}(\theta^2+\sigma^2-\overline{\mathbf{Z}^2})^2 + \frac{n_2(\overline{\mathbf{Z}^2}+\sigma^{*2}+\overline{\boldsymbol{\mu}^{*2}})}{4(\sigma^{*2}+\overline{\boldsymbol{\mu}^{*2}})^2\overline{\mathbf{Z}^2}^2}(\overline{\mathbf{Z}^2}-\sigma^{*2}-\overline{\boldsymbol{\mu}^{*2}})(\theta^2+\sigma^2-\overline{\mathbf{Z}^2})^2\\
&\quad + \frac{n_2}{6(\sigma^2+s^2)^4}(3\overline{\mathbf{Z}^2}-\sigma^2-s^2)(\theta^2+\sigma^2-\overline{\mathbf{Z}^2})^3,
\end{aligned}$$

for some s^2 between θ^2 and $\overline{\mathbf{Z}^2} - \sigma^2$. If $(\sigma^2,\theta^2) \in B_1 \times B_2$, then, on A_n, both $\overline{\mathbf{Z}^2} - \sigma^2$ and s^2 are in B_2'. With (1.A.29) and (1.A.31), we have $\sup_{u,v\in B_2'} |u - v| = O(\zeta_n)$ and $(\sigma^{*2} + \overline{\boldsymbol{\mu}^{*2}})/2 \leq \sigma^2 + s^2 \leq 2(\sigma^{*2} + \overline{\boldsymbol{\mu}^{*2}})$ for sufficiently large n. Together with the reasoning for (1.A.32), this leads to

$$\sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} \sup_{\sigma^2\in B_1, \theta^2 \in B_2} \left| \ell(\sigma^2+\theta^2|\mathbf{Z}) - \ell(\overline{\mathbf{Z}^2}|\mathbf{Z}) + \frac{n_2}{4(\sigma^{*2}+\overline{\boldsymbol{\mu}^{*2}})^2}(\theta^2+\sigma^2-\overline{\mathbf{Z}^2})^2 \right|$$

being bounded by $\lesssim n\zeta_n^3$ and thus converging to zero.

Proof of (v): Define $\Pi_3(\cdot|\mathbf{Y},\mathbf{Z})$ as the distribution on $(0,\infty)^2$, with density (1.5.14), that is,

$$\pi_3(\sigma^2,\theta^2|\mathbf{Y},\mathbf{Z}) \propto \mathbf{1}(\sigma^2 \geq 0, \theta^2 \geq 0) e^{-\frac{n_1}{4\sigma^{*4}}(\sigma^2 - \overline{\mathbf{Y}^2})^2} e^{-\frac{n_2}{4(\sigma^{*2}+\overline{\boldsymbol{\mu}^{*2}})^2}(\theta^2+\sigma^2-\overline{\mathbf{Z}^2})^2}.$$

and $\widetilde{\Pi}_3(\cdot|\mathbf{Y},\mathbf{Z})$ as the localization of $\Pi_3(\cdot|\mathbf{Y},\mathbf{Z})$ on $B_1 \times B_2$, that is, the distribution with density

$$\widetilde{\pi}_3(\sigma^2,\theta^2|\mathbf{Y},\mathbf{Z}) \propto \mathbf{1}(\sigma^2 \in B_1, \theta^2 \in B_2) e^{-\frac{n_1}{4\sigma^{*4}}(\sigma^2 - \overline{\mathbf{Y}^2})^2} e^{-\frac{n_2}{4}(\sigma^{*2}+\overline{\boldsymbol{\mu}^{*2}})^{-2}(\theta^2+\sigma^2-\overline{\mathbf{Z}^2})^2}.$$

Here B_1, B_2 are as defined in (1.A.15). The marginal distributions of $\widetilde{\Pi}_3(\cdot|\mathbf{Y},\mathbf{Z})$ and $\Pi_3(\cdot|\mathbf{Y},\mathbf{Z})$ with respect to σ^2 are $\Pi_2(\cdot|\mathbf{Y},\mathbf{Z})$ and $\Pi_\infty(\cdot|\mathbf{Y},\mathbf{Z})$, respectively. Applying (1.A.27) yields

$$\begin{aligned}\big\|\Pi_2(\cdot|\mathbf{Y},\mathbf{Z})-\Pi_\infty(\cdot|\mathbf{Y},\mathbf{Z})\big\|_{\mathrm{TV}} &\leq \big\|\widetilde{\Pi}_3(\cdot|\mathbf{Y},\mathbf{Z})-\Pi_3(\cdot|\mathbf{Y},\mathbf{Z})\big\|_{\mathrm{TV}} \\ &\leq 2\Pi_3\big(\big\{\sigma^2\notin B_1\big\}\cup\big\{\theta^2\notin B_2\big\}\big|\mathbf{Y},\mathbf{Z}\big).\end{aligned} \tag{1.A.34}$$

To prove (v), it remains to show that for $n\to\infty$,

$$\sup_{\sigma^{*2}\in K,\mu_i^*\in K',\forall i}\Pi_3\big(\big\{\sigma^2\notin B_1\big\}\cup\big\{\theta^2\notin B_2\big\}\big|\mathbf{Y},\mathbf{Z}\big)\mathbf{1}\big((\mathbf{Y},\mathbf{Z})\in A_n\big)\to 0. \tag{1.A.35}$$

By Lemma 1.5.4, it is enough to prove that on A_n,

$$\sup_{\sigma^{*2}\in K,\mu_i^*\in K',\forall i} P\big(\xi\notin B_1\big|(0\leq\xi\leq\eta)\big)+P\big(\eta-\xi\notin B_2\big|(0\leq\xi\leq\eta)\big)\to 0, \tag{1.A.36}$$

for independent $\xi\sim\mathcal{N}(\overline{\mathbf{Y}^2},2\sigma^{*4}/n_1),\eta\sim\mathcal{N}(\overline{\mathbf{Z}^2},2(\sigma^{*2}+\overline{\boldsymbol{\mu}^{*2}})^2/n_2)$. Recall that this and all the following statements in (v) should be understood conditionally on $\mathbf{Y},\mathbf{Z}$.

To bound the terms, we heavily rely on the exponential bounds for tail probabilities of Gaussian variables given by Mill's ratio [45]

$$\Big(\frac{x^2}{1+x^2}\Big)\frac{e^{-x^2/2}}{\sqrt{2\pi}x}\leq P\Big(\mathcal{N}(0,1)>x\Big)\leq\frac{e^{-x^2/2}}{\sqrt{2\pi}x},\quad \forall x>0. \tag{1.A.37}$$

In a first step we derive a lower bound on $P(0\leq\xi\leq\eta)$. Using that on A_n, $\overline{\mathbf{Y}^2}/(1+\delta_n/2)\leq\overline{\mathbf{Z}^2}=E[\eta]$, the definition of ξ, the symmetry properties of the $\mathcal{N}(0,1)$ distribution, $\sigma^{*2}/2\leq\overline{\mathbf{Y}^2}\leq 2\sigma^{*2}$ on A_n, and Mill's ratio, we find

$$\begin{aligned}P\big(0\leq\xi\leq\eta\big)&\geq P\Big(0\leq\xi\leq\frac{\overline{\mathbf{Y}^2}}{1+\delta_n/2}\Big)P\big(\overline{\mathbf{Z}^2}\leq\eta\big)\\ &=\frac{1}{2}P\Big(\mathcal{N}(0,1)\in\Big[-\frac{\sqrt{n_1}\overline{\mathbf{Y}^2}}{\sqrt{2}\sigma^{*2}},-\frac{\sqrt{n_1}\delta_n\overline{\mathbf{Y}^2}}{2\sqrt{2}(1+\delta_n/2)\sigma^{*2}}\Big]\Big)\\ &=\frac{1}{2}P\Big(\mathcal{N}(0,1)\in\Big[\frac{\sqrt{n_1}\delta_n\overline{\mathbf{Y}^2}}{2\sqrt{2}(1+\delta_n/2)\sigma^{*2}},\frac{\sqrt{n_1}\overline{\mathbf{Y}^2}}{\sqrt{2}\sigma^{*2}}\Big]\Big)\\ &\geq\frac{1}{2}P\Big(\mathcal{N}(0,1)\in\Big[\frac{\sqrt{n_1}\delta_n}{\sqrt{2}},\frac{\sqrt{n_1}}{2\sqrt{2}}\Big]\Big)\\ &=P\Big(\mathcal{N}(0,1)\geq\frac{\sqrt{n_1}\delta_n}{\sqrt{2}}\Big)-P\Big(\mathcal{N}(0,1)\geq\frac{\sqrt{n_1}}{2\sqrt{2}}\Big)\\ &\geq\frac{1}{2\sqrt{\pi n_1}\delta_n}e^{-\frac{n_1\delta_n^2}{4}}-\frac{2}{\sqrt{\pi n_1}}e^{-\frac{n_1}{16}}.\end{aligned} \tag{1.A.38}$$

where in the last inequality we used that $x^2/(1+x^2)>\frac{1}{2}$ for $x>1$.

We now derive an upper bound for $P(\xi\notin B_1)$. Using the definition of ξ, $\zeta_n\leq 1$,

$\overline{\mathbf{Y}^2} \geq \sigma^{*2}/2$, and Mill's ratio (1.A.37),

$$\begin{aligned} P(\xi \notin B_1) &= P\Big(\mathcal{N}(0,1) \notin \Big[-\frac{\sqrt{n_1}\zeta_n\overline{\mathbf{Y}^2}}{\sqrt{2}(1+\zeta_n)\sigma^{*2}}, \frac{\sqrt{n_1}\zeta_n\overline{\mathbf{Y}^2}}{\sqrt{2}(1-\zeta_n)\sigma^{*2}}\Big]\Big) \\ &\leq 2P\Big(\mathcal{N}(0,1) > \frac{\sqrt{n_1}\zeta_n\overline{\mathbf{Y}^2}}{\sqrt{2}(1+\zeta_n)\sigma^{*2}}\Big) \\ &\leq 2P\Big(\mathcal{N}(0,1) > \frac{\sqrt{n_1}\zeta_n}{4\sqrt{2}}\Big) \\ &\leq \frac{8}{\sqrt{\pi n_1}\zeta_n}e^{-\frac{n_1\zeta_n^2}{64}}. \end{aligned} \tag{1.A.39}$$

Next, we obtain a similar bound for $P(\eta-\xi \notin B_2, \xi \leq \eta, \xi \in B_1)$. If we define the difference of two sets U, V as $U - V := \{u - v : u \in U, v \in V\}$, then, $B_2 = ([\overline{\mathbf{Z}^2}/(1+\zeta_n), \overline{\mathbf{Z}^2}/(1-\zeta_n)] - B_1) \cap \mathbb{R}_+$. On the event $\xi \leq \eta, \xi \in B_1$, we have that $\eta \in [\overline{\mathbf{Z}^2}/(1+\zeta_n), \overline{\mathbf{Z}^2}/(1-\zeta_n)]$ implies that $\eta - \xi \in B_2$, which is equivalent to saying that $\eta - \xi \notin B_2$ implies $\eta \notin [\overline{\mathbf{Z}^2}/(1+\zeta_n), \overline{\mathbf{Z}^2}/(1-\zeta_n)]$. On A_n, $|\overline{\mathbf{Z}^2} - \overline{\boldsymbol{\mu}^{*2}} - \sigma^{*2}| \leq \sigma^{*2}\delta_n$ by definition. Because of $\delta_n \leq 1/2$, we obtain $\overline{\mathbf{Z}^2} \geq (\overline{\boldsymbol{\mu}^{*2}} + \sigma^{*2})/2$. Together with the symmetry properties of the normal distribution, $\zeta_n \leq 1$, and Mill's ratio (1.A.37), this yields

$$\begin{aligned} &P\big(\eta - \xi \notin B_2, \xi \leq \eta, \xi \in B_1\big) \\ &\leq P\Big(\eta \notin \Big[\frac{\overline{\mathbf{Z}^2}}{1+\zeta_n}, \frac{\overline{\mathbf{Z}^2}}{1-\zeta_n}\Big]\Big) \\ &= P\Big(\mathcal{N}(0,1) \notin \Big[-\frac{\sqrt{n_2}\zeta_n\overline{\mathbf{Z}^2}}{\sqrt{2}(\overline{\boldsymbol{\mu}^{*2}}+\sigma^{*2})(1+\zeta_n)}, \frac{\sqrt{n_2}\zeta_n\overline{\mathbf{Z}^2}}{\sqrt{2}(\overline{\boldsymbol{\mu}^{*2}}+\sigma^{*2})(1-\zeta_n)}\Big]\Big) \\ &\leq 2P\Big(\mathcal{N}(0,1) > \frac{\sqrt{n_2}\zeta_n}{4\sqrt{2}}\Big) \\ &\leq \frac{8}{\sqrt{\pi n_2}\zeta_n}e^{-\frac{n_2\zeta_n^2}{64}}. \end{aligned} \tag{1.A.40}$$

To prove (1.A.36), we bound

$$P\big(\xi \notin B_1 \big| (0 \leq \xi \leq \eta)\big) \leq \frac{P(\xi \notin B_1)}{P(0 \leq \xi \leq \eta)}$$

and

$$P\big(\eta - \xi \notin B_2 \big| (0 \leq \xi \leq \eta)\big) \leq \frac{P(\eta - \xi \notin B_2, \xi \in B_1, 0 \leq \xi \leq \eta) + P(\xi \notin B_1)}{P(0 \leq \xi \leq \eta)}.$$

Now (1.A.36) (and therefore (1.A.35)) follow from the inequalities (1.A.38), (1.A.39), (1.A.40) and the definition of δ_n. This completes the proof of (v).

Proof of (vi): Recall the definitions of the densities

$$\begin{aligned} \pi_\infty(\sigma^2|\mathbf{Y},\mathbf{Z}) &\propto \mathbf{1}(\sigma^2 \geq 0)\exp\Big(-\frac{n_1}{4\sigma^{*4}}(\sigma^2 - \overline{\mathbf{Y}^2})^2\Big)\Big(1 - \Phi\Big(\frac{\sqrt{n_2}(\sigma^2 - \overline{\mathbf{Z}^2})}{\sqrt{2}(\sigma^{*2} + \overline{\boldsymbol{\mu}^{*2}})}\Big)\Big), \\ \widetilde{\pi}_\infty(\sigma^2|\mathbf{Y}) &\propto \mathbf{1}(\sigma^2 \geq 0)\exp\Big(-\frac{n_1}{4\sigma^{*4}}(\sigma^2 - \overline{\mathbf{Y}^2})^2\Big), \end{aligned}$$

and let

$$\pi_{\infty,B_1}(\sigma^2|\mathbf{Y},\mathbf{Z}) \propto \pi_\infty(\sigma^2|\mathbf{Y},\mathbf{Z})\mathbf{1}(\sigma^2 \in B_1),$$
$$\widetilde{\pi}_{\infty,B_1}(\sigma^2|\mathbf{Y}) \propto \widetilde{\pi}_\infty(\sigma^2|\mathbf{Y})\mathbf{1}(\sigma^2 \in B_1),$$

be their localised versions on B_1. It is enough to show that, on A_n,

$$\sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} \left\|\Pi_\infty(\cdot|\mathbf{Y},\mathbf{Z}) - \Pi_{\infty,B_1}(\cdot|\mathbf{Y},\mathbf{Z})\right\|_{\mathrm{TV}} \xrightarrow{n\to\infty} 0, \tag{1.A.41}$$

$$\sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} \left\|\widetilde{\Pi}_\infty(\cdot|\mathbf{Y}) - \widetilde{\Pi}_{\infty,B_1}(\cdot|\mathbf{Y})\right\|_{\mathrm{TV}} \xrightarrow{n\to\infty} 0, \tag{1.A.42}$$

$$\sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} \left\|\Pi_{\infty,B_1}(\cdot|\mathbf{Y},\mathbf{Z}) - \widetilde{\Pi}_{\infty,B_1}(\cdot|\mathbf{Y})\right\|_{\mathrm{TV}} \xrightarrow{n\to\infty} 0. \tag{1.A.43}$$

For (1.A.41), we apply (1.A.27) and the fact that $\Pi_\infty(\cdot|\mathbf{Y},\mathbf{Z})$ is the marginal distribution of $\Pi_3(\cdot|\mathbf{Y},\mathbf{Z})$, finding

$$\left\|\Pi_\infty(\cdot|\mathbf{Y},\mathbf{Z}) - \Pi_{\infty,B_1}(\cdot|\mathbf{Y},\mathbf{Z})\right\|_{\mathrm{TV}} \le 2\Pi_\infty(B_1^c|\mathbf{Y},\mathbf{Z}) = 2\Pi_3(B_1^c|\mathbf{Y},\mathbf{Z}).$$

In (v) we proved that the right hand side converges to zero uniformly over $\sigma^{*2} \in K, \mu_i^* \in K', \forall i$. For (1.A.42), we argue similarly, using that

$$\left\|\widetilde{\Pi}_\infty(\cdot|\mathbf{Y}) - \widetilde{\Pi}_{\infty,B_1}(\cdot|\mathbf{Y})\right\|_{\mathrm{TV}} \le 2\widetilde{\Pi}_\infty(B_1^c|\mathbf{Y}) = 2P(\xi \notin B_1),$$

with $\xi \sim \mathcal{N}(\overline{\mathbf{Y}^2}, 2\sigma^{*4}/n_1)$. Using (1.A.39), we see that the right hand side converges to zero, uniformly over $\sigma^{*2} \in K, \mu_i^* \in K', \forall i$.

For (1.A.43), we apply Lemma 1.A.2. On A_n, the likelihood ratio of $\Pi_{\infty,B_1}(\cdot|\mathbf{Y},\mathbf{Z})$ and $\widetilde{\Pi}_{\infty,B_1}(\cdot|\mathbf{Y})$ is given by

$$h(\sigma^2|\mathbf{Y},\mathbf{Z}) := \left(1 - \Phi\left(\frac{\sqrt{n_2}(\sigma^2 - \overline{\mathbf{Z}^2})}{\sqrt{2}(\overline{\boldsymbol{\mu}^{*2}} + \sigma^{*2})}\right)\right)\mathbf{1}(\sigma^2 \in B_1).$$

On A_n,

$$\sup_{\sigma^2\in B_1} \sigma^2 - \overline{\mathbf{Z}^2} = \frac{\overline{\mathbf{Y}^2}}{1-\zeta_n} - \overline{\mathbf{Z}^2} \le \frac{\sigma^{*2}(1+\delta_n)}{1-\zeta_n} - \overline{\boldsymbol{\mu}^{*2}} - \sigma^{*2}(1-\delta_n).$$

Uniformly over $\sigma^{*2} \in K$ and $\inf_{\mu_i^*\in K'} |\mu_i^*|^2 \gg \zeta_n$, the right hand side can be further upper bounded by $-\overline{\boldsymbol{\mu}^{*2}}/2$ for sufficiently large n. Thus,

$$\begin{aligned}|h(\sigma^2|\mathbf{Y},\mathbf{Z}) - 1| &= P\left(\mathcal{N}(0,1) \le \frac{\sqrt{n_2}(\sigma^2 - \overline{\mathbf{Z}^2})}{\sqrt{2}(\overline{\boldsymbol{\mu}^{*2}} + \sigma^{*2})}\right)\\ &\le P\left(\mathcal{N}(0,1) \ge \frac{\sqrt{n_2}\overline{\boldsymbol{\mu}^{*2}}}{2\sqrt{2}(\overline{\boldsymbol{\mu}^{*2}} + \sigma^{*2})}\right).\end{aligned}$$

Since $n\overline{\boldsymbol{\mu}^{*2}} \gg n\zeta_n \to \infty$ for $n \to \infty$,

$$\sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} P\left(\mathcal{N}(0,1) \ge \frac{\sqrt{n_2}\overline{\boldsymbol{\mu}^{*2}}}{2\sqrt{2}(\overline{\boldsymbol{\mu}^{*2}} + \sigma^{*2})}\right) \xrightarrow{n\to\infty} 0.$$

This concludes the proof of *(vi)*. □

Proof of Theorem 1.5.2. We insert $1 = \mathbf{1}((\mathbf{Y},\mathbf{Z}) \in A_n) + \mathbf{1}((\mathbf{Y},\mathbf{Z}) \notin A_n)$ in the expectation. Since the total variation distance of probability measures is bounded, the result follows from Proposition 1.A.1. □

Proof of Corollary 1.5.3. Recall that the posterior is the marginal distribution of $\widetilde{\Pi}(\cdot|\mathbf{Y},\mathbf{Z})$ with respect to σ^2. By Proposition 1.A.1 (ii), we have that

$$\sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} \Pi\Big(\sigma^2 \notin \Big[\frac{\overline{\mathbf{Y}^2}}{1+\zeta_n}, \frac{\overline{\mathbf{Y}^2}}{1-\zeta_n}\Big]\Big|\mathbf{Y},\mathbf{Z}\Big)\mathbf{1}\big((\mathbf{Y},\mathbf{Z}) \in A_n\big) \to 0.$$

Using that on A_n, $\sigma^{*2}(1-\delta_n) \leq \overline{\mathbf{Y}^2} \leq \sigma^{*2}(1+\delta_n)$, and $\delta_n = C^{-1}\zeta_n = O(\sqrt{\log n/n})$, we obtain

$$\sup_{\sigma^{*2}\in K, \mu_i^* \in K', \forall i} \Pi\Big(\Big|\frac{\sigma^2}{\sigma^{*2}} - 1\Big| \geq M\sqrt{\frac{\log n}{n}}\Big|\mathbf{Y},\mathbf{Z}\Big)\mathbf{1}\big((\mathbf{Y},\mathbf{Z}) \in A_n\big) \to 0$$

for a constant $M = M(\alpha)$ that is chosen to be sufficiently large. The claim follows by splitting the expected posterior, inserting $1 = \mathbf{1}((\mathbf{Y},\mathbf{Z}) \in A_n) + \mathbf{1}((\mathbf{Y},\mathbf{Z}) \notin A_n)$ in the expectation and using Proposition 1.A.1 (i). □

Proof of Lemma 1.5.4. To prove the result, we derive an expression for the joint density of $(\xi, \eta-\xi)\big|(0 \leq \xi \leq \eta)$. Observe that

$$\begin{aligned} P\big(\xi \leq s, \eta - \xi \leq t\big|(0 \leq \xi \leq \eta)\big) &= \frac{P(\xi \leq s, \eta-\xi \leq t, 0 \leq \xi \leq \eta)}{P(0 \leq \xi \leq \eta)} \\ &\propto P\big((\eta - t) \vee 0 \leq \xi \leq \eta \wedge s\big). \end{aligned}$$

The right hand side is zero if $s \leq 0$. Suppose now that $0 \leq s \leq t$. Conditioning on η, the right hand side can be rewritten as

$$\begin{aligned} &= \int_0^s P\big(0 \leq \xi \leq u\big) f_\eta(u)du + \int_s^t P\big(0 \leq \xi \leq s\big) f_\eta(u)du \\ &\quad + \int_t^{t+s} P\big(u - t \leq \xi \leq s\big) f_\eta(u)du. \end{aligned}$$

Taking derivatives $\partial_s\partial_t$, the density of $(\xi, \eta-\xi)\big|(0 \leq \xi \leq \eta)$ at point (s,t) equals up to a multiplicative constant $f_\xi(s)f_\eta(t+s)$. Which completes the proof for the case $0 \leq s \leq t$.

The case $0 \leq t \leq s$ is similar and the proof for this case therefore omitted.

Since the posterior limit distribution is the marginal over the first component of the joint distribution in (1.5.14), it must coincide with the distribution of $\xi\big|(0 \leq \xi \leq \eta)$. □

Chapter 2

Posterior contraction for deep Gaussian process priors

This chapter is based on:
G. Finocchio and J. Schmidt-Hieber. Posterior contraction for deep Gaussian process priors. *Arxiv preprint, arXiv:2105.07410 (2021).*

Abstract

We study posterior contraction rates for a class of deep Gaussian process priors applied to the nonparametric regression problem under a general composition assumption on the regression function. It is shown that the contraction rates can achieve the minimax convergence rate (up to $\log n$ factors), while being adaptive to the underlying structure and smoothness of the target function. The proposed framework extends the Bayesian nonparametrics theory for Gaussian process priors.

2.1 Introduction

In the multivariate nonparametric regression model with random design supported on $[-1,1]^d$, we observe n i.i.d. pairs $(\mathbf{X}_i, Y_i) \in [-1,1]^d \times \mathbb{R}$, $i = 1, \ldots, n$, with

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.1.1}$$

and ε_i independent and standard normal random variables that are independent of the design vectors $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$. We aim to recover the true regression function $f : [-1,1]^d \to \mathbb{R}$ from the sample. Here it is assumed that the regression function f itself is a composition of a number of unknown simpler functions. This comprises several important cases including (generalized) additive models. In [76] it has been shown that sparsely connected deep neural networks are able to pick up the underlying composition structure and achieve near minimax estimation rates. On the contrary, wavelet thresholding methods are shown to be unable to adapt to the underlying structure resulting in potentially much slower convergence rates.

Deep Gaussian process priors (DGPs), cf. [68, 33, 32], can be viewed as a Bayesian analogue of deep networks. While deep nets are build on a hierarchy of individual network layers, DGPs are based on iterations of Gaussian processes. Compared to neural networks, DGPs have moreover the advantage that the posterior can be used for uncertainty quantification. This makes them potentially attractive for AI applications with a strong safety aspect, such as automated driving and health.

In the classical Bayesian nonparametric regression setting, Gaussian process priors are a natural choice and a comprehensive literature is available, see for instance [82] or Section 11 in [41]. In this work we extend the theory of Gaussian process priors to derive posterior contraction rates for DGPs. Inspired by model selection priors, we construct classes of DGP priors in a hierarchical manner by first assigning a prior to possible composition structures and smoothness indices. Given a composition structure with corresponding smoothness indices, we then generate the prior distribution by putting suitable Gaussian processes on all functions in this structure. It is shown that for such a DGP prior construction the posterior contraction rate matches nearly the minimax estimation rate. In particular, if there is some low-dimensional structure in the composition, the posterior will not suffer from the curse of dimensionality.

Stabilization enhancing methods such as dropout and batch normalization are crucial for the performance of deep learning. In particular, batch normalization guarantees that the signal sent through a trained network cannot explode. We argue that for deep Gaussian processes similar effects play a role. In Figure 2.1 below we visualize the effect of composing independent copies of a Gaussian process, the resulting trajectories are rougher and more versatile than those generated by the original process alone. This may however lead to wild behavior of the sample paths. As we aim for a fully Bayesian approach, the only possibility is to induce stability through the selection of the prior. We enforce stability by conditioning each individual Gaussian process to lie in a set of 'stable' paths. To achieve near optimal contraction rates, these sets have to be carefully selected and depend on the optimal contraction rate itself.

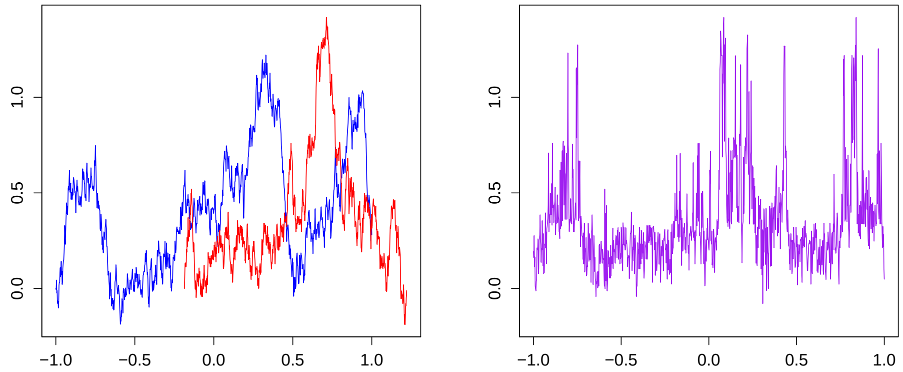

Figure 2.1: Composition of Gaussian processes results in rougher and more versatile sample paths. On the left: trajectories of two independent copies of a standard Brownian motion. On the right: the composition (red ∘ blue) of the trajectories.

Compared to the well-established nonparametric Bayes theory for Gaussian processes, posterior contraction for compositions of Gaussian processes raises some theoretical challenges, such as bounding the decentered small ball probabilities of the DGPs. We show that this can be done by using an extension of the concentration function for Gaussian processes introduced in [82]. To our knowledge, the closest results in the literature are bounds on the centred small ball probabilities of iterated processes. They have been obtained for self-similar processes in [4] and for time-changed self-similar processes in [52]. A good reference on the literature of iterated processes is given by [3]. In a different line of research, iterated Brownian motions (IBMs) occur in [38] as solutions of high-order parabolic stochastic differential equations (SDEs) and the path properties are studied in [15]. The composition of general processes in relation with high-order parabolic and hyperbolic SDEs has been studied in [48]. More recently, the *ad libitum* (infinite) iteration of Brownian motions has been studied in [31, 16].

The article is structured as follows. In Section 2.2 we formalize the model and give an explicit parametrization of the underlying graph and the smoothness index. Section 2.3 provides a detailed construction of the deep Gaussian process prior. In Section 2.4 we state the main posterior contraction results. In Section 2.5 we present a construction achieving optimal contraction rates and provide explicit examples in Section 2.6. Section 2.7 compares Bayes with DGPs and deep learning. All proofs are deferred to Section 2.A.

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 $\mathbf{x}_S = (x_i)_{i \in S}$ and $|S|$ for the cardinality of S. As usual, we define $|\mathbf{x}|_p := (\sum_{i=1}^d |\mathbf{x}_i|^p)^{1/p}$, $|\mathbf{x}|_\infty := \max_i |\mathbf{x}_i|$, $|\mathbf{x}|_0 := \sum_{i=1}^d \mathbf{1}(\mathbf{x}_i \neq 0)$, 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 $\|\cdot\|_p$. 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 $a_n \leq C b_n$ for all n. Moreover, $a_n \asymp b_n$ means that $(a_n)_n \lesssim (b_n)_n$ and $(b_n)_n \lesssim (a_n)_n$. For positive sequences $(a_n)_n$ and $(b_n)_n$ we write $a_n \ll b_n$ if a_n/b_n tends to zero when n tends to infinity.

2.2 Composition structure on the regression function

We assume that the regression function f in the nonparametric regression model (2.1.1) can be written as the composition of $q+1$ functions, that is, $f = g_q \circ g_{q-1} \circ \ldots \circ g_1 \circ g_0$, for functions $g_i : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}}$ with $d_0 = d$ and $d_{q+1} = 1$. It should be clear that we are interested in reconstruction of f but not of the individual components $g_0, \ldots, g_q$.

If f takes values in the interval $[-1, 1]$, rescaling $h_i = g_i(\|g_{i-1}\|_\infty \cdot)/\|g_i\|_\infty$ with $\|g_{-1}\|_\infty := 1$ leads to the alternative representation

$$f = h_q \circ h_{q-1} \circ \ldots \circ h_1 \circ h_0 \tag{2.2.1}$$

for functions $h_i : [-1, 1]^{d_i} \to [-1, 1]^{d_{i+1}}$. We also write $h_i = (h_{ij})_{j=1,\ldots,d_{i+1}}^\top$, with $h_{ij} : [-1, 1]^{d_i} \to [-1, 1]$. The representation can be modified if f takes values outside $[-1, 1]$, but to avoid unnecessary technical complications, we do not consider this case here. Although the function h_i in the representation of f is defined on $[-1,1]^{d_i}$, we allow each component function h_{ij} to possibly only depend on a subset of t_i variables for some $t_i \leq d_i$.

To define suitable function classes and priors on composition functions, it is natural to first associate to each composition structure a directed graph. The nodes in the graph are arranged in $q+2$ layers with $q+1$ the number of components in (2.2.1). The number of nodes in each layer is given by the integer vector $\mathbf{d} := (d, d_1, \ldots, d_q, 1) \in \mathbb{N}^{q+2}$ storing the dimensions of the components h_i appearing in (2.2.1). As mentioned above, each component function h_{ij} might only depend on a subset $\mathcal{S}_{ij} \subseteq \{1, \ldots, d_i\}$ of all d_i variables. We call $\mathcal{S}_{ij}$ the active set of the function h_{ij}. In the graph, we draw an edge between the j-th node in the $i+1$-st layer and the k-th node in the i-th layer iff $k \in \mathcal{S}_{ij}$. For any i, the subsets corresponding to different nodes $j = 1, \ldots, d_{i+1}$, are combined into $\mathcal{S}_i := (\mathcal{S}_{i1}, \ldots, \mathcal{S}_{id_{i+1}})$ and $\mathcal{S} := (\mathcal{S}_0, \ldots, \mathcal{S}_q)$. By definition of t_i, we have $t_i := \max_{j=1,\ldots,d_{i+1}} |\mathcal{S}_{ij}|$ and we define $\mathbf{t} := (t_0, \ldots, t_q, 1)$. We summarize the previous quantities into the hyper-parameter

$$\lambda := (q, \mathbf{d}, \mathbf{t}, \mathcal{S}), \tag{2.2.2}$$

which we refer to as the graph of the function f in (2.2.1). The set of all possible graphs is denoted by Λ.

As an example consider the function $f(x_1, \ldots, x_5) = h(g_1(x_1, x_3, x_4), g_2(x_1, x_4, x_5), g_3(x_2))$ with corresponding graph representation displayed in Figure 2.2. In this case, we have $q = 1$, $d_0 = 5, d_1 = 3, d_2 = 1$ and $t_0 = t_1 = 3$. The active sets are $\mathcal{S}_{11} = \{1,3,4\}, \mathcal{S}_{12} = \{1,4,5\}, \mathcal{S}_{13} = \{2\}$, and $\mathcal{S}_{21} = \{1,2,3\}$.

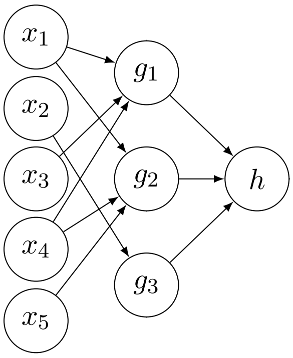

Figure 2.2: Graph representation of the example function f.

We assume that all functions in the composition are Hölder smooth. A function has Hölder smoothness index $\beta > 0$ if all partial derivatives up to order $\lfloor \beta \rfloor$ exist and are bounded, and the partial derivatives of order $\lfloor \beta \rfloor$ are $(\beta - \lfloor \beta \rfloor)$-Hölder. Here, the ball of β-smooth Hölder functions of radius K is defined as

$$\mathcal{C}_r^\beta(K) = \Big\{ f : [-1,1]^r \to [-1,1] : \\ 2r \sum_{\boldsymbol{\alpha}: |\boldsymbol{\alpha}| < \lfloor \beta \rfloor} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + 2^{\beta - \lfloor \beta \rfloor} \sum_{\boldsymbol{\alpha}: |\boldsymbol{\alpha}| = \lfloor \beta \rfloor} \sup_{\substack{\mathbf{x}, \mathbf{y} \in [-1,1]^r \\ \mathbf{x} \neq \mathbf{y}}} \frac{|\partial^{\boldsymbol{\alpha}} f(\mathbf{x}) - \partial^{\boldsymbol{\alpha}} f(\mathbf{y})|}{|\mathbf{x} - \mathbf{y}|_\infty^{\beta - \lfloor \beta \rfloor}} \leq K \Big\}, \tag{2.2.3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r) \in \mathbb{N}^r$ is a multi-index, $|\boldsymbol{\alpha}| := |\boldsymbol{\alpha}|_1$ and $\partial^{\boldsymbol{\alpha}} = \partial^{\alpha_1} \ldots \partial^{\alpha_r}$. The factors $2r$ and $2^{\beta - \lfloor \beta \rfloor}$ guarantee the embedding $\mathcal{C}_r^\beta(K) \subseteq \mathcal{C}_r^{\beta'}(K)$ whenever $\beta' \leq \beta$, see Lemma 2.A.1.

For any subset of indexes S, we write $(\cdot)_S : \mathbf{x} \mapsto \mathbf{x}_S = (x_i)_{i \in S}$ and $(\cdot)_S^{-1}$ for the inverse. Since h_{ij} depends on at most t_i variables, we can think of the function $\overline{h}_{ij} := h_{ij} \circ (\cdot)_{\mathcal{S}_{ij}}^{-1}$ as a function mapping $[-1,1]^{t_i}$ to $[-1,1]$ and assume that $\overline{h}_{ij} \in \mathcal{C}^{\beta_i}_{t_i}(K)$, with $\beta_i \in [\beta_-, \beta_+]$, for some known and fixed $0 < \beta_- \leq \beta_+ < +\infty$. The smoothness indexes of the $q+1$ components are collected into the vector

$$\boldsymbol{\beta} := (\beta_0, \dots, \beta_q) \in [\beta_-, \beta_+]^{q+1} =: I(\lambda). \tag{2.2.4}$$

Combined with the graph parameter λ in (2.2.2), the function composition is completely described by

$$\eta := (\lambda, \boldsymbol{\beta}) = (q, \mathbf{d}, \mathbf{t}, \mathcal{S}, \boldsymbol{\beta}). \tag{2.2.5}$$

We refer to η as the composition structure of the regression function f in (2.2.1). The set of all possible choices of $\eta = (\lambda, \boldsymbol{\beta})$ with $\lambda \in \Lambda$ and $\boldsymbol{\beta} \in I(\lambda)$ is denoted by Ω.

Throughout the following, we assume that the true regression function belongs to the function space $\mathcal{F}(\eta, K)$ where

$$\begin{aligned}\mathcal{F}\big(\eta, K\big) := \Big\{ f = h_q \circ h_{q-1} \circ \ldots \circ h_1 \circ h_0 :\ & h_i = (h_{ij})_j : [-1,1]^{d_i} \to [-1,1]^{d_{i+1}}, \\ & h_{ij} \circ (\cdot)^{-1}_{S_{ij}} \in \mathcal{C}^{\beta_i}_{t_i}\big(K\big) \Big\},\end{aligned} \tag{2.2.6}$$

for some known $K > 0$ and unknown $\eta \in \Omega$. In fact, the regression function f might belong to the space $\mathcal{F}(\eta, K)$ for several choices of η. Since different choices of η lead to different posterior contraction rates, the regression function will always be associated with an η that leads to the fastest contraction rate.

2.3 Deep Gaussian process prior

In this section, we construct the deep Gaussian process prior as prior on composition functions. Because of the complexity of the underlying graph structure, the construction is split into several steps. The final DGP prior consists of a prior on the graph describing the composition structure and, given the graph, a prior on all the individual functions that occur in the representation. To achieve fast contraction rates, the prior weight assigned to a specific composition structure depends on the smoothness properties and the sample size. Therefore, the composition structure can not be decoupled from the estimation problem.

STEP 0. Choice of Gaussian processes. For a centered Gaussian process $X = (X_t)_{t \in T}$, the covariance operator viewed as a function on $T \times T$, that is, $(s,t) \mapsto k(s,t) = \mathbb{E}[X_s X_t]$ is a positive semidefinite function. The reproducing kernel Hilbert space (RKHS) generated by k is called the RKHS corresponding to the Gaussian process X, see [83] for more details.

For any $r = 1, 2, \dots,$ and any $\beta > 0$, let $\widetilde{G}^{(\beta,r)} = (\widetilde{G}^{(\beta,r)}(\mathbf{u}))_{\mathbf{u} \in [-1,1]^r}$ be a centered Gaussian process on the Banach space of continuous functions from $[-1,1]^r$ to $\mathbb{R}$ equipped with the supremum norm. Write $\|\cdot\|_{\mathbb{H}^{(\beta,r)}}$ for the RKHS-norm of the reproducing kernel Hilbert space $\mathbb{H}^{(\beta,r)}$ corresponding to $\widetilde{G}^{(\beta,r)}$. For positive Hölder radius K, we call

$$\varphi^{(\beta,r,K)}(u) := \sup_{f \in \mathcal{C}^{\beta}_r(K)} \ \inf_{g : \|g - f\|_\infty \leq u} \|g\|^2_{\mathbb{H}^{(\beta,r)}} - \log \mathbb{P}\big(\big\|\widetilde{G}^{(\beta,r)}\big\|_\infty \leq u\big), \tag{2.3.1}$$

the concentration function over $\mathcal{C}_r^\beta(K)$. This is the global version of the local concentration function appearing in the posterior contraction theory for Gaussian process priors [82]. For any $0 < \alpha \leq 1$, let $\varepsilon_n(\alpha, \beta, r)$ be such that

$$\varphi^{(\beta,r,K)}\big(\varepsilon_n(\alpha,\beta,r)^{1/\alpha}\big) \leq n\varepsilon_n(\alpha,\beta,r)^2. \tag{2.3.2}$$

STEP 1. Deep Gaussian processes. We now define a corresponding DGP $G^{(\eta)}$ on a given composition structure $\eta = (q, \mathbf{d}, \mathbf{t}, \mathcal{S}, \boldsymbol{\beta})$. Let $\mathbb{B}_\infty(R) := \{f : \sup_{\mathbf{x}\in[-1,1]^r} |f(\mathbf{x})| \leq R\}$ be the supremum unitary ball with radius R. For simplicity, we suppress the dependence on r. Recall that K is assumed to be known. With $\alpha_i := \prod_{\ell=i+1}^q (\beta_\ell \wedge 1)$, we define the subset of paths

$$\mathcal{D}_i(\eta, K) := \mathbb{B}_\infty(1) \cap \Big(\mathcal{C}_{t_i}^{\beta_i}(K) + \mathbb{B}_\infty\big(2\varepsilon_n(\alpha_i,\beta_i,t_i)^{1/\alpha_i}\big)\Big), \tag{2.3.3}$$

containing all functions that belong to the supremum unitary ball $\mathbb{B}_\infty(1)$ and are at most $2\varepsilon_n(\alpha_i,\beta_i,t_i)^{1/\alpha_i}$-away in supremum norm from the Hölder-ball $\mathcal{C}_{t_i}^{\beta_i}(K)$. With $\widetilde{G}^{(\beta,r)}$ the centred Gaussian process in Step 0, write $\overline{G}_i^{(\beta_i,t_i)}$ for the process $\widetilde{G}^{(\beta_i,t_i)}$ conditioned on the event $\{\widetilde{G}^{(\beta_i,t_i)} \in \mathcal{D}_i(\eta, K)\}$. Recall that for an index set S, the function $(\cdot)_S$ maps a vector to the components in S. For each $i = 0, \ldots, q$, $j = 1, \ldots, d_{i+1}$, define the component functions $G_{ij}^{(\eta)}$ to be independent copies of the processes $\overline{G}_i^{(\beta_i,t_i)} \circ (\cdot)_{S_{ij}} : [-1,1]^{d_i} \to [-1,1]$. Finally, set $G_i^{(\eta)} := (G_{ij}^{(\eta)})_{j=1}^{d_{i+1}}$ and define the deep Gaussian process $G^{(\eta)} := G_q^{(\eta)} \circ \ldots \circ G_0^{(\eta)} : [-1,1]^d \to [-1,1]$. We denote by $\Pi(\cdot|\eta)$ the distribution of $G^{(\eta)}$.

STEP 2. Structure prior. We now construct a hyper-prior on the underlying composition structure. For each graph λ, we have $\boldsymbol{\beta} \in I(\lambda) = [\beta_-, \beta_+]^{q+1}$ with known lower and upper bounds $0 < \beta_- \leq \beta_+ < +\infty$. For any function $a(\eta) = a(\lambda, \beta)$, it is convenient to define

$$\int a(\eta)\, d\eta := \sum_{\lambda\in\Lambda} \int_{I(\lambda)} a(\lambda, \beta)\, d\beta.$$

Let γ be a probability density on the possible composition structures, that is, $\int \gamma(\eta)\, d\eta = 1$. We can construct such a measure γ by first choosing a distribution on the number of compositions q. Given q one can then select distributions on the ambient dimensions $\mathbf{d}$, the efficient dimensions $\mathbf{t}$, the active sets $\mathcal{S}$ and finally the smoothness $\boldsymbol{\beta} \in [\beta_-, \beta_+]^{q+1}$ via the conditional density formula $\gamma(\eta) = \gamma(\lambda)\gamma(\boldsymbol{\beta}|\lambda) = \gamma(q)\gamma(\mathbf{d}|q)\gamma(\mathbf{t}|\mathbf{d}, q)\gamma(\mathcal{S}|\mathbf{t}, \mathbf{d}, q)\gamma(\boldsymbol{\beta}|\mathcal{S}, \mathbf{t}, \mathbf{d}, q)$. For a sequence $\varepsilon_n(\eta)$ satisfying

$$\varepsilon_n(\eta) \geq \max_{i=0,\ldots,q} \varepsilon_n(\alpha_i,\beta_i,t_i), \quad \text{with} \quad \alpha_i := \prod_{\ell=i+1}^q \big(\beta_\ell \wedge 1\big), \tag{2.3.4}$$

and $|\mathbf{d}|_1 = 1 + \sum_{i=0}^q d_i$, consider the hyper-prior

$$\pi(\eta) := \frac{e^{-\Psi_n(\eta)}\gamma(\eta)}{\int e^{-\Psi_n(\eta)}\gamma(\eta)\, d\eta}, \quad \text{with} \quad \Psi_n(\eta) := n\varepsilon_n(\eta)^2 + e^{e^{|\mathbf{d}|_1}}. \tag{2.3.5}$$

The denominator is positive and finite, since $0 < e^{-\Psi_n(\eta)} \leq 1$ and $\int \gamma(\eta)\, d\eta = 1$.

STEP 3. DGP prior. We define the deep Gaussian process prior as

$$\Pi(df) := \int_{\Omega} \Pi(df|\eta)\pi(\eta)\, d\eta, \tag{2.3.6}$$

where Ω is the set of all valid composition structures, $\Pi(\cdot|\eta)$ is the distribution of the DGP $G^{(\eta)}$ and $\pi(\eta)$ is the structure prior on η.

Lemma 2.5.1 shows that it is often enough to check (2.3.2) for $\alpha = 1$ only. Conditioning the Gaussian process to the set $\mathcal{D}_i(\eta, K)$ is well-defined since $\mathcal{D}_i(\eta, K) \supset \mathbb{B}_\infty\big(2\varepsilon_n(\alpha_i, \beta_i, t_i)^{1/\alpha_i}\big)$ and Gaussian processes with continuous sample paths give positive mass to $\mathbb{B}_\infty(R)$ for any $R > 0$. The inequality in (2.3.4) provides the flexibility to choose sequences that also satisfy Assumption 2.4.2. The prior $\pi(\eta)$ on the composition structure η should be viewed as a model selection prior, see also Section 10 in [41]. As always, some care is required to avoid that the posterior concentrates on models that are too large and consequently leads to sub-optimal posterior contraction rates. This is achieved by the carefully chosen exponent Ψ_n in (2.3.5), which depends on the sample size and penalizes large composition structures.

2.4 Main results

Denote by $\Pi\big(\cdot|\mathbf{X}, \mathbf{Y}\big)$ the posterior distribution corresponding to a DGP prior Π constructed as above and $(\mathbf{X}, \mathbf{Y}) = (\mathbf{X}_i, Y_i)_i$ a sample from the nonparametric regression model (2.1.1). For normalizing factor $Z_n := \int p_f/p_{f^*}(\mathbf{X}, \mathbf{Y})\, \Pi(df)$ and any Borel measurable $\mathcal{A}$ in the Banach space of continuous functions on $[-1, 1]^d$,

$$\Pi\big(\mathcal{A}|\mathbf{X}, \mathbf{Y}\big) = Z_n^{-1} \int_{\mathcal{A}} \frac{p_f}{p_{f^*}}(\mathbf{X}, \mathbf{Y})\, \Pi(df), \quad \Pi(df) := \int_{\Omega} \Pi(df|\eta)\pi(\eta)\, d\eta, \tag{2.4.1}$$

where $(p_f/p_{f^*})(\mathbf{X}, \mathbf{Y})$ denotes the likelihood ratio. With a slight abuse of notation, for any subset of composition structures $\mathcal{M} \subseteq \Omega$, we set

$$\Pi\big(\eta \in \mathcal{M}|\mathbf{X}, \mathbf{Y}\big) := Z_n^{-1} \int \frac{p_f}{p_{f^*}}(\mathbf{X}, \mathbf{Y}) \int_{\mathcal{M}} \Pi(df|\eta)\pi(\eta)\, d\eta, \tag{2.4.2}$$

which is the contribution of the composition structures $\mathcal{M} \subseteq \Omega$ to the posterior mass.

Before we can state the results, we first need to impose some conditions. The first condition is on the graph prior from Step 2 in the DGP prior construction. It states that all graphs have to be charged with non-negative mass and also requires $\int \sqrt{\gamma(\eta)}\, d\eta$ to be finite. The latter somehow requires that the prior mass decreases quickly enough as the graphs become more complex.

Assumption 2.4.1. *We assume that, for any graph λ, the measure $\gamma(\cdot|\lambda)$ is the uniform distribution on the hypercube of possible smoothness indices $I(\lambda) = [\beta_-, \beta_+]^{q+1}$. Furthermore, we assume that the distribution γ is independent of n, that it assigns positive mass $\gamma(\eta) > 0$ to all composition structures η, and that it satisfies $\int \sqrt{\gamma(\eta)}\, d\eta < +\infty$.*

The second assumption guarantees that the rates $\varepsilon_n(\alpha, \beta, r)$ are not too fast and controls the local changes of the rates under perturbations on the smoothness indices.

Assumption 2.4.2. *We assume the following on the rates appearing in the construction of the prior.*

(i) For any positive integer r, any $\beta > 0$, let $Q_1(\beta, r, K)$ the constant from Lemma 2.A.5. Then, the sequences $\varepsilon_n(\alpha, \beta, r)$ solving the concentration function inequality (2.3.2) *are chosen in such a way that*

$$\varepsilon_n(\alpha, \beta, r) \geq Q_1(\beta, r, K)^{\frac{\beta}{2\beta+r}} n^{-\frac{\beta\alpha}{2\beta\alpha+r}}. \tag{2.4.3}$$

(ii) There exists a constant $Q \geq 1$ such that the following holds. For any $n > 1$, any graph $\lambda = (q, \mathbf{d}, \mathbf{t}, \mathcal{S})$ with $|\mathbf{d}|_1 = 1 + \sum_{i=0}^{q} d_i \leq \log(2\log n)$, and any $\boldsymbol{\beta}' = (\beta_0', \ldots, \beta_q'), \boldsymbol{\beta} = (\beta_0, \ldots, \beta_q) \in I(\lambda)$ satisfying $\beta_i' \leq \beta_i \leq \beta_i' + 1/\log^2 n$ for all $i = 0, \ldots, q$, the rates relative to the composition structures $\eta = (\lambda, \boldsymbol{\beta})$ and $\eta' = (\lambda, \boldsymbol{\beta}')$ satisfy

$$\varepsilon_n(\eta) \leq \varepsilon_n(\eta') \leq Q\varepsilon_n(\eta). \tag{2.4.4}$$

The rate $\varepsilon_n(\eta)$ associated to a composition structure η can be viewed as measure of the complexity of this structure, where larger rates $\varepsilon_n(\eta)$ correspond to more complex models. Our first result states that the posterior concentrates on small models in the sense that all posterior mass is asymptotically allocated on a set

$$\mathcal{M}_n(C) := \big\{\eta : \varepsilon_n(\eta) \leq C\varepsilon_n(\eta^*)\big\} \cap \big\{\eta : |\mathbf{d}|_1 \leq \log(2\log n)\big\} \tag{2.4.5}$$

with sufficiently large constant C. This shows that the posterior not only concentrates on models with fast rates $\varepsilon_n(\eta)$ but also on graph structures with number of nodes in each layer bounded by $\log(2\log n)$. The proof is given in Section 2.A.1.

Theorem 2.4.3 (Model selection). *Let $\Pi\big(\cdot\,|\mathbf{X}, \mathbf{Y}\big)$ be the posterior distribution corresponding to a DGP prior Π constructed as in Section 2.3 and satisfying Assumptions 2.4.1 and 2.4.2. Let $\eta^* = (\lambda^*, \boldsymbol{\beta}^*)$ for some $\boldsymbol{\beta}^* \in [\beta_-, \beta_+]^{q^*+1}$ and suppose $\varepsilon_n(\eta^*) \leq 1/(4Q)$. Then, for a positive constant $C = C(\eta^*)$,*

$$\sup_{f^* \in \mathcal{F}(\eta^*, K)} \mathbb{E}_{f^*}\big[\Pi\big(\eta \notin \mathcal{M}_n(C)\big|\mathbf{X}, \mathbf{Y}\big)\big] \xrightarrow{n\to\infty} 0,$$

where $\mathbb{E}_{f^*}$ denotes the expectation with respect to $\mathbb{P}_{f^*}$, the true distribution of the sample $(\mathbf{X}, \mathbf{Y})$.

Denote by μ the distribution of the covariate vector $\mathbf{X}_1$ and write $L^2(\mu)$ for the weighted L^2-space with respect to the measure μ. The next result shows that the posterior distribution achieves contraction rate $\varepsilon_n(\eta^*)$ up to a $\log n$ factor. The proof is given in Section 2.A.1.

Theorem 2.4.4 (Posterior contraction). *Let $\Pi\big(\cdot\,|\mathbf{X}, \mathbf{Y}\big)$ be the posterior distribution corresponding to a DGP prior Π constructed as in Section 2.3 and satisfying Assumptions 2.4.1* *and 2.4.2. Let $\eta^* = (\lambda^*, \boldsymbol{\beta}^*)$ for some $\boldsymbol{\beta}^* \in [\beta_-, \beta_+]^{q^*+1}$ and suppose $\varepsilon_n(\eta^*) \leq 1/(4Q)$. Then, for a positive constant $L = L(\eta^*)$,*

$$\sup_{f^* \in \mathcal{F}(\eta^*, K)} \mathbb{E}_{f^*} \left[\Pi \left(\|f - f^*\|_{L^2(\mu)} \geq L(\log n)^{1+\log K} \varepsilon_n(\eta^*) \big| \mathbf{X}, \mathbf{Y} \right) \right] \xrightarrow{n\to\infty} 0,$$

where $\mathbb{E}_{f^*}$ denotes the expectation with respect to $\mathbb{P}_{f^*}$, the true distribution of the sample $(\mathbf{X}, \mathbf{Y})$.

Remark 2.4.5. *Our proving strategy allows for the following modification to the construction of the DGP prior. The concentration functions $\varphi^{(\beta,r,K)}$ in* (2.3.1) *are defined globally over the Hölder-ball $\mathcal{C}_r^\beta(K)$. The concentration function inequality in* (2.3.2) *essentially requires that the closure $\overline{\mathbb{H}}^{(\beta,r)}$ of the RKHS of the underlying Gaussian process $\widetilde{G}^{(\beta,r)}$ contains the whole Hölder ball. There are classical examples for which this is too restrictive, and one might want to weaken the construction by considering a subset $\mathcal{H}_r^\beta(K) \subseteq \mathcal{C}_r^\beta(K)$. This can be done by replacing $\mathcal{C}_{t_i}^{\beta_i}(K)$ with the corresponding subset $\mathcal{H}_{t_i}^{\beta_i}(K)$ in the definition of the conditioning sets $\mathcal{D}_i(\eta, K)$ in* (2.3.3) *and the function class $\mathcal{F}(\eta, K)$ in* (2.2.6)*. As a consequence, this also reduces the class of functions for which the posterior contraction rates derived in Theorem 2.4.4 hold.*

We would like to stress, that we do not impose an a-priori known upper bound on the complexity of the underlying composition structure (2.2.1). While we think that this is natural in practice, it causes some extra technical complications. If we additionally assume that the true composition structure satisfies $|\mathbf{d}|_1 \leq D$ for a known upper bound D, then the factor $e^{e^{|\mathbf{d}|_1}}$ in (2.3.5) can be avoided. Moreover, the $(\log n)^{1+\log K}$-factor occurring in the posterior contraction rate is somehow an artifact of the proof, and could be replaced by K^D, see the proof of Lemma 2.A.6 for more details. A trade-off regarding the choice of K appears. To allow for larger classes of functions and a weaker constraint induced by the conditioning on (2.3.3), we want to select a large K. On the contrary, large K results in slower posterior contraction guarantees.

We view the proposed Bayesian analysis rather as a proof of concept than something that is straightforward implementable or computationally efficient. The main obstacles towards a scalable Bayesian method are the combinatorial nature of the set of graphs as well as conditioning the sample paths to neighborhoods of Hölder functions. Regarding the first point, considerable progress has been made recently to construct fast Bayesian methods for model selection priors in high-dimensional settings with theoretical guarantees, see for instance [13, 74]. To avoid a large number of composition graphs, it might be sufficient to restrict to small structures with number of compositions q below five, say, and $|\mathbf{d}|_1$ in the tens. Moreover, in view of the achievable contraction rates, there are plenty of redundant composition structures. Lemma 2.5.4 below shows this in a very specific setting.

Concerning the conditioning of the sample paths in Step 1 of the deep Gaussian process prior construction, it is, in principle, possible to incorporate the conditioning into an accept/reject framework, where we always reject if the generated path is outside the conditioned set. To avoid that the acceptance probability of the algorithm becomes too small, one needs to ensure that a path of the Gaussian process falls into the conditioned set with positive probability that does not vanish as the sample size increases. Lemma 2.6.3 establishes this for a Gaussian wavelet series prior.

2.5 On nearly optimal contraction rates

Theorem 2.5 in [39] ensures existence of a frequentist estimator converging to the true parameter with the posterior contraction rate. This implies that the fastest possible posterior contraction rate is the minimax estimation rate. For the prediction loss $\|f-g\|_{L^2(\mu)}$, the minimax estimation rate over the class $\mathcal{F}(\eta, K)$ is, up to some logarithmic factors,

$$\mathfrak{r}_n(\eta) = \max_{i=0,\ldots,q} n^{-\frac{\beta_i\alpha_i}{2\beta_i\alpha_i+t_i}}, \quad \text{with} \quad \alpha_i := \prod_{\ell=i+1}^{q} \big(\beta_\ell \wedge 1\big), \tag{2.5.1}$$

see [76]. This rate is attained by suitable estimators based on sparsely connected deep neural networks. It is also shown in [76] that wavelet estimators do not achieve this rate and can be sub-optimal by a large polynomial factor in the sample size. Below we derive conditions that are simpler than Assumption 2.4.2 and ensure posterior contraction rate $\mathfrak{r}_n(\eta)$ up to $\log n$-factors. Inspired by the negative result for wavelet estimators, we conjecture moreover that there are composition structures such that any Gaussian process prior will lead to a posterior contraction rate that is suboptimal by a polynomial factor compared with $\mathfrak{r}_n(\eta)$.

The first result shows that the solution to the concentration function inequality for arbitrary $0<\alpha\leq 1$ can be deduced from the solution for $\alpha=1$. The proof is in Section 2.A.2.

Lemma 2.5.1. *Let $\varepsilon_n(1,\beta,r)$ be a solution to the concentration function inequality* (2.3.2) *for $\alpha=1$. Then, any sequence $\varepsilon_n(\alpha,\beta,r)\geq \varepsilon_{m_n}(1,\beta,r)^\alpha$ where m_n is chosen such that $m_n\varepsilon_{m_n}(1,\beta,r)^{2-2\alpha}\leq n$, solves the concentration function inequality for arbitrary $\alpha\in(0,1]$.*

The following result makes the construction in Lemma 2.5.1 explicit, when the solution $\varepsilon_n(1,\beta,r)$ for $\alpha=1$ is at most by a $\log n$ factor larger than $n^{-\beta/(2\beta+r)}$. The proof is given in Section 2.A.2.

Lemma 2.5.2. *Let $n\geq 3$.*

(i) If the sequence $\varepsilon_n(1,\beta,r)=C_1(\log n)^{C_2}n^{-\beta/(2\beta+r)}$ solves the concentration function inequality (2.3.2) *for $\alpha=1$, with constants $C_1\geq 1$ and $C_2\geq 0$, then, any sequence*

$$\varepsilon_n(\alpha,\beta,r)\geq C_1^2(2\beta+1)^{2C_2}(\log n)^{C_2(2\beta+2)}\, n^{-\frac{\beta\alpha}{2\beta\alpha+r}}$$

solves the concentration function inequality for arbitrary $\alpha\in(0,1]$.

(ii) If there are constant $C_1'\geq 1$ and $C_2'\geq 0$ such that the concentration function satisfies

$$\varphi^{(\beta,r,K)}(\delta)\leq C_1'(\log\delta^{-1})^{C_2'}\delta^{-\frac{r}{\beta}}, \quad \textit{for all } 0<\delta\leq 1, \tag{2.5.2}$$

then, any sequence

$$\varepsilon_n(\alpha,\beta,r)\geq C_1'(\log n)^{C_2'}\, n^{-\frac{\beta\alpha}{2\beta\alpha+r}}$$

solves the concentration function inequality (2.3.2) *for arbitrary $\alpha\in(0,1]$.*

To verify Assumption 2.4.2 (ii), we need to pick suitable sequences $\varepsilon_n(\eta) \geq \max_{i=0,\ldots,q} \varepsilon_n(\alpha_i, \beta_i, t_i)$. If $\varepsilon_n(\alpha, \beta, r) = C_1(\beta, r)(\log n)^{C_2(\beta,r)} n^{-\beta\alpha/(2\beta\alpha+r)}$ for some constants $C_1(\beta, r) \geq 1$ and $C_2(\beta, r) \geq 0$, then, a suitable choice is

$$\varepsilon_n(\eta) = \widetilde{C}_1(\eta)(\log n)^{\widetilde{C}_2(\eta)} \mathfrak{r}_n(\eta), \tag{2.5.3}$$

provided the constants $\widetilde{C}_j(\eta) := \max_{i=0,\ldots,q} \sup_{\beta\in[\beta_-,\beta_+]} C_j(\beta, t_i)$, $j \in \{1, 2\}$ are finite. In the subsequent examples, this will be checked by verifying that $\beta \mapsto C_j(\beta, r)$, $j \in \{1, 2\}$ are bounded functions on $[\beta_-, \beta_+]$.

Lemma 2.5.3. *The rates* $\varepsilon_n(\eta)$ *in* (2.5.3) *satisfy condition (ii) in Assumption 2.4.2 with* $Q = e^{\beta_+}$.

Let Π be a DGP prior constructed with the Gaussian processes and rates given in this section. Then, the corresponding posterior satisfies Theorem 2.4.4 and contracts with rate $\mathfrak{r}_n(\eta^*)$ up to the multiplicative factor $L(\eta^*)\widetilde{C}_1(\lambda^*, K)(\log n)^{1+\log K+\widetilde{C}_2(\lambda^*,K)}$.

To complete this section, the next lemma shows that there are redundant composition structures. More precisely, conditions are provided such that the number of compositions q can be reduced by one, while still achieving nearly optimal posterior contraction rates.

Lemma 2.5.4. *Suppose that* f *is a function with composition structure* $\eta = (q, \mathbf{d}, \mathbf{t}, \mathcal{S}, \boldsymbol{\beta})$ *and assume that* $\beta_+ = K = 1$. *If there exists an index* $j \in \{1, \ldots, q\}$ *with* $t_j = t_{j-1} = 1$, *then,* f *can also be written as a function with composition structure* $\eta' := (q-1, \mathbf{d}_{-j}, \mathbf{t}_{-j}, \mathcal{S}_{-j}, \boldsymbol{\beta}')$, *where* $\mathbf{d}_{-j}, \mathbf{t}_{-j}, \mathcal{S}_{-j}$ *denote* $\mathbf{d}, \mathbf{t}, \mathcal{S}$ *with entries* $d_j, t_j, \mathcal{S}_j$ *removed, respectively, and* $\boldsymbol{\beta}' := (\beta_0, \ldots, \beta_{j-2}, \beta_{j-1}\beta_j, \beta_{j+1}, \ldots, \beta_q)$. *Moreover the induced posterior contraction rates agree, that is,* $\mathfrak{r}_n(\eta) = \mathfrak{r}_n(\eta')$.

A more complete notion of equivalence classes on composition graphs that maintain the optimal posterior contraction rates is beyond the scope of this work.

2.6 Examples of DGP priors

The construction of the deep Gaussian process prior requires the choice of a family of Gaussian processes $\{\widetilde{G}^{(\beta,r)} : \beta \in [\beta_-, \beta_+]\}$. In this section we show that standard families appearing in the Gaussian process prior literature achieve near optimal posterior contraction rates. To show this, we rely on the conditions derived in the previous section.

2.6.1 Lévy's fractional Brownian motion

Assume that the upper bound β_+ on the possible range of smoothness indices is bounded by one. A zero-mean Gaussian process X^β is called a Lévy fractional Brownian motion of order $\beta \in (0, 1)$ if

$$X^\beta(0) = 0, \quad \mathbb{E}\left[|X^\beta(\mathbf{u}) - X^\beta(\mathbf{u}')|^2\right] = |\mathbf{u} - \mathbf{u}'|_2^{2\beta}, \quad \forall \mathbf{u}, \mathbf{u}' \in [-1, 1]^r.$$

The covariance function of the process is $\mathbb{E}[X^\beta(\mathbf{u})X^\beta(\mathbf{u}')] = \frac{1}{2}(|\mathbf{u}|_2^{2\beta} + |\mathbf{u}'|_2^{2\beta} - |\mathbf{u}-\mathbf{u}'|_2^{2\beta})$. Chapter 3 in [27] provides the following representation for a r-dimensional β-fractional Brownian motion. Denote by $\widehat{f}(\boldsymbol{\xi}) := (2\pi)^{-r/2}\int_{\mathbb{R}^r} e^{i\mathbf{u}^\top\boldsymbol{\xi}} f(\mathbf{u})d\mathbf{u}$ the Fourier transform of the function f. For $W = (W(\mathbf{u}))_{\mathbf{u}\in[-1,1]^r}$ a multidimensional Brownian motion, and C_β a positive constant depending only on β, r,

$$X^\beta(\mathbf{u}) = \int_{\mathbb{R}^r} \frac{e^{-i\mathbf{u}^\top\boldsymbol{\xi}}-1}{C_\beta^{1/2}|\boldsymbol{\xi}|_2^{\beta+r/2}}\widehat{W}(d\boldsymbol{\xi}),$$

in distribution, where $\widehat{W}(d\boldsymbol{\xi})$ is the Fourier transform of the Brownian random measure $W(d\mathbf{u})$, see Section 2.1.6 in [27] for definitions and properties. The same reference defines, for all $\varphi \in L^2([-1,1]^r)$, the integral operator

$$(I^\beta\varphi)(\mathbf{u}) := \int_{\mathbb{R}^r} \overline{\widehat{\varphi}(\boldsymbol{\xi})}\frac{e^{-i\mathbf{u}^\top\boldsymbol{\xi}}-1}{C_\beta^{1/2}|\boldsymbol{\xi}|_2^{\beta+r/2}}\frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}}.$$

As a corollary, the RKHS $\mathbb{H}^\beta$ of X^β is given in Section 3.3 in [27] as

$$\mathbb{H}^\beta = \Big\{I^\beta\varphi : \varphi \in L^2([-1,1]^r)\Big\}, \quad \langle I^\beta\varphi, I^\beta\varphi'\rangle_{\mathbb{H}^\beta} = \langle\varphi,\varphi'\rangle_{L^2([-1,1]^r)}.$$

Since the process X^β is always zero at $\mathbf{u}=0$, we release it at zero. That is, let $Z \sim \mathcal{N}(0,1)$ independent of X^β and consider the process $\mathbf{u} \mapsto Z + X^\beta(\mathbf{u})$. The RKHS of the constant process $\mathbf{u} \mapsto Z$ is the set $\mathbb{H}^Z$ of all constant functions and, by Lemma I.18 in [41], the RKHS of $Z + X^\beta$ is the direct sum $\mathbb{H}^Z \oplus \mathbb{H}^\beta$.

The next result is proved in Section 2.A.3 and can be viewed as the multidimensional extension of the RKHS bounds in Theorem 4 in [17]. Whereas the original proof relies on kernel smoothing and Taylor approximations, we use a spectral approach. Write

$$\mathcal{W}_r^\beta(K) := \Big\{h : [-1,1]^r \to [-1,1] : \int_{\mathbb{R}^r} |\widehat{h}(\boldsymbol{\xi})|^2(1+|\boldsymbol{\xi}|_2)^{2\beta}\frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}} \le K\Big\}, \qquad (2.6.1)$$

for the β-Sobolev ball of radius K.

Lemma 2.6.1. *Let $\beta \in [\beta_-, \beta_+]$ and $Z+X^\beta = (Z+X^\beta(\mathbf{u}))_{\mathbf{u}\in[-1,1]^r}$ the fractional Brownian motion of order β released at zero. Fix $h \in \mathcal{C}_r^\beta(K) \cap \mathcal{W}_r^\beta(K)$. Set $\phi_\sigma = \sigma^{-r}\phi(\cdot/\sigma)$ with ϕ a suitable regular kernel and $\sigma < 1$. Then, $\|h * \phi_\sigma - h\|_\infty \le KR_\beta\sigma^\beta$ and $\|h*\phi_\sigma\|_{\mathbb{H}^Z\oplus\mathbb{H}^\beta}^2 \le K^2L_\beta^2\sigma^{-r}$ for some constants R_β, L_β that depend only on β, r.*

The next lemma shows that for Lévy's fractional Brownian motion released at zero near optimal posterior contraction rates can be obtained. For that we need to restrict the definition of the global concentration function to the smaller class $\mathcal{C}_r^\beta(K) \cap \mathcal{W}_r^\beta(K)$. The proof of the lemma is in Section 2.A.3.

Lemma 2.6.2. *Let $\beta_+ \le 1$ and work on the reduced function spaces $\mathcal{H}_r^\beta(K) = \mathcal{C}_r^\beta(K) \cap \mathcal{W}_r^\beta(K)$ as outlined in Remark 2.4.5. For $\{\widetilde{G}^{(\beta,r)} : \beta \in [\beta_-,\beta_+]\}$ the family of Levy's fractional Brownian motions $Z + X^\beta$ released at zero, there exist sequences $\varepsilon_n(\eta) = C_1(\eta)(\log n)^{C_2(\eta)}\mathfrak{r}_n(\eta)$ such that Assumption 2.4.2 holds.*

2.6.2 Truncated wavelet series

Let $\{\psi_{j,k} : j \in \mathbb{N}_+,\ k = 1, \dots 2^{jr}\}$ be an orthonormal wavelet basis of $L^2([-1,1]^r)$. For any $\varphi \in L^2([-1,1]^r)$, we denote by $\varphi = \sum_{j=1}^{\infty} \sum_{k=1}^{2^{jr}} \lambda_{j,k}(\varphi)\psi_{j,k}$ its wavelet expansion. The quantities $\lambda_{j,k}(\varphi)$ are the corresponding real coefficients. For any $\beta > 0$, we denote by $\mathcal{B}_{\infty,\infty,\beta}$ the Besov space of functions φ with finite

$$\|\varphi\|_{\infty,\infty,\beta} := \sup_{j\in\mathbb{N}} 2^{j(\beta+\frac{r}{2})} \max_{k=1,\dots 2^{jr}} |\lambda_{j,k}(\varphi)|.$$

We assume that the wavelet basis is s-regular with $s > \beta_+$. For i.i.d. random variables $Z_{j,k} \sim \mathcal{N}(0,1)$, consider the Gaussian process induced by the truncated series expansion

$$X^\beta(\mathbf{u}) := \sum_{j=1}^{J_\beta} \sum_{k=1}^{2^{jr}} \frac{2^{-j(\beta+\frac{r}{2})}}{\sqrt{jr}} Z_{j,k}\psi_{j,k}(\mathbf{u}),$$

where the maximal resolution J_β is chosen as the integer closest to the solution J of the equation $2^J = n^{1/(2\beta+r)}$, see Section 4.5 in [82]. The RKHS of the process X^β is given in the proof of Theorem 4.5 in [82] as the set $\mathbb{H}^\beta$ of functions $\varphi = \sum_{j=1}^{J_\beta} \sum_{k=1}^{2^{jr}} \lambda_{j,k}(\varphi)\psi_{j,k}$ with coefficients $\lambda_{j,k}(\varphi)$ satisfying $\|\varphi\|_{\mathbb{H}^\beta}^2 := \sum_{j=1}^{J_\beta} \sum_{k=1}^{2^{jr}} jr2^{2j(\beta+r/2)}\lambda_{j,k}(\varphi)^2 < \infty$.

For this family of Gaussian processes, it is rather straightforward to verify that conditioning on a neighbourhood of β-smooth functions as in Step 1 of the deep Gaussian process prior construction is not a restrictive constraint. The next result shows that, with high probability, the process X^β belong to the $\mathcal{B}_{\infty,\infty,\beta}$-ball of radius $(1+K')\sqrt{2\log 2}$, with $K' > \sqrt{3}$. Since the Besov space $\mathcal{B}_{\infty,\infty,\beta}$ contains the Hölder space $\mathcal{C}_r^\beta$ for any $\beta > 0$, the process belongs to the Hölder-ball $\mathcal{C}_r^\beta(K)$ for some suitable K' only depending on K.

Lemma 2.6.3. *Let X^β be the truncated wavelet process. Then, for any $K' > \sqrt{3}$,*

$$\mathbb{P}\Big(\|X^\beta\|_{\infty,\infty,\beta} \leq (1+K')\sqrt{2\log 2}\Big) \geq 1 - \frac{4}{2^{rK'^2} - 4}.$$

The proof is postponed to Section 2.A.3. The probability in the latter display converges quickly to one. As an example consider $K' = 2$. Since $r \geq 1$, the bound implies that more than 2/3 of the simulated sample paths $\mathbf{u} \mapsto X^\beta(\mathbf{u})$ lie in the Hölder ball $\mathcal{B}_{\infty,\infty,\beta}(3\sqrt{2\log 2})$.

The next lemma shows that for the truncated series expansion, near optimal posterior contraction rates can be achieved. The proof of the lemma is deferred to Section 2.A.3.

Lemma 2.6.4. *For $\{\widetilde{G}^{(\beta,r)} : \beta \in [\beta_-, \beta_+]\}$ the family of truncated Gaussian processes X^β there exist sequences $\varepsilon_n(\eta) = C_1(\eta)(\log n)^{C_2(\eta)}\mathfrak{r}_n(\eta)$ such that Assumption 2.4.2 holds.*

2.6.3 Stationary process

A zero-mean Gaussian process $X^\nu = (X^\nu(\mathbf{u}))_{\mathbf{u}\in[-1,1]^r}$ is called stationary if its covariance function can be represented by a spectral density measure ν on $\mathbb{R}^r$ as

$$\mathbb{E}[X^\nu(\mathbf{u})X^\nu(\mathbf{u}')] = \int_{\mathbb{R}^r} e^{-i(\mathbf{u}-\mathbf{u}')^\top \boldsymbol{\xi}}\nu(\boldsymbol{\xi})d\boldsymbol{\xi},$$

see Example 11.8 in [41]. We consider stationary Gaussian processes with radially decreasing spectral measures that have exponential moments, that is, $\int e^{c|\boldsymbol{\xi}|_2}\nu(\boldsymbol{\xi})d\boldsymbol{\xi} < +\infty$ for some $c > 0$. Such processes have smooth sample paths thanks to Proposition I.4 in [41]. An example is the square-exponential process with spectral measure $\nu(\boldsymbol{\xi}) = 2^{-r}\pi^{-r/2}e^{-|\boldsymbol{\xi}|_2^2/4}$. For any $\varphi \in L^2(\nu)$, set $(H^\nu\varphi)(\mathbf{u}) := \int_{\mathbb{R}^r} e^{i\boldsymbol{\xi}^\top \mathbf{u}}\varphi(\boldsymbol{\xi})\nu(\boldsymbol{\xi})d\boldsymbol{\xi}$. The RKHS of X^ν is given in Lemma 11.35 in [41] as $\mathbb{H}^\nu = \{H^\nu\varphi : \varphi \in L^2(\nu)\}$ with inner product $\langle H^\nu\varphi, H^\nu\varphi'\rangle_{\mathbb{H}^\nu} = \langle \varphi, \varphi'\rangle_{L^2(\nu)}$.

For every $\beta \in [\beta_-, \beta_+]$, take $\widetilde{G}^{(\beta,r)}$ to be the rescaled process $X^\nu(a\cdot) = (X^\nu(a\mathbf{u}))_{\mathbf{u}\in[-1,1]^r}$ with scaling

$$a = a(\beta, r) = n^{\frac{1}{2\beta+r}}(\log n)^{-\frac{1+r}{2\beta+r}}. \tag{2.6.2}$$

The process $\widetilde{G}^{(\beta,r)}$ thus depends on n. We prove the next result in Section 2.A.3.

Lemma 2.6.5. *For $\{\widetilde{G}^{(\beta,r)} : \beta \in [\beta_-, \beta_+]\}$ the family of rescaled stationary processes $X^\nu(a\cdot)$, there exist sequences $\varepsilon_n(\eta) = C_1(\eta)(\log n)^{C_2(\eta)}\mathfrak{r}_n(\eta)$ such that Assumption 2.4.2 holds.*

2.7 DGP priors, wide neural networks and regularization

In this section, we explore similarities and differences between deep learning and the Bayesian analysis based on (deep) Gaussian process priors. Both methods are based on the likelihood. It is moreover known that standard random initialization schemes in deep learning converge to Gaussian processes in the wide limit. Since the initialization is crucial for the success of deep learning, this suggests that the initialization could act in a similar way as a Gaussian prior in the Bayesian world. Next to a proper initialization scheme, stability enhancing regularization techniques such as batch normalization are widely studied in deep learning and a comparison might help us to identify conditions that constraint the potentially wild behavior of deep Gaussian process priors. Below we investigate these aspects in more detail.

It has been argued in the literature that Bayesian neural networks and regression with Gaussian process priors are intimately connected. In Bayesian neural networks, we generate a function valued prior distribution by using a neural network and drawing the network weights randomly. Recall that a neural network with a single hidden layer is called shallow, and a neural network with a large number of units in all hidden layers is called wide. If the network weights in a shallow and wide neural network are drawn i.i.d., and the scaling of the variances is such that the prior does not become degenerate, then, it has been argued in [68] that the prior will converge in the wide limit to a Gaussian process prior and expressions for the covariance structure of the limiting process are known. One might be tempted to believe that for a deep neural network one should obtain a deep Gaussian process as a limit distribution. If the width of all hidden layers tends simultaneously to infinity, [63] proves that this is not true and that one still obtains a Gaussian limit. The covariance of the limiting process is, however, more complicated and can be given via a recursion formula, where each step in the recursion describes the change of the covariance by a hidden layer. [63] shows moreover in a simulation study that Bayesian neural networks and Gaussian process priors with appropriate choice of the covariance structure behave indeed similarly.

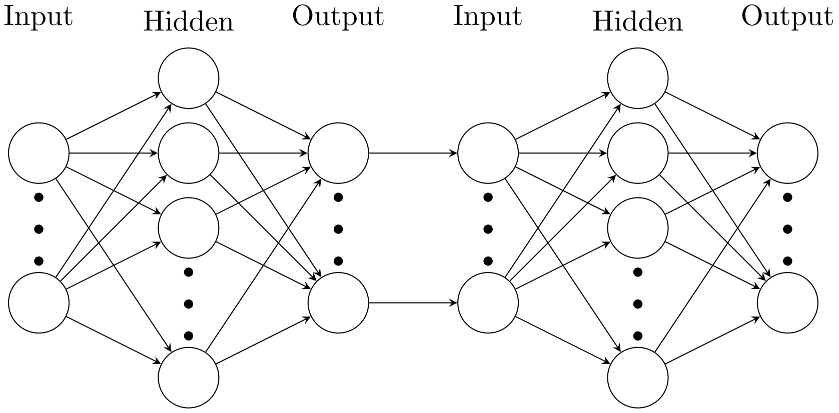

Figure 2.3: Schematic stacking of two shallow neural networks.

It is conceivable that if one keeps the width of some hidden layers fixed and let the width of all other hidden layers tend to infinity, the Bayesian neural network prior will converge, if all variances are properly scaled, to a deep Gaussian process. By stacking for instance two shallow networks as indicated in Figure 2.3 and making the first and last hidden layer wide, the limit is the composition of two Gaussian process and thus, a deep Gaussian process. In the hierarchical deep Gaussian prior construction in Section 2.3, we pick in a first step a prior on composition structures. For Bayesian neural networks this is comparable with selecting first a hyperprior on neural network architectures.

Even more recently, [71, 46] studied the behaviour of neural networks with random weights when both depth and width tend to infinity.

While the discussion so far indicates that Bayesian neural networks and Bayes with (deep) Gaussian process priors are similar methods, the question remains whether deep learning with randomly initialized network weights behaves similarly as a Bayes estimator with respect to a (deep) Gaussian process prior. The random network initialization means that the deep learning algorithm is initialized approximately by a (deep) Gaussian process. Since it is well-known that the initialization is crucial for the success of deep learning, this suggests that the initialization indeed acts as a prior. Denote by $-\ell$ the negative log-likelihood/cross entropy. Whereas in deep learning we fit a function by iteratively decreasing the cross-entropy using gradient descent method, the posterior is proportional to $\exp(\ell)\times$prior and concentrates on elements in the support of the prior with small cross entropy. Gibbs sampling of the posterior has moreover a similar flavor as coordinate-wise descent methods for the cross-entropy. The only theoretical result that we are aware of examining the relationship between deep learning and Bayesian neural networks is [72]. It proves that for a neural network prior with network weights drawn i.i.d. from a suitable spike-and-slab prior, the full posterior behaves similarly as the global minimum of the cross-entropy (that is, the empirical risk minimizer) based on sparsely connected deep neural networks.

As a last point we now compare stabilization techniques for deep Gaussian process priors and deep learning. In Step 1 of the deep Gaussian process prior construction, we have conditioned the individual Gaussian processes to map to $[-1,1]$ and to generate sample paths in small neighborhoods of a suitable Hölder ball. This induces regularity in the prior and avoids the wild behaviour of the composed sample paths due to bad realizations of individual components. We argue that this form of regularization has a similar flavor as batch normalization, cf. Section 8.7 in [44]. The purpose of batch normalization is to avoid vanishing or exploding gradients due to the composition of several functions in deep neural networks. The main idea underlying batch normalization is to normalize the outputs from a fixed hidden layer in the neural network before they become the input of the next hidden layer. The normalization step is now different than the conditioning proposed for the compositions of Gaussian processes. In fact, for batch normalization, the mean and the variance of the outputs are estimated based on a subsample and an affine transformation is applied such that the outputs are approximately centered and have variance one. One of the key differences is that this normalization invokes the distribution of the underlying design, while the conditioning proposed for deep Gaussian processes is independent of the distribution of the covariates. One suggestion that we can draw from this comparison is that instead of conditioning the processes to have sample paths in $[-1,1]$, it might also be interesting to apply the normalization $f \mapsto f(t)/\sup_{t\in[-1,1]^r}|f(t)|$ between any two compositions. This also ensures that the output maps to $[-1,1]$ and is closer to batch normalization. A data-dependent normalization of the prior cannot be incorporated in the fully Bayesian framework considered here and would result in an empirical Bayes method.

Appendix 2.A Proofs

2.A.1 Proofs for Section 2.4

Information geometry in the nonparametric regression model. The following results are fairly standard in the nonparametric Bayes literature. As we are aiming for a self-contained presentation of the material, these facts are reproduced here. Let P_f be the law of *one* observation $(\mathbf{X}_i, Y_i)$. The Kullback-Leibler divergence in the nonparametric regression model is

$$\mathrm{KL}\left(P_f, P_g\right) = \int (f(\mathbf{x}) - g(\mathbf{x}))^2 d\mu(\mathbf{x}) \leq \|f-g\|_{L^\infty([-1,1]^d)}$$

with μ the distribution of the covariates $\mathbf{X}_1$. Using that $\mathbb{E}_f[\log dP_f/dP_g] = \mathrm{KL}(P_f, P_g)$, $\mathrm{Var}(Z) \leq \mathbb{E}[Z^2]$, $\mathbb{E}_f[Y|\mathbf{X}] = f(\mathbf{X})$ and $\mathbb{E}_f[Y^2|\mathbf{X}] = 1$, we also have that

$$V_2(P_f, P_g) := \mathbb{E}_f\Big[\Big|\log\frac{dP_f}{dP_g} - \mathrm{KL}(P_f, P_g)\Big|^2\Big]$$

$$\begin{aligned}
&\leq \mathbb{E}_f\Big[\Big|\log \frac{dP_f}{dP_g}\Big|^2\Big] \\
&= \mathbb{E}_f\Big[\Big(Y\big(f(\mathbf{X}) - g(\mathbf{X})\big) - \frac{1}{2}f(\mathbf{X})^2 + \frac{1}{2}g(\mathbf{X})^2\Big)^2\Big] \\
&= \int \big(f(\mathbf{x}) - g(\mathbf{x})\big)^2 + \frac{1}{4}\big(f(\mathbf{x}) - g(\mathbf{x})\big)^4 \, d\mu(\mathbf{x}).
\end{aligned}$$

In particular, for $\varepsilon \leq 1$, $\|f - g\|_\infty \leq \varepsilon/2$ implies that $V_2(P_f, P_g) \leq \varepsilon^2$ and therefore

$$B_2(P_f, \varepsilon) = \Big\{g : \mathrm{KL}(P_f, P_g) < \varepsilon^2, V_2(P_f, P_g) < \varepsilon^2\Big\} \supseteq \Big\{g : \|f - g\|_\infty \leq \frac{\varepsilon}{2}\Big\}. \tag{2.A.1}$$

We derive posterior contraction rates for the Hellinger distance. This can then be related to the $\|\cdot\|_{L^2(\mu)}$-norm as explained below. Using the moment generating function of a standard normal distribution, the Hellinger distance for one observation $(\mathbf{X}, Y)$ becomes

$$\begin{aligned}
d_H(P_f, P_g) = 1 - \int \sqrt{dP_f dP_g} &= 1 - \int \sqrt{dP_f/dP_g}\, dP_g \\
&= 1 - \mathbb{E}_g\Big[e^{\frac{1}{4}(Y - g(\mathbf{X}))^2 - \frac{1}{4}(Y - f(\mathbf{X}))^2}\Big] \\
&= 1 - \mathbb{E}\Big[\mathbb{E}_g\Big[e^{\frac{1}{2}(Y - g(\mathbf{X}))(f(\mathbf{X}) - g(\mathbf{X}))}\Big|\mathbf{X}\Big] e^{-\frac{1}{4}(f(\mathbf{X}) - g(\mathbf{X}))^2}\Big] \\
&= 1 - \mathbb{E}\Big[e^{-\frac{1}{8}(f(\mathbf{X}) - g(\mathbf{X}))^2}\Big] \\
&= 1 - \int e^{-\frac{1}{8}(f(\mathbf{x}) - g(\mathbf{x}))^2}\, d\mu(\mathbf{x}).
\end{aligned}$$

Since $1 - e^{-x} \leq x$ and μ is a probability measure, we have that $d_H(P_f, P_g) \leq \frac{1}{8}\int (f(\mathbf{x}) - g(\mathbf{x}))^2\, d\mu(\mathbf{x})$. Due to $1 - e^{-x} \geq e^{-x}x$, we also find

$$d_H(P_f, P_g) \geq \frac{e^{-Q^2/2}}{8} \int \big(f(\mathbf{x}) - g(\mathbf{x})\big)^2\, d\mu(\mathbf{x}), \quad \text{for all } f, g, \text{ with } \|f\|_\infty, \|g\|_\infty \leq Q. \tag{2.A.2}$$

By Proposition D.8 in [41], for any f, g, there exists a test such that $\mathbb{E}_f \phi \leq \exp(-\frac{n}{8} d_H(P_f, P_g)^2)$ and $\sup_{h: d_H(P_h, P_g) < d_H(P_f, P_g)/2} \mathbb{E}_h[1 - \phi] \leq \exp(-\frac{n}{8} d_H(P_f, P_g)^2)$. This means that for the Hellinger distance, the test condition in (8.2) in [41] holds for $\xi = 1/2$.

Function spaces. The next result shows that the Hölder-balls defined in this paper are nested.

Lemma 2.A.1. *If $0 < \beta' \leq \beta$, then, for any positive integer r and any $K > 0$, we have $\mathcal{C}_r^\beta(K) \subseteq \mathcal{C}_r^{\beta'}(K)$.*

Proof of Lemma 2.A.1. If $\lfloor \beta' \rfloor = \lfloor \beta \rfloor$ the embedding follows from the definition of the Hölder-ball and the fact that $\sup_{\mathbf{x}, \mathbf{y} \in [-1,1]^r} |\mathbf{x} - \mathbf{y}|_\infty^{\beta - \beta'} = 2^{\beta - \beta'}$. If $\lfloor \beta' \rfloor < \lfloor \beta \rfloor$, it remains to prove $\mathcal{C}^\beta(K) \subseteq \mathcal{C}^{\lfloor \beta' \rfloor + 1}(K)$. This follows from first order Taylor expansion,

$$\begin{aligned}
2 \sum_{\boldsymbol{\alpha}: |\boldsymbol{\alpha}| = \lfloor \beta' \rfloor} \sup_{\substack{\mathbf{x}, \mathbf{y} \in [-1,1]^r \\ \mathbf{x} \neq \mathbf{y}}} \frac{|\partial^{\boldsymbol{\alpha}} f(\mathbf{x}) - \partial^{\boldsymbol{\alpha}} f(\mathbf{y})|}{|\mathbf{x} - \mathbf{y}|_\infty} &\leq 2 \sum_{\boldsymbol{\alpha}: |\boldsymbol{\alpha}| = \lfloor \beta' \rfloor} \big\| |\nabla(\partial^{\boldsymbol{\alpha}} f)|_1 \big\|_\infty \\
&\leq 2r \sum_{\boldsymbol{\alpha}: |\boldsymbol{\alpha}| = \lfloor \beta' \rfloor + 1} \big\| \partial^{\boldsymbol{\alpha}} f \big\|_\infty,
\end{aligned}$$

and the definition of the Hölder-ball in (2.2.3). □

The following is a slight variation of Lemma 3 in [76].

Lemma 2.A.2. *Let* $h_{ij} : [-1,1]^{t_i} \to [-1,1]$ *be as in* (2.2.1)*. Assume that, for some* $K \geq 1$ *and* $\eta_i \geq 0$, $|h_{ij}(\mathbf{x}) - h_{ij}(\mathbf{y})|_\infty \leq \eta_i + K|\mathbf{x} - \mathbf{y}|_\infty^{\beta_i \wedge 1}$ *for all* $\mathbf{x}, \mathbf{y} \in [-1,1]^{t_i}$*. Then, for any functions* $\widetilde{h}_i = (\widetilde{h}_{ij})_j^\top$ *with* $\widetilde{h}_{ij} : [-1,1]^{t_i} \to [-1,1]$,

$$\big\|h_q \circ \ldots \circ h_0 - \widetilde{h}_q \circ \ldots \circ \widetilde{h}_0\big\|_{L^\infty[-1,1]^d} \leq K^q \sum_{i=0}^{q} \eta_i^{\alpha_i} + \big\||h_i - \widetilde{h}_i|_\infty\big\|_\infty^{\alpha_i}.$$

with $\alpha_i = \prod_{\ell=i+1}^{q} \beta_\ell \wedge 1$.

Proof of Lemma 2.A.2. We prove the assertion by induction over q. For $q = 0$, the result is trivially true. Assume now that the statement is true for a positive integer k. To show that the assertion also holds for $k+1$, define $H_k = h_k \circ \ldots \circ h_0$ and $\widetilde{H}_k = \widetilde{h}_k \circ \ldots \circ \widetilde{h}_0$. By triangle inequality,

$$\begin{aligned}
&\big|h_{k+1} \circ H_k(\mathbf{x}) - \widetilde{h}_{k+1} \circ \widetilde{H}_k(\mathbf{x})\big|_\infty \\
&\leq |h_{k+1} \circ H_k(\mathbf{x}) - h_{k+1} \circ \widetilde{H}_k(\mathbf{x})\big|_\infty + |h_{k+1} \circ \widetilde{H}_k(\mathbf{x}) - \widetilde{h}_{k+1} \circ \widetilde{H}_k(\mathbf{x})\big|_\infty \\
&\leq \eta_{k+1} + K\big|H_k(\mathbf{x}) - \widetilde{H}_k(\mathbf{x})\big|_\infty^{\beta_{k+1} \wedge 1} + \||h_{k+1} - \widetilde{h}_{k+1}|_\infty\|_\infty.
\end{aligned}$$

Together with the induction hypothesis and the inequality $(y+z)^\alpha \leq y^\alpha + z^\alpha$ which holds for all $y, z \geq 0$ and all $\alpha \in [0,1]$, the induction step follows. □

The next result is a corollary of Theorem 8.9 in [41].

Lemma 2.A.3. *Denote the data by* $\mathcal{D}_n$ *and the (generic) posterior by* $\Pi(\cdot|\mathcal{D}_n)$*. Let* $(A_n)_n$ *be a sequence of events and* $B_2(P_{f^*}, \varepsilon)$ *as in* (2.A.1)*. Assume that*

$$e^{2na_n^2} \frac{\Pi(A_n)}{\Pi(B_2(P_{f^*}, a_n))} \xrightarrow{n \to \infty} 0, \tag{2.A.3}$$

for some positive sequence $(a_n)_n$*. Then,*

$$\mathbb{E}_{f^*}\big[\Pi(A_n|\mathcal{D}_n)\big] \xrightarrow{n \to \infty} 0,$$

where $\mathbb{E}_{f^*}$ *is the expectation with respect to* P_{f^*}.

We now can prove Theorem 2.4.3.

Proof of Theorem 2.4.3. By definition (2.4.2), the quantity $\Pi\big(\eta \notin \mathcal{M}_n(C)|\mathbf{X}, \mathbf{Y}\big)$ denotes the posterior mass of the functions whose models are in the complement of $\mathcal{M}_n(C)$. In view of Lemma 2.A.3, it is sufficient to show condition (2.A.3) for $A_n = \{\eta \notin \mathcal{M}_n(C)\} =: \mathcal{M}_n^c(C)$ and a_n proportional to $\varepsilon_n(\eta^*)$. We now prove that

$$e^{2na_n^2} \frac{\int_{\mathcal{M}_n^c(C)} \pi(\eta)\, d\eta}{\Pi(B_2(P_{f^*}, a_n))} \to 0, \tag{2.A.4}$$

for $a_n = 4K^{q^*}(q^*+1)Q\varepsilon_n(\eta^*)$ and Π the deep Gaussian process prior. The next result deals with the lower bound on the denominator. For any hypercube I, we introduce the notation $\operatorname{diam}(I) := \sup_{\boldsymbol{\beta}, \boldsymbol{\beta}' \in I} |\boldsymbol{\beta} - \boldsymbol{\beta}'|_\infty$.

Lemma 2.A.4. *Let Π be a DGP prior that satisfies the assumptions of Theorem 2.4.3. With Q the universal constant from Assumption 2.4.2, $R^* := 4K^{q^*}(q^*+1)Q$ and sufficiently large sample size n, we have*

$$\Pi\Big(B_2\big(P_{f^*}, 2R^*\varepsilon_n(\eta^*)\big)\Big) \geq e^{-|\mathbf{d}^*|_1 Q^2 n\varepsilon_n(\eta^*)^2}\pi(\lambda^*, I_n^*),$$

where $I_n^* \subset [\beta_-, \beta_+]^{q^*+1}$ is a hypercube containing $\boldsymbol{\beta}^*$ with $\operatorname{diam}(I_n^*) = 1/\log^2 n$.

Proof of Lemma 2.A.4. By construction (2.A.1), the set $B_2(P_{f^*}, 2R^*\varepsilon_n(\eta^*))$ is a superset of $\{g : \|f^* - g\|_\infty \leq R^*\varepsilon_n(\eta^*)\}$, so that

$$\Pi\Big(B_2\big(P_{f^*}, 2R^*\varepsilon_n(\eta^*)\big)\Big) \geq \Pi\Big(f^* + \mathbb{B}_\infty\left(R^*\varepsilon_n(\eta^*)\right)\Big).$$

We then localize the probability in the latter display in a neighborhood around the true $\boldsymbol{\beta}^* = (\beta_0^*, \ldots, \beta_q^*)$. More precisely, let $I_n^* := \{\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{q^*}) : \beta_i \in [\beta_i^* - b_n, \beta_i^*], \forall i\}$ with $b_n := 1/\log^2 n$. Since $\boldsymbol{\beta} \in I(\lambda^*) = [\beta_-, \beta_+]^{q^*+1}$, we can always choose n large enough such that $I_n^* \subseteq I(\lambda^*)$. With $f^* = h_{q^*}^* \circ \ldots \circ h_0^*$ and $R^* = 4K^{q^*}(q^*+1)Q$,

$$\begin{aligned}
&\Pi\Big(\{g : \|f^* - g\|_\infty \leq R^*\varepsilon_n(\eta^*)\}\Big) \\
&\quad \geq \int_{I_n^*} \mathbb{P}\Big(\big\|h_{q^*}^* \circ \ldots \circ h_0^* - G_{q^*}^{(\lambda^*,\boldsymbol{\beta})} \circ \ldots \circ G_0^{(\lambda^*,\boldsymbol{\beta})}\big\|_\infty \leq R^*\varepsilon_n(\eta^*)\Big)\pi(\lambda^*, \boldsymbol{\beta})d\boldsymbol{\beta}.
\end{aligned} \tag{2.A.5}$$

Fix any $\boldsymbol{\beta} \in I_n^*$. Both $G_{ij}^{(\lambda^*,\boldsymbol{\beta})}$ and h_{ij}^* map $[-1,1]^{d_i^*}$ into $[-1,1]$. They also depend on the same subset of variables $\mathcal{S}_{ij}^*$. By construction, the process $\overline{G}_{ij}^{(\lambda^*,\boldsymbol{\beta})} := G_{ij}^{(\lambda^*,\boldsymbol{\beta})} \circ (\cdot)_{S_{ij}^*}^{-1}$ is an independent copy of $\overline{G}_i^{(\beta_i,t_i^*)} : [-1,1]^{t_i^*} \to [-1,1]$ and $\overline{G}_i^{(\beta_i,t_i^*)}$ is the conditioned Gaussian process $\widetilde{G}^{(\beta_i,t_i^*)}|\{\widetilde{G}^{(\beta_i,t_i^*)} \in \mathcal{D}_i(\lambda^*, \boldsymbol{\beta}, K)\}$. The function $\overline{h}_{ij}^* := h_{ij}^* \circ (\cdot)_{S_{ij}^*}^{-1}$ belongs by definition to the space $\mathcal{C}_{t_i^*}^{\beta_i^*}(K)$ and satisfies $|\overline{h}_{ij}^*(\mathbf{x}) - \overline{h}_{ij}^*(\mathbf{y})| \leq K|\mathbf{x} - \mathbf{y}|_\infty^{\beta_i^* \wedge 1}$ for all $\mathbf{x}, \mathbf{y} \in [-1,1]^{t_i^*}$ and all $i = 0, \ldots, q^*$; $j = 1, \ldots, d_{i+1}^*$. By Lemma 2.A.2 with $\eta_i = 0$, we thus find

$$\|f^* - G^{(\lambda^*,\boldsymbol{\beta})}\|_\infty \leq K^{q^*}\sum_{i=0}^{q^*} \max_{j=1,\ldots,d_{i+1}^*} \|\overline{h}_{ij}^* - \overline{G}_{ij}^{(\lambda^*,\boldsymbol{\beta})}\|_\infty^{\alpha_i^*},$$

where $\alpha_i^* = \prod_{\ell=i+1}^{q^*} \beta_\ell^* \wedge 1$. Since $\alpha_i^* \geq \alpha_i = \prod_{\ell=i+1}^{q^*}(\beta_\ell \wedge 1)$ for $\boldsymbol{\beta} \in I_n^*$, if $\|\overline{h}_{ij}^* - \overline{G}_{ij}^{(\lambda^*,\boldsymbol{\beta})}\|_\infty$ is smaller than one, the latter display is bounded above by

$$\|f^* - G^{(\lambda^*,\boldsymbol{\beta})}\|_\infty \leq K^{q^*}\sum_{i=0}^{q^*} \max_{j=1,\ldots,d_{i+1}^*} \|\overline{h}_{ij}^* - \overline{G}_{ij}^{(\lambda^*,\boldsymbol{\beta})}\|_\infty^{\alpha_i}.$$

Set $\delta_{in} := \varepsilon_n(\alpha_i, \beta_i, t_i^*)^{1/\alpha_i}$. By Assumption 2.4.2, we have $\delta_{in} \leq (Q\varepsilon_n(\alpha_i^*, \beta_i^*, t_i^*))^{1/\alpha_i}$ and so $4\delta_{in} < 1$ since we are also assuming $\varepsilon_n(\eta^*) < 1/(4Q)$. Together with the definition of $\varepsilon_n(\eta^*)$ in (2.3.4), imposing $\|\overline{h}_{ij}^* - \overline{G}_{ij}^{(\lambda^*,\boldsymbol{\beta})}\|_\infty \leq 4\delta_{in}$ for all $i = 0, \ldots, q^*$ and $j = 1, \ldots, d_{i+1}^*$ implies $\|f^* - G^{(\lambda^*,\boldsymbol{\beta})}\|_\infty \leq R^*\varepsilon_n(\eta^*)$. Consequently,

$$\begin{aligned}
&\mathbb{P}\Big(\big\|h_{q^*}^* \circ \ldots \circ h_0^* - G_{q^*}^{(\lambda^*,\boldsymbol{\beta})} \circ \ldots \circ G_0^{(\lambda^*,\boldsymbol{\beta})}\big\|_\infty \leq R^*\varepsilon_n(\eta^*)\Big) \\
&\quad \geq \prod_{i=0}^{q^*}\prod_{j=1}^{d_{i+1}^*} \mathbb{P}\Big(\|\overline{h}_{ij}^* - \overline{G}_{ij}^{(\lambda^*,\boldsymbol{\beta})}\|_\infty \leq 4\delta_{in}\Big).
\end{aligned} \tag{2.A.6}$$

We now lower bound the probabilities on the right hand side. If $\overline{h}_{ij}^* \in \mathcal{C}_{t_i^*}^{\beta_i^*}(K)$, then, $(1-2\delta_{in})\overline{h}_{ij}^* \in \mathcal{C}_{t_i^*}^{\beta_i^*}((1-2\delta_{in})K)$ and by the embedding property in Lemma 2.A.1, we obtain $(1-2\delta_{in})\overline{h}_{ij}^* \in \mathcal{C}_{t_i^*}^{\beta_i}(K)$.

When $\|(1-2\delta_{in})\overline{h}_{ij}^* - \widetilde{G}^{(\beta_i,t_i^*)}\|_\infty \leq 2\delta_{in}$, the Gaussian process $\widetilde{G}^{(\beta_i,t_i^*)}$ is at most $2\delta_{in}$-away from $(1-2\delta_{in})\overline{h}_{ij}^* \in \mathcal{C}_{t_i^*}^{\beta_i}(K)$. Consequently, $\{\|(1-2\delta_{in})\overline{h}_{ij}^* - \widetilde{G}^{(\beta_i,t_i^*)}\|_\infty \leq 2\delta_{in}\} \subseteq \{\widetilde{G}^{(\beta_i,t_i^*)} \in \mathcal{D}_i(\lambda^*,\boldsymbol{\beta},K)\}$, where the bound for the unitary ball follows from the triangle inequality. Since $\|(1-2\delta_{in})\overline{h}_{ij}^* - \widetilde{G}^{(\beta_i,t_i^*)}\|_\infty \leq 2\delta_{in}$ implies $\|\overline{h}_{ij}^* - \widetilde{G}^{(\beta_i,t_i^*)}\|_\infty \leq 4\delta_{in}$, by the concentration function property in Lemma I.28 in [41] and the concentration function inequality in (2.3.2), we find

$$\begin{aligned}\mathbb{P}\Big(\|\overline{h}_{ij}^* - \overline{G}_{ij}^{(\lambda^*,\boldsymbol{\beta})}\|_\infty \leq 4\delta_{in}\Big) &\geq \frac{\mathbb{P}(\|(1-\delta_{in})\overline{h}_{ij}^* - \widetilde{G}^{(\beta_i,t_i^*)}\|_\infty \leq 2\delta_{in})}{\mathbb{P}\big(\widetilde{G}^{(\beta_i,t_i^*)} \in \mathcal{D}_i(\lambda^*,\boldsymbol{\beta},K)\big)} \\ &\geq \mathbb{P}\Big(\big\|(1-\delta_{in})\overline{h}_{ij}^* - \widetilde{G}^{(\beta_i,t_i^*)}\big\|_\infty \leq 2\delta_{in}\Big) \\ &\geq \exp\Big(-\varphi^{(\beta_i,t_i^*,K)}(\delta_{in})\Big) \\ &\geq \exp\Big(-n\varepsilon_n(\alpha_i,\beta_i,t_i^*)^2\Big) \\ &\geq \exp\Big(-Q^2 n\varepsilon_n(\eta^*)^2\Big),\end{aligned}$$

here Q is the universal constant from Assumption 2.4.2. With $|\mathbf{d}^*|_1 = 1+\sum_{j=0}^{q^*} d_j^*$, (2.A.5), (2.A.6) and the previous display we recover the claim. □

The latter result shows that

$$\Pi\Big(B_2\big(P_{f^*}, 4K^{q^*}(q^*+1)Q\varepsilon_n(\eta^*)\big)\Big) \geq e^{-|\mathbf{d}^*|_1 Q^2 n\varepsilon_n(\eta^*)^2}\int_{I_n^*}\pi(\lambda^*,\boldsymbol{\beta})\,d\boldsymbol{\beta}, \tag{2.A.7}$$

with $I_n^* = \{\boldsymbol{\beta} = (\beta_0,\ldots,\beta_{q^*}) : \beta_i \in [\beta_i^* - b_n, \beta_i^*], \forall i\}$ and $b_n = 1/\log^2 n$. Recall that, by construction (2.3.5), $\pi(\eta) \propto e^{-\Psi_n(\eta)}\gamma(\eta)$ with $\Psi_n(\eta) = n\varepsilon_n(\eta)^2 + e^{e^{|\mathbf{d}|_1}}$. For any $\boldsymbol{\beta} \in I_n^*$, we have $\Psi_n(\lambda^*,\boldsymbol{\beta}) \leq \Psi_n(\lambda^*,\boldsymbol{\beta}^*)$, and Assumption 2.4.1 gives $\gamma(\lambda^*,\boldsymbol{\beta}) = \gamma(\lambda^*)\gamma(\boldsymbol{\beta}|\lambda^*)$ with $\gamma(\lambda^*) > 0$ independent of n and $\gamma(\cdot|\lambda^*)$ the uniform distribution over $I(\lambda^*) = [\beta_-,\beta_+]^{q^*+1}$. Thus, $\gamma(I_n^*|\lambda^*) = |I_n^*|/|I(\lambda^*)|$ and $|I_n^*| = (1/\log^2 n)^{q^*+1}$, so that

$$\frac{\pi(\eta)}{\int_{I_n^*}\pi(\lambda^*,\boldsymbol{\beta})\,d\boldsymbol{\beta}} \leq \frac{e^{\Psi_n(\eta^*)-\Psi_n(\eta)}\gamma(\eta)}{\gamma(\lambda^*)\gamma(I_n^*|\lambda^*)} = \frac{|I(\lambda^*)|}{\gamma(\lambda^*)}e^{\Psi_n(\eta^*)-\Psi_n(\eta)}e^{2(q^*+1)\log\log n}\gamma(\eta). \tag{2.A.8}$$

Both $\gamma(\lambda^*)$ and $|I(\lambda^*)| = (\beta_+ - \beta_-)^{q^*+1}$ are constants independent of n. Furthermore, $e^{\Psi_n(\eta^*)} = \exp(e^{e^{|\mathbf{d}^*|_1}})e^{n\varepsilon_n(\eta^*)^2}$ and the quantity $\exp(e^{e^{|\mathbf{d}^*|_1}})$ is independent of n as well. We can finally verify condition (2.A.4) by showing that, with $a_* = 4K^{q^*}(q^*+1)Q$,

$$e^{2a_*^2 n\varepsilon_n(\eta^*)^2}e^{n\varepsilon_n(\eta^*)^2+2(q^*+1)\log\log n}\sum_\lambda\int_{\boldsymbol{\beta}:(\lambda,\boldsymbol{\beta})\notin\mathcal{M}_n(C)} e^{-\Psi_n(\eta)}\gamma(\eta)\,d\boldsymbol{\beta} \to 0.$$

By the lower bound in Assumption 2.4.2 (i), we have $n\varepsilon_n(\eta^*)^2 \geq n\mathfrak{r}_n(\eta^*)^2 \gg 2(q^*+1)\log\log n$, since the quantity $n\mathfrak{r}_n(\eta^*)^2$ is a positive power of n by definition (2.5.1). The complement of the set $\mathcal{M}_n(C)$ is the union of $\{\eta : \varepsilon_n(\eta) > C\varepsilon_n(\eta^*)\}$ and $\{\eta : |\mathbf{d}|_1 > \log(2\log n)\}$. Over these sets, by construction (2.3.5), we have either $\Psi_n(\eta) > C^2 n\varepsilon_n(\eta^*)^2$ or $\Psi_n(\eta) > n^2$. Therefore, the term $e^{-\Psi_n(\eta)}$ decays faster than either $e^{-C^2 n\varepsilon_n(\eta^*)^2}$ or e^{-n^2}. In the first case, the latter display converges to zero for sufficiently large $C > 0$. In the second case, the latter display converges to zero since $\varepsilon_n(\eta^*) < 1$ and $n\varepsilon_n(\eta^*)^2 < n \ll n^2$. This completes the proof. □

We need some preliminary notation and results before proving Theorem 2.4.4.

Entropy bounds. Previous bounds for the metric entropy of Hölder-balls, e.g. Proposition C.5 in [41], are of the form $\log\mathcal{N}\big(\delta, \mathcal{C}_r^\beta(K), \|\cdot\|_\infty\big) \le Q_1(\beta, r, K)\delta^{-r/\beta}$ for some constant $Q_1(\beta, r, K)$ that is hard to control. An exception is Theorem 8 in [14] that, however, only holds for $\beta \le 1$. We derive an explicit bound on the constant $Q_1(\beta, r, K)$ for all $\beta > 0$. The proof is given in Section 2.A.1.

Lemma 2.A.5. *For any positive integer r, any $\beta > 0$ and $0 < \delta < 1$, we have*

$$\begin{aligned}\mathcal{N}\left(\delta, C_r^\beta(K), \|\cdot\|_\infty\right) &\le \left(\frac{4eK^2r^\beta}{\delta} + 1\right)^{(\beta+1)^r} \left(2^{\beta+2}eKr^\beta + 1\right)^{4^r(\beta+1)^r r^r (2eK)^{\frac{r}{\beta}}\delta^{-\frac{r}{\beta}}} \\ &\le e^{Q_1(\beta,r,K)\delta^{-\frac{r}{\beta}}}\end{aligned}$$

with $Q_1(\beta, r, K) := (1 + eK)4^{r+1}(\beta + 3)^{r+1}r^{r+1}(8eK^2)^{r/\beta}$. For any $0 < \alpha \le 1$ and any sequence $\delta_n \ge Q_1(\beta, r, K)^{\beta/(2\beta+r)}n^{-\beta\alpha/(2\beta\alpha+r)}$, we also have

$$\log\mathcal{N}\left(\delta_n^{1/\alpha}, C_r^\beta(K), \|\cdot\|_\infty\right) \le n\delta_n^2. \tag{2.A.9}$$

Support of DGP prior and local complexity. For any graph $\lambda = (q, \mathbf{d}, \mathbf{t}, \mathcal{S})$ and any $\boldsymbol{\beta} \in [\beta_-, \beta_+]^{q+1}$, denote by $\Theta_n(\lambda, \boldsymbol{\beta}, K)$ the space of functions $f : [-1, 1]^d \to [-1, 1]$ for which there exists a decomposition $f = h_q \circ \ldots \circ h_0$ such that $h_{ij} : [-1, 1]^{d_i} \to [-1, 1]$ and $\overline{h}_{ij} = h_{ij} \circ (\cdot)^{-1}_{\mathcal{S}_{ij}} \in \mathcal{D}_i(\lambda, \boldsymbol{\beta}, K)$, for all $i = 0, \ldots, q$; $j = 1, \ldots, d_{i+1}$ and $\mathcal{D}_i(\lambda, \boldsymbol{\beta}, K)$ as defined in (2.3.3). Differently speaking

$$\Theta_n(\lambda, \boldsymbol{\beta}, K) := \Theta_{q,n}(\lambda, \boldsymbol{\beta}, K) \circ \cdots \circ \Theta_{0,n}(\lambda, \boldsymbol{\beta}, K) \tag{2.A.10}$$

with

$$\Theta_{i,n}(\lambda, \boldsymbol{\beta}, K) := \left\{h_i : [-1, 1]^{d_i} \to [-1, 1]^{d_{i+1}} : h_{ij} \circ (\cdot)^{-1}_{\mathcal{S}_{ij}} \in \mathcal{D}_i(\lambda, \boldsymbol{\beta}, K), j = 1, \ldots, d_{i+1}\right\}.$$

By construction, the support of the deep Gaussian process $G^{(\eta)} \sim \Pi(\cdot|\eta)$ is contained in $\Theta_n(\lambda, \boldsymbol{\beta}, K)$. For a subset $B \subseteq [\beta_-, \beta_+]^{q+1}$ we also set $\Theta_n(\lambda, B, K) := \cup_{\boldsymbol{\beta}\in B}\Theta_n(\lambda, \boldsymbol{\beta}, K)$. The next lemma provides a bound for the covering number of $\Theta_n(\lambda, B, K)$. Recall that $\operatorname{diam}(B) = \sup_{\boldsymbol{\beta},\boldsymbol{\beta}'\in B} |\boldsymbol{\beta} - \boldsymbol{\beta}'|_\infty$. We postpone the proof to Section 2.A.1.

Lemma 2.A.6. *Suppose that Assumption 2.4.2 holds and let λ be a graph such that $|\mathbf{d}|_1 \le \log(2\log n)$. Let $B \subseteq [\beta_-, \beta_+]^{q+1}$ with $\operatorname{diam}(B) \le 1/\log^2 n$. Then, with $R_n := 5Q(2\log n)^{1+\log K}$,*

$$\sup_{\boldsymbol{\beta}\in B} \frac{\log\mathcal{N}\left(R_n\varepsilon_n(\lambda, \boldsymbol{\beta}), \Theta_n(\lambda, B, K), \|\cdot\|_\infty\right)}{n\varepsilon_n(\lambda, \boldsymbol{\beta})^2} \le \frac{R_n^2}{25}.$$

We now prove Theorem 2.4.4 by following a classical argument in [41], namely Theorem 8.14, which recovers the posterior contraction rates by means of partition entropy.

Proof of Theorem 2.4.4. For any $\rho > 0$, introduce the complement Hellinger ball $\mathcal{H}^c(f^*, \rho) := \{f : d_H(P_f, P_{f^*}) > \rho\}$. The convergence with respect to the Hellinger distance d_H implies convergence in $L^2(\mu)$ thanks to (2.A.2). As a consequence, we show that

$$\sup_{f^* \in \mathcal{F}(\eta^*, K)} \mathbb{E}_{f^*}\Big[\Pi\Big(\mathcal{H}^c(f^*, L_n\varepsilon_n(\eta^*))\Big|(\mathbf{X}, \mathbf{Y})\Big)\Big] \to 0,$$

with $L_n := MR_n$, $R_n := 10QC(2\log n)^{1+\log K}$ and $M > 0$ a sufficiently large universal constant to be determined. With $\mathcal{M}_n(C) = \{\eta : \varepsilon_n(\eta) \leq C\varepsilon_n(\eta^*)\} \cap \{\eta : |\mathbf{d}|_1 \leq \log(2\log n)\}$ the set of good composition structures, and the notation in (2.4.2), we denote by $\Pi(\cdot \cap \mathcal{M}_n(C)|\mathbf{X}, \mathbf{Y})$ the contribution of the good structures to the posterior mass.

We denote by $\mathcal{L}_n(C)$ the set of graphs that are realized by some good composition structure, that is,

$$\mathcal{L}_n(C) := \big\{\lambda \in \Lambda : \exists \boldsymbol{\beta} \in I(\lambda),\ \eta = (\lambda, \boldsymbol{\beta}) \in \mathcal{M}_n(C)\big\}. \tag{2.A.11}$$

Condition (ii) in Assumption 2.4.2 holds for all the graphs in $\mathcal{L}_n(C)$. For any $\lambda \in \mathcal{L}_n(C)$, partition $I(\lambda) = [\beta_-, \beta_+]^{q+1}$ into hypercubes of diameter $1/\log^2 n$ and let $B_1(\lambda), \ldots, B_{N(\lambda)}$ be the $N(\lambda)$ blocks that contain at least one $\boldsymbol{\beta} \in I(\lambda)$ that is realized by some composition structure in $\mathcal{M}_n(C)$. The blocks may contain also values of $\boldsymbol{\beta}$ for which $(\lambda, \boldsymbol{\beta}) \notin \mathcal{M}_n(C)$. Then, the set of good composition structures is contained in the enlargement

$$\mathcal{M}_n(C) \subseteq \widetilde{\mathcal{M}}_n(C) := \bigcup_{\lambda \in \mathcal{L}_n(C)} \bigcup_{k=1}^{N(\lambda)} \Big(\{\lambda\} \times B_k(\lambda)\Big).$$

Thanks to Theorem 2.4.3 and the enlarged set of structures $\widetilde{\mathcal{M}}_n(C)$, it is enough to show, for sufficiently large constants M, C,

$$\sup_{f^* \in \mathcal{F}(\eta^*, K)} \mathbb{E}_{f^*}\Big[\Pi\Big(\mathcal{H}^c(f^*, MR_n\varepsilon_n(\eta^*)) \cap \widetilde{\mathcal{M}}_n(C)\Big|(\mathbf{X}, \mathbf{Y})\Big)\Big] \to 0. \tag{2.A.12}$$

Fix any $f^* \in \mathcal{F}(\eta^*, K)$. Since there is no ambiguity, we shorten the notation to $\mathcal{H}_n^c = \mathcal{H}^c(f^*, MR_n\varepsilon_n(\eta^*))$ and rewrite

$$\mathbb{E}_{f^*}\Big[\Pi\Big(\mathcal{H}_n^c \cap \widetilde{\mathcal{M}}_n(C)\Big|(\mathbf{X}, \mathbf{Y})\Big)\Big] = \mathbb{E}_{f^*}\left[\frac{\int_{\mathcal{H}_n^c} \Pi(df \cap \widetilde{\mathcal{M}}_n(C)|\mathbf{X}, \mathbf{Y})}{\int \Pi(df|\mathbf{X}, \mathbf{Y})}\right].$$

We follow the steps of the proof of Theorem 8.14 in [41]. In their notation we use $\varepsilon_n = R_n\varepsilon_n(\eta^*)$ and $\xi = 1/2$ for contraction with respect to Hellinger loss. Set

$$A_n^* = \Big\{\int \Pi(df|\mathbf{X}, \mathbf{Y}) \geq \Pi\Big(B_2(f^*, R_n\varepsilon_n(\eta^*))\Big)e^{-2R_n^2 n\varepsilon_n(\eta^*)^2}\Big\}.$$

Then $\mathbb{P}_{f^*}(A_n^*)$ tends to 1, thanks to Lemma 8.10 in [41] applied with $D = 1$. Since $1 = \mathbf{1}(A_n^*) + \mathbf{1}(A_n^{*,c})$, we have

$$\mathbb{E}_{f^*}\left[\Pi\Big(\mathcal{H}_n^c \cap \widetilde{\mathcal{M}}_n(C)\Big|(\mathbf{X},\mathbf{Y})\Big)\right] \leq \mathbb{P}_{f^*}(A_n^{*,c}) + \mathbb{E}_{f^*}\left[\mathbf{1}(A_n^*)\frac{\int_{\mathcal{H}_n^c}\Pi(df \cap \widetilde{\mathcal{M}}_n(C)|\mathbf{X},\mathbf{Y})}{\int \Pi(df|\mathbf{X},\mathbf{Y})}\right]$$

and $\mathbb{P}_{f^*}(A_n^{*,c}) \to 0$ when $n \to +\infty$. It remains to show that the second terms on the right side tends to zero.

Let $\phi_{n,k}(\lambda)$ be arbitrary statistical tests to be chosen later. Test are to be understood as $\phi_{n,k}(\lambda) = \phi_{n,k}(\lambda)(\mathbf{X},\mathbf{Y})$ measurable functions of the sample $(\mathbf{X},\mathbf{Y})$, taking values in $[0,1]$. Then, $1 = \phi_{n,k}(\lambda) + (1 - \phi_{n,k}(\lambda))$. Using the definition of $\Pi(df \cap \widetilde{\mathcal{M}}_n(C)|\mathbf{X},\mathbf{Y})$ and Fubini's theorem, we find

$$\mathbb{E}_{f^*}\left[\Pi\Big(\mathcal{H}_n^c \cap \widetilde{\mathcal{M}}_n(C)\Big|(\mathbf{X},\mathbf{Y})\Big)\right] \leq \mathbb{P}_{f^*}(A_n^{*,c}) + \mathbb{E}_{f^*}\left[T_1 + T_2\right], \tag{2.A.13}$$

where

$$T_1 := \mathbf{1}(A_n^*)\frac{\sum_{\lambda\in\mathcal{L}_n(C)}\sum_{k=1}^{N(\lambda)}\phi_{n,k}(\lambda)\int_{B_k(\lambda)}\pi(\lambda,\boldsymbol{\beta})\left(\int_{\mathcal{H}_n^c}\frac{p_f}{p_{f^*}}(\mathbf{X},\mathbf{Y})\Pi(df|\lambda,\boldsymbol{\beta})\right)d\boldsymbol{\beta}}{\int \Pi(df|\mathbf{X},\mathbf{Y})},$$

$$T_2 := \mathbf{1}(A_n^*)\frac{\sum_{\lambda\in\mathcal{L}_n(C)}\sum_{k=1}^{N(\lambda)}\int_{B_k(\lambda)}\pi(\lambda,\boldsymbol{\beta})\left(\int_{\mathcal{H}_n^c}(1-\phi_{n,k}(\lambda))\frac{p_f}{p_{f^*}}(\mathbf{X},\mathbf{Y})\Pi(df|\lambda,\boldsymbol{\beta})\right)d\boldsymbol{\beta}}{\int \Pi(df|\mathbf{X},\mathbf{Y})}.$$

We bound T_1 by using $\mathbf{1}(A_n^*) \leq 1$, together with

$$\frac{\int_{B_k(\lambda)}\pi(\lambda,\boldsymbol{\beta})\left(\int_{\mathcal{H}_n^c}\frac{p_f}{p_{f^*}}(\mathbf{X},\mathbf{Y})\Pi(df|\lambda,\boldsymbol{\beta})\right)d\boldsymbol{\beta}}{\int \Pi(df|\mathbf{X},\mathbf{Y})} \leq 1,$$

so that

$$\mathbb{E}_{f^*}\left[T_1\right] \leq \sum_{\lambda\in\mathcal{L}_n(C)}\sum_{k=1}^{N(\lambda)}\mathbb{E}_{f^*}\left[\phi_{n,k}(\lambda)\right]. \tag{2.A.14}$$

We bound T_2 using the definition of A_n^*, and obtain

$$T_2 \leq \frac{\sum_{\lambda\in\mathcal{L}_n(C)}\sum_{k=1}^{N(\lambda)}\int_{B_k(\lambda)}\pi(\lambda,\boldsymbol{\beta})\left(\int_{\mathcal{H}_n^c}(1-\phi_{n,k}(\lambda))\frac{p_f}{p_{f^*}}(\mathbf{X},\mathbf{Y})\Pi(df|\lambda,\boldsymbol{\beta})\right)d\boldsymbol{\beta}}{\Pi\left(B_2\big(f^*, R_n\varepsilon_n(\eta^*)\big)\right)e^{-2R_n^2 n\varepsilon_n(\eta^*)^2}}.$$

For large n, $R_n = 10QC(2\log n)^{1+\log K} \geq 4QK^{q^*}(q^*+1)$. By Lemma 2.A.4, we find

$$\Pi\left(B_2\big(f^*, R_n\varepsilon_n(\eta^*)\big)\right) \geq e^{-R_n^2 n\varepsilon_n(\eta^*)^2}\pi(\lambda^*, I_n^*), \tag{2.A.15}$$

with $I_n^* := \{\boldsymbol{\beta} = (\beta_0,\dots,\beta_{q^*}) : \beta_i \in [\beta_i^* - b_n, \beta_i^*], \forall i\}$ and $b_n := 1/\log^2 n$. By the construction of the prior π in (2.3.5), the denominator term $e^{-\Psi_n(\eta)}$ is bounded above by 1 and $\int_\Omega \gamma(\eta)\,d\eta = 1$, thus

$$\pi(\lambda^*, I_n^*) = \frac{\int_{I_n^*} e^{-\Psi_n(\lambda^*,\boldsymbol{\beta})}\gamma(\lambda^*,\boldsymbol{\beta})d\boldsymbol{\beta}}{\int_\Omega e^{-\Psi_n(\eta)}\gamma(\eta)\,d\eta} \geq \int_{I_n^*} e^{-\Psi_n(\lambda^*,\boldsymbol{\beta})}\gamma(\lambda^*,\boldsymbol{\beta})d\boldsymbol{\beta}.$$

Furthermore, by condition (ii) in Assumption 2.4.2, we have $\varepsilon_n(\lambda^*, \boldsymbol{\beta}) \leq Q\varepsilon_n(\lambda^*, \boldsymbol{\beta}^*)$ for any $\boldsymbol{\beta} \in I_n^*$, which gives

$$\pi(\lambda^*, I_n^*) \geq \exp(-e^{e^{|\mathbf{d}^*|_1}})\, e^{-Q^2 n\varepsilon_n(\eta^*)^2} \gamma(\lambda^*, I_n^*),$$

with proportionality constant $\exp(-e^{e^{|\mathbf{d}^*|_1}}) > 0$ independent of n. Again by construction, the measure $\gamma(\lambda^*, I_n^*)$ can be split into the product $\gamma(\lambda^*)\gamma(I_n^*|\lambda^*)$ where $\gamma(\cdot|\lambda^*)$ is the uniform measure on $I(\lambda^*) = [\beta_-, \beta_+]^{q^*+1}$ and the quantity $\gamma(\lambda^*) > 0$ is a constant independent of n. Thus, with $c(\eta^*) := \exp(-e^{e^{|\mathbf{d}^*|_1}})\gamma(\lambda^*)/|I(\lambda^*)|$, $\pi(\lambda^*, I_n^*) \geq c(\eta^*)|I_n^*|e^{-Q^2 n\varepsilon_n(\eta^*)^2}$. In the discussion after (2.A.8) we have shown that $n\varepsilon_n(\eta^*)^2$ is a positive power of n and so, for a large enough n depending only on η^*, one has $|I_n^*| = (1/\log^2 n)^{q^*+1} > e^{-n\varepsilon_n(\eta^*)^2}$. This results in

$$\pi(\lambda^*, I_n^*) \geq c(\eta^*)e^{-(Q^2+1)n\varepsilon_n(\eta^*)^2} \geq c(\eta^*)e^{-R_n^2 n\varepsilon_n(\eta^*)^2}. \tag{2.A.16}$$

By putting together the small ball probability bound (2.A.15) and the hyperprior bound (2.A.16), we recover

$$T_2 \leq c(\eta^*)^{-1} \frac{\sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \int_{B_k(\lambda)} \pi(\lambda, \boldsymbol{\beta}) \left(\int_{\mathcal{H}_n^c} (1 - \phi_{n,k}(\lambda)) \frac{p_f}{p_{f^*}}(\mathbf{X}, \mathbf{Y}) \Pi(df|\lambda, \boldsymbol{\beta}) \right) d\boldsymbol{\beta}}{e^{-4R_n^2 n\varepsilon_n(\eta^*)^2}}, \tag{2.A.17}$$

with proportionality constant $c(\eta^*)^{-1}$ independent of n and depending only on η^*, β_-, β_+. We now bound the numerator of T_2 using Fubini's theorem and the inequality $\mathbb{E}_{f^*}[(1 - \phi_{n,k}(\lambda))(p_f/p_{f_*})(\mathbf{X}, \mathbf{Y})] \leq \mathbb{E}_f[1 - \phi_{n,k}(\lambda)]$,

$$\begin{aligned}
&\mathbb{E}_{f^*}\left[\sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \int_{B_k(\lambda)} \pi(\lambda, \boldsymbol{\beta}) \left(\int_{\mathcal{H}_n^c} (1 - \phi_{n,k}(\lambda)) \frac{p_f}{p_{f^*}}(\mathbf{X}, \mathbf{Y}) \Pi(df|\lambda, \boldsymbol{\beta}) \right) d\boldsymbol{\beta} \right] \\
&\quad = \sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \int_{B_k(\lambda)} \pi(\lambda, \boldsymbol{\beta}) \left(\int_{\mathcal{H}_n^c} \mathbb{E}_{f^*}\left[(1 - \phi_{n,k}(\lambda)) \frac{p_f}{p_{f^*}}(\mathbf{X}, \mathbf{Y}) \right] \Pi(df|\lambda, \boldsymbol{\beta}) \right) d\boldsymbol{\beta} \\
&\quad \leq \sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \int_{B_k(\lambda)} \pi(\lambda, \boldsymbol{\beta}) \int_{\mathcal{H}_n^c} \mathbb{E}_f\left[(1 - \phi_{n,k}(\lambda)) \right] \Pi(df|\lambda, \boldsymbol{\beta}) d\boldsymbol{\beta}. \qquad (2.A.18)
\end{aligned}$$

With the supports $\Theta_n(\lambda, \boldsymbol{\beta}, K)$ in (2.A.10) and any $k = 1, \ldots, N(\lambda)$, consider $\Theta_n(\lambda, B_k(\lambda), K) = \cup_{\boldsymbol{\beta} \in B_k(\lambda)} \Theta_n(\lambda, \boldsymbol{\beta}, K)$. Now, for any fixed λ, k choose tests $\phi_{n,k}(\lambda)$ according to Theorem D.5 in [41], so that for $f \in \Theta_n(\lambda, B_k(\lambda), K) \cap \mathcal{H}^c(f^*, MR_n\varepsilon_n(\eta^*))$ we have, for some universal constant $\widetilde{K} > 0$,

$$\begin{aligned}
\mathbb{E}_{f^*}[\phi_{n,k}(\lambda)] &\leq c_k(\lambda) \mathcal{N}\left(\frac{R_n\varepsilon_n(\eta^*)}{2}, \Theta_n(\lambda, B_k(\lambda), K), \|\cdot\|_\infty \right) \frac{e^{-\widetilde{K}M^2R_n^2 n\varepsilon_n(\eta^*)^2}}{1 - e^{-\widetilde{K}M^2R_n^2 n\varepsilon_n(\eta^*)^2}}, \\
\mathbb{E}_f[1 - \phi_{n,k}(\lambda)] &\leq c_k(\lambda)^{-1} e^{-\widetilde{K}M^2R_n^2 n\varepsilon_n(\eta^*)^2},
\end{aligned}$$

with choice of coefficients

$$c_k(\lambda)^2 := \frac{\pi(\lambda, B_k(\lambda))}{\mathcal{N}\left(\frac{R_n\varepsilon_n(\eta^*)}{2}, \Theta_n(\lambda, B_k(\lambda), K), \|\cdot\|_\infty \right)}.$$

Let us denote by $\rho_k(\lambda)$ the local complexities

$$\rho_k(\lambda) := \sqrt{\pi(\lambda, B_k(\lambda))} \cdot \sqrt{\mathcal{N}\left(\frac{R_n \varepsilon_n(\eta^*)}{2}, \Theta_n(\lambda, B_k(\lambda), K), \|\cdot\|_\infty\right)}.$$

Combining this with the bound on T_1 in (2.A.14) and the bounds on T_2 in (2.A.17)–(2.A.18), gives

$$\mathbb{E}_{f^*}[T_1] \leq \frac{e^{-\widetilde{K}M^2 R_n^2 n\varepsilon_n(\eta^*)^2}}{1 - e^{-\widetilde{K}M^2 R_n^2 n\varepsilon_n(\eta^*)^2}} \sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \rho_k(\lambda),$$

$$\mathbb{E}_{f^*}[T_2] \leq c(\eta^*)^{-1} e^{(4-\widetilde{K}M^2) R_n^2 n\varepsilon_n(\eta^*)^2} \sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \rho_k(\lambda).$$

It remains to show that both expectations in the latter display tend to zero when $n \to +\infty$.

Since $M > 0$ can be chosen arbitrarily large, we choose it in such a way that $\widetilde{K}M^2 > 5$ and the proof is complete if we can show that

$$\sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \rho_k(\lambda) \lesssim e^{R_n^2 n\varepsilon_n(\eta^*)^2}, \tag{2.A.19}$$

for some proportionality constant independent of n. Fix $\lambda \in \mathcal{L}_n(C)$ and $k = 1, \ldots, N(\lambda)$. By construction, there exists $\boldsymbol{\beta} \in B_k(\lambda)$ such that $(\lambda, \boldsymbol{\beta}) \in \mathcal{M}_n(C)$. Since $R_n = 10QC(2\log n)^{1+\log K}$, using $C\varepsilon_n(\eta^*) \geq \varepsilon_n(\eta)$ together with Lemma 2.A.6, we find

$$\log \mathcal{N}\left(\frac{R_n \varepsilon_n(\eta^*)}{2}, \Theta_n(\lambda, B, K), \|\cdot\|_\infty\right) \leq \frac{R_n^2}{25} n\varepsilon_n(\eta^*)^2.$$

This results in

$$\begin{aligned}\sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \rho_k(\lambda) &\leq \exp\left(\frac{1}{50} R_n^2 n\varepsilon_n(\eta^*)^2\right) \sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \sqrt{\pi(\lambda, B_k(\lambda))} \\ &= \exp\left(\frac{1}{50} R_n^2 n\varepsilon_n(\eta^*)^2\right) z_n^{-\frac{1}{2}} \sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \sqrt{\gamma(\lambda, B_k(\lambda))},\end{aligned}$$

where $z_n = \sum_\lambda \int_{I(\lambda)} e^{-\Psi_n(\lambda, \boldsymbol{\beta})} \gamma(\lambda, \boldsymbol{\beta}) d\boldsymbol{\beta}$ is the normalization term in (2.3.5). By the localization argument in (2.A.16), we know that $z_n \geq \pi(\lambda^*, I_n^*) \gtrsim e^{-R_n^2 n\varepsilon_n(\eta^*)^2}$, with proportionality constant independent of n. Thus,

$$\sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \rho_k(\lambda) \lesssim \exp\left(\frac{26}{50} R_n^2 n\varepsilon_n(\eta^*)^2\right) \sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \sqrt{\gamma(\lambda, B_k(\lambda))}.$$

Since $R_n = 10QC(2\log n)^{1+\log K} \gg 1$, it is sufficient for (2.A.19) that

$$\sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \sqrt{\gamma(\lambda, B_k(\lambda))} \lesssim e^{\frac{1}{2} n\varepsilon_n(\eta^*)^2}, \tag{2.A.20}$$

for some proportionality constant that can be chosen independent of n. To see this, observe that by Assumption 2.4.1, we have $\gamma(\lambda, \boldsymbol{\beta}) = \gamma(\lambda)\gamma(\boldsymbol{\beta}|\lambda)$ with $\gamma(\cdot|\lambda)$ the uniform distribution over $I(\lambda)$. Thus

$$\sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \sqrt{\gamma(\lambda, B_k(\lambda))} = \sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \sqrt{|B_k(\lambda)|}\sqrt{\gamma(\lambda, \boldsymbol{\beta}_k(\lambda))},$$

where $\boldsymbol{\beta}_k(\lambda)$ is the center point of the hypercube $B_k(\lambda)$. Since $|B_k(\lambda)| = (1/\log^2 n)^{q+1}$ and $(q+1) \leq |\mathbf{d}|_1 \leq \log(2\log n)$, we have $\log(|B_k(\lambda)|^{-1}) = 2(q+1)(\log n) \leq 4\log^2 n \ll n\varepsilon_n(\eta^*)^2$. Therefore, for a sufficiently large n depending only on η^*, we find $|B_k(\lambda)|^{-1} \leq e^{n\varepsilon_n(\eta^*)^2}$. The above discussion yields the following bound on the latter display,

$$\begin{aligned}\sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \sqrt{\gamma(\lambda, B_k(\lambda))} &= \sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \frac{1}{\sqrt{|B_k(\lambda)|}}|B_k(\lambda)|\sqrt{\gamma(\lambda, \boldsymbol{\beta}_k(\lambda))} \\ &\leq e^{\frac{1}{2}n\varepsilon_n(\eta^*)^2} \sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} |B_k(\lambda)|\sqrt{\gamma(\lambda, \boldsymbol{\beta}_k(\lambda))}.\end{aligned}$$

This is enough to obtain (2.A.20) since, by Assumption 2.4.1,

$$\sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} |B_k(\lambda)|\sqrt{\gamma(\lambda, \boldsymbol{\beta}_k(\lambda))} \leq \sum_{\lambda\in\Lambda} \int_{I(\lambda)} \sqrt{\gamma(\lambda, \boldsymbol{\beta})}d\boldsymbol{\beta} = \int_{\Omega} \sqrt{\gamma(\eta)}\, d\eta$$

is a finite constant independent of n.

Since all bounds are independent of the particular choice of f^* and only depend on the function class $\mathcal{F}(\eta^*, K)$, this concludes the proof of the uniform statement (2.A.12). □

Proofs of auxiliary results

Proof of Lemma 2.A.5. We follow the proof of Theorem 2.7.1 in [84] and provide explicit expressions for all constants. We start by covering the interval $[-1,1]^r$ with a grid of width $\tau = (\delta/c(\beta))^{1/\beta}$, where $c(\beta) := er^\beta + 2K$. The grid consists of M points $\mathbf{x}_1, \ldots, \mathbf{x}_M$ with

$$M \leq \frac{vol([-2,2]^r)}{\tau^r} = 4^r c(\beta)^{\frac{r}{\beta}}\delta^{-\frac{r}{\beta}}. \tag{2.A.21}$$

For any $h \in \mathcal{C}_r^\beta(K)$ and any $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r) \in \mathbb{N}^r$ with $|\boldsymbol{\alpha}|_1 = \alpha_1 + \ldots + \alpha_r \leq \lfloor\beta\rfloor$, set

$$A^{\boldsymbol{\alpha}}h := \left(\left\lfloor \frac{\partial^{\boldsymbol{\alpha}}h(\mathbf{x}_1)}{\tau^{\beta-|\boldsymbol{\alpha}|_1}}\right\rfloor, \ldots, \left\lfloor \frac{\partial^{\boldsymbol{\alpha}}h(\mathbf{x}_M)}{\tau^{\beta-|\boldsymbol{\alpha}|_1}}\right\rfloor\right). \tag{2.A.22}$$

The vector $\tau^{\beta-|\boldsymbol{\alpha}|_1}A^{\boldsymbol{\alpha}}h$ consists of the values $\partial^{\boldsymbol{\alpha}}h(\mathbf{x}_i)$ discretized on a grid of mesh-width $\tau^{\beta-|\boldsymbol{\alpha}|_1}$. Since $\partial^{\boldsymbol{\alpha}}h \in \mathcal{C}_r^{\beta-|\boldsymbol{\alpha}|_1}(K)$ and $\tau < 1$ by construction, the entries of the vector in the latter display are integers bounded in absolute value by

$$\left\lfloor \frac{|\partial^{\boldsymbol{\alpha}}h(\mathbf{x}_i)|}{\tau^{\beta-|\boldsymbol{\alpha}|_1}}\right\rfloor \leq \left\lfloor \frac{K}{\tau^{\beta-|\boldsymbol{\alpha}|_1}}\right\rfloor \leq \left\lfloor \frac{K}{\tau^{\beta}}\right\rfloor. \tag{2.A.23}$$

Let $h, \widetilde{h} \in \mathcal{C}_r^{\beta}(K)$ be two functions such that $A^{\boldsymbol{\alpha}} h = A^{\boldsymbol{\alpha}} \widetilde{h}$ for all $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}|_1 \leq \lfloor \beta \rfloor$. We now show that $\|h - \widetilde{h}\|_\infty \leq \delta$. For any $\mathbf{x} \in [-1,1]^r$, let $\mathbf{x}_i$ be the closest grid vertex, so that $|\mathbf{x} - \mathbf{x}_i|_\infty \leq \tau$. Taylor expansion around $\mathbf{x}_i$ gives

$$\begin{aligned}(h - \widetilde{h})(\mathbf{x}) &= \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1 \leq \lfloor \beta \rfloor} \partial^{\boldsymbol{\alpha}}(h - \widetilde{h})(\mathbf{x}_i) \frac{(\mathbf{x} - \mathbf{x}_i)^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} + R, \\ R &= \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1 = \lfloor \beta \rfloor} \Big[\partial^{\boldsymbol{\alpha}}(h - \widetilde{h})(\mathbf{x}_\xi) - \partial^{\boldsymbol{\alpha}}(h - \widetilde{h})(\mathbf{x}_i)\Big] \frac{(\mathbf{x} - \mathbf{x}_i)^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!},\end{aligned} \tag{2.A.24}$$

with $\mathbf{x}_{\xi_i} = \mathbf{x}_i + \xi_i(\mathbf{x} - \mathbf{x}_i)$ and a suitable $\xi_i \in [0,1]$. With $|\partial^{\boldsymbol{\alpha}}(h - \widetilde{h})|_{\beta - |\boldsymbol{\alpha}|_1}$ the Hölder seminorm of the function $\partial^{\boldsymbol{\alpha}}(h - \widetilde{h}) \in \mathcal{C}_r^{\beta - |\boldsymbol{\alpha}|_1}(2K)$, we bound the remainder R by

$$\begin{aligned}|R| &\leq \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1 = \lfloor \beta \rfloor} \Big|\partial^{\boldsymbol{\alpha}}(h - \widetilde{h})(\mathbf{x}_\xi) - \partial^{\boldsymbol{\alpha}}(h - \widetilde{h})(\mathbf{x}_i)\Big| \frac{\tau^{|\boldsymbol{\alpha}|_1}}{\boldsymbol{\alpha}!} \\ &\leq \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1 = \lfloor \beta \rfloor} \Big|\partial^{\boldsymbol{\alpha}}(h - \widetilde{h})\Big|_{\beta - |\boldsymbol{\alpha}|_1} \tau^{\beta - |\boldsymbol{\alpha}|_1} \frac{\tau^{|\boldsymbol{\alpha}|_1}}{\boldsymbol{\alpha}!} \\ &\leq 2K\tau^{\beta}.\end{aligned}$$

Plugging this into the bound (2.A.24) gives

$$\begin{aligned}|(h - \widetilde{h})(\mathbf{x})| &\leq \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1 \leq \lfloor \beta \rfloor} \Big|\partial^{\boldsymbol{\alpha}}(h - \widetilde{h})(\mathbf{x}_i)\Big| \frac{\tau^{|\boldsymbol{\alpha}|_1}}{\boldsymbol{\alpha}!} + K\tau^{\beta} \\ &= \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1 \leq \lfloor \beta \rfloor} \tau^{\beta - |\boldsymbol{\alpha}|_1} \left|\frac{\partial^{\boldsymbol{\alpha}}(h - \widetilde{h})(\mathbf{x}_i)}{\tau^{\beta - |\boldsymbol{\alpha}|_1}}\right| \frac{\tau^{|\boldsymbol{\alpha}|_1}}{\boldsymbol{\alpha}!} + 2K\tau^{\beta}.\end{aligned}$$

In view of definition (2.A.22), we denote $A^{\boldsymbol{\alpha}}(h - \widetilde{h})(\mathbf{x}_i) = \lfloor \partial^{\boldsymbol{\alpha}}(h - \widetilde{h})(\mathbf{x}_i)/\tau^{\beta - |\boldsymbol{\alpha}|_1} \rfloor$ and $B^{\boldsymbol{\alpha}}(h - \widetilde{h})(\mathbf{x}_i) = \partial^{\boldsymbol{\alpha}}(h - \widetilde{h})(\mathbf{x}_i)/\tau^{\beta - |\boldsymbol{\alpha}|_1} - A^{\boldsymbol{\alpha}}(h - \widetilde{h})(\mathbf{x}_i)$. Thus $A^{\boldsymbol{\alpha}}(h - \widetilde{h})(\mathbf{x}_i) = 0$ by assumption on $h, \widetilde{h}$ and $|B^{\boldsymbol{\alpha}}(h - \widetilde{h})(\mathbf{x}_i)| < 1$. We now prove

$$\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1 \leq \lfloor \beta \rfloor} 1/\boldsymbol{\alpha}! \leq e r^{\beta}. \tag{2.A.25}$$

In fact, for any positive integer k, consider the multinomial distribution induced by a fair r-sided die over k independent rolls. The corresponding p.m.f. is $\boldsymbol{\alpha} \mapsto r^{-k} k!/\boldsymbol{\alpha}!$ and is supported on $\{\boldsymbol{\alpha} : |\boldsymbol{\alpha}|_1 = k\}$. Since the p.m.f. sums to one, we have $\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1 = k} 1/\boldsymbol{\alpha}! = r^k/k!$. By summing over $k = 0, \ldots, \lfloor \beta \rfloor$, one finds $\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1 \leq \lfloor \beta \rfloor} 1/\boldsymbol{\alpha}! = \sum_{k=0}^{\lfloor \beta \rfloor} \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1 = k} 1/\boldsymbol{\alpha}! \leq e r^{\beta}$, proving (2.A.25). Combining the discussion above together with the previous bounds, yields

$$|(h - \widetilde{h})(\mathbf{x})| \leq \tau^{\beta} \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1 \leq \lfloor \beta \rfloor} \frac{1}{\boldsymbol{\alpha}!} + 2K\tau^{\beta} \leq \tau^{\beta}\Big(e r^{\beta} + 2K\Big) = \delta, \tag{2.A.26}$$

proving that, if two functions $h, \widetilde{h} \in \mathcal{C}_r^{\beta}(K)$ have $A^{\boldsymbol{\alpha}} h = A^{\boldsymbol{\alpha}} \widetilde{h}$ for all $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}|_1 \leq \lfloor \beta \rfloor$, then $\|h - \widetilde{h}\|_\infty \leq \delta$.

The quantity $\mathcal{N}(\delta, \mathcal{C}_r^\beta(K), \|\cdot\|)$ is bounded above by cardinality $\#\mathcal{A}$ of the set of matrices

$$\mathcal{A} = \Big\{ Ah = (A^{\boldsymbol{\alpha}} h)^\top_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1 \leq \lfloor \beta \rfloor} : h \in \mathcal{C}_r^\beta(K) \Big\}.$$

The rows of the matrix Ah consist of the row vectors $A^{\boldsymbol{\alpha}} h$. Since we consider $\boldsymbol{\alpha} : |\boldsymbol{\alpha}|_1 \leq \lfloor \beta \rfloor$, the matrix Ah can have at most $(\lfloor \beta \rfloor + 1)^r$ rows. Any matrix Ah has moreover M columns.

To complete the counting argument, we first explain the underlying idea. If two neighboring grid points $\mathbf{x}_i, \mathbf{x}_j$ are selected such that $|\mathbf{x}_i - \mathbf{x}_j|_\infty < 2\tau$, say, then, $\partial^{\boldsymbol{\alpha}} h(\mathbf{x}_i) \approx \partial^{\boldsymbol{\alpha}} h(\mathbf{x}_j)$, whenever $|\boldsymbol{\alpha}|_1 < \beta$. Since the ℓ-th column of Ah contains the discretized entries $(\partial^{\boldsymbol{\alpha}} h(\mathbf{x}_\ell))_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1 \leq \lfloor \beta \rfloor}$, the number of possible realizations of the i-th and j-th column vector can be bounded by the possible realizations of the i-th column vector times a factor that describes the number of possible deviations of the values in the j-th column vector.

We now show that this factor is bounded by $2^{\beta+1} c(\beta)$. To see this, observe that Taylor expansion gives

$$\partial^{\boldsymbol{\alpha}} h(\mathbf{x}_j) = \sum_{\mathbf{k}:|\boldsymbol{\alpha}|_1 + |\mathbf{k}|_1 < \lfloor \beta \rfloor} \partial^{\boldsymbol{\alpha}+\mathbf{k}} h(\mathbf{x}_i) \frac{(\mathbf{x}_j - \mathbf{x}_i)^{\mathbf{k}}}{\mathbf{k}!} + \sum_{\mathbf{k}:|\boldsymbol{\alpha}|_1 + |\mathbf{k}|_1 = \lfloor \beta \rfloor} \partial^{\boldsymbol{\alpha}+\mathbf{k}} h(\mathbf{x}_\xi) \frac{(\mathbf{x}_j - \mathbf{x}_i)^{\mathbf{k}}}{\mathbf{k}!},$$

for some $\mathbf{x}_\xi$ on the line with endpoints $\mathbf{x}_i, \mathbf{x}_j$. By replacing $\widetilde{h}$ by 0, h by $\partial^{\boldsymbol{\alpha}} h$, β by $\beta - |\boldsymbol{\alpha}|_1$ and τ by 2τ, we can argue as for (2.A.26) to find

$$\begin{aligned}\left| \partial^{\boldsymbol{\alpha}} h(\mathbf{x}_j) - \sum_{\mathbf{k}:|\boldsymbol{\alpha}|_1 + |\mathbf{k}|_1 \leq \lfloor \beta \rfloor} \tau^{\beta - |\boldsymbol{\alpha}|_1 - |\mathbf{k}|_1} A^{\boldsymbol{\alpha}+\mathbf{k}} h(\mathbf{x}_i) \frac{(\mathbf{x}_j - \mathbf{x}_i)^{\mathbf{k}}}{\mathbf{k}!} \right| &\leq 2^{\beta - |\boldsymbol{\alpha}|_1} c(\beta - |\boldsymbol{\alpha}|_1) \tau^{\beta - |\boldsymbol{\alpha}|_1} \\ &\leq 2^\beta c(\beta) \tau^{\beta - |\boldsymbol{\alpha}|_1}.\end{aligned}$$

This shows that, if the i-th column of Ah is fixed, the values $\partial^{\boldsymbol{\alpha}} h(\mathbf{x}_j)$ range over an interval of length at most $2 \cdot 2^\beta c(\beta) \tau^{\beta - |\boldsymbol{\alpha}|_1}$. The entry $\lfloor \partial^{\boldsymbol{\alpha}} h(\mathbf{x}_j) / \tau^{\beta - |\boldsymbol{\alpha}|_1} \rfloor$ can attain therefore at most $2^{\beta+1} c(\beta) \tau^{\beta - |\boldsymbol{\alpha}|_1} / \tau^{\beta - |\boldsymbol{\alpha}|_1} + 1 = 2^{\beta+1} c(\beta) + 1$ different values. As there are at most $(\beta + 1)^r$ many rows, for fixed i-th column of Ah, the j-th column of Ah can attain at most $(2^{\beta+1} c(\beta) + 1)^{(\beta+1)^r}$ different values.

Without loss of generality, assume that the points $\mathbf{x}_1, \dots, \mathbf{x}_M$ are ordered in such a way that for each $j > 1$, there exists $i < j$, such that $|\mathbf{x}_i - \mathbf{x}_j|_\infty < 2\tau$. This determines then also the ordering of the columns of the matrix Ah. In view of Equation (2.A.23), the first column of Ah can attain at most $(2K\tau^{-\beta} + 1)^{(\beta+1)^r}$ different values. For each of the $M - 1$ remaining columns, we can use the argument above and find

$$\mathcal{N}(\delta, \mathcal{C}_r^\beta(K), \|\cdot\|_\infty) \leq \#\mathcal{A} \leq (2K\tau^{-\beta} + 1)^{(\beta+1)^r} \cdot (2^{\beta+1} c(\beta) + 1)^{(M-1)(\beta+1)^r}.$$

Since $x + y \leq xy$ for all $x, y \geq 2$, using $c(\beta) = er^\beta + 2K \leq 2eKr^\beta$, the bound on M in (2.A.21) and the definition of $\tau = (\delta / c(\beta))^{1/\beta}$, the first assertion of the lemma follows.

For the bound on the constant $Q_1(\beta, r, K)$, we take the logarithm. With $\log(x + 1) \leq \log(2x)$ for all $x > 1$, we get

$$\log \left(\left(\frac{4eK^2 r^\beta}{\delta} + 1 \right)^{(\beta+1)^r} \left(2^{\beta+2} eKr^\beta + 1 \right)^{4^r (\beta+1)^r r^r (2eK)^{\frac{r}{\beta}} \delta^{-\frac{r}{\beta}}} \right)$$

$$
\begin{aligned}
&\le (\beta+1)^r \log\left(\frac{8eK^2r^\beta}{\delta}\right) + 4^r(\beta+1)^r r^r (2eK)^{\frac{r}{\beta}}\delta^{-\frac{r}{\beta}} \log\left(2^{\beta+3}eKr^\beta\right)\\
&=: A_1 + A_2.
\end{aligned}
$$

Observe that $\log(x) < x^a/a$ for all $a, x > 0$, then

$$
\log\left(\frac{8eK^2r^\beta}{\delta}\right) \le \frac{\beta}{r}(8eK^2r^\beta)^{\frac{r}{\beta}}\delta^{-\frac{r}{\beta}} = \beta r^{r-1}(8eK^2)^{\frac{r}{\beta}}\delta^{-\frac{r}{\beta}},
$$

which yields $A_1 \le (\beta+1)^{r+1}r^{r-1}(8eK^2)^{r/\beta}\delta^{-r/\beta}$. Furthermore, using that $\log x < x$ for all $x > 0$,

$$
\log\left(2^{\beta+3}eKr^\beta\right) \le 2(\beta+3) + \beta r + eK \le (r+2)(\beta+3) + eK.
$$

Since $r > 1$, the latter display is smaller than $4(\beta+3)r + eK \le 4eK(\beta+3)r$ and so $A_2 \le 4eK(\beta+3)r \cdot 4^r(\beta+1)^r r^r(2eK)^{r/\beta}\delta^{-r/\beta}$. Putting together the bounds on A_1 and A_2, we find

$$
\begin{aligned}
\frac{A_1 + A_2}{\delta^{-\frac{r}{\beta}}} &\le (\beta+1)^{r+1}r^{r-1}(8eK^2)^{\frac{r}{\beta}} + eK(\beta+3)4^{r+1}r^{r+1}(\beta+1)^r(2eK)^{\frac{r}{\beta}}\\
&\le (1+eK)4^{r+1}(\beta+3)^{r+1}r^{r+1}(8eK^2)^{\frac{r}{\beta}},
\end{aligned}
$$

which matches the definition of $Q_1(\beta, r, K)$ in the statement.

We now prove the entropy bound in (2.A.9). Let $Q_1 = Q_1(\beta, r, K)$, $C_1 = Q_1^{\beta\alpha/(2\beta\alpha+r)}$ and $\mathfrak{r}_n = n^{-\beta\alpha/(2\beta\alpha+r)}$. By construction, $\mathfrak{r}_n^{-r/\beta\alpha} = n\mathfrak{r}_n^2$ and $Q_1C_1^{-r/\beta\alpha} = C_1^2$. For any sequence $\delta_n \ge C_1\mathfrak{r}_n$, the first part of the proof gives

$$
\begin{aligned}
\log\mathcal{N}\left(\delta_n^{\frac{1}{\alpha}}, \mathcal{C}_r^\beta(K), \|\cdot\|_\infty\right) &\le \log\mathcal{N}\left((C_1\mathfrak{r}_n)^{\frac{1}{\alpha}}, \mathcal{C}_{t_i}^{\beta_i}(K), \|\cdot\|_\infty\right)\\
&\le Q_1C_1^{-\frac{r}{\beta\alpha}}\mathfrak{r}_n^{-\frac{r}{\beta\alpha}}\\
&= C_1^2 n\mathfrak{r}_n^2\\
&\le n\delta_n^2.
\end{aligned}
$$

The proof is complete. □

Proof of Lemma 2.A.6. Fix any $\boldsymbol{\beta} \in [\beta_-, \beta_+]^{q+1}$. In a first step we show that, with $R := 5K^q(q+1)$ and $\delta_{in}(\lambda, \boldsymbol{\beta}) = \varepsilon_n(\alpha_i, \beta_i, t_i)^{1/\alpha_i}$, we have

$$
\mathcal{N}\left(R\varepsilon_n(\lambda, \boldsymbol{\beta}), \Theta_n(\lambda, \boldsymbol{\beta}, K), \|\cdot\|_\infty\right) \le \prod_{i=0}^{q} \mathcal{N}\left(3\delta_{in}(\lambda, \boldsymbol{\beta}), \Theta_{i,n}(\lambda, \boldsymbol{\beta}, K), \|\cdot\|_\infty\right). \tag{2.A.27}
$$

For any $i = 0, \ldots, q$, let $g_{i,1}, \ldots, g_{i,N_i}$ be the centers of a $3\delta_{in}(\lambda, \boldsymbol{\beta})$-covering of $\Theta_{i,n}(\lambda, \boldsymbol{\beta}, K)$. Then, any function $g_q \circ \cdots \circ g_0 \in \Theta_n(\lambda, \boldsymbol{\beta}, K)$ belongs to a ball around a composition of centers $g_{q,k_q} \circ \cdots \circ g_{0,k_0}$ for some $\mathbf{k} = (k_0, \ldots, k_q)$ and such that $\|g_i - g_{i,k_i}\|_\infty \le 3\delta_{in}(\lambda, \boldsymbol{\beta})$. By definition of $\Theta_{i,n}(\lambda, \boldsymbol{\beta}, K)$, the components $(g_{ij,k_i})_j$ of g_{i,k_i} satisfy $|g_{ij,k_i}(\mathbf{x}) - g_{ij,k_i}(\mathbf{y})| \le$

$2\delta_{in}(\lambda,\boldsymbol{\beta}) + K|\mathbf{x}-\mathbf{y}|_\infty^{\beta_i\wedge 1}$, for all $\mathbf{x},\mathbf{y}\in[-1,1]^{d_i}$. Using Lemma 2.A.2, the definition of $\delta_{in}(\lambda,\boldsymbol{\beta})$ and the fact that $\alpha_i\le 1$, gives

$$\begin{aligned}\left\|g_q\circ\cdots\circ g_0 - g_{q,k_q}\circ\cdots\circ g_{0,k_0}\right\|_\infty &\le K^q\sum_{i=0}^{q}(2\delta_{in}(\lambda,\boldsymbol{\beta}))^{\alpha_i}+(3\delta_{in}(\lambda,\boldsymbol{\beta}))^{\alpha_i}\\ &\le 5K^q(q+1)\varepsilon_n(\lambda,\boldsymbol{\beta})\\ &= R\varepsilon_n(\lambda,\boldsymbol{\beta}).\end{aligned}$$

Since there are $N_0\times\cdots\times N_q$ centers, this concludes the first part of the proof.

We now focus on the set $\Theta_n(\lambda,B,K)=\cup_{\boldsymbol{\beta}\in B}\Theta_n(\lambda,\boldsymbol{\beta},K)$. Set $\underline{\boldsymbol{\beta}}$ the infimum of B, that is, for any $\boldsymbol{\beta}\in B$ we have $\underline{\beta}_i\le\beta_i$, for all $i=0,\dots,q$. Since B is contained in the closed hypercube $[\beta_-,\beta_+]^{q+1}$, we have $\underline{\boldsymbol{\beta}}\in[\beta_-,\beta_+]^{q+1}$. We now show that, for any $i=0,\dots,q$ and $\boldsymbol{\beta}\in B$, the set $\Theta_{i,n}(\lambda,\boldsymbol{\beta},K)$ is contained in the set $\Theta_{i,n}(\lambda,\underline{\boldsymbol{\beta}},K)$. In fact, by definition, any function $h_i\in\Theta_{i,n}(\lambda,\boldsymbol{\beta},K)$ satisfies $h_{ij}\circ(\cdot)^{-1}_{\mathcal{S}_{ij}}\in\mathcal{D}_i(\lambda,\boldsymbol{\beta},K)$ and, as a consequence of the embedding in Lemma 2.A.1 and the fact that $\delta_{in}(\lambda,\boldsymbol{\beta})\le\delta_{in}(\lambda,\underline{\boldsymbol{\beta}})$ by the rate comparison condition (ii) in Assumption 2.4.2, one has $\mathcal{C}^{\beta_i}_{t_i}(K)+\mathbb{B}_\infty(2\delta_{in}(\lambda,\boldsymbol{\beta}))\subseteq\mathcal{C}^{\underline{\beta}_i}_{t_i}(K)+\mathbb{B}_\infty(2\delta_{in}(\lambda,\underline{\boldsymbol{\beta}}))$. Thus, $\mathcal{D}_i(\lambda,\boldsymbol{\beta},K)\subseteq\mathcal{D}_i(\lambda,\underline{\boldsymbol{\beta}},K)$ and $\Theta_{i,n}(\lambda,\boldsymbol{\beta},K)\subseteq\Theta_{i,n}(\lambda,\underline{\boldsymbol{\beta}},K)$. Together with (2.A.27), we obtain

$$\begin{aligned}\mathcal{N}\left(R\varepsilon_n(\lambda,\underline{\boldsymbol{\beta}}),\Theta_n(\lambda,B,K),\|\cdot\|_\infty\right)&\le\mathcal{N}\left(R\varepsilon_n(\lambda,\underline{\boldsymbol{\beta}}),\Theta_n(\lambda,\underline{\boldsymbol{\beta}},K),\|\cdot\|_\infty\right)\\ &\le\prod_{i=0}^{q}\mathcal{N}\left(3\delta_{in}(\lambda,\underline{\boldsymbol{\beta}}),\Theta_{i,n}(\lambda,\underline{\boldsymbol{\beta}},K),\|\cdot\|_\infty\right).\end{aligned}$$

We now use the definition of $\Theta_{i,n}(\lambda,\underline{\boldsymbol{\beta}},K)$ and upper bound the metric entropy by removing the constraint $\mathbb{B}_\infty(1)$ in the definition of $\mathcal{D}_i(\lambda,\underline{\boldsymbol{\beta}},K)$. This gives

$$\mathcal{N}\left(3\delta_{in}(\lambda,\underline{\boldsymbol{\beta}}),\Theta_{i,n}(\lambda,\underline{\boldsymbol{\beta}},K),\|\cdot\|_\infty\right)\le\prod_{j=i}^{d_{i+1}}\mathcal{N}\left(3\delta_{in}(\lambda,\underline{\boldsymbol{\beta}}),\mathcal{C}^{\underline{\beta}_i}_{t_i}(K)+\mathbb{B}_\infty\left(2\delta_{in}(\lambda,\underline{\boldsymbol{\beta}})\right),\|\cdot\|_\infty\right).$$

Any function in $\mathcal{C}^{\underline{\beta}_i}_{t_i}(K)+\mathbb{B}_\infty(2\delta_{in}\left(\lambda,\underline{\boldsymbol{\beta}}\right))$ is at most, in sup-norm distance, $2\delta_{in}(\lambda,\underline{\boldsymbol{\beta}})$-away from some function in $\mathcal{C}^{\underline{\beta}_i}_{t_i}(K)$. Therefore, by applying Lemma 2.A.5 with $r=t_i$, $\beta=\underline{\beta}_i$, $\alpha=\underline{\alpha}_i$, and $\delta_n=\delta_{in}(\eta)$,

$$\begin{aligned}\mathcal{N}\left(3\delta_{in}(\lambda,\underline{\boldsymbol{\beta}}),\mathcal{C}^{\underline{\beta}_i}_{t_i}(K)+\mathbb{B}_\infty\left(2\delta_{in}(\lambda,\underline{\boldsymbol{\beta}})\right),\|\cdot\|_\infty\right)&\le\mathcal{N}\left(\delta_{in}(\lambda,\underline{\boldsymbol{\beta}}),\mathcal{C}^{\underline{\beta}_i}_{t_i}(K),\|\cdot\|_\infty\right)\\ &\le e^{n\varepsilon_n(\lambda,\underline{\boldsymbol{\beta}})^2}.\end{aligned}$$

Assumption 2.4.2 ensures that $\varepsilon_n(\lambda,\underline{\boldsymbol{\beta}})\le Q\varepsilon_n(\lambda,\boldsymbol{\beta})$, thus combining the last inequalities gives

$$\begin{aligned}\log\mathcal{N}\left(RQ\varepsilon_n(\lambda,\boldsymbol{\beta}),\Theta_n(\lambda,B,K),\|\cdot\|_\infty\right)&\le\log\mathcal{N}\left(R\varepsilon_n(\lambda,\underline{\boldsymbol{\beta}}),\Theta_n(\lambda,B,K),\|\cdot\|_\infty\right)\\ &\le\sum_{i=0}^{q}\sum_{j=1}^{d_{i+1}}n\varepsilon_n(\lambda,\underline{\boldsymbol{\beta}})^2\end{aligned}$$

$$\begin{aligned}&= |\mathbf{d}|_1 n\varepsilon_n(\lambda, \underline{\boldsymbol{\beta}})^2\\ &\leq Q^2|\mathbf{d}|_1 n\varepsilon_n(\lambda, \boldsymbol{\beta})^2.\end{aligned}$$

Since $R = 5K^q(q+1)$, we use that $(q+1) \leq |\mathbf{d}|_1 \leq \log(2\log n)$, together with $\log(2\log n) \leq 2\log n$. Thus, $K^q(q+1) \leq (2\log n)^{1+\log K}$ and, with $R_n = 5Q(2\log n)^{1+\log K}$,

$$\log \mathcal{N}\left(R_n\varepsilon_n(\lambda, \boldsymbol{\beta}), \Theta_n(\lambda, B, K), \|\cdot\|_\infty\right) \leq \frac{R_n^2}{25} n\varepsilon_n(\lambda, \boldsymbol{\beta})^2,$$

which concludes the proof. □

2.A.2 Proofs for Section 2.5

Proof of Lemma 2.5.1. Fix β, r and let ε_T be such that $\varphi^{(\beta,r,K)}(\varepsilon_T) \leq T\varepsilon_T^2$ for all $T \geq 1$. For this choice of (β, r), we show that the concentration function inequality (2.3.2) holds for any $0 < \alpha \leq 1$ with $\varepsilon_n(\alpha, \beta, r) := \varepsilon_{m_n}^\alpha$, where the sequence m_n is chosen such that $m_n\varepsilon_{m_n}^{2-2\alpha} \leq n$. To see this, observe that

$$\varphi^{(\beta,r,K)}\big(\varepsilon_n(\alpha,\beta,r)^{1/\alpha}\big) = \varphi^{(\beta,r,K)}\big(\varepsilon_{m_n}\big) \leq m_n\varepsilon_{m_n}^2 \leq n\varepsilon_{m_n}^{2\alpha} = n\varepsilon_n(\alpha,\beta,r)^2.$$

By Lemma 3 in [17], the function $u \mapsto \varphi^{(\beta,r,K)}(u)$ is strictly decreasing on $u \in (0, +\infty)$, thus any $\overline{\varepsilon}_n(\alpha,\beta,r) \geq \varepsilon_n(\alpha,\beta,r)$ satisfies

$$\varphi^{(\beta,r,K)}\big(\overline{\varepsilon}_n(\alpha,\beta,r)^{1/\alpha}\big) \leq \varphi^{(\beta,r,K)}\big(\varepsilon_n(\alpha,\beta,r)^{1/\alpha}\big) \leq n\varepsilon_n(\alpha,\beta,r)^2 \leq n\overline{\varepsilon}_n(\alpha,\beta,r)^2,$$

which concludes the proof. □

Proof of Lemma 2.5.2. (i): By Lemma 2.5.1, the sequence $\varepsilon_n(\alpha,\beta,r)$ can be obtained from $\varepsilon_n(1,\beta,r)$ via $\varepsilon_n(\alpha,\beta,r) = \varepsilon_{m_n}(1,\beta,r)^\alpha$, for any sequence m_n such that $m_n\varepsilon_{m_n}(1,\beta,r)^{2-2\alpha} \leq n$. We verify this for the sequence $m_n = C_3(\log n)^{-C_4}n^{(2\beta+r)/(2\beta\alpha+r)}$ with

$$C_3 := \left[C_1(2\beta+1)^{C_2}\right]^{-\frac{(2-2\alpha)(2\beta+r)}{2\beta\alpha+r}} \quad \text{and} \quad C_4 := \frac{(2-2\alpha)(2\beta+r)}{2\beta\alpha+r}C_2.$$

Since $C_1 \geq 1$ and $n \geq 3$, we must have $C_3 \leq 1$, $(\log n)^{-C_4} \leq 1$ and thus also $\log(m_n) \leq (2\beta+1)\log(n)$. Consequently,

$$\begin{aligned}m_n\varepsilon_{m_n}(1,\beta,r)^{2-2\alpha} &= m_n\left[C_1(\log m_n)^{C_2}m_n^{-\frac{\beta}{2\beta+r}}\right]^{2-2\alpha}\\ &= C_1^{2-2\alpha}(\log m_n)^{C_2(2-2\alpha)}m_n^{\frac{2\beta+r-2\beta+2\beta\alpha}{2\beta+r}}\\ &\leq C_1^{2-2\alpha}\big((2\beta+1)\log(n)\big)^{C_2(2-2\alpha)}m_n^{\frac{2\beta\alpha+r}{2\beta+r}}\\ &\leq C_1^{2-2\alpha}(2\beta+1)^{C_2(2-2\alpha)}C_3^{\frac{2\beta\alpha+r}{2\beta+r}}(\log n)^{C_2(2-2\alpha)-C_4\frac{2\beta\alpha+r}{2\beta+r}}n\\ &= n.\end{aligned}$$

By Lemma 2.5.1, any sequence $\overline{\varepsilon}_n(\alpha,\beta,r) \geq \varepsilon_n(\alpha,\beta,r)$ is still a solution to the concentration function inequality (2.3.2). We now derive a simple upper bound for $\varepsilon_n(\alpha,\beta,r)$. Using that $\alpha \leq 1$ and $\log(m_n) \leq (2\beta+1)\log(n)$, we find

$$\begin{aligned}\varepsilon_n(\alpha,\beta,r) &= \varepsilon_{m_n}(1,\beta,r)^{\alpha}\\ &\leq C_1^{\alpha}\log(m_n)^{\alpha C_2} m_n^{-\frac{\beta\alpha}{2\beta+r}}\\ &\leq C_1(2\beta+1)^{C_2} C_3^{-\frac{\beta\alpha}{2\beta+r}}(\log n)^{C_2+C_4\frac{\beta\alpha}{2\beta\alpha+r}} n^{-\frac{\beta\alpha}{2\beta\alpha+r}}.\end{aligned}$$

Using the definition of C_3 together with $0<\alpha\leq 1$, $(2-2\alpha)\leq 2$ and $2\beta\alpha/(2\beta\alpha+r)\leq 1$ yields

$$C_3^{-\frac{\beta\alpha}{2\beta+r}} = \left[C_1(2\beta+1)^{C_2}\right]^{\frac{(2-2\alpha)\beta\alpha}{2\beta\alpha+r}} \leq C_1(2\beta+1)^{C_2}.$$

Similarly, we get

$$C_4\frac{\beta\alpha}{2\beta\alpha+r} \leq C_2\frac{(2-2\alpha)(2\beta+r)}{2\beta\alpha+r}\cdot\frac{1}{2} \leq \frac{2\beta+r}{2\beta\alpha+r} \leq C_2(2\beta+1).$$

The two previous displays recover the first assertion.

(ii): By assumption, for any $\delta\in(0,1)$, we have $\varphi^{(\beta,r,K)}(\delta) \leq C_1'(\log\delta^{-1})^{C_2'}\delta^{-\frac{r}{\beta}}$. We now choose $\delta = \varepsilon_n(\alpha,\beta,r)^{1/\alpha}$ and $\varepsilon_n(\alpha,\beta,r) = C_1'(\log n)^{C_2'} n^{-\beta\alpha/(2\beta\alpha+r)}$. Since $C_1'\geq 1$, $C_2'\geq 0$, and $\log n \geq 1$,

$$\log \varepsilon_n(\alpha,\beta,r)^{-\frac{1}{\alpha}} \leq -\frac{1}{\alpha}\log n^{-\frac{\beta\alpha}{2\beta\alpha+r}} \leq \frac{\beta}{2\beta+r}\log n \leq \log n.$$

Similarly,

$$\varepsilon_n(\alpha,\beta,r)^{-\frac{r}{\beta\alpha}} \leq \left(n^{-\frac{\beta\alpha}{2\beta\alpha+r}}\right)^{-\frac{r}{\beta\alpha}} = n\cdot n^{-\frac{2\beta\alpha}{2\beta\alpha+r}} \leq \frac{n\varepsilon_n(\alpha,\beta,r)^2}{(C_1')^2(\log n)^{2C_2'}},$$

and therefore,

$$\varphi^{(\beta,r,K)}\left(\varepsilon_n(\alpha,\beta,r)^{\frac{1}{\alpha}}\right) \leq C_1'\left(\log\varepsilon_n(\alpha,\beta,r)^{-\frac{1}{\alpha}}\right)^{C_2'}\varepsilon_n(\alpha,\beta,r)^{-\frac{r}{\beta\alpha}} \leq n\varepsilon_n(\alpha,\beta,r)^2.$$

□

Proof of Lemma 2.5.3. It is sufficient to show that for $n>1$, any composition graph $\lambda = (q,\mathbf{d},\mathbf{t},\mathcal{S})$, and any $\boldsymbol{\beta'} = (\beta_0',\ldots,\beta_q'), \boldsymbol{\beta} = (\beta_0,\ldots,\beta_q)\in I(\lambda)$ satisfying $\beta_i'\leq\beta_i\leq\beta_i' + 1/\log^2 n$ for all $i=0,\ldots,q$, the rates relative to the composition structures $\eta = (\lambda,\boldsymbol{\beta})$ and $\eta' = (\lambda,\boldsymbol{\beta'})$ satisfy $\varepsilon_n(\eta)\leq\varepsilon_n(\eta')\leq e^{\beta_+}\varepsilon_n(\eta)$.

Since $\varepsilon_n(\eta) = \widetilde{C}_1(\eta)(\log n)^{\widetilde{C}_2(\eta)}\mathfrak{r}_n(\eta)$ with $\widetilde{C}_j(\eta) := \max_{i=0,\ldots,q}\sup_{\beta\in[\beta_-,\beta_+]} C_j(\beta,t_i)$, $j\in\{1,2\}$, we have that $\widetilde{C}_j(\eta) = \widetilde{C}_j(\eta')$ and it is thus sufficient to prove $\mathfrak{r}_n(\eta)\leq\mathfrak{r}_n(\eta')\leq e^{\beta_+}\mathfrak{r}_n(\eta)$.

Using that $\mathfrak{r}_n(\eta) = \max_{i=0,\ldots,q} n^{-\beta_i\alpha_i/(2\beta_i\alpha_i+t_i)}$ and the fact that the function $x\mapsto x/(2x+t_i)$ is strictly increasing for $x>0$ (its derivative is $x\mapsto t_i/(2x+t_i)^2$), the first inequality $\mathfrak{r}_n(\eta) \leq \mathfrak{r}_n(\eta')$ follows. For the second inequality, rewriting the expressions and simplifying the exponents gives

$$\frac{\mathfrak{r}_n(\eta')}{\mathfrak{r}_n(\eta)} \leq \max_{i=0,\ldots,q} \min_{j=0,\ldots,q} n^{-\frac{\beta_i'\alpha_i'}{2\beta_i'\alpha_i'+t_i}+\frac{\beta_j\alpha_j}{2\beta_j\alpha_j+t_j}} \leq \max_{i=0,\ldots,q} n^{-\frac{\beta_i'\alpha_i'}{2\beta_i'\alpha_i'+t_i}+\frac{\beta_i\alpha_i}{2\beta_i\alpha_i+t_i}} \leq \max_{i=0,\ldots,q} n^{\beta_i\alpha_i-\beta_i'\alpha_i'}.$$

We conclude the proof by showing that $|\beta_i\alpha_i - \beta_i'\alpha_i'| \leq \beta_+/\log n$. For $u, u', v, v' \geq 0$, we have that $|uv - u'v'| \leq u|v - v'| + v'|u - u'|$. In particular, if $u, v' \leq 1$, then also $|uv - u'v'| \leq |v - v'| + |u - u'|$. By iterating this argument, we find that

$$|\alpha_i - \alpha_i'| \leq \sum_{\ell=i+1}^{q} \big|(1 \wedge \beta_\ell) - (1 \wedge \beta_\ell')\big| \leq \sum_{\ell=i+1}^{q} \big|\beta_\ell - \beta_\ell'\big| \leq \frac{q-i}{\log^2 n}$$

and thus, $\beta_i\alpha_i - \beta_i'\alpha_i' \leq \beta_+|\alpha_i - \alpha_i'| + \alpha_i'|\beta_i - \beta_i'| \leq \beta_+(q+1)/\log^2 n$. Since we are restricting ourselves to graphs λ such that $|\mathbf{d}|_1 = 1 + \sum_{i=0}^q d_i \leq \log(2\log n)$, we have as well $q + 1 \leq \log(2\log n)$. Since $\log x \leq x/2$ for all $x > 0$, we find $(q + 1) \leq \log n$. This gives $(q+1)/\log^2 n \leq 1/\log n$ and thus Assumption 2.4.2 (ii) holds with $Q = e^{\beta_+}$. □

Proof of Lemma 2.5.4. For two functions $g_k \in \mathcal{C}_1^{\beta_k}(1)$, $k = 1, 2$, and $\beta_1, \beta_2 \leq 1$, we have that $|g_2(g_1(x)) - g_2(g_1(y))| \leq |g_1(x) - g_1(y)|^{\beta_2} \leq |x - y|^{\beta_1\beta_2}$. Hence, $g_2 \circ g_1 \in \mathcal{C}_1^{\beta_1\beta_2}(1)$. We now write

$$f = h_q \circ \cdots \circ h_0 = h_q \circ \cdots \circ h_{j+1} \circ \widetilde{h}_j \circ h_{j-2} \circ \cdots \circ h_0,$$

with $\widetilde{h}_j := h_j \circ h_{j-1}$. The right hand side can be written as composition structure $\eta' := (q - 1, \mathbf{d}_{-j}, \mathbf{t}_{-j}, \mathcal{S}_{-j}, \boldsymbol{\beta}')$, with $\mathbf{d}_{-j}, \mathbf{t}_{-j}, \mathcal{S}_{-j}, \boldsymbol{\beta}'$ as defined in the statement of the lemma. Due to $\beta_+ \leq 1$, we have $\mathfrak{r}_n(\eta) = \max_{i=0,\ldots,q} n^{-\frac{\gamma_i}{2\gamma_i+t_i}}$, with $\gamma_i = \prod_{\ell=i}^q \beta_\ell$ and it follows that $\mathfrak{r}_n(\eta) = \mathfrak{r}_n(\eta')$. □

2.A.3 Proofs for Section 2.6

Proof of Lemma 2.6.1. For the first part of the proof, take a kernel ϕ with $R_\beta := \int_{\mathbb{R}^r} |\mathbf{v}|_\infty^\beta \phi(\mathbf{v})d\mathbf{v} < +\infty$. Using that $h \in \mathcal{C}_r^\beta(K)$ and the change of variable $\mathbf{v}' = \mathbf{v}/\sigma$, we immediately get

$$\begin{aligned}
|(h * \phi_\sigma)(\mathbf{u}) - h(\mathbf{u})| &\leq \int_{\mathbb{R}^r} \phi_\sigma(\mathbf{v})|h(\mathbf{u} - \mathbf{v}) - h(\mathbf{u})|\, d\mathbf{v} \\
&\leq K \int_{\mathbb{R}^r} |\mathbf{v}|_\infty^\beta \phi_\sigma(\mathbf{v})\, d\mathbf{v} \\
&= K \int_{\mathbb{R}^r} |\mathbf{v}|_\infty^\beta \sigma^{-r}\phi(\mathbf{v}/\sigma)\, d\mathbf{v} \\
&= K \int_{\mathbb{R}^r} |\sigma\mathbf{v}'|_\infty^\beta \phi(\mathbf{v}')\, d\mathbf{v}' \\
&\leq K R_\beta \sigma^\beta.
\end{aligned}$$

This shows $\|h * \phi_\sigma - h\|_\infty \leq K R_\beta \sigma^\beta$ and concludes the first part of the proof.

We now deal with the RKHS norm. Notice that

$$\begin{aligned}(h*\phi_\sigma)(\mathbf{u}) &= \int_{\mathbb{R}^r} \widehat{h}(\boldsymbol{\xi})\widehat{\phi_\sigma}(\boldsymbol{\xi})e^{-i\mathbf{u}^\top\boldsymbol{\xi}}\frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}}\\ &= C_\beta^{1/2}\int_{\mathbb{R}^r} \widehat{h}(\boldsymbol{\xi})\widehat{\phi_\sigma}(\boldsymbol{\xi})|\boldsymbol{\xi}|_2^{\beta+r/2}\frac{e^{-i\mathbf{u}^\top\boldsymbol{\xi}}-1}{C_\beta^{1/2}|\boldsymbol{\xi}|_2^{\beta+r/2}}\frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}} + (h*\phi_\sigma)(0).\end{aligned}$$

The RKHS of $Z+X^\beta$ is the direct sum of the space of constant functions $\mathbb{H}^Z$ and $\mathbb{H}^\beta$. If the term $(h*\phi_\sigma)(0)$ is finite, it is a constant and thus belongs to $\mathbb{H}^Z$. Then, the function $h*\phi_\sigma$ is a candidate element of $\mathbb{H}^Z\oplus\mathbb{H}^\beta$ since $h*\phi_\sigma-(h*\phi_\sigma)(0)$ has been represented as a potential element of the RKHS of X^β. We now bound their norm using the isometry property of the norm $\|\cdot\|_{\mathbb{H}^\beta}$, so that

$$\|h*\phi_\sigma\|^2_{\mathbb{H}^Z\oplus\mathbb{H}^\beta} \le 2|(h*\phi_\sigma)(0)|^2 + 2C_\beta\int_{\mathbb{R}^r}|\widehat{h}(\boldsymbol{\xi})|^2|\widehat{\phi_\sigma}(\boldsymbol{\xi})|^2|\boldsymbol{\xi}|_2^{2\beta+r}\frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}}.$$

By the change of variable $\boldsymbol{\xi}'=\sigma\boldsymbol{\xi}$, the fact that $\boldsymbol{\xi}\mapsto(1+|\boldsymbol{\xi}|_2)^\beta\widehat{h}(\boldsymbol{\xi})$ has L^2-norm bounded by K, the property $\widehat{\phi_\sigma}(\boldsymbol{\xi})=\widehat{\phi}(\sigma\boldsymbol{\xi})$, and choosing ϕ such that $M^2:=\sup_{\boldsymbol{\xi}\in\mathbb{R}^r}|\widehat{\phi}(\boldsymbol{\xi})|^2|\boldsymbol{\xi}|_2^r<+\infty$, we can bound

$$\begin{aligned}&\int|\widehat{h}(\boldsymbol{\xi})|^2|\widehat{\phi_\sigma}(\boldsymbol{\xi})|^2|\boldsymbol{\xi}|_2^{2\beta+r}\frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}}\\ &= \sigma^{-2\beta-2r}\int_{\mathbb{R}^r}|\widehat{h}(\boldsymbol{\xi}/\sigma)|^2|\widehat{\phi}(\boldsymbol{\xi})|^2|\boldsymbol{\xi}|_2^{2\beta+r}\frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}}\\ &= \sigma^{-2\beta-2r}\int_{\mathbb{R}^r}\frac{(1+|\boldsymbol{\xi}/\sigma|_2)^{2\beta}}{(1+|\boldsymbol{\xi}/\sigma|_2)^{2\beta}}|\widehat{h}(\boldsymbol{\xi}/\sigma)|^2|\widehat{\phi}(\boldsymbol{\xi})|^2|\boldsymbol{\xi}|_2^{2\beta+r}\frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}}\\ &\le \sigma^{-2\beta-2r}\sup_{\boldsymbol{\xi}\in\mathbb{R}^r}\frac{|\widehat{\phi}(\boldsymbol{\xi})|^2|\boldsymbol{\xi}|_2^{2\beta+r}}{(1+|\boldsymbol{\xi}/\sigma|_2)^{2\beta}}\int_{\mathbb{R}^r}(1+|\boldsymbol{\xi}/\sigma|_2)^{2\beta}|\widehat{h}(\boldsymbol{\xi}/\sigma)|^2\frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}}\\ &\le \sigma^{-2\beta-2r}\sup_{\boldsymbol{\xi}\in\mathbb{R}^r}\frac{|\widehat{\phi}(\boldsymbol{\xi})|^2|\boldsymbol{\xi}|_2^{2\beta+r}}{|\boldsymbol{\xi}/\sigma|_2^{2\beta}}\int_{\mathbb{R}^r}\sigma^r(1+|\boldsymbol{\xi}|_2)^{2\beta}|\widehat{h}(\boldsymbol{\xi})|^2\frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}}\\ &\le K^2M^2\sigma^{-r}.\end{aligned}$$

Similarly, by choosing ϕ such that $N^2:=(2\pi)^{-r/2}\int_{\mathbb{R}^r}|\widehat{\phi}(\boldsymbol{\xi})|^2\,d\boldsymbol{\xi}<+\infty$, we obtain

$$\begin{aligned}|(h*\phi_\sigma)(0)|^2 &\le \Big(\int_{\mathbb{R}^r}|\widehat{h}(\boldsymbol{\xi})||\widehat{\phi_\sigma}(\boldsymbol{\xi})|\frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}}\Big)^2\\ &= \Big(\int_{\mathbb{R}^r}|\widehat{h}(\boldsymbol{\xi})|(1+|\boldsymbol{\xi}|_2)^\beta\frac{|\widehat{\phi}(\sigma\boldsymbol{\xi})|}{(1+|\boldsymbol{\xi}|_2)^\beta}\frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}}\Big)^2\\ &\le \Big(\int_{\mathbb{R}^r}|\widehat{h}(\boldsymbol{\xi})|^2(1+|\boldsymbol{\xi}|_2)^{2\beta}\frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}}\Big)\Big(\int_{\mathbb{R}^r}\frac{|\widehat{\phi}(\sigma\boldsymbol{\xi})|^2}{(1+|\boldsymbol{\xi}|_2)^{2\beta}}\frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}}\Big)\\ &\le K^2\sigma^{-r}\int_{\mathbb{R}^r}\frac{|\widehat{\phi}(\boldsymbol{\xi})|^2}{(1+|\boldsymbol{\xi}/\sigma|_2)^{2\beta}}\frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}}\\ &\le K^2N^2\sigma^{-r}.\end{aligned}$$

The proof is complete by taking $L_\beta^2:=2(C_\beta+1)(M^2\vee N^2)$. Since this will be useful for the proof of Lemma 2.6.2, the explicit form of the constant C_β is given in (3.67) in [27] as

$$C_\beta=\frac{\pi^{1/2}\Gamma(\beta+1/2)}{2^{r/2}\beta\Gamma(2\beta)\sin(\pi\beta)\Gamma(\beta+r/2)},$$

and depends only on $\beta,\ r$. □

Proof of Lemma 2.6.2. We show that:

(1.) Lemma 2.5.2 (ii) holds for some $C_1'(\beta,r) \geq 1$ and $C_2'(\beta,r) = 0$.

(2.) Any sequence $\varepsilon_n(\alpha,\beta,r) \geq C_1'(\beta,r)n^{-\beta\alpha/(2\beta\alpha+r)}$ solves (2.3.2).

(3.) Assumption 2.4.2 (i) holds for $\varepsilon_n(\alpha,\beta,r) = C_1(\beta,r)n^{-\beta\alpha/(2\beta\alpha+r)}$ with $C_1(\beta,r) := C_1'(\beta,r) \vee Q_1(\beta,r,K)^{\beta/(2\beta+r)}$.

(4.) $\sup_{\beta\in[\beta_-,\beta_+]} C_1(\beta,r) < +\infty$.

(5.) Assumption 2.4.2 (ii) holds for $\varepsilon_n(\eta)$ of the form (2.5.3).

By Lemma 2.5.2, (1.) $\implies$ (2.) $\implies$ (3.) and by Lemma 2.5.3, (4.) $\implies$ (5.). Thus it remains to prove (1.) and (4.).

Proof of (1.): Fix $\beta \in [\beta_-,\beta_+]$. We have to show that, for all $\delta \in (0,1)$, $\varphi^{(\beta,r,K)}(\delta) \leq C_1'(\beta,r)\delta^{-r/\beta}$, for some constant $C_1'(\beta,r) \geq 1$ depending only on $\beta,\ r,\ K$. We denote the small ball probability term by $\varphi_0^{(\beta,r)}(\delta) := -\log \mathbb{P}(\|Z+X^\beta\|_\infty \leq \delta)$ and show that

$$(A): \sup_{\delta\in(0,1)} \delta^{r/\beta}\varphi_0^{(\beta,r)}(\delta) < +\infty, \quad (B): \sup_{\delta\in(0,1)} \delta^{r/\beta}\big(\varphi^{(\beta,r,K)}(\delta) - \varphi_0^{(\beta,r)}(\delta)\big) < +\infty.$$

To prove (A), observe that the process $Z+X^\beta$ is the sum of two independent processes, thus its small ball probability can be bounded by $\log \mathbb{P}(\|Z+X^\beta\|_\infty < \delta) \geq \log \mathbb{P}(\|Z\|_\infty < \delta/2) + \log \mathbb{P}(\|X^\beta\|_\infty < \delta/2)$. It is then sufficient to study the small ball probabilities of Z and X^β, separately. We now show the following condition, which implies (A),

$$\sup_{\delta\in(0,1)} -\delta^{r/\beta}\log \mathbb{P}(\|X^\beta\|_\infty < \delta) < +\infty, \quad \sup_{\delta\in(0,1)} -\delta^{r/\beta}\log \mathbb{P}(|Z| < \delta) < +\infty. \tag{2.A.28}$$

Sharp bounds are known, see Theorem 5.1 in [59], for the small ball probability of the fractional Brownian motion X^β. In particular, for $0<\delta<1$, we have $-\log \mathbb{P}(\|X^\beta\|_\infty \leq \delta) \leq c_X(\beta,r)\delta^{-r/\beta}$ for a finite constant $c_X(\beta,r)$ depending only on $\beta,\ r$. Since Z is a standard normal, we have $\mathbb{P}(|Z| \leq \delta) = (2\pi)^{-1/2}\int_{-\delta}^{\delta} e^{-x^2/2}dx \geq 2\delta e^{-\delta^2/2}/\sqrt{2\pi}$. With the universal constant $c := 2/\sqrt{2\pi}$, this gives $-\log \mathbb{P}(|Z| \leq \delta) \leq \log(c^{-1}\delta^{-1}) + \delta^2/2$. Therefore, using $\log(x) \leq x^a/a$ for all $x>1, a>0$, we get

$$\sup_{\delta\in(0,1)} -\delta^{r/\beta}\log \mathbb{P}(|Z|<\delta) \leq \sup_{\delta\in(0,1)} \delta^{r/\beta}\Big(\frac{\beta}{r}\left(\frac{1}{c\delta}\right)^{r/\beta} + \frac{\delta^2}{2}\Big) = \frac{\beta}{c^{r/\beta}r} + \frac{1}{2} =: c_Z(\beta,r).$$

This concludes the proof of (A).

To prove (B), we apply Lemma 2.6.1. In particular, with finite constants $R(\beta,r), L(\beta,r)$ depending only on $\beta,\ r$, take $\sigma = (KR(\beta,r))^{-1/\beta}\delta^{1/\beta}$. Then, any function $h \in \mathcal{C}_r^\beta(K) \cap \mathcal{W}_r^\beta(K)$ can be well approximated by the convolution $h * \phi_\sigma$ in such a way that $\|h - h * \phi_\sigma\|_\infty \leq \delta$ and $\|h*\phi_\sigma\|^2_{\mathbb{H}^Z\oplus\mathbb{H}^\beta} \leq K^2L(\beta,r)^2\delta^{-r/\beta}$. This proves (B) because it gives

$$\sup_{\delta\in(0,1)} \frac{\varphi^{(\beta,r,K)}(\delta) - \varphi_0^{(\beta,r)}(\delta)}{\delta^{-r/\beta}} \leq \sup_{\delta\in(0,1)} \frac{K^2L(\beta,r)^2\delta^{-r/\beta}}{\delta^{-r/\beta}} = K^2L(\beta,r)^2.$$

We have thus concluded the proof of (1.), that is, for any $\beta \in [\beta_-, \beta_+]$, condition (ii) in Lemma 2.5.2 holds with finite constants

$$C_1'(\beta, r) := c_X(\beta, r) + c_Z(\beta, r) + K^2 L(\beta, r)^2, \quad C_2'(\beta, r) = 0.$$

Proof of (4.): With the definition of $C_1(\beta, r)$, we want to show that

$$\sup_{\beta \in [\beta_-, \beta_+]} C_1'(\beta, r) \vee Q_1(\beta, r, K)^{\beta/(2\beta+r)} < +\infty.$$

The constant $Q_1(\beta, r, K)$ is given explicitly in Lemma 2.A.5 and depends continuously on $\beta > 0$. Thus, $\sup_{\beta \in [\beta_-, \beta_+]} Q_1(\beta, r, K) =: \widetilde{Q}_1 < +\infty$. Since the function $\beta \mapsto \beta/(2\beta + r)$ is increasing for $\beta > 0$, we also have $Q_1(\beta, r, K)^{\beta/(2\beta+r)} \le \widetilde{Q}_1^{\beta_+/(2\beta_+ + r)}$.

In the previous part of the proof we have found $C_1'(\beta, r) = c_Z(\beta, r) + c_X(\beta, r) + K^2 L(\beta, r)^2$, thus it remains to prove

$$\sup_{\beta \in [\beta_-, \beta_+]} c_Z(\beta, r) + c_X(\beta, r) + K^2 L(\beta, r)^2 < +\infty. \tag{2.A.29}$$

By examining the proof of Lemma 2.6.1, we know that $\sup_{\beta \in [\beta_-, \beta_+]} K^2 L(\beta, r)^2 < +\infty$. The explicit form of $c_Z(\beta)$ is given in (2.A.28) and so, with $c = 2/\sqrt{2\pi}$,

$$\sup_{\beta \in [\beta_-, \beta_+]} c_Z(\beta, r) \le \frac{\beta_+}{r c^{\beta_+}} + \frac{1}{2} < +\infty.$$

We now show that the properties of $c_X(\beta, r)$ can be deduced from Theorem 5.2 in [59]. We observe that $\mathbb{E}[|X^\beta(\mathbf{u}) - X^\beta(\mathbf{u}')|^2] = |\mathbf{u} - \mathbf{u}'|_2^{2\beta}$. Furthermore, the function $\beta \mapsto \mathbb{E}[X^\beta(\mathbf{u}) X^\beta(\mathbf{u}')]$ is continuous for all fixed $\mathbf{u}, \mathbf{u}' \in [-1, 1]^r$. In the notation of Theorem 5.2 in [59], we can take $\sigma_\beta(\delta) := \delta^\beta$ and check that, with $c_1 := 1/2^{\beta_+}$ and $c_2 := 1$, $c_1 \sigma_\beta(2\delta \wedge 1) \le \sigma_\beta(\delta) \le c_2 \sigma_\beta(2\delta \wedge 1)$ for all $\delta \in (0, 1)$. The constants c_1, c_2 are chosen to be independent of β in the compact interval $[\beta_-, \beta_+]$. From this, one obtains a constant $c_X(r)$ that only depends on c_1, c_2 and such that $-\log \mathbb{P}(\|X^\beta\|_\infty < \delta) \le c_X(r) \delta^{-r/\beta}$ for all $\delta > 0$. This shows that the quantity $c_X(\beta, r)$ in (2.A.29) can be replaced by $c_X(r)$ and thus is bounded, concluding the proof of (4.). □

Proof of Lemma 2.6.3. Using the definition of X^β yields

$$\|X^\beta\|_{\infty,\infty,\beta} = \max_{j=1,\dots,J_\beta} \frac{1}{\sqrt{jr}} \max_{k=1,\dots,2^{jr}} |Z_{j,k}|.$$

It is known that $\mathbb{E}[\max_{k=1,\dots,2^{jr}} Z_{j,k}] \le \sqrt{2 \log(2^{jr})}$, a reference is Lemma 2.3 in [62]. For $K' > 1$, using symmetry of $Z_{j,k}$ and the Borell-TIS inequality, e.g. Theorem 2.1.1 in [1],

$$\begin{aligned}
\mathbb{P}\left(\max_{k=1,\dots,2^{jr}} |Z_{j,k}| \ge (1 + K')\sqrt{2 \log(2^{jr})}\right) &\le 2\mathbb{P}\left(\max_{k=1,\dots,2^{jr}} Z_{j,k} \ge (1 + K')\sqrt{2 \log(2^{jr})}\right) \\
&\le 4 \exp\left(-K'^2 \log(2^{jr})\right).
\end{aligned}$$

Combining this with the union bound and the formula for the geometric sum, we obtain for any $K' > 2/\sqrt{r}$,

$$\mathbb{P}\left(\exists j = 1, \dots, J_\beta, \max_{k=1,\dots,2^{jr}} |Z_{j,k}| \geq (1+K')\sqrt{2\log(2^{jr})}\right) \leq \sum_{j=1}^{J_\beta} 2^{j(2-rK'^2)} \leq \frac{1}{1-2^{2-rK'^2}} - 1.$$

Therefore, with $K' > \sqrt{3}$, on an event with probability at least $1 - 4/(2^{rK'^2} - 4)$, we find

$$\|X^\beta\|_{\infty,\infty,\beta} \leq \max_{j=1,\dots,J_\beta} \frac{(1+K')\sqrt{2\log(2^{jr})}}{\sqrt{jr}} = (1+K')\sqrt{2\log 2}.$$

□

Proof of Lemma 2.6.4. In view of Section 4.3.6 in [42], the Besov space $\mathcal{B}_{\infty,\infty,\beta}$ contains the Hölder space $\mathcal{C}_r^\beta$ for any $\beta > 0$, and they coincide whenever $\beta \notin \mathbb{N}$. Thus there exists K' such that $\mathcal{C}_r^\beta(K) \subseteq \mathcal{B}_{\infty,\infty,\beta}(K')$. We show that:

(1.) Lemma 2.5.2 (i) holds for some $C_1'(\beta, r) \geq 1$ and $C_2'(\beta, r) = 3/2$.

(2.) Any sequence $\varepsilon_n(\alpha, \beta, r) \geq C_1'(\beta, r)^2(2\beta+1)^3(\log n)^{3(\beta+1)} n^{-\beta\alpha/(2\beta\alpha+r)}$ solves the concentration function inequality (2.3.2).

(3.) Assumption 2.4.2 (i) holds by taking $\varepsilon_n(\alpha, \beta, r) = C_1(\beta, r)(\log n)^{C_2(\beta,r)} n^{-\beta\alpha/(2\beta\alpha+r)}$ with $C_1(\beta, r) := C_1'(\beta, r)^2(2\beta+1)^3 \vee Q_1(\beta, r, K)^{\beta/(2\beta+r)}$ and $C_2(\beta, r) := 3(\beta+1)$.

(4.) $\sup_{\beta \in [\beta_-, \beta_+]} C_1(\beta, r) < +\infty$ and $\sup_{\beta \in [\beta_-, \beta_+]} C_2(\beta, r) < +\infty$.

(5.) Assumption 2.4.2 (ii) holds for an $\varepsilon_n(\eta)$ of the form (2.5.3).

By Lemma 2.5.2, (1.) $\implies$ (2.) $\implies$ (3.) and by Lemma 2.5.3, (4.) $\implies$ (5.). Thus it remains to prove (1.) and (4.).

Proof of (i.): We denote the small ball probability term by $\varphi_0^{(\beta,r)}(\delta) := -\log \mathbb{P}(\|X^\beta\|_\infty \leq \delta)$ and the RKHS term by $\varphi^{(\beta,r,K)}(\delta) - \varphi_0^{(\beta,r)}(\delta)$. We start with the RKHS term. The proof of Theorem 4.5 in [82] shows that any function $h \in \mathcal{B}_{\infty,\infty,\beta}(K')$ can be well approximated by its projection h^{J_β} at truncation level J_β. In fact, one has $\|h - h^{J_\beta}\|_\infty \leq K' 2^{-J_\beta \beta}/(2^\beta - 1)$ and, with coefficients $\omega_j = 2^{-j(\beta+r/2)}/\sqrt{jr}$,

$$\|h^{J_\beta}\|_{\mathbb{H}^\beta}^2 = \sum_{j=1}^{J_\beta} \sum_{k=1}^{2^{jr}} \lambda_{j,k}(h)^2 \omega_j^{-2} \leq K'^2 r J_\beta \sum_{j=1}^{J_\beta} 2^{jr} \leq K'^2 r J_\beta^2 2^{J_\beta r}.$$

Recall that J_β is defined as the closest integer to the solution J of $2^J = n^{1/(2\beta+r)}$. By definition, we always have $J_\beta \leq 1 + \log_2 n/(2\beta + r)$ and so $2^{J_\beta} \leq 2n^{1/(2\beta+r)}$, $2^{-J_\beta\beta} \geq 2^{-\beta} n^{-\beta/(2\beta+r)}$. With all the above, we choose

$$\delta_n := K' \frac{(2^\beta+1)^2}{(2^\beta-1)} \sqrt{r2^r} J_\beta^{3/2} 2^{-J_\beta \beta},$$

implying $\|h - h^{J_\beta}\|_\infty < \delta_n$ and

$$\begin{aligned}\varphi^{(\beta,r,K)}(\delta_n) - \varphi_0^{(\beta,r)}(\delta_n) &\leq K'^2 r J_\beta^2 2^{J_\beta r} \\ &\leq K'^2 r 2^r J_\beta^2 n^{\frac{r}{2\beta+r}}\end{aligned}$$

$$
\begin{aligned}
&\leq K'^2 \frac{(2^\beta+1)^2}{(2^\beta-1)^2} r 2^r J_\beta^2 n^{\frac{r}{2\beta+r}} \\
&\leq n K'^2 \frac{(2^\beta+1)^4}{(2^\beta-1)^2} r 2^r J_\beta^2 2^{-2J_\beta \beta} \\
&\leq n\delta_n^2.
\end{aligned}
$$

We now study the small ball probability. The proof of Theorem 4.5 in [82] shows that, for any sequence $\delta_n \in (0,1)$,

$$
\varphi_0^{(\beta,r)}(\delta_n) \leq -\sum_{j=1}^{J_\beta} 2^{jr} \log\left(2\Phi\left(\frac{\delta_n 2^{j\beta}}{\widetilde{K}(\beta)+j^2r^2}\right)-1\right),
$$

where $\widetilde{K}(\beta)$ is chosen in such a way that the function $x \mapsto x^{\beta/r}/(\widetilde{K}(\beta)+\log_2^2(x))$ is increasing for $x \geq 1$. Taking the derivative and imposing it to be positive for $x > 1$ yields $\beta\widetilde{K}(\beta)/r + \log_2^2(x) - 2\log(x)/\log_2^2(2) > 0$, which is solved for any $\widetilde{K}(\beta) \geq 4r/\beta$. [82] also shows that the function $f(y) = -\log(2\Phi(y)-1)$ is decreasing and can be bounded above by $f(y) \leq 1 + |\log y|$ on any interval $y \in [0,c]$. Thus we find

$$
\varphi_0^{(\beta,r)}(\delta_n) \leq \sum_{j=1}^{J_\beta} 2^{jr}\left(1+\left|\log\left(\frac{\delta_n 2^{j\beta}}{\widetilde{K}(\beta)+j^2r^2}\right)\right|\right).
$$

Now, assume that our sequence satisfies $\delta_n \leq (\widetilde{K}(\beta)+J_\beta^2 r^2)2^{-J_\beta\beta}$, then

$$
\begin{aligned}
\varphi_0^{(\beta,r)}(\delta_n) &\leq \sum_{j=1}^{J_\beta} 2^{jr}\left(1+\log\left(\frac{\widetilde{K}(\beta)+j^2r^2}{\delta_n 2^{j\beta}}\right)\right) \\
&\leq J_\beta 2^{J_\beta r}\left(1+\log\left(\frac{\widetilde{K}(\beta)+r^2}{\delta_n 2^{\beta}}\right)\right) \\
&\leq 2J_\beta 2^{J_\beta r}\left[\log\left(\frac{\widetilde{K}(\beta)+r^2}{2^\beta}\right)+\log\left(\frac{1}{\delta_n}\right)\right].
\end{aligned}
$$

We now show that indeed $\delta_n \leq (\widetilde{K}(\beta)+J_\beta^2 r^2)2^{-J_\beta\beta}$. In fact one has to check that

$$
K'\frac{(2^\beta+1)^2}{(2^\beta-1)}\sqrt{r2^r}J_\beta^{3/2} \leq \widetilde{K}(\beta)+J_\beta^2r^2,
$$

which holds for sufficiently large n since $J_\beta^{3/2} \ll J_\beta^2$. Then, with the definition of δ_n,

$$
\begin{aligned}
\varphi_0^{(\beta,r)}(\delta_n) &\leq 2\left[\log\left(\frac{\widetilde{K}(\beta)+r^2}{2^\beta}\right)+\log\left(\frac{1}{\delta_n}\right)\right]J_\beta 2^{J_\beta r} \\
&\leq 2\left[\log\left(\frac{\widetilde{K}(\beta)+r^2}{2^\beta}\right)+\log\left(\frac{2^{J_\beta\beta}}{K'\frac{(2^\beta+1)^2}{(2^\beta-1)}\sqrt{r2^rJ_\beta^3}}\right)\right]J_\beta 2^{J_\beta r} \\
&\leq 2\left[\log\left(\frac{\widetilde{K}(\beta)+r^2}{2^\beta}\right)+\log\left(2^{J_\beta\beta}\right)\right]J_\beta 2^{J_\beta r}
\end{aligned}
$$

$$
\begin{aligned}
&\leq 2\left[\log\left(\frac{\widetilde{K}(\beta)+r^2}{2^\beta}\right)+\beta\log(2)\right]J_\beta^2 2^{J_\beta r}\\
&\leq 2\left[\log\left(\frac{\widetilde{K}(\beta)+r^2}{2^\beta}\right)+\beta\log(2)\right]2^r J_\beta^2 n^{\frac{r}{2\beta+r}}.
\end{aligned}
$$

Since $J_\beta^2 \ll J_\beta^3$, for large enough n the latter display is smaller than $K'^2\frac{(2^\beta+1)^2}{(2^\beta-1)^2}2^r J_\beta^3 n^{\frac{r}{2\beta+r}} \leq n\delta_n^2$.

This concludes the proof of (1.), since we have shown that the sequence

$$\varepsilon_n(1,\beta,r) := K'\frac{(2^\beta+1)^2}{(2^\beta-1)}\sqrt{r2^r}J_\beta^{3/2}2^{-J_\beta\beta},$$

solves the concentration function inequality (2.3.2) for $\alpha = 1$.

Proof of (4.): The constant $Q_1(\beta,r,K)$ is given explicitly in Lemma 2.A.5 and depends continuously on $\beta > 0$, so $\sup_{\beta\in[\beta_-,\beta_+]} Q_1(\beta,r,K) =: \widetilde{Q}_1 < +\infty$. Since the function $\beta \mapsto \beta/(2\beta+r)$ is increasing for $\beta > 0$, we also have $Q_1(\beta,r,K)^{\beta/(2\beta+r)} \leq \widetilde{Q}_1^{\beta_+/(2\beta_++r)}$.

All the quantities involved in the construction of the rate $\varepsilon_n(1,\beta,r)$ are explicit. It is immediate to see that they are all bounded on the compact interval $\beta \in [\beta_-,\beta_+]$ since $\beta_- > 0$.

This concludes the proof of (4.). □

Proof of Lemma 2.6.5. We show that:

(1.) Lemma 2.5.2 (i) holds for some $C_1'(\beta,r) \geq 1$ and $C_2'(\beta,r) = (1+r)\beta/(2\beta+r)$.

(2.) Any sequence $\varepsilon_n(\alpha,\beta,r) \geq C_1'(\beta,r)^2(2\beta+1)^{2C_2'(\beta,r)}(\log n)^{(2\beta+2)C_2'(\beta,r)}n^{-\beta\alpha/(2\beta\alpha+r)}$ solves the concentration function inequality (2.3.2).

(3.) Assumption 2.4.2 (i) holds for any $\varepsilon_n(\alpha,\beta,r) = C_1(\beta,r)(\log n)^{C_2(\beta,r)}n^{-\beta\alpha/(2\beta\alpha+r)}$ with $C_1(\beta,r) := C_1'(\beta,r)^2(2\beta+1)^{2C_2'(\beta,r)} \vee Q_1(\beta,r,K)^{\beta/(2\beta+r)}$ and $C_2(\beta,r) := (2\beta+2)(1+r)\beta/(2\beta+r)$.

(4.) $\sup_{\beta\in[\beta_-,\beta_+]} C_1(\beta,r) < +\infty$ and $\sup_{\beta\in[\beta_-,\beta_+]} C_2(\beta,r) < +\infty$.

(5.) Assumption 2.4.2 (ii) holds for an $\varepsilon_n(\eta)$ of the form (2.5.3).

By Lemma 2.5.2, (1.) $\Longrightarrow$ (2.) $\Longrightarrow$ (3.) and by Lemma 2.5.3, (4.) $\Longrightarrow$ (5.). Thus it remains to prove (1.) and (4.).

Proof of (1.): Let $\varphi_a^{(\beta,r,K)}$ the concentration function of the rescaled process $X^\nu(a\cdot)$, then Lemma 11.55 and Lemma 11.56 in [41] show that, for all $0 < \delta < 1$,

$$\varphi_a^{(\beta,r,K)}(\delta) \leq \Big(C(r)\big(\log(a\delta^{-1})\big)^{1+r} + D(r)\Big)a^r,$$

where $C(r)$ and $D(r)$ are constants that only depend on r and the spectral measure ν of X^ν. It is sufficient to solve the concentration function inequality (2.3.2) for $\alpha = 1$. The solution is given in Section 11.5.2 in [41] as $\varepsilon_n(1,\beta,r) = C_1'(\beta,r)(\log n)^{(1+r)\beta/(2\beta+r)}n^{-\beta/(2\beta+r)}$, for some constant $C_1'(\beta,r) \geq 1$ depending on β, r, K.

Proof of (4.): The constant $Q_1(\beta,r,K)$ is given explicitly in Lemma 2.A.5 and depends continuously on $\beta > 0$, so $\sup_{\beta\in[\beta_-,\beta_+]} Q_1(\beta,r,K) =: \widetilde{Q}_1 < +\infty$. Since the function $\beta \mapsto \beta/(2\beta+r)$ is increasing for $\beta > 0$, we also have $Q_1(\beta,r,K)^{\beta/(2\beta+r)} \leq \widetilde{Q}_1^{\beta_+/(2\beta_++r)}$.

The dependence on β in the concentration function bound only appears in the scaling $a = a(\beta, r)$, since the constants $C(r)$ and $D(r)$ are independent of β. The right side of the latter display depends continuously on the scaling $a = a(\beta, r)$, which in turn is continuous in $\beta \in [\beta_-, \beta_+]$ by construction (2.6.2). This gives

$$\sup_{\beta\in\beta\in[\beta_-,\beta_+]} C_1(\beta, r) < +\infty, \qquad \sup_{\beta\in\beta\in[\beta_-,\beta_+]} C_2(\beta, r) \leq (2\beta_+ + 1)(1 + r)\frac{\beta_+}{2\beta_+ + r}.$$

□

Chapter 3

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

This chapter is based on:
G. Finocchio, A. Derumigny and K. Proksch. Robust-to-outliers square-root Lasso, simultaneous inference with a MOM approach. *Arxiv preprint, arXiv:2103.10420 (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 [55] 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 known, 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.

3.1 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 [55] 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, provided that the number of outliers is not too large. 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 [37, Section 4.1]. Some of the key ideas behind the median-of-means construction can be traced back to the work on stochastic optimization [69, 58], sampling from large discrete structures [50], and sketching algorithms [2].

Our work builds on the MOM method introduced in [55], 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 $\boldsymbol{\beta}$. There, the MOM-Lasso method [55] yields a robust version of the Lasso estimator, which is known to be minimax optimal, see [7, 9, 8], 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 [10], which was shown to be minimax optimal by [35], 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 level $\sigma_- > 0$ bounding above and below the noise standard deviation as in [36], whereas in this article, we only require a known (or estimated) upper bound σ_+.

Our main results deal with the simultaneous estimation of the oracle function f^* and standard deviation σ^* of the residual $\zeta := Y - f^*(\mathbf{X})$. 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 [11, 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 [11] correspond to a different penalty level than the one used in [35] that allows to derive optimal rates for the estimation of $\boldsymbol{\beta}^*$. A related paper is [28], 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 [9]	-	-	-
Aggreg. Lasso	Optimal [9]	Yes	-	-
Square-root Lasso	Optimal [35]	-	Yes, complicated rate [11]	-
Aggreg. Square-root Lasso	Optimal [35]	Yes	Yes, but no rate	-
MOM-Lasso	Optimal [55]	-	-	Yes
Aggreg. MOM-Lasso	Optimal [55]	Yes	-	Yes
Robust SR-Lasso	Optimal (Th. 3.4.4)	-	$\sqrt{\frac{s^*}{n} \log\left(\frac{ed}{s^*}\right)}$ (Th. 3.4.4)	Yes
Aggreg. Robust SR-Lasso	Optimal (Th. 3.4.7)	Yes	$\sqrt{\frac{s^*}{n} \log\left(\frac{ed}{s^*}\right)}$ (Th. 3.4.7)	Yes

Table 3.1: 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 3.1 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 $\mathcal{D} = (\mathbf{X}_i, Y_i)_{i=1,\dots,n}$ that might be contaminated by a set of outliers $(\mathbf{X}_i, Y_i)_{i \in \mathcal{O}}$ (for some $\mathcal{O} \subset \{1, \dots, n\}$) 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

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

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 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,

$$\begin{aligned}
\mathcal{A}_{\sigma^*, \boldsymbol{\beta}^*, s^*}(\mathcal{D}) &:= \bigcap_{\mathcal{D}' \in \mathcal{D}\left(c s^* \log(ed/s^*)\right)} \mathcal{A}_{\sigma^*}(\mathcal{D}') \cap \mathcal{A}_{\boldsymbol{\beta}^*}(\mathcal{D}') \cap \mathcal{A}_{s^*}(\mathcal{D}'), \\
\mathcal{A}_{\sigma^*}(\mathcal{D}') &:= \left\{ \left| \widetilde{\sigma}(\mathcal{D}') - \sigma^* \right| \leq C \sigma_+ \sqrt{\frac{s^*}{n} \log\left(\frac{ed}{s^*}\right)} \right\}, \\
\mathcal{A}_{\boldsymbol{\beta}^*}(\mathcal{D}') &:= \left\{ \left| \widetilde{\boldsymbol{\beta}}(\mathcal{D}') - \boldsymbol{\beta}^* \right|_p \leq C \sigma_+ s^{*1/p} \sqrt{\frac{1}{n} \log\left(\frac{ed}{s^*}\right)} \right\}, \\
\mathcal{A}_{s^*}(\mathcal{D}') &:= \Big\{ \widetilde{s}(\mathcal{D}') \leq s^* \Big\},
\end{aligned}$$

and $(\widetilde{\boldsymbol{\beta}}(\mathcal{D}'), \widetilde{\sigma}(\mathcal{D}'), \widetilde{s}(\mathcal{D}'))$ is the aggregated estimator obtained from the perturbed dataset $\mathcal{D}'$. Our method only requires the knowledge of the upper bounds (σ_+, s_+).

The manuscript is organized as follows. In Section 3.2, we introduce the main framework and notation, as well as the step-by-step construction of the MOM estimator. In Section 3.3 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 3.4. In Section 3.5 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.

3.2 Notation and framework

3.2.1 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 $PX = \mathbb{E}_P[X]$, or $\mathbb{E}[X]$ when the measure P is fixed. 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 $a_n \leq C b_n$ for all n. Moreover, $a_n \asymp b_n$ means that $(a_n)_n \lesssim (b_n)_n$ and $(b_n)_n \lesssim (a_n)_n$.

3.2.2 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. This 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 $\overline{f}(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$, 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

$$f^* \in \mathcal{F}^* := \underset{f \in \mathcal{F}}{\arg\min}\, \mathbb{E}\big[(Y - f(\mathbf{X}))^2\big], \tag{3.2.1}$$

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

$$Y = f^*(\mathbf{X}) + \zeta, \quad \zeta := Y - f^*(\mathbf{X}), \tag{3.2.2}$$

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

Assumption 3.2.1. *We make the following assumptions on the residual ζ,*

$$\mathbb{E}[\zeta] = 0, \quad \sigma^* := \mathbb{E}[\zeta^2]^{\frac{1}{2}} \leq \sigma_+, \quad \mathfrak{m}^* := \mathbb{E}[\zeta^4]^{\frac{1}{4}} \leq \mathfrak{m}_+ := \sigma_+\kappa_+, \quad \kappa^* := \frac{\mathfrak{m}^{*4}}{\sigma^{*4}} \leq \kappa_+, \tag{3.2.3}$$

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 [61]. 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 3.2.2. *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 $\mathcal{I} \subseteq \{1, \dots, n\}$ are independent and distributed as $(\mathbf{X}, Y)$ in the regression model* (3.2.2).
- ***Outliers.*** *Nothing is assumed on the pairs $(\mathbf{X}_i, Y_i)_{i\in\mathcal{O}} =: \mathcal{D}_{\mathcal{O}}$ with $\mathcal{O} \subseteq \{1, \dots, n\}$. 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 [55], 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, $\mathcal{I} \cup \mathcal{O} = \{1, \dots, n\}$ and $\mathcal{I} \cap \mathcal{O} = \emptyset$, but the statistician does not know whether any fixed index $i \in \{1, \dots, n\}$ 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 $1 \le s \le d$, we define

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

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

3.2.3 Convex-concave formulation

We follow the formalization made in [55]. 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

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

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

$$f^* \in \operatorname*{arg\,min}_{f\in\mathcal{F}} \sup_{g\in\mathcal{F}} \mathbb{E}\big[\ell_f(\mathbf{X}, Y) - \ell_g(\mathbf{X}, Y)\big], \quad \sigma^* = \mathbb{E}\big[\ell_{f^*}(\mathbf{X}, Y)\big]^{\frac{1}{2}}, \tag{3.2.4}$$

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^*.

3.2.4 Construction of the estimator

The starting point of our approach is the regularized median-of-means (MOM) tournament introduced in [60], which has been proposed as a procedure to outperform the regularized empirical risk minimizer (RERM)

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

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 [55] 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

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

by replacing the empirical average $n^{-1} \sum_{i=1}^n \big(\ell_f(\mathbf{X}_i, Y_i) - \ell_g(\mathbf{X}_i, Y_i)\big)$ 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

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

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 3.5 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 (3.2.1)–(3.2.3). We start with a preliminary definition.

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

$$\mathcal{Q}_\alpha[\mathbf{x}] := \Big\{u \in \mathbb{R} : \big|\{k = 1, \dots, K : x_k \geq u\}\big| \geq (1-\alpha)K, \\ \text{and } \big|\{k = 1, \dots, K : x_k \leq u\}\big| \geq \alpha K\Big\}. \tag{3.2.5}$$

This means that $Q_\alpha[\mathbf{x}]$ is a *α-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 $\mathcal{D} = \{(\mathbf{X}_i, Y_i) : i = 1, \dots, n\}$ into K blocks $\mathcal{D}_1, \dots, \mathcal{D}_K$ of size n/K (assumed to be an integer). This corresponds to a partition of $\{1. \dots, n\}$ into blocks $B_1, \dots, B_K$.

STEP 2. Local criterion.

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

$$R_c(\ell_g, \chi, \ell_f, \sigma) := (\sigma - \chi)\left(1 - 2\frac{\ell_f + \ell_g}{(\sigma + \chi)^2}\right) + 2c\frac{\ell_f - \ell_g}{\sigma + \chi}. \tag{3.2.6}$$

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 $\mathcal{X} \times \mathbb{R}$. For each $k \in [K]$, 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,

$$\mathbb{P}_{B_k}\Big(R_c(\ell_g, \chi, \ell_f, \sigma)\Big) := \frac{1}{|B_k|}\sum_{i \in B_k} R_c\Big(\ell_g(\mathbf{X}_i, Y_i), \chi, \ell_f(\mathbf{X}_i, Y_i), \sigma\Big), \tag{3.2.7}$$

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

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

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

$$MOM_K\Big(R_c(\ell_g, \chi, \ell_f, \sigma)\Big) := Q_{1/2,K}\Big[R_c(\ell_g, \chi, \ell_f, \sigma)\Big], \tag{3.2.8}$$

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

$$T_{K,\mu}(g, \chi, f, \sigma) := MOM_K\Big(R_c(\ell_g, \chi, \ell_f, \sigma)\Big) + \mu(\|f\| - \|g\|), \tag{3.2.9}$$

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 (3.2.3), we define the *MOM–K* estimator of (f^*, σ^*) as

$$(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+}) := \underset{f\in\mathcal{F},\ \sigma\leq\sigma_+}{\arg\min}\ \max_{g\in\mathcal{F},\ \chi\leq\sigma_+} T_{K,\mu}(g,\chi,f,\sigma), \tag{3.2.10}$$

where $T_{K,\mu}$ is the penalized functional in (3.2.9). Furthermore, set

$$\mathcal{C}_{K,\mu}(f,\sigma) := \max_{g\in\mathcal{F},\ \chi\leq\sigma_+} T_{K,\mu}(g,\chi,f,\sigma). \tag{3.2.11}$$

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.

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

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

Assumption 3.3.1. *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 [55]. 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 3.2.1, helps controlling the dependence between ζ and $\mathbf{X}$; this also matches Assumption 3.1 in [61]. 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 [55] with $\theta_m^2 = \theta_1^2\mathfrak{m}^{*2}$, whereas in our setting this is a consequence of Assumption 3.2.1 and Assumption 3.3.1.

3.3.1 Complexity parameters

With the introduction of MOM tournaments procedures, see [61] 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 denote 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 denote the quadratic and multiplier empirical processes on J as

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

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

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

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

$$\begin{aligned} r_P(\rho, \gamma_P) &:= \inf\left\{ r > 0 : \sup_{J \subset \mathcal{I}, |J| \geq \frac{n}{2}} \mathbb{E}\left[\sup_{f \in \mathbb{B}(f^*, \rho, r)} \left| \frac{1}{|J|} \sum_{i \in J} \xi_i (f - f^*)(\mathbf{X}_i) \right| \right] \leq \gamma_P r \right\}, \\ r_Q(\rho, \gamma_Q) &:= \inf\left\{ r > 0 : \sup_{J \subset \mathcal{I}, |J| \geq \frac{n}{2}} \mathbb{E}\left[\sup_{f \in \mathbb{B}(f^*, \rho, r)} \left| \frac{1}{|J|} \sum_{i \in J} \xi_i (f - f^*)^2(\mathbf{X}_i) \right| \right] \leq \gamma_Q r^2 \right\}, \\ r_M(\rho, \gamma_M) &:= \inf\left\{ r > 0 : \sup_{J \subset \mathcal{I}, |J| \geq \frac{n}{2}} \mathbb{E}\left[\sup_{f \in \mathbb{B}(f^*, \rho, r)} \left| \frac{1}{|J|} \sum_{i \in J} \xi_i \zeta_i (f - f^*)(\mathbf{X}_i) \right| \right] \leq \gamma_M r^2 \right\}. \end{aligned} \tag{3.3.1}$$

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

$$r(\rho) \geq \max\left\{ r_P(\rho, \gamma_P), r_M(\rho, \gamma_M) \right\}, \tag{3.3.2}$$

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 [55]. We refer to Section 3.5 for a detailed discussion on the role of complexity parameters, here we only mention that in the sub-Gaussian setting of [56], 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 $\mathbb{B}(f^*, \rho)$.

3.3.2 Sparsity equation

We follow the setup of [55], 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 3.2.4 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 (3.3.2), we denote $H_\rho = \{f \in \mathcal{F} : \|f - f^*\| = \rho, \ \|f - f^*\|_{2,\mathbf{X}} \leq r(\rho)\}$. Furthermore, we set

$$\begin{aligned} \Gamma_{f^*}(\rho) &:= \bigcup_{f \in \mathcal{F} \colon \|f - f^*\| \leq \rho/20} (\partial\|\cdot\|)_f, \\ \Delta(\rho) &:= \inf_{f \in H_\rho} \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(f - f^*). \end{aligned} \tag{3.3.3}$$

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

$$\Delta(\rho) \geq \frac{4}{5}\rho, \quad \rho^* := \inf\left\{\rho > 0 : \Delta(\rho) \geq \frac{4\rho}{5}\right\}. \tag{3.3.4}$$

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

3.3.3 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 3.2.4. Fix any constant $c > 2$ in the definition on the functional R_c in (3.2.6) and, with $\sigma_+, \mathfrak{m}_+, \kappa_+$ the known bounds on the moments of the noise $\zeta = Y - f^*(\mathbf{X})$, set

$$\begin{aligned} c_\mu &:= 200(c+2)\kappa_+^{1/2}, \\ \varepsilon &:= \frac{c-2}{192\,\theta_0^2(c+2)\big(8+134\,\kappa_+^{1/2}((1+\frac{\sigma_+}{\sigma^*})\vee\frac{6}{5})\big)}, \\ c_\alpha^2 &:= \frac{3(c-2)}{5\theta_0^2}, \end{aligned} \tag{3.3.5}$$

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 (3.3.4) and $r(\cdot)$ any function such that $r(\rho) \geq \max\{r_P(\rho,\gamma_P), r_M(\rho,\gamma_M)\}$ as in (3.3.2). Define K^* as the smallest integer satisfying

$$K^* \geq \frac{n\varepsilon^2 r^2(\rho^*)}{384\,\theta_1^2\mathfrak{m}^{*2}}, \tag{3.3.6}$$

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

$$r^2(\rho_K) = \frac{384\,\theta_1^2\mathfrak{m}^{*2}K}{n\varepsilon^2}. \tag{3.3.7}$$

Assumption 3.3.2. *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 3.5.4.

Theorem 3.3.3. *With the notation above, let Assumptions 3.2.1–3.3.2 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 (3.2.10) with K blocks and penalization parameter*

$$\mu := \iota_\mu c_\mu \varepsilon \frac{r^2(\rho_K)}{\mathfrak{m}^*\rho_K}, \tag{3.3.8}$$

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

$$\|\widehat{f}_{K,\mu,\sigma_+} - f^*\| \leq 2\,\rho_K, \quad \|\widehat{f}_{K,\mu,\sigma_+} - f^*\|_{2,\mathbf{X}} \leq r(2\rho_K), \quad |\widehat{\sigma}_{K,\mu,\sigma_+} - \sigma^*| \leq c_\alpha r(2\rho_K), \tag{3.3.9}$$

$$\begin{aligned} R(\widehat{f}_{K,\mu,\sigma_+}) \leq R(f^*) &+ \left(2 + 2c_\alpha + (44+5c_\mu)\,\varepsilon + \frac{25\kappa^{*1/2}}{8\theta_1^2}\varepsilon^2\right) r^2(2\rho_K) \\ &+ 4\,\theta_1^2\varepsilon\left(r^2(2\rho_K) \vee r_Q^2(2\rho_K,\gamma_Q)\right). \end{aligned} \tag{3.3.10}$$

The proof of Theorem 3.3.3 is given in Appendix 3.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 3.3.3 extends the scope of Theorem 1 in [55] 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, [55] 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.

3.4 The high-dimensional sparse linear regression

3.4.1 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 3.4.1 (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 [56] and the references therein for a discussion and a list of examples.

Definition 3.4.2 (Contaminated datasets)**.** *For a dataset $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1,\dots,n} \in \mathbb{R}^{(d+1)\times n}$ and for $N \in [n]$, we denote by $\mathcal{D}(N)$ the set of all datasets $\mathcal{D}' = (\mathbf{x}'_i, y'_i)_{i=1,\dots,n} \in \mathbb{R}^{(d+1)\times n}$*

that differ from $\mathcal{D}$ by at most N observations, i.e.

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

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 3.4.3. *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 rates of convergence for $\boldsymbol{\beta}^* \in \mathcal{F}_{s^*}$ defined for $p \in [1,2]$ by $\mathfrak{r}_p := s^{*1/p}\sqrt{(1/n)\log(ed/s^*)}$. The allowed maximum number of outliers is defined by $\mathfrak{r}_{\mathcal{O}} := s^* \log(ed/s^*) = n\mathfrak{r}_2^2$.

Theorem 3.4.4. *Assume that $\mathfrak{r}_2 < 1$. For every $\theta_0, \theta_1, c_0, L, \kappa_+ > 1$, there exists universal constants $\widetilde{c}_1, \dots, \widetilde{c}_5 > 0$ 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

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

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

This theorem is proved in Section 3.B.1. Theorem 3.4.4 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 [67] and Lemma 5.3 in [57]. 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 [9], 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 [35] 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 3.4.4, we allow for any distribution of ζ with finite fourth moment. The MOM-Lasso [55] 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 [28] 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^*, obtained in [11, Corollary 2] as $\left|(\hat{\sigma}^{SR\text{-}Lasso})^2 - \sigma^2\right| \lesssim \sigma^2\Big(\frac{s^*\log(n \vee d\log n)}{n} + \sqrt{\frac{s^*\log(d \vee n)}{n}} + \frac{1}{\sqrt{n}}\Big)$.

Remark 3.4.5. *When $\boldsymbol{\beta}^*$ is not sparse but very close to a sparse vector, that is, $|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{**}|_1 \lesssim \sigma^*\sqrt{s^*\log(ed/s^*)/n}$ 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 3.4.4.*

In practice, it may not be obvious to choose a good value for σ_+. 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 $\widehat{\sigma}^2_{K,+} := Q_{1/2,K}\left[Y^2\right] - \left(Q_{1/2,K}[Y]\right)^2$. 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 (3.2.10). The following corollary, proved in Section 3.B.3, gives a bound on the performance of this estimator on the larger class $\mathcal{P}_{[0,\,+\infty]}$.

Corollary 3.4.6 (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*

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

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 $\sqrt{\mathrm{Var}\left[\mathbf{X}^\top \beta\right]}\mathfrak{r}_p$.

3.4.2 Adaptation to the unknown sparsity

We now provide an adaptive to s version of Theorem 3.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 [35, 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 [55] 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 $\widehat{\boldsymbol{\beta}}_{(s),\sigma_+} := \widehat{\boldsymbol{\beta}}_{K_s,\mu_s,\sigma_+}$, $\widehat{\sigma}_{(s),\sigma_+} := \widehat{\sigma}_{K_s,\mu_s,\sigma_+}$ and

$$K_s := \left\lceil \widetilde{c}_2 s \log\left(\frac{ed}{s}\right) \right\rceil, \quad \mu_s := \widetilde{c}_\mu \sqrt{\frac{1}{n}\log\left(\frac{ed}{s}\right)}. \tag{3.4.1}$$

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 (3.4.1).
- For $u \in \{1, \dots, 2s_+\}$, let $\mathfrak{r}_p(u) = u^{1/p}\sqrt{(1/n)\log(ed/u)}$ and

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

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

- 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 3.C.2 and gives uniform bounds for the performance of the aggregated estimator $(\widetilde{\boldsymbol{\beta}}_{\sigma_+}, \widetilde{\sigma}_{\sigma_+}, \widetilde{s}_{\sigma_+})$.

Theorem 3.4.7. *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*

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

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

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 $s^* \in \{1, \dots, s_+\}$, 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 3.4.6 can be aggregated in the same way. It will satisfy the same bounds up to some small degradation in the probability of the event.

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

Our construction in Section 3.2.4 produces a family of MOM estimators

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

where R_c is a carefully chosen functional in (3.2.6). As mentioned in Section 3.2.3, this extends the scope of the MOM estimator in [55]

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

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 3.5.1 we motivate our choice by showing that, in the sparse linear setting, we recover a robust version of the square-root Lasso. In Section 3.5.2 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 3.5.3 and Section 3.5.4 we reproduce the main results on complexity parameters in the sub-Gaussian and sparse linear case respectively.

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

Since we implement the same proving strategy as in [55], we introduce the following properties as natural assumptions that the functional R_c should satisfy.

P1. Anti-symmetry. For all $f, g \in \mathcal{F}$, $\chi, \sigma \in \mathbb{R}_+$ 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 $\boldsymbol{\chi}$, given $\boldsymbol{f = g}$. 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 $\boldsymbol{g}$, given $\boldsymbol{\sigma = \chi}$. For any fixed $f \in \mathcal{F}$ and $\chi = \sigma \in \mathbb{R}_+$, the problems of maximizing the functionals

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

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

The latter condition requires that our functional $R_c(\ell_g, \sigma, \ell_f, \sigma)$ 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 $\boldsymbol{\sigma}$, given $\boldsymbol{f = g}$. 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 $\boldsymbol{f}$, given $\boldsymbol{\sigma = \chi}$. For any fixed $g \in \mathcal{F}$ and $\chi = \sigma \in \mathbb{R}_+$, the problems of minimizing the functionals

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

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

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

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

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 [55] 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 [8, 7, 9], but its optimal tuning parameter λ is proportional to σ^*. An adaptive version of the Lasso is the square-root Lasso introduced in [10], which is also minimax optimal, as shown in [35]. This adaptive method uses

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

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 [43], 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

$$\begin{aligned} \widetilde{R}_c(\ell_g, \chi, \ell_f, \sigma) &= \frac{\ell_f}{\sigma} + \sigma - \frac{\ell_g}{\chi} - \chi, \\ &= (\sigma - \chi)\left(1 - \frac{\ell_f}{\sigma\chi}\right) + \frac{\ell_f - \ell_g}{\chi}, \\ \widetilde{T}_{K,\mu}(g, \chi, f, \sigma) &= MOM_K\left(\widetilde{R}_c(\ell_g, \chi, \ell_f, \sigma)\right) + \mu\big(\|f\| - \|g\|\big), \end{aligned}$$

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 $\chi \in \mathbb{R}_+$.

- We compare $MOM_K(\ell_f - \ell_g)$ to $MOM_K(\widetilde{R}_c(\ell_g, \chi, \ell_f, \sigma))$ when $\sigma = \chi$,

$$MOM_K(\ell_f - \ell_g) = \frac{1}{|B_k|} \sum_{i \in B_k} \Big(\ell_f(\mathbf{X}_i, Y_i) - \ell_g(\mathbf{X}_i, Y_i) \Big),$$
$$MOM_K\Big(\widetilde{R}_c(\ell_g, \sigma, \ell_f, \sigma)\Big) = \frac{1}{|B_k|} \sum_{i \in B_k} \left(\frac{\ell_f(\mathbf{X}_i, Y_i) - \ell_g(\mathbf{X}_i, Y_i)}{\sigma} \right),$$

  where B_k is the block realizing the median. The block B_k is the same in both cases because the multiplicative factor σ^{-1} is positive and does not depend on the observations. Therefore, for any fixed $f \in \mathcal{F}$, maximizing the functionals in the latter display over $g \in \mathcal{F}$ are equivalent problems.

This choice comes with a drawback. The proof of our main result is based on the argument proposed in [55], 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 (3.2.3). 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

$$g \mapsto MOM_K\left(R_c(\ell_g, \chi, \ell_f, \sigma)\right)$$

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

and the quantity $c+(\sigma-\chi)/(\sigma+\chi)$ belongs to the interval $[c-1,c+1]$ and $c>1$ by construction. We check property P3 by fixing $\sigma=\chi$, this gives

$$MOM_K\left(R_c(\ell_g,\chi,\ell_f,\sigma)\right)=\frac{1}{|B_k|}\sum_{i\in B_k}\frac{c}{\sigma}\Big(\ell_f(\mathbf{X}_i,Y_i)-\ell_g(\mathbf{X}_i,Y_i)\Big),$$

with B_k the block realizing the median. The block B_k is the same for $MOM_K(\ell_f-\ell_g)$ since the factor c/σ is positive and independent of the observations. Therefore, maximizing the two functionals over $g\in\mathcal{F}$ are equivalent problems.

3.5.2 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

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

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 $\text{Risk}(\widehat{f}_{K,\mu})-\text{Risk}(f^*)$ is as small as possible.

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

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

with appropriate radius ρ and complexity measure $r(2\rho)$. In the proof of Theorem 3.3.3, we show that this can be achieved with $\rho=\rho_K$ and any $r(\rho)\ge\max\{r_P(\rho,\gamma_P),\ r_M(\rho,\gamma_M)\}$, 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 [61] and the robust MOM method in [55]. 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 $B_{1,1}>0$,

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

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 $-T_{K,\mu}(\widehat{f}_{K,\mu},\widehat{\sigma}_{K,\mu},f^*,\sigma^*)\le B_{1,1}$.
- By the anti-symmetry property P1 of R_c, together with the quantile properties in Lemma 3.D.2, we have $-T_{K,\mu}(f,\sigma,f^*,\sigma^*)\le T_{K,\mu}(f^*,\sigma^*,f,\sigma)$ and it is sufficient to find

$$T_{K,\mu}(f^*,\sigma^*,\widehat{f}_{K,\mu,\sigma_+},\widehat{\sigma}_{K,\mu,\sigma_+})\le B_{1,1}.$$

- 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 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 3.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 [55], 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 quadratic and multiplier empirical processes, which in turn can be obtained by means of the complexities r_P and r_M alone. These facts have been established in [57, 60] and we provide them in Lemma 3.D.5, Lemma 3.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

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

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 3.D.7, and adding the quadratic term $(\widehat{f}_{K,\mu,\sigma_+} - f^*)^2$, the expectation term becomes

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

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 3.D.9 we find

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

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 3.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 3.D.8 we find $Q_{15/16,K}[(\widehat{f}_{K,\mu,\sigma_+} - f^*)^2] \leq r^2(2\rho_K) + \alpha_Q^2 \lesssim r^2(2\rho_K) \vee r_Q^2(2\rho_K, \gamma_Q)$.

3.5.3 Complexity parameters in the sub-Gaussian setting

We follow the construction presented in [56]. 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

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

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

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

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

$$\begin{aligned} s_n^*(\gamma) &:= \inf\{r > 0 : \mathbb{E}\big[\|G\|_{\mathbb{B}_2(r)\cap(\mathcal{F}-\mathcal{F})}\big] \leq \gamma r^2 \sqrt{n}\}, \\ r_n^*(\gamma') &:= \inf\{r > 0 : \mathbb{E}\big[\|G\|_{\mathbb{B}_2(r)\cap(\mathcal{F}-\mathcal{F})}\big] \leq \gamma' r \sqrt{n}\}. \end{aligned} \tag{3.5.1}$$

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

Lemma 3.5.1. *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* (3.3.1) *satisfy*

$$r_P(\rho, \gamma_P) \leq r_n^*\left(\frac{\gamma_P}{c_2 L^2}\right), \quad r_Q(\rho, \gamma_Q) \leq r_n^*\left(\frac{\gamma_Q}{c_2 L^2}\right), \quad r_M(\rho, \gamma_M) \leq s_n^*\left(\frac{\gamma_M}{c_3 L \mathfrak{m}^*}\right). \tag{3.5.2}$$

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 (3.3.2).

Proof of Lemma 3.5.1. We invoke Lemma 3.5.2, Lemma 3.5.3 and Lemma 3.5.4 below. Their proofs are based on a symmetrization argument in [66], which controls the processes

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

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

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 $(\xi_i)_{i=1,\ldots,n}$. The latter play a role in the definition of the complexities in (3.3.1).

Lemma 3.5.2 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 3.5.3 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 3.5.4 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 $\gamma = \gamma_M/(c_3 L\mathfrak{m}^*)$ and $h = f^*$ gives, for all $r > s_n^*(\gamma_M/(c_3 L\mathfrak{m}^*))$,

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

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 $r_M(\rho,\gamma_M) \le s_n^*(\gamma_M/(c_3 L\mathfrak{m}^*))$. □

Lemma 3.5.2 (Corollary 1.8 in [66]). *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*

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

Lemma 3.5.3 (Lemma 2.6 in [56]). *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*

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

Lemma 3.5.4 (Corollary of Theorem 2.7 in [56]). *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*

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

3.5.4 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 (3.5.2).

Theorem 3.5.5 (Theorem 1.6 in [67]). *There exists an absolute constant c_1 and for $K \ge 1$, $L \ge 1$ 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 $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$;*
- $\mathfrak{m}^* = \mathbb{E}\big[|\zeta|^{q_0}\big]^{1/q_0} < +\infty$;

- *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* $\mathbf{y} \in \mathbb{R}^d$, $1 \leq p \leq c_0 \log(ed)$, $1 \leq j \leq d$,
$$\mathbb{E}\big[|\mathbf{X}^\top \mathbf{e}_j|^p\big]^{\frac{1}{p}} \leq L\sqrt{p}\,\mathbb{E}\big[|\mathbf{X}^\top \mathbf{e}_j|^2\big]^{\frac{1}{2}}.$$

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

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

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 3.5.1 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 (3.3.2). Our Definition 3.4.1 restricts our analysis to settings where the assumptions of the previous theorem are satisfied.

By following Section 4 in [56], we provide bounds for the complexity parameters r_n^*, s_n^* in (3.5.2). 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 $\mathcal{F} = \big\{f_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathbb{R}^d\big\}$ and, for any $\rho > 0$,

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

Assume that $\mathbf{X}$ is an isotropic random vector that satisfies the weak moment condition of Theorem 3.5.5, 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 $2\rho/\sqrt{d} \leq r$,

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

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

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

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,

$$r^2(\rho) = C^2_{\gamma_P,\gamma_M} \begin{cases} \max\left\{\rho\mathfrak{m}^*\sqrt{\frac{\log d}{n}},\ \frac{\rho^2}{n}\log\left(\frac{ed}{n}\right)\right\}, & \text{if } \rho \le \frac{\mathfrak{m}^*\sqrt{\log d}}{\sqrt{n}}, \\ \max\left\{\rho\mathfrak{m}^*\sqrt{\frac{1}{n}\log\left(\frac{ed^2\mathfrak{m}^{*2}}{\rho^2 n}\right)},\ \frac{\rho^2}{n}\log\left(\frac{ed}{n}\right)\right\}, & \text{if } \frac{\mathfrak{m}^*\sqrt{\log d}}{\sqrt{n}} \le \rho \le \frac{\mathfrak{m}^* d}{\sqrt{n}}, \end{cases} \tag{3.5.3}$$

which coincides with the function obtained in Section 4.4 in [55].

Solution of the sparsity equation. We study the case $n \ge s\log(ed/s)$ and assume there exists a s-sparse vector in $\boldsymbol{\beta}^* + \mathbb{B}_1(\rho/20)$. In the proof of Theorem 1.4 in [57], it is shown that the smallest solution of the sparsity equation (3.3.4) 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 $1 < s^* \le d$. By definition, we have

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

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 $d \gg n \ge s^*\log(ed/s^*)$. 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^*.

Appendix 3.A Proof of Theorem 3.3.3

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 3.A.2 to Lemma 3.A.10. Then, we compare the bounds over different slices in Lemma 3.A.11 and show that the leading contribution comes from a bounded ball of the form

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

In Lemma 3.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 3.A.13.

In the notation of Theorem 3.3.3, for any $c > 2$ we have

$$\begin{aligned} c_\mu &:= 200(c+2)\kappa_+^{1/2}, \\ \varepsilon &:= \frac{c-2}{192\theta_0^2(c+2)\big(8 + 134\kappa_+^{1/2}((1+\frac{\sigma_+}{\sigma^*}) \vee \frac{6}{5})\big)}, \\ c_\alpha^2 &:= \frac{3(c-2)}{5\theta_0^2}, \end{aligned}$$

furthermore, we use the auxiliary parameters

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

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 3.3.2, 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 3.D.4, the event $\Omega_2(K)$ in Lemma 3.D.7 and the event $\Omega_3(K)$ in Lemma 3.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

$$\alpha_{K,c_\rho} := c_\alpha r(c_\rho \rho_K), \quad \delta_{K,n}^2 := \frac{25\mathfrak{m}^{*4}K}{n}, \quad r^2(\rho_K) = \frac{384\theta_1^2\delta_{K,n}^2}{25\mathfrak{m}^{*2}\varepsilon^2}, \tag{3.A.1}$$

the last equation rewrites the implicit definition of ρ_K in (3.3.7).

The next lemma checks that the choices made in Theorem 3.3.3 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 3.A.1. *The assumptions of Theorem 3.3.3 imply, with $c_K^2 = 384$ and any $\iota_\mu \in [1/4, 4]$,*

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

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

$$\frac{c-2}{24\theta_0^2} > \frac{800\kappa_+^{1/2}\varepsilon^2}{c_K^2\theta_1^2} + 16(c+2)\varepsilon + \left(\frac{1+\frac{\sigma_+}{\sigma^*}}{3} \vee \frac{36}{10}\right)\iota_\mu c_\mu \varepsilon, \tag{3.A.4}$$

$$c_\alpha^2 > \frac{1800\kappa_+^{1/2}\varepsilon^2}{c_K^2\theta_1^2} + 108(c+2)\varepsilon + \frac{144\iota_\mu c_\mu \varepsilon}{10}. \tag{3.A.5}$$

Conditions (3.A.2) *and* (3.A.5) *imply* $4\delta_{K,n}/\sigma^* < \alpha_{K,c_\rho} < \sigma^*$. *Condition* (3.A.4) *implies both*

$$\frac{1}{16\theta_0^2} > 4\varepsilon + \frac{(\sigma^* + \sigma_+)\iota_\mu c_\mu \varepsilon}{2(c-2)\mathfrak{m}^*}, \tag{3.A.6}$$

$$\frac{c-2}{24\theta_0^2} > \frac{800\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 16(c+2)\varepsilon + \frac{36\iota_\mu c_\mu\varepsilon}{10}. \tag{3.A.7}$$

Proof of Lemma 3.A.1. Condition (3.A.2) is equivalent to the upper bound $K \leq n\varepsilon^2/C^2$ on the number of blocks, which is itself an assumption of Theorem 3.3.3.

We now show that (3.A.2) implies $\alpha_{K,c_\rho} < \sigma^*$. By Assumption 3.3.2, we have $r^2(\rho_K) = c_K^2\theta_1^2\mathfrak{m}^{*2}K/(\varepsilon^2 n)$ and $r^2(2\rho_K) \leq c_r^2 r^2(\rho_K)$. Since $\alpha_{K,2} = c_\alpha r(2\rho_K)$, then also $\alpha_{K,2} \leq c_r\alpha_{K,1}$ and, by using condition (3.A.2) in the last inequality,

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

thus $\alpha_{K,c_\rho} < \sigma^*$. Condition (3.A.5) implies $4\delta_{K,n}/\sigma^* < \alpha_{K,c_\rho}$, since it gives the inequality

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

We now show (3.A.3). By definition of c_μ in (3.3.5), 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 (3.A.3) holds since, by construction (3.3.5),

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

We now deal with (3.A.4), which we rewrite using $c_K^2 = 384$,

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

With the definition of c_μ in (3.3.5) and $\iota_\mu = 4$, this becomes

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

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 (3.3.5) coincides with imposing $\varepsilon = c_\varepsilon \cdot \min\{1/\sqrt{2A},\ 1/2B\} = c_\varepsilon/2B$, with $c_\varepsilon = 1/2$ and

$$\begin{aligned}\frac{1}{\sqrt{2A}} &= \sqrt{c-2}\frac{\theta_1}{10\theta_0\kappa_+^{1/4}},\\ \frac{1}{2B} &= \frac{c-2}{96\theta_0^2(c+2)\big(8+134\kappa_+^{1/2}((1+\frac{\sigma_+}{\sigma^*}) \vee \frac{6}{5})\big)},\end{aligned}$$

we have used that $400/3 < 134$. Thus, condition (3.A.4) is satisfied. It is immediate to verify that this implies both (3.A.6) and (3.A.7).

We conclude by showing (3.A.5). With $c_K^2 = 384$, the definition of c_μ in (3.3.5) and $\iota_\mu = 4$, we rewrite this 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 (3.A.5) 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 (3.3.5), since $3/5 > 441/768$. The proof is complete. □

3.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

$$\begin{aligned}
\mathcal{F}_1^{(c_\rho)} &:= \{(g,\chi)\in\mathcal{F}\times I_+ : \|g-f^*\|\leq c_\rho\rho_K, \|g-f^*\|_{2,\mathbf{X}}\leq r(c_\rho\rho_K),\ |\sigma^*-\chi|\leq\alpha_{K,c_\rho}\}\\
\mathcal{F}_2^{(c_\rho)} &:= \{(g,\chi)\in\mathcal{F}\times I_+ : \|g-f^*\|\leq c_\rho\rho_K, \|g-f^*\|_{2,\mathbf{X}}> r(c_\rho\rho_K),\ |\sigma^*-\chi|\leq\alpha_{K,c_\rho}\}\\
\mathcal{F}_3^{(c_\rho)} &:= \{(g,\chi)\in\mathcal{F}\times I_+ : \|g-f^*\|> c_\rho\rho_K,\ |\sigma^*-\chi|\leq\alpha_{K,c_\rho}\}\\
\mathcal{F}_4^{(c_\rho)} &:= \{(g,\chi)\in\mathcal{F}\times I_+ : \|g-f^*\|\leq c_\rho\rho_K, \|g-f^*\|_{2,\mathbf{X}}\leq r(c_\rho\rho_K),\ \chi>\sigma^*+\alpha_{K,c_\rho}\}\\
\mathcal{F}_5^{(c_\rho)} &:= \{(g,\chi)\in\mathcal{F}\times I_+ : \|g-f^*\|\leq c_\rho\rho_K, \|g-f^*\|_{2,\mathbf{X}}> r(c_\rho\rho_K),\ \chi>\sigma^*+\alpha_{K,c_\rho}\}\\
\mathcal{F}_6^{(c_\rho)} &:= \{(g,\chi)\in\mathcal{F}\times I_+ : \|g-f^*\|> c_\rho\rho_K,\ \chi>\sigma^*+\alpha_{K,c_\rho}\}\\
\mathcal{F}_7^{(c_\rho)} &:= \{(g,\chi)\in\mathcal{F}\times I_+ : \|g-f^*\|\leq c_\rho\rho_K, \|g-f^*\|_{2,\mathbf{X}}\leq r(c_\rho\rho_K),\ \chi<\sigma^*-\alpha_{K,c_\rho}\}\\
\mathcal{F}_8^{(c_\rho)} &:= \{(g,\chi)\in\mathcal{F}\times I_+ : \|g-f^*\|\leq c_\rho\rho_K, \|g-f^*\|_{2,\mathbf{X}}> r(c_\rho\rho_K),\ \chi<\sigma^*-\alpha_{K,c_\rho}\}\\
\mathcal{F}_9^{(c_\rho)} &:= \{(g,\chi)\in\mathcal{F}\times I_+ : \|g-f^*\|> c_\rho\rho_K,\ \chi<\sigma^*-\alpha_{K,c_\rho}\}.
\end{aligned}$$

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 $i=1,\dots,9$, 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,

$$\sup_{(g,\chi)\in\mathcal{F}_i^{(c_\rho)}} T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq B_{i,c_\rho}, \tag{3.A.8}$$

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 (3.2.6), and $\ell_g = \ell_{f^*} + \ell_g - \ell_{f^*}$, we find

$$\begin{aligned} R_c(\ell_g, \chi, \ell_{f^*}, \sigma^*) &= (\sigma^* - \chi)\bigg(1 - 2\frac{\ell_{f^*} + \ell_g}{(\sigma^* + \chi)^2}\bigg) + 2c\frac{\ell_{f^*} - \ell_g}{\sigma^* + \chi} \\ &= (\sigma^* - \chi)\bigg(1 - \frac{4\ell_{f^*}}{(\sigma^* + \chi)^2}\bigg) + 2\bigg(c + \frac{\sigma^* - \chi}{\sigma^* + \chi}\bigg)\frac{\ell_{f^*} - \ell_g}{\sigma^* + \chi} \\ &= R_c(\ell_{f^*}, \chi, \ell_{f^*}, \sigma^*) + 2\Delta_c(\chi, \sigma^*)\frac{\ell_{f^*} - \ell_g}{\sigma^* + \chi}, \end{aligned}$$

with

$$\Delta_c(\chi, \sigma) := \bigg(c + \frac{\sigma - \chi}{\sigma + \chi}\bigg) \in [c-1, c+1], \quad \forall \sigma, \chi \in (0, +\infty),$$

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

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

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 3.D.2, we can write

$$\begin{aligned} &T_{K,\mu}(g, \chi, f^*, \sigma^*) \\ &\quad= Q_{1/2,K}\bigg[R_c(\ell_{f^*}, \chi, \ell_{f^*}, \sigma^*) + 2\Delta_c(\chi, \sigma^*)\frac{\ell_{f^*} - \ell_g}{\sigma^* + \chi}\bigg] + \mu(\|f^*\| - \|g\|) \\ &\quad\le Q_{3/4,K}\big[R_c(\ell_{f^*}, \chi, \ell_{f^*}, \sigma^*)\big] + \frac{2\Delta_c(\chi, \sigma^*)}{(\sigma^* + \chi)}Q_{3/4,K}\Big[2\zeta(g - f^*) - (g - f^*)^2\Big] \\ &\qquad+ \mu(\|f^*\| - \|g\|). \end{aligned} \tag{3.A.9}$$

By arguing as in the proof of Lemma 3.D.9, see discussion after (3.D.2) for bounding (3.D.3), the quantity

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

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

$$Q_{3/4,K}\big[R_c(\ell_{f^*}, \chi, \ell_{f^*}, \sigma^*)\big] \le (\chi - \sigma^*)\Big(\frac{4\sigma^{*2} + 4\delta_{K,n}}{(\sigma^* + \chi)^2} - 1\Big), \tag{3.A.10}$$

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

$$Q_{3/4,K}\big[R_c(\ell_{f^*}, \chi, \ell_{f^*}, \sigma^*)\big] \le (\sigma^* - \chi)\Big(1 - \frac{4\sigma^{*2} - 4\delta_{K,n}}{(\sigma^* + \chi)^2}\Big). \tag{3.A.11}$$

The following lemmas give, on the event $\Omega(K)$, the bounds B_{i,c_ρ} in (3.A.8) for $i = 1, \dots, 9$ and $c_\rho \in \{1, 2\}$.

Lemma 3.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*

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

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 3.A.2. Let $(g,\chi) \in \mathcal{F}_1^{(c_\rho)}$. Using the bound obtained in (3.A.9), the inequality $(g-f^*)^2 \ge 0$ and the triangular inequality, the quantity $T_{K,\mu}(g,\chi,f^*,\sigma^*)$ is bounded above by

$$\begin{aligned}
&Q_{3/4,K}\big[R_c(\ell_{f^*},\chi,\ell_{f^*},\sigma^*)\big] + \frac{2\Delta_c(\chi,\sigma^*)}{\sigma^*+\chi} Q_{3/4,K}\big[2\zeta(g-f^*)-(g-f^*)^2\big] + \mu(\|f^*\|-\|g\|)\\
&\qquad \le Q_{3/4,K}\big[R_c(\ell_{f^*},\chi,\ell_{f^*},\sigma^*)\big] + \frac{2\Delta_c(\chi,\sigma^*)}{\sigma^*+\chi} Q_{3/4,K}\big[2\zeta(g-f^*)\big] + \mu\|f^*-g\|.
\end{aligned}$$

By Lemma 3.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 $\Delta_c(\chi,\sigma^*) \le c+2$ and our choice $\mu = (c_\mu\varepsilon/\mathfrak{m}^*)r^2(\rho_K)/\rho_K$, we find

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

We now bound the quantile term appearing in the latter display. Directly from (3.A.10) and (3.A.11), we get

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

By arguing as in the proof of Lemma 3.D.9, see bound (3.D.4) 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. □

Lemma 3.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 3.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 3.A.2, we know already that

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

The quantile properties in Lemma 3.D.2 give $Q_{3/4,K}[2\zeta(g-f^*)-(g-f^*)^2] \le Q_{7/8,K}[2\zeta(g-f^*)] - Q_{1/8,K}[(g-f^*)^2]$. An application of Lemma 3.D.7 bounds from above $Q_{7/8,K}[2\zeta(g-f^*)] \le \alpha_M^2 \le 4\varepsilon r^2(c_\rho\rho_K)$ and from below $Q_{1/8,K}[(g-f^*)^2] \ge (4\theta_0)^{-2}\|g-f^*\|_{2,\mathbf{X}}^2$. Since $4\varepsilon < 1/(4\theta_0)^2$ by condition (3.A.6), we have $\alpha_M^2 - \|g-f^*\|_{2,\mathbf{X}}^2 (4\theta_0)^{-2} \le (4\varepsilon - (4\theta_0)^{-2}) r^2(c_\rho\rho_K)$. Together with $\Delta_c(\chi,\sigma^*) \ge c-2$,

$$\begin{aligned} &T_{K,\mu}(g,\chi,f^*,\sigma^*) \\ &\quad \le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + \frac{2\Delta_c(\chi,\sigma^*)}{(\sigma^*+\chi)}\big(\alpha_M^2 - (4\theta_0)^{-2}\|g-f^*\|_{2,\mathbf{X}}^2\big) + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*} r^2(\rho_K) \\ &\quad \le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + 2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* + \alpha_{K,c_\rho}} r^2(c_\rho\rho_K) + \frac{c_\mu\varepsilon c_\rho}{\mathfrak{m}^*} r^2(\rho_K), \end{aligned}$$

which concludes the proof. □

Lemma 3.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

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

Proof of Lemma 3.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 3.A.2 and Lemma 3.A.3, together with an application of Lemma 3.D.1 with $\rho = \rho_K$, the bound in (3.A.9) becomes

$$\begin{aligned} &T_{K,\mu}(g,\chi,f^*,\sigma^*) \\ &\quad \le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + \frac{2\Delta_c(\chi,\sigma^*)}{(\sigma^*+\chi)} Q_{3/4,K}\big[2\zeta(g-f^*) - (g-f^*)^2\big] + \mu(\|f^*\| - \|g\|) \\ &\quad \le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + \frac{2\Delta_c(\chi,\sigma^*)}{(\sigma^*+\chi)} Q_{3/4,K}\big[2\zeta(g-f^*) - (g-f^*)^2\big] \\ &\qquad - \mu \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g-f^*) + \frac{\mu\rho_K}{10}. \end{aligned}$$

We follow now the proof of Lemma 5 in [55]. 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,

$$\begin{aligned} &T_{K,\mu}(g, \chi, f^*, \sigma^*) \\ &\quad \le \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta^2_{K,n} + \frac{2\Upsilon\Delta_c(\chi, \sigma^*)}{(\sigma^* + \chi)} Q_{3/4,K}\left[2\zeta(f - f^*) - (f - f^*)^2\right] \\ &\quad - \mu\Upsilon \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) + \frac{\mu\rho_K}{10}. \end{aligned}$$

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 3.3.2. 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

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

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 3.D.7, all together yield

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

By condition (3.A.3), the term multiplied by Υ is negative. This is true because $\kappa_+^{1/4} \ge \kappa^{*1/4} = \mathfrak{m}^*/\sigma^* > 1$ and $\alpha_{K,c_\rho} < \sigma^*$, by Lemma 3.A.1, so that

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

Since $\Upsilon > c_\rho$, we have

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

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 3.D.7 bounds from above the quantiles of $2\zeta(g - f^*)$ and from below the quantiles of $(g - f^*)^2$, this gives

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

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

using that $\Upsilon > c_\rho$ and that, as we show below, the term multiplied by Υ is negative. In fact, by condition (3.A.6) one has

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

This concludes the second part of the proof. □

Lemma 3.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*

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

is bounded above by

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

Proof of Lemma 3.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 3.A.2 and using that $\chi > \sigma^*+\alpha_{K,c_\rho}$, from (3.A.10) we get

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

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. □

Lemma 3.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 3.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 3.A.2, Lemma 3.A.3 and Lemma 3.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. □

Lemma 3.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

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

Proof of Lemma 3.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 3.A.4 and Lemma 3.A.6, we find

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

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 3.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 with $(f - f^*)^2 \geq 0$, 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 (3.A.3), 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,

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

using that the quantity multiplied by Υ is negative by condition (3.A.6), and $\Upsilon > c_\rho$. This concludes the proof. □

Lemma 3.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

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

Proof of Lemma 3.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 $\|g - f^*\| \leq c_\rho\rho_K$ and $\|g - f^*\|_{2,\mathbf{X}} \leq r(c_\rho\rho_K)$. By arguing as in the proof of Lemma 3.A.2 and using $\chi < \sigma^* - \alpha_{K,c_\rho}$, from (3.A.11) we get

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

Since $4\delta_{K,n}/\sigma^* < \alpha_{K,c_\rho} < \sigma^*$ by Lemma 3.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

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

which is sufficient to conclude. □

Lemma 3.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 3.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 3.A.2, Lemma 3.A.3 and Lemma 3.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. □

Lemma 3.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

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

Proof of Lemma 3.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 3.A.7 and Lemma 3.A.8, we get

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

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

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

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

We combine this with $(f - f^*)^2 \geq 0$ and get

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) &\leq -\frac{2\sigma^*}{(2\sigma^* - \alpha^2_{K,c_\rho})^2}\alpha^2_{K,c_\rho} + \Upsilon\bigg(\frac{8(c+2)\varepsilon}{\sigma^*} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}\bigg)r^2(\rho_K) + \frac{\mu\rho_K}{10} \\ &\leq -\frac{2\sigma^*}{(2\sigma^* - \alpha^2_{K,c_\rho})^2}\alpha^2_{K,c_\rho} + c_\rho\bigg(\frac{8(c+2)\varepsilon}{\sigma^*} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}\bigg)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K), \end{aligned}$$

using that the quantity multiplied by Υ is negative by condition (3.A.3), 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

$$\begin{aligned} &T_{K,\mu}(g,\chi,f^*,\sigma^*) \\ &\quad\leq -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho})^2}\alpha^2_{K,c_\rho} + \Upsilon\bigg(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,c_\rho}}r^2(\rho_K) + \mu\rho_K\bigg) + \frac{\mu\rho_K}{10} \\ &\quad\leq -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,c_\rho})^2}\alpha^2_{K,c_\rho} + c_\rho\bigg(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,c_\rho}} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\bigg)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K), \end{aligned}$$

using that the quantity multiplied by Υ is negative by condition (3.A.6), and $\Upsilon > c_\rho$. This concludes the proof. □

3.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 3.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 3.A.11. We start by showing that $B_{1,1}$ is bigger than the other $B_{i,1}$, $i = 2, \ldots, 9$. In Lemma 3.A.2 we have found

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

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

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

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

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

which is always true since $4\varepsilon - (4\theta_0)^{-2} < 0$, by condition (3.A.6).

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

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

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

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

$$\frac{8(c+2)\varepsilon}{2\sigma^*-\alpha_{K,1}}-\frac{17c_\mu\varepsilon}{10\mathfrak{m}^*}\leq\frac{8(c+2)\varepsilon}{2\sigma^*-\alpha_{K,1}},$$
$$2(c-2)\frac{4\varepsilon-(4\theta_0)^{-2}}{2\sigma^*+\alpha_{K,1}}+\frac{c_\mu\varepsilon}{10\mathfrak{m}^*}\leq\frac{8(c+2)\varepsilon}{2\sigma^*-\alpha_{K,1}}.$$

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

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

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

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

By Lemma 3.A.1, we have $0<\alpha_{K,1}<\sigma^*$ and the above display is satisfied if

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

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

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

which holds by condition (3.A.7).

Take $i=4$. By Lemma 3.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 3.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 (3.A.6).

Take $i = 6$. By Lemma 3.A.7, we have

$$B_{6,1} = \max\Bigg\{ -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,1})^2}\alpha_{K,1}^2 + \left(\frac{8(c+2)\varepsilon}{2\sigma^* + \alpha_{K,1}} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}\right) r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*} r^2(\rho_K), \\ -\frac{2\sigma^*}{(2\sigma^* + \alpha_{K,1})^2}\alpha_{K,1}^2 + \left(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{\sigma^* + \sigma_+} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\right) r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*} r^2(\rho_K)\Bigg\},$$

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 (3.A.6). With the ratio $\delta_{K,n}^2/r^2(\rho_K)$ in (3.A.1), 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 3.A.1, we have $0 < \alpha_{K,1} < \sigma^*$, so it is enough that

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

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 (3.A.5).

Take $i = 7$. By Lemma 3.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 (3.A.1) 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 (3.A.5).

Take $i = 8$. By Lemma 3.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 (3.A.6).

Take $i = 9$. By Lemma 3.A.10, we have

$$B_{9,1} = \max\Bigg\{ -\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,1})^2}\alpha_{K,1}^2 + \left(\frac{8(c+2)\varepsilon}{\sigma^*} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*} + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}\right) r^2(\rho_K),$$
$$-\frac{2\sigma^*}{(2\sigma^* - \alpha_{K,1})^2}\alpha_{K,1}^2 + \left(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* - \alpha_{K,1}} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\right) r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*} r^2(\rho_K)\Bigg\},$$

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

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

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 (3.A.6), 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 3.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 3.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

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

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 3.A.1, we have $0 < \alpha_{K,2} < \sigma^*$ and, with the ratio $\delta_{K,n}^2/r^2(\rho_K)$ in (3.A.1), 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 (3.A.7).

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

$$\begin{aligned}B_{3,2} = \max\Bigg\{&\frac{16}{\sigma^*(2\sigma^* - \alpha_{K,2})^2}\delta_{K,n}^2 + 2\bigg(\frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,2}} - \frac{4c_\mu\varepsilon}{5\mathfrak{m}^*}\bigg)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K),\\ &\frac{16}{\sigma^*(2\sigma^* - \alpha_{K,2})^2}\delta_{K,n}^2 + 2\bigg(2(c-2)\frac{4\varepsilon - (4\theta_0)^{-2}}{2\sigma^* + \alpha_{K,2}} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\bigg)r^2(\rho_K) + \frac{c_\mu\varepsilon}{10\mathfrak{m}^*}r^2(\rho_K)\Bigg\},\end{aligned}$$

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

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

As for the previous point, we use that $r^2(2\rho_K) \geq r^2(\rho_K)$, $\alpha_{K,2} \geq \alpha_{K,1}$ and $0 < \alpha_{K,2} < \sigma^*$ by Lemma 3.A.1, and the ratio $\delta_{K,n}^2/r^2(\rho_K)$ in (3.A.1). It is sufficient that both

$$\begin{aligned}\frac{800\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 16(c+2)\varepsilon + 8(c+2)\varepsilon &< \frac{c_\mu\varepsilon}{2\kappa^{*1/4}},\\ \frac{800\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 8(c-2)\varepsilon + 8(c+2)\varepsilon + \frac{31c_\mu\varepsilon}{10\kappa^{*1/4}} &< \frac{c-2}{12\theta_0^2}.\end{aligned}$$

The first bound holds by condition (3.A.3), so we plug it into the second line using $\kappa^* \geq 1$, we obtain the sufficient condition $36c_\mu\varepsilon/10 < (c-2)/(12\theta_0^2)$, which follows from condition (3.A.7).

Take $i = 4$. By Lemma 3.A.5, we have

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

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

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

Again, we use that $r^2(2\rho_K) \geq r^2(\rho_K)$, $\alpha_{K,2} \geq \alpha_{K,1}$ and $0 < \alpha_{K,2} < \sigma^*$ by Lemma 3.A.1, and the ratio $\delta_{K,n}^2/r^2(\rho_K)$ in (3.A.1). It is sufficient that

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

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

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

which follows from condition (3.A.5).

Take $i = 5$. By Lemma 3.A.6, we have

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

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

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

The second term in the latter display is negative by condition (3.A.6). We use that $r^2(2\rho_K) \geq r^2(\rho_K)$, $\alpha_{K,2} \geq \alpha_{K,1}$ and $0 < \alpha_{K,2} < \sigma^*$ by Lemma 3.A.1, and the ratio $\delta_{K,n}^2/r^2(\rho_K)$ in (3.A.1). It is sufficient that

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

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

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

which is true thanks to condition (3.A.5).

Take $i = 6$. By Lemma 3.A.7, we have

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

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 (3.A.6), the last terms on the left side of both equations are negative. As for the previous point, we use that $r^2(2\rho_K)\geq r^2(\rho_K)$, $\alpha_{K,2}\geq\alpha_{K,1}$ and $0<\alpha_{K,2}<\sigma^*$ by Lemma 3.A.1, and the ratio $\delta_{K,n}^2/r^2(\rho_K)$ in (3.A.1). 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 (3.A.5).

Take $i=7$. By Lemma 3.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

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

Again, we use that $r^2(2\rho_K)\geq r^2(\rho_K)$, $\alpha_{K,2}\geq\alpha_{K,1}$ and $0<\alpha_{K,2}<\sigma^*$ by Lemma 3.A.1, and the ratio $\delta_{K,n}^2/r^2(\rho_K)$ in (3.A.1). 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 (3.A.5).

Take $i=8$. By Lemma 3.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 (3.A.6), the last term on the left side is negative. We use that $r^2(2\rho_K) \geq r^2(\rho_K)$, $\alpha_{K,2} \geq \alpha_{K,1}$ and $0 < \alpha_{K,2} < \sigma^*$ by Lemma 3.A.1, and the ratio $\delta_{K,n}^2/r^2(\rho_K)$ in (3.A.1). 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 (3.A.5).

Take $i = 9$. By Lemma 3.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 (3.A.6), the last terms on the left side in the latter display are negative. One last time, we use that $r^2(2\rho_K) \geq r^2(\rho_K)$, $\alpha_{K,2} \geq \alpha_{K,1}$ and $0 < \alpha_{K,2} < \sigma^*$ by Lemma 3.A.1, and the ratio $\delta_{K,n}^2/r^2(\rho_K)$ in (3.A.1). It is sufficient that

$$\frac{400\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 16(c+2)\varepsilon + 8(c+2)\varepsilon < \frac{2c_\alpha^2}{9},$$
$$\frac{400\kappa^{*1/2}\varepsilon^2}{c_K^2\theta_1^2} + 8(c+2)\varepsilon + \frac{32c_\mu\varepsilon}{10\kappa^{*1/4}} < \frac{2c_\alpha^2}{9}.$$

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 (3.A.5). □

3.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 3.3.2, 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 $K \in \left[K^* \vee 32|\mathcal{O}|,\ n\varepsilon^2/C^2\right]$. We denote by $\Omega(K)$ the intersection of the event $\Omega_1(K)$ in Lemma 3.D.4, the event $\Omega_2(K)$ in Lemma 3.D.7 and the event $\Omega_3(K)$ in Lemma 3.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 3.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 (3.3.9).

Proof of Lemma 3.A.12. By definition (3.2.11), Lemma 3.A.11 gives the last inequality of

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

Then, with the property $Q_{1/2}[\mathbf{x}] \geq -Q_{1/2}[-\mathbf{x}]$ from Lemma 3.D.2,

$$\begin{aligned} B_{1,1} \geq \mathcal{C}_{K,\mu}(\widehat{f}_{K,\mu}, \widehat{\sigma}_{K,\mu}) &= \sup_{g\in\mathcal{F},\, \chi<\sigma_+} T_{K,\mu}(g, \chi, \widehat{f}_{K,\mu}, \widehat{\sigma}_{K,\mu}) \\ &\geq T_{K,\mu}(f^*, \sigma^*, \widehat{f}_{K,\mu}, \widehat{\sigma}_{K,\mu}) \geq -T_{K,\mu}(\widehat{f}_{K,\mu}, \widehat{\sigma}_{K,\mu}, f^*, \sigma^*). \end{aligned}$$

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 3.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. □

Lemma 3.A.13. *On the event $\Omega(K)$ defined above, the MOM−K estimator $(\widehat{f}_{K,\mu,\sigma_+}, \widehat{\sigma}_{K,\mu,\sigma_+})$ satisfies*

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

thus recovering the excess risk bound in (3.3.10).

Proof of Lemma 3.A.13. We apply Lemma 3.D.9 with $\rho = 2\rho_K$ and $\alpha_{K,c_\rho} = \alpha_{K,2}$, which gives

$$
\begin{aligned}
R(\widehat{f}_{K,\mu}) - R(f^*) &= \|\widehat{f}_{K,\mu} - f^*\|^2_{2,\mathbf{X}} + \mathbb{E}[-2\zeta(\widehat{f}_{K,\mu} - f^*)(\mathbf{X})] \\
&\leq r^2(2\rho_K) + \frac{2\sigma^* + \alpha_{K,2}}{2c} T_{K,\mu}(f^*, \sigma^*, \widehat{f}_{K,\mu}, \widehat{\sigma}_{K,\mu}) + \frac{2\sigma^* + \alpha_{K,2}}{c}\mu\rho_K + \alpha_M^2 \\
&\quad + \frac{8(2\sigma^* + \alpha_{K,2})}{c\sigma^*(2\sigma^* - \alpha_{K,2})^2}\delta^2_{K,n} + \frac{\alpha_{K,2}}{c(2\sigma^* - \alpha_{K,2})}\left(2\sigma^* r(2\rho_K) + r^2(2\rho_K) + \alpha_Q^2 + \alpha_M^2\right).
\end{aligned}
$$

In the proof of Lemma 3.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 3.A.2, the ratio $\delta^2_{K,n}/r^2(2\rho_K)$ in (3.A.1), $\mathfrak{m}^* \geq \sigma^*$ and $\alpha_{K,c_\rho} < \sigma^*$ by Lemma 3.A.1, we have

$$
\begin{aligned}
B_{1,1} &= \frac{16}{\sigma^*(2\sigma^* - \alpha_{K,1})^2}\delta^2_{K,n} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} r^2(\rho_K) + \frac{c_\mu\varepsilon}{\mathfrak{m}^*} r^2(\rho_K) \\
&= \left(\frac{25\mathfrak{m}^{*2}\varepsilon^2}{24\theta_1^2\sigma^*(2\sigma^* - \alpha_{K,1})^2} + \frac{8(c+2)\varepsilon}{2\sigma^* - \alpha_{K,1}} + \frac{c_\mu\varepsilon}{\mathfrak{m}^*}\right) r^2(\rho_K) \\
&\leq \left(\frac{25\kappa^{*1/2}\varepsilon^2}{24\theta_1^2\sigma^*} + \frac{8(c+2)\varepsilon}{\sigma^*} + \frac{c_\mu\varepsilon}{\sigma^*}\right) r^2(\rho_K).
\end{aligned}
$$

This gives

$$
\begin{aligned}
\frac{2\sigma^* + \alpha_{K,2}}{2c} B_{1,1} &\leq \frac{3\sigma^*}{2c}\left(\frac{25\kappa^{*1/2}\varepsilon^2}{24\theta_1^2\sigma^*} + \frac{8(c+2)\varepsilon}{\sigma^*} + \frac{c_\mu\varepsilon}{\sigma^*}\right) r^2(2\rho_K) \\
&= \left(\frac{25\kappa^{*1/2}\varepsilon^2}{16\theta_1^2 c} + \frac{12(c+2)\varepsilon}{c} + \frac{3c_\mu\varepsilon}{2c}\right) r^2(2\rho_K).
\end{aligned}
$$

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

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

By Lemma 3.D.7 we have $\alpha_M^2 \leq 4\varepsilon r^2(2\rho_K)$, whereas by Lemma 3.D.8 we bound

$$
\begin{aligned}
\alpha_Q^2 &\leq \varepsilon \max\left(\|f - f^*\|^2_{2,\mathbf{X}} \frac{1488\theta_1^4}{\varepsilon^2}\frac{K}{n},\ r_Q^2(\rho, \gamma_Q),\ \|f - f^*\|^2_{2,\mathbf{X}}\right) \\
&\leq \varepsilon\left(r^2(2\rho_K) \vee r_Q^2(2\rho_K, \gamma_Q)\right) \max\left(\frac{1488\theta_1^4 K}{n\varepsilon^2},\ 1\right) \\
&\leq 4\theta_1^2\varepsilon\left(r^2(2\rho_K) \vee r_Q^2(2\rho_K, \gamma_Q)\right),
\end{aligned}
$$

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

With $\alpha_{K,2} < \sigma^*$ and the ratio $\delta^2_{K,n}/r^2(\rho_K)$ in (3.A.1), we find

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

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

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

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

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

which completes the proof. □

Appendix 3.B Proofs for the high-dimensional sparse linear regression

3.B.1 Proof of Theorem 3.4.4

In Section 3.B.2, we prove the following Theorem 3.B.1. We show now how this theorem can be used to derive our Theorem 3.4.4.

Theorem 3.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* $|\mathcal{I}| \ge n/2$, $|\mathcal{O}| \le \widetilde{c}_0 s^* \log(ed/s^*)$, $n \ge s^* \log(ed/s^*)$ *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* (3.2.10) *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*

$$|\widehat{\boldsymbol{\beta}}_{K,\mu,\sigma_+} - \boldsymbol{\beta}^*|_p \le \widetilde{c}_3 \varepsilon^{-1} \kappa^* \sigma^* s^{*\frac{1}{p}} \sqrt{\frac{1}{n}\log\left(\frac{ed}{s^*}\right)},$$
$$|\widehat{\sigma}_{K,\mu,\sigma_+} - \sigma^*| \le c_\alpha \widetilde{c}_3 \varepsilon^{-1} \kappa^* \sigma^* s^{*\frac{1}{2}} \sqrt{\frac{1}{n}\log\left(\frac{ed}{s^*}\right)}, \qquad (3.B.1)$$

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

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

for a constant $C > 0$, and therefore

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

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}.

3.B.2 Proof of Theorem 3.B.1

In this section we use the results in Theorem 3.3.3 and the computations in Section 3.5.4 for the sparse linear setting. For any fixed $\varepsilon \in (0,1)$, the function

$$r_\varepsilon^2(\rho) = C_{\gamma_P,\gamma_M}^2 \begin{cases} \max\left\{\rho \mathfrak{m}^* \sqrt{\frac{\log d}{n\varepsilon^2}},\ \frac{\rho^2}{n\varepsilon^2} \log\left(\frac{ed}{n\varepsilon^2}\right)\right\}, & \text{if } \rho \leq \frac{\mathfrak{m}^* \sqrt{\log d}}{\sqrt{n\varepsilon^2}}, \\ \max\left\{\rho \mathfrak{m}^* \sqrt{\frac{1}{n\varepsilon^2} \log\left(\frac{ed^2 \mathfrak{m}^{*2}}{\rho^2 n\varepsilon^2}\right)},\ \frac{\rho^2}{n\varepsilon^2} \log\left(\frac{ed}{n\varepsilon^2}\right)\right\}, & \text{if } \frac{\mathfrak{m}^* \sqrt{\log d}}{\sqrt{n\varepsilon^2}} \leq \rho \leq \frac{\mathfrak{m}^* d}{\sqrt{n\varepsilon^2}}, \end{cases} \tag{3.B.2}$$

is a strict upper bound on $r^2(\rho)$ defined in (3.5.3). The smallest solution of the sparsity equation is of the form

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

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

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

this matches definition (3.3.6) in Theorem 3.3.3 with $C^2 = 384\theta_1^2$ and $r = r_\varepsilon$. By definition, this is equivalent to

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

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

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

which is a rearrangement of definition (3.3.7) in Theorem 3.3.3. For all $\rho^* \leq \rho_K \lesssim \mathfrak{m}^*\sqrt{n\varepsilon^2}$, we have

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

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 $K \lesssim n\varepsilon^2$. To check this, let us compute

$$\begin{aligned}\frac{n\varepsilon^2}{K\mathfrak{m}^{*2}} r_\varepsilon^2(\rho_K) &= C_{\gamma_P,\gamma_M}^2 C_K \sqrt{\left[\log\left(\frac{ed^2}{K^2}\right)\right]^{-1} \log\left(\frac{ed^2}{C_K^2 K^2}\log\left(\frac{ed^2}{K^2}\right)\right)} \\ &= C_{\gamma_P,\gamma_M}^2 C_K \sqrt{\frac{\log\left(\frac{ed^2}{K^2}\right) + \log\log\left(\frac{ed^2}{K^2}\right) - \log\left(C_K^2\right)}{\log\left(\frac{ed^2}{K^2}\right)}},\end{aligned}$$

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 $C_K \in [C_{\gamma_P,\gamma_M}^2/(\sqrt{2}C^2), \sqrt{2}C_{\gamma_P,\gamma_M}^2/C^2]$ recovering the solution.

As mentioned earlier, we can write $K^* = \lceil \widetilde{c}s^*\log(ed/s^*)\rceil$ with $\widetilde{c} = C_{\gamma_P,\gamma_M}^{*2}/(384\theta_1^2)$ 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 3.3.3. Then set $\widetilde{c}_2 = 2\widetilde{c}$ and apply Theorem 3.3.3 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

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

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

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

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 $\iota_K \widetilde{c}_2 \leq (ed/s^*)^{1/4}$,

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

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 3.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 3.3.3 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 (3.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.

3.B.3 Proof of Corollary 3.4.6

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 3.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 3.B.2 and Theorem 3.4.4 by replacing σ_+ by $\widehat{\sigma}_{K,+}$ and reasoning on the intersection of both events yields the conclusion.

We now prove Proposition 3.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 3.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 $(Q_{1/2,K}\left[Y\right])^2 \in [(\mathbb{E}[Y] - b_{K,n})^2, (\mathbb{E}[Y] + b_{K,n})^2]$.

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

$$\max\left\{\frac{2a_{K,n}}{\sigma^{*2}/4}, \ \frac{4\mathbb{E}[Y]b_{K,n}}{\sigma^{*2}/4}, \ \frac{2b_{K,n}^2}{\sigma^{*2}/4}\right\} \leq \frac{1}{3}\left(\frac{2\operatorname{Var}(Y)}{\sigma^{*2}/4} - 1\right).$$

By construction, we have

$$\begin{aligned} \frac{8a_{K,n}}{\sigma^{*2}} &= \frac{\sqrt{\operatorname{Var}(Y^2)}}{\sigma^{*2}}\sqrt{\frac{960K}{n}}, \\ \frac{16\mathbb{E}[Y]b_{K,n}}{\sigma^{*2}} &= \frac{\mathbb{E}[Y]\sqrt{\operatorname{Var}(Y)}}{\sigma^{*2}}\sqrt{\frac{3840K}{n}}, \\ \frac{8b_{K,n}^2}{\sigma^{*2}} &= \frac{\operatorname{Var}(Y)}{\sigma^{*2}}\frac{120K}{n}, \end{aligned}$$

and the quantities A_Y, B_Y are defined in such a way that $\sqrt{\operatorname{Var}(Y^2)} = A_Y \operatorname{Var}(Y)$ and $\mathbb{E}[Y] = B_Y\sqrt{\operatorname{Var}(Y)}$. Therefore, it is enough that

$$\begin{aligned} A_Y(SNR+1)\sqrt{\frac{8640K}{n}} &\leq 8(SNR+1) - 1, \\ B_Y(SNR+1)\sqrt{\frac{34560K}{n}} &\leq 8(SNR+1) - 1, \\ (SNR+1)\frac{360K}{n} &\leq 8(SNR+1) - 1. \end{aligned}$$

We now divide by $(SNR+1)$ and use $1/(SNR+1) \leq 1$, the stronger condition

$$\begin{aligned} A_Y\sqrt{\frac{8640K}{n}} &\leq 7, \\ B_Y\sqrt{\frac{34560K}{n}} &\leq 7, \\ \frac{360K}{n} &\leq 7, \end{aligned}$$

is then satisfied if $K \leq n/\max\{177A_Y^2, \ 706B_Y^2, \ 52\}$, which is true by assumption on the upper bound on the number of blocks. This completes the proof of $\sigma^{*2} \leq 8\widehat{\sigma}_{K,+}^2$ on the event $\Omega(K)$.

We now deal with $2\widehat{\sigma}_{K,+}^2 \leq 4\operatorname{Var}(Y)$. Since the quantity $-b_{K,n}^2$ is negative, it is sufficient that $2\operatorname{Var}(Y) + 2a_{K,n} + 2\mathbb{E}[Y]b_{K,n} \leq 2\operatorname{Var}(Y)$ and, dividing by σ^{*2},

$$\frac{2a_{K,n}}{\sigma^{*2}} + \frac{2\mathbb{E}[Y]b_{K,n}}{\sigma^{*2}} \leq \frac{2\operatorname{Var}(Y)}{\sigma^{*2}}.$$

We show the stronger inequalities

$$\begin{aligned} \frac{2a_{K,n}}{\sigma^{*2}} &\leq \frac{\operatorname{Var}(Y)}{\sigma^{*2}}, \\ \frac{2\mathbb{E}[Y]b_{K,n}}{\sigma^{*2}} &\leq \frac{\operatorname{Var}(Y)}{\sigma^{*2}}, \end{aligned}$$

by arguing as for the previous step. It is sufficient that

$$A_Y(SNR+1)\sqrt{\frac{60K}{n}} \leq (SNR+1),$$

$$B_Y(SNR+1)\sqrt{\frac{60K}{n}} \leq (SNR+1),$$

which holds if $K \leq n/\max\{60A_Y^2,\ 60B_Y^2\}$, and the latter is true by assumption on the upper bound on the number of blocks. This completes the proof of $2\widehat{\sigma}_{K,+}^2 \leq 4\operatorname{Var}(Y)$ on the event $\Omega(K)$. □

Appendix 3.C Proofs for adaptivity to the sparsity level

3.C.1 A general algorithm for simultaneous adaptivity

In this section, we prove a more general theorem, that will yield Theorem 3.4.7 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 $m \in \{1, \dots, M+1\}$, compute $(\widehat{\boldsymbol{\beta}}_{(2^m)}, \widehat{\sigma}_{(2^m)})$.
- Set
$$\mathcal{M} := \Big\{ m \in \{1, \dots, M\} : \text{ for all } k \geq m,\ |\widehat{\boldsymbol{\beta}}_{(2^{k-1})} - \widehat{\boldsymbol{\beta}}_{(2^k)}|_1 \leq C_1 \widehat{\sigma} w_1(2^k),$$
$$|\widehat{\boldsymbol{\beta}}_{(2^{k-1})} - \widehat{\boldsymbol{\beta}}_{(2^k)}|_2 \leq C_2 \widehat{\sigma} w_2(2^k) \text{ and } |\widehat{\sigma}_{(2^{k-1})} - \widehat{\sigma}_{(2^k)}| \leq C_3 \widehat{\sigma} w_3(2^k) \Big\}.$$
- 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 3.C.1. *Let Θ be a subset of $\mathbb{R}^d \times \mathbb{R}_+$ and $\|\cdot\|$ a norm on Θ. For a given $s \in \{2, \dots, d/(2e)\}$, 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*

$$\inf_{\boldsymbol{\beta}^* \in \mathcal{F}_s,\, \sigma^* > 0} P^{\otimes n}_{\beta^*, P_{\mathbf{X}}, \zeta} \Big(\forall \mathcal{D}' \in \mathcal{D}(N), \|\widehat{\theta}_{(s)}(\mathcal{D}') - \theta^*\| \leq C_1 \sigma^* w(s) \Big) \geq 1 - \widetilde{c}_6 C_2 \left(\frac{s}{ed}\right)^{\widetilde{c}_5 s} - u_n, \tag{3.C.1}$$

$$\inf_{\boldsymbol{\beta}^* \in \widetilde{\mathcal{F}}_{2s},\, \sigma^* > 0} P^{\otimes n}_{\beta^*, P_{\mathbf{X}}, \zeta} \Big(\forall \mathcal{D}' \in \mathcal{D}(N), \|\widehat{\theta}_{(s)}(\mathcal{D}') - \theta^*\| \leq C_1 \sigma^* w(s) \Big) \geq 1 - \widetilde{c}_6 C_2 \left(\frac{2s}{ed}\right)^{2\widetilde{c}_5 s} - u_n, \tag{3.C.2}$$

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 $m = 1, \dots, \lfloor \log_2(s_+) \rfloor$, 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 $b = 1, \dots, s_+$,

$$w(2b) \leq C'' w(b).$$

Theorem 3.C.2 (Joint adaptation of $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})$ to s)**.** *Let $s_+ \in \{2, \dots, d/(2e)\}$ and for $s = 1, \dots, 2s_+$, 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 $s^* \in \{1, \dots, s_+\}$ and $\boldsymbol{\beta}^* \in \widetilde{\mathcal{F}}_{s^*}$, the aggregated estimator $(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}, \widetilde{s})$ satisfies

$$P^{\otimes n}_{\beta^*, P_{\mathbf{X}}, \zeta}\Big(\forall \mathcal{D}' \in \mathcal{D}(N), |\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1 \leq \widetilde{C}_1 \sigma^* w_1(s^*),\ |\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2 \leq \widetilde{C}_2 \sigma^* w_2(s^*),\ |\widetilde{\sigma} - \sigma^*| \leq \widetilde{C}_3 \sigma^* w_3(s^*)\Big)$$
$$\geq 1 - 21(\log_2(s_+) + 1)^2 \left(\widetilde{c}_5 \left(\frac{2s^*}{d}\right)^{2\widetilde{c}_6 s^*} + u_n \right) - 21\, \widetilde{c}_6 \left(\frac{2^{M+1}}{d}\right)^{\widetilde{c}_5 2^{M+1}} - 21 u_n$$

and

$$\mathbb{P}_{\boldsymbol{\beta}^*}\left(\forall \mathcal{D}' \in \mathcal{D}(N), \widetilde{s} \leq s^*\right) \geq 1 - 6(\log_2(s_+) + 1)^2 \left(\widetilde{c}_6 \left(\frac{2s^*}{d}\right)^{2\widetilde{c}_5 s^*} + u_n \right) - 6\, \widetilde{c}_6 \left(\frac{2^{M+1}}{d}\right)^{\widetilde{c}_5 2^{M+1}} - 6u_n.$$

We adapt the proof given in [35, Section 7.3.1] to this new setting where the adaptation is done on both estimators simultaneously. Proof of Theorem 3.C.2 is given in Section 3.C.3.

3.C.2 Proof of Theorem 3.4.7

To prove Theorem 3.4.7, we will apply Theorem 3.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 [35], w_1, w_2 and w_3 satisfy the three conditions in Definition 3.C.1.

It remains to check that the following bounds in probability (3.C.1) and (3.C.2) hold for all $s^* = 1, \dots, s_+$. Applying Theorem 3.4.4 gives

$$\inf_{\boldsymbol{\beta}^* \in \mathcal{F}_{s^*}, \sigma^* > 0} P^{\otimes n}_{\beta^*, P_{\mathbf{X}}, \zeta}\left(\sup_{\mathcal{D}' \in \mathcal{D}(\widetilde{c}_3 \mathfrak{r}_{\mathcal{O}})} \left\{ \mathfrak{r}_2^{-1} \left|\widehat{\sigma}(\mathcal{D}') - \sigma^*\right| \vee \sup_{p \in [1,2]} \mathfrak{r}_p^{-1} \left|\widehat{\boldsymbol{\beta}}(\mathcal{D}') - \boldsymbol{\beta}^*\right|_p \right\} \leq \widetilde{c}_4 \sigma_+ \right) \geq 1 - 4\Big(\frac{s^*}{ed}\Big)^{\widetilde{c}_5 s^*},$$

proving that the bound (3.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 3.4.4 with $2s^*$ instead of s^*, yields

$$\inf_{\boldsymbol{\beta}^* \in \mathcal{F}_{2s^*},\, \sigma^* > 0} P^{\otimes n}_{\beta^*, P_{\mathbf{X}, \zeta}} \Bigg(\forall \mathcal{D}' \in \mathcal{D}(\widetilde{c}_3 \mathfrak{r}_{\mathcal{O}}),\ \Bigg\{ \big|\widehat{\sigma}(\mathcal{D}') - \sigma^*\big| \leq \widetilde{c}_4 \sigma_+ \sqrt{\frac{2s^*}{n} \log\left(\frac{ed}{2s^*}\right)}$$
$$\text{and } \forall p \in [1,2],\ \big|\widehat{\boldsymbol{\beta}}(\mathcal{D}') - \boldsymbol{\beta}^*\big|_1 \leq \widetilde{c}_4 \sigma_+ (2s^*)^{1/p} \sqrt{\frac{1}{n} \log\left(\frac{ed}{2s^*}\right)} \Bigg) \geq 1 - 4\Big(\frac{2s^*}{ed}\Big)^{\widetilde{c}_5 2s^*},$$

proving that the bound (3.C.2) is satisfied with $\widetilde{c}_4$ multiplied by 4.

3.C.3 Proof of Theorem 3.C.2

We choose $s \in [1, s_+]$ and assume that $\boldsymbol{\beta}^* \in \mathcal{F}_s$. Define $\mathbb{P} := \mathbb{P}_{\beta^*, \sigma^*}$ 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 $C_3 w_3(2s_+) \leq 1/2$ and that $\widehat{\sigma}_{(s)}$ robustly converges to σ^* in $|\cdot|$-norm with bound $C_3 \sigma^* w_3(s)$, we get

$$\mathbb{P}\left(\forall \mathcal{D}' \in \mathcal{D}(N), \sigma^*/2 \leq \widehat{\sigma} \leq (3/2)\sigma^*\right) \geq 1 - \widetilde{c}_6 \left(\frac{2^{M+1}}{d}\right)^{\widetilde{c}_5 2^{M+1}} - u_n. \tag{3.C.3}$$

Bound on the probability $\mathbb{P}(\exists \mathcal{D}' \in \mathcal{D}(N),\, \tilde{m} \geq m_0 + 1)$. We have

$$\begin{aligned}
&\mathbb{P}(\exists \mathcal{D}' \in \mathcal{D}(N),\, \tilde{m} \geq m_0 + 1) \leq \sum_{m=m_0+1}^{M} \mathbb{P}(\exists \mathcal{D}' \in \mathcal{D}(N), \tilde{m} = m_0 + 1) \\
&\leq \sum_{m=m_0+1}^{M} \sum_{k=m}^{M} \mathbb{P}\Big(\exists \mathcal{D}' \in \mathcal{D}(N), |\widehat{\boldsymbol{\beta}}_{(2^{k-1})} - \widehat{\boldsymbol{\beta}}_{(2^k)}|_1 > 4C_1 \widehat{\sigma} w_1(2^k) \\
&\qquad\qquad \text{or } |\widehat{\boldsymbol{\beta}}_{(2^{k-1})} - \widehat{\boldsymbol{\beta}}_{(2^k)}|_2 > 4C_2 \widehat{\sigma} w_2(2^k) \text{ or } |\widehat{\sigma}_{(2^{k-1})} - \widehat{\sigma}_{(2^k)}| > 4C_3 \widehat{\sigma} w_3(2^k) \Big) \\
&\leq \sum_{m=m_0+1}^{M} \sum_{k=m}^{M} \mathbb{P}\Big(\exists \mathcal{D}' \in \mathcal{D}(N), \exists p \in [3], d_p\big(\widehat{\theta}^{(p)}_{(2^{k-1})}, \widehat{\theta}^{(p)}_{(2^k)}\big) > 4C_p \widehat{\sigma} w_p(2^k) \Big) \\
&\leq \sum_{p=1}^{3} \sum_{m=m_0+1}^{M} \sum_{k=m}^{M} \mathbb{P}\Big(\exists \mathcal{D}' \in \mathcal{D}(N), d_p\big(\widehat{\theta}^{(p)}_{(2^{k-1})}, \widehat{\theta}^{(p)}_{(2^k)}\big) > 4C_p \widehat{\sigma} w_p(2^k) \Big) \\
&\leq \sum_{p=1}^{3} \sum_{m=m_0+1}^{M} \sum_{k=m}^{M} \mathbb{P}\Big(\exists \mathcal{D}' \in \mathcal{D}(N), d_p\big(\widehat{\theta}^{(p)}_{(2^{k-1})}, \theta^{(p),*}\big) > 4C_p \widehat{\sigma} w_p(2^k) \Big) \\
&\qquad + \mathbb{P}\Big(\exists \mathcal{D}' \in \mathcal{D}(N), d_p\big(\widehat{\theta}^{(p)}_{(2^k)}, \theta^{(p),*}\big) > 4C_p \widehat{\sigma} w_p(2^k) \Big)
\end{aligned}$$

$$\leq 2\sum_{p=1}^{3}\sum_{m=m_0+1}^{M}\sum_{k=m-1}^{M}\mathbb{P}\Big(\exists \mathcal{D}'\in\mathcal{D}(N), d_p\big(\widehat{\theta}^{(p)}_{(2^{k-1})},\theta^{(p),*}\big) > 4C_p\widehat{\sigma}w_p(2^k)\Big)$$

$$\leq 2\sum_{p=1}^{3}\sum_{m=m_0+1}^{M}\sum_{k=m-1}^{M}\mathbb{P}\Big(\exists \mathcal{D}'\in\mathcal{D}(N), d_p\big(\widehat{\theta}^{(p)}_{(2^{k-1})},\theta^{(p),*}\big) > 4C_p\widehat{\sigma}w_p(2^k), \widehat{\sigma}\geq\frac{\sigma}{2}\Big)$$
$$+ 6\,\mathbb{P}\Big(\exists \mathcal{D}'\in\mathcal{D}(N), \hat{\sigma}<\frac{\sigma}{2}\Big).$$

Combining the previous equation with Equation (3.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

$$\begin{aligned}
&\mathbb{P}(\exists \mathcal{D}'\in\mathcal{D}(N),\, \widetilde{m}\geq m_0+1)\\
&\quad\leq 2\sum_{p=1}^{3}\sum_{m=m_0+1}^{M}\sum_{k=m-1}^{M}\mathbb{P}\Big(\exists \mathcal{D}'\in\mathcal{D}(N), d_p\big(\widehat{\theta}^{(p)}_{(2^{k-1})},\theta^{(p),*}\big) > 2C_p\widehat{\sigma}w_p(2^k)\Big)\\
&\qquad\qquad - 6\widetilde{c}_6\left(\frac{2^{M+1}}{d}\right)^{\widetilde{c}_5 2^{M+1}} - 6u_n\\
&\quad\leq 6M^2\widetilde{c}_6\left(\left(\frac{2s}{p}\right)^{2\widetilde{c}_5 s}+u_n\right) - 6\widetilde{c}_6\left(\frac{2^{M+1}}{d}\right)^{2^{M+1}\widetilde{c}_5} - 6u_n\\
&\quad\leq 6(\log_2(s_+)+1)^2\widetilde{c}_6\left(\left(\frac{2s}{p}\right)^{2\widetilde{c}_6 s}+u_n\right) - 6\widetilde{c}_6\left(\frac{2^{M+1}}{d}\right)^{\widetilde{c}_5 2^{M+1}} - 6u_n. \qquad (3.C.4)
\end{aligned}$$

This gives the bound on $\tilde{s}$ as claimed.

Bound on the deviation probability of $\widetilde{\theta}^{(p)}$. For any $a>0$, we have

$$\mathbb{P}\big(\exists \mathcal{D}'\in\mathcal{D}(N), d_p(\widetilde{\theta}^{(p)},\theta^{(p),*})\geq a\big) \leq \mathbb{P}\big(\exists \mathcal{D}'\in\mathcal{D}(N), d_p(\widetilde{\theta}^{(p)},\theta^{(p),*})\geq a, \widetilde{m}\leq m_0\big)$$
$$+\,\mathbb{P}(\exists \mathcal{D}'\in\mathcal{D}(N),\widetilde{m}\geq m_0+1). \qquad (3.C.5)$$

On the event $\{\widetilde{m}\leq m_0\}$, we have the decomposition

$$d_p(\widetilde{\theta}^{(p)},\theta^{(p),*}) \leq \sum_{k=\widetilde{m}+1}^{m_0} d_p\left(\widehat{\theta}^{(p)}_{(2^{k-1})},\widehat{\theta}^{(p)}_{(2^k)}\right) + d_p(\widehat{\theta}^{(p)}_{(2^{m_0})},\theta^{(p),*}). \qquad (3.C.6)$$

Using the assumption on the function w_p, we get that

$$\begin{aligned}
\sum_{k=\widetilde{m}+1}^{m_0} d_p\left(\widehat{\theta}^{(p)}_{(2^{k-1})},\widehat{\theta}^{(p)}_{(2^k)}\right) &\leq \sum_{k=\widetilde{m}+1}^{m_0} 4\hat{\sigma}C_0 w(2^k)\\
&\leq 4\hat{\sigma}C_pC'w_p(2^{m_0}) \leq 4\hat{\sigma}C_pC'C''w_p(s). \qquad (3.C.7)
\end{aligned}$$

We have $2^{m_0}\leq 2s$, therefore applying Assumption (3.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 $\mathcal{D}'\in\mathcal{D}(N)$,

$$d_p(\widehat{\theta}^{(p)}_{(2^{m_0})},\theta^{(p),*}) \leq C_p\widehat{\sigma}w(2s) \leq C_pC''\widehat{\sigma}w(s). \qquad (3.C.8)$$

Combining Equations (3.C.6), (3.C.7), (3.C.8) and (3.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 $\mathcal{D}'\in\mathcal{D}(N)$,

$$d_p(\widetilde{\theta}^{(p)},\theta^{(p),*}) \leq \big(4C_pC'C''+(3/2)C_pC''\big)\,\sigma w(s). \qquad (3.C.9)$$

Combining Equation (3.C.4) with Equations (3.C.5) and (3.C.9), we finally get that

$$\mathbb{P}\left(\exists \mathcal{D}' \in \mathcal{D}(N), d_p(\widetilde{\theta}^{(p)}, \theta^{(p),*}) \geq \left(4C_pC'C'' + (3/2)C_pC''\right)\sigma w_p(s)\right)$$
$$\leq 7(\log_2(s_+)+1)^2\left(\widetilde{c}_6\left(\frac{2s}{p}\right)^{2\widetilde{c}_5 s} + u_n\right) - 7\,\widetilde{c}_6\left(\frac{2^{M+1}}{d}\right)^{\widetilde{c}_5 2^{M+1}} - 7u_n.$$

By a union bound, we then obtain

$$\mathbb{P}_{\boldsymbol{\beta}^*,\sigma^*}\left(\forall \mathcal{D}' \in \mathcal{D}(N), \forall p = 1,2,3,\ d_p(\widetilde{\theta}^{(p)}, \theta^{(p),*}) \geq \left(4C'C'' + (3/2)C''\right)C_p\sigma w_p(s)\right)$$
$$\geq 1 - 21(\log_2(s_+)+1)^2\left(\widetilde{c}_6\left(\frac{2s}{d}\right)^{2\widetilde{c}_5 s} + u_n\right) - 21\,\widetilde{c}_6\left(\frac{2^{M+1}}{d}\right)^{\widetilde{c}_5 2^{M+1}} - 21u_n,$$

as claimed.

Appendix 3.D Auxiliary results

In this section we give auxiliary results that are used in the proofs of the main results.

Lemma 3.D.1 (Lemma 6 in suppl. mat. of [55]). *Let $\rho \geq 0$ and denote $\Gamma_{f^*}(\rho) := \bigcup_{f\in\mathcal{F}:\,\|f-f^*\|\leq\rho/20}\left(\partial\|\cdot\|\right)_f$. 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 3.2.4. For any $K \in \mathbb{N}$, set $[K] = \{1,\dots,K\}$. For all $\alpha \in (0,1)$ the α-quantile of a vector $\mathbf{x} = (x_1,\dots,x_K) \in \mathbb{R}^K$ 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 3.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* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^K$, $Q_\alpha[\mathbf{x}+\mathbf{y}] \le Q_{\alpha+\beta}[\mathbf{x}] + Q_{1-\beta}[\mathbf{y}]$.

Proof of Lemma 3.D.2. We prove property 1. Write $\mathbf{x} = (x_j)_{j\in[K]}$. The property $Q_\beta[\mathbf{x}] \le Q_\alpha[\mathbf{x}]$ is true by construction, because $Q_\alpha[\mathbf{x}] \le u$ implies that there are at least $\alpha K \ge \beta K$ components such that $x_j \le u$.

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 $x_j \ge u$ and at least αK components such that $x_j \le u$. We now show that $u \ge -Q_{1-\alpha}[-\mathbf{x}]$. This is equivalent to $Q_{1-\alpha}[-\mathbf{x}] \ge -u$, which requires at least αK components such that $-x_j \ge -u$, that is, $x_j \le u$. 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 $ax_j \ge u$ and at least αK components such that $ax_j \le u$. Thus, the same components also satisfy $ax_j + b \ge u+b$ or $ax_j+b \le u+b$. 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 $\mathrm{sgn}(a)x_j \ge u/|a|$ and at least αK components $\mathrm{sgn}(a)x_j \le u/|a|$. The latter conditions are equivalent to $|a|\,\mathrm{sgn}(a)x_j \ge u$ and $|a|\,\mathrm{sgn}(a)x_j \le u$. 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 $x_j \ge u$;
- there are at least $(\alpha+\beta)K$ components $x_j \le u$;
- there are at least $(1-\beta)K$ components $y_i \ge l$;
- there are at least βK components $y_i \le l$.

With $(\mathbf{x}-\mathbf{y}) = (x_k - y_k)_{k\in[K]}$, we want to show that $Q_\alpha[\mathbf{x}-\mathbf{y}] \le u-l$, which means there are αK components $x_k - y_k \le u-l$. We now count how many times this inequality fails. In order for a component to be $x_k - y_k \ge u-l$, it is necessary that either $x_k \ge u$, which can happen at most $(1-\alpha-\beta)K$ times, or $y_k \le l$, which can happen at most βK times. Therefore, the inequality $x_k - y_k \ge u-l$ is satisfied by at most $(1-\alpha-\beta)K + \beta K = (1-\alpha)K$ components, leaving at least αK components where $x_k - y_k \le u - l$. This is enough to conclude.

Property 5 is a consequence of property 4 and property 2. □

In the following, we use the notation $[K] = \{1, \dots, K\}$ and $[K]_I := \{k \in [K] : B_k \subset \mathcal{I}\}$. We denote by K_I the cardinality of $[K]_I$.

Lemma 3.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* $\gamma(1 - KVar(Z)/(n\delta_{K,n}^2) - x) \ge \max\{\eta, 1-\eta\}$. *Let* $K \in [|\mathcal{O}|/(1-\gamma), n]$. *There exists an event* $\Omega = \Omega(Z, K)$ *with* $\mathbb{P}(\Omega) \ge 1 - \exp(-K\gamma x^2/2)$ *such that, on this event*

$$\left|\{k \in [K] : |\mathbb{P}_{B_k}(Z) - \mathbb{E}[Z]| \le \delta_{K,n}\}\right| \ge \max\{\eta, 1-\eta\}K,$$

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 3.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

$$\sum_{k\in[K]_I} \Big(\mathbf{1}\{|\mathbb{P}_{B_k}(Z) - \mathbb{E}[Z]| \geq \delta_{K,n}\} - \mathbb{P}_{\mathbf{X}}\{|\mathbb{P}_{B_k}(Z) - \mathbb{E}[Z]| \geq \delta_{K,n}\}\Big) \leq xK_I,$$

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 $\Omega(Z, K)$,

$$|\{k \in [K]_I : |\mathbb{P}_{B_k}(Z) - \mathbb{E}[Z]| \leq \delta_{K,n}\}| \geq K_I \left(1 - \frac{KVar[Z]}{n\delta_{K,n}^2} - x\right) \geq K\gamma \left(1 - \frac{KVar[Z]}{n\delta_{K,n}^2} - x\right),$$

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 $[\mathbb{E}[Z] - \delta_{K,n}, \mathbb{E}[Z] + \delta_{K,n}]$. □

Lemma 3.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* (3.A.1).

Proof of Lemma 3.D.4. We use Lemma 3.D.3 with $\eta = 1/8$, $Z = \zeta^2$, $\mathrm{Var}(Z) = \mathbb{E}[\zeta^4] - \mathbb{E}[\zeta^2]^2 = \sigma^{*4}(\kappa^* - 1)$, $\eta = 1/8$, $\gamma = 15/16$, $x = 1/45$, and $\delta_{K,n}^2 \geq 25(K/n)\,\mathrm{Var}(Z)$. Then,

$$\gamma \left(1 - x - \frac{KVar(Z)}{n\delta_{K,n}^2}\right) \geq \frac{15}{16}\left(1 - \frac{1}{45} - \frac{1}{25}\right) = \frac{15}{16} - \frac{7}{120} > \frac{7}{8} = 1 - \eta.$$

The probability of the corresponding event is $\mathbb{P}(\Omega(K)) \geq 1 - \exp(-K\gamma x^2/2) = 1 - \exp(-K/4320)$. □

Lemma 3.D.5 (Lemma 3 in suppl. mat. of [55])**.** *Grant Assumption 3.3.1. 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*

$$\left|\left\{k \in [K] : \mathbb{P}_{B_k}(f - f^*)^2 \geq (4\theta_0)^{-2}\|f - f^*\|_{2,\mathbf{X}}^2\right\}\right| \geq (1-\eta)K$$

In particular, $Q_{\eta,K}[(f - f^*)^2] \geq (4\theta_0)^{-2}\|f - f^*\|_{2,\mathbf{X}}^2$.

Lemma 3.D.6 (Lemma 4 in suppl. mat. of [55])**.** *Grant Assumption 3.3.1. 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 $\|f-f^*\| \leq \rho$,*

$$\left|\left\{k \in [K] : |(\mathbb{P}_{B_k} - \mathbb{E})(2\zeta(f-f^*)| \leq \alpha_M^2\right\}\right| \geq (1-\eta)K,$$

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 3.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 3.3.3 and $\|f-f^*\|_{2,\mathbf{X}} \leq r(\rho)$, we find $\alpha_M^2 \leq 4\varepsilon r^2(\rho)$.

Proof of Lemma 3.D.7. The first property follows from applying Lemma 3.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 3.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 $r^2(\rho_K) = 384\theta_m^2 K/(n\varepsilon^2)$,

$$\frac{\alpha_M^2}{r^2(2\rho_K)} \leq \frac{\alpha_M^2}{r^2(\rho_K)} = \varepsilon \max\left(\frac{1488\theta_m^2}{\varepsilon^2}\frac{K}{n}\frac{1}{r^2(\rho_K)},\ 1\right) = \varepsilon\frac{1488}{384} < 4\varepsilon.$$

□

Lemma 3.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 3.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 the supplementary material of [55]. 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

$$\begin{aligned}
g_f(\mathcal{D}_k) &:= (\mathbb{P}_{B_k} - \mathbb{E})[(f - f^*)^2],\\
\alpha_Q^2(f) &:= \varepsilon \max\left(\|f - f^*\|_{2,\mathbf{X}}^2 \frac{4\theta_1^4}{\varepsilon^2\alpha} \cdot \frac{K}{n}, r_Q^2(\rho, \gamma_Q), \|f - f^*\|_{2,\mathbf{X}}^2\right).
\end{aligned}$$

Let $[K]_I = \{k \in [K] : B_k \subset \mathcal{I}\}$ and consider any $k \in [K]_I$. An application of Markov inequality gives

$$\mathbb{P}\big(2|g_f(\mathcal{D}_k)| \geq \alpha_Q^2(f)\big) \leq \frac{4\mathbb{E}\Big[|g_f(\mathcal{D}_k)|^2\Big]}{\alpha_Q^2(f) \cdot \alpha_Q^2(f)}.$$

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 $\alpha_Q^2(f) \geq \|f - f^*\|_{2,\mathbf{X}}^2 4\theta_1^4 K/(\varepsilon\alpha n)$. Since $\|f - f^*\|_{4,\mathbf{X}} \leq \theta_1\|f - f^*\|_{2,\mathbf{X}}$ by Assumption 3.3.1, this gives

$$\begin{aligned}
\mathbb{P}\big(2|g_f(\mathcal{D}_k)| \geq \alpha_Q^2(f)\big) &\leq \frac{4\mathbb{E}\Big[\big((\mathbb{P}_{B_k} - \mathbb{P}_{\mathbf{X}})(f - f^*)^2\big)^2\Big]}{\|f - f^*\|_{2,\mathbf{X}}^2 \frac{4\theta_1^4}{\alpha}\frac{K}{n}\|f - f^*\|_{2,\mathbf{X}}^2}\\
&\leq \frac{\sum_{i \in B_k} \operatorname{Var}\big((f - f^*)^2(\mathbf{X}_i)\big)}{|B_k|^2 \frac{\theta_1^4}{\alpha}\frac{K}{n}\|f - f^*\|_{2,\mathbf{X}}^4}\\
&\leq \frac{\mathbb{E}[(f - f^*)^4(\mathbf{X})]}{|B_k|\frac{\theta_1^4}{\alpha}\frac{K}{n}\|f - f^*\|_{2,\mathbf{X}}^4}\\
&\leq \frac{\alpha\|f - f^*\|_{4,\mathbf{X}}^4}{\theta_1^4\|f - f^*\|_{2,\mathbf{X}}^4}
\end{aligned}$$

$$\leq \alpha.$$

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 $\|f - f^*\|_{2,\mathbf{X}} \leq r_Q(\rho)$. With the argument in the proof of Lemma 4 in the supplementary material of [55], one finds

$$\mathbb{E}\left[\sup_{f\in\mathbb{B}(f^*,\rho)} \sum_{k\in[K]_I} \xi_k \frac{g_f(\mathcal{D}_k)}{\alpha_Q^2(f)}\right] \leq \frac{2}{\varepsilon r_Q^2(\rho)} \mathbb{E}\left[\sup_{f\in\mathbb{B}(f^*,\rho,r_Q(\rho))} \left|\sum_{k\in[K]_I} \xi_k(\mathbb{P}_{B_k} - \mathbb{E})(f - f^*)^2\right|\right],$$

and, with the definition of $r_Q(\cdot)$ and the symmetrization argument in the same reference,

$$\mathbb{E}\left[\sup_{f\in\mathbb{B}(f^*,\rho)} \sum_{k\in[K]_I} \xi_k \frac{g_f(\mathcal{D}_k)}{\alpha_Q^2(f)}\right] \leq \frac{4K}{\varepsilon n}\gamma_Q |[K]_I| \frac{n}{K} = \frac{4\gamma_Q}{\varepsilon}|[K]_I|.$$

In the same proof, the authors define a suitable function ψ such that, on an event $\Omega(K)$ with probability at least $1 - \exp(-K\gamma x^2/2) = 1 - \exp(-K/8928)$,

$$\begin{aligned}
&\sum_{k\in[K]_I} \mathbf{1}\big(|g_f(\mathcal{D}_k)| < \alpha_Q^2(f)\big) \\
&\quad\geq (1-\alpha)|[K]_I| - 2\mathbb{E}\left[\sup_{f\in\mathbb{B}(f^*,\rho)} \sum_{k\in[K]_I} \psi\left(\frac{|g_f(\mathcal{D}_k)|}{\alpha_Q^2(f)}\right)\right] + |[K]_I|x \\
&\quad\geq (1-\alpha)|[K]_I| - 2\mathbb{E}\left[\sup_{f\in\mathbb{B}(f^*,\rho)} \sum_{k\in[K]_I} \xi_k \frac{|g_f(\mathcal{D}_k)|}{\alpha_Q^2(f)}\right] - |[K]_I|x \\
&\quad\geq |[K]_I|\left(1 - \alpha - x - \frac{4\gamma_Q}{\varepsilon}\right) \\
&\quad\geq \gamma K\left(1 - \alpha - x - \frac{4\gamma_Q}{\varepsilon}\right).
\end{aligned}$$

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, $Q_{15/16,K}[(f-f^*)^2] \leq \|f - f^*\|_{2,\mathbf{X}}^2 + \alpha_Q^2(f)$. □

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 3.D.4, Lemma 3.D.7 and Lemma 3.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 3.D.9. *Let* $C^2 = 384\theta_1^2 c_r^2 c_\alpha^2 \kappa_+^{1/2}$ *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 $|\sigma - \sigma^*| \leq \alpha_{K,c_\rho}$,

$$\begin{aligned}\mathbb{E}[-2\zeta(f-f^*)(\mathbf{X})] \leq & \frac{2\sigma^* + \alpha_{K,c_\rho}}{2c} T_{K,\mu}(f^*, \sigma^*, f, \sigma) + \frac{2\sigma^* + \alpha_{K,c_\rho}}{2c}\mu\rho + \alpha_M^2 \\ & + \frac{8(2\sigma^* + \alpha_{K,c_\rho})}{c\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + \frac{\alpha_{K,c_\rho}}{c(2\sigma^* - \alpha_{K,c_\rho})}\left(2\sigma^* r(c_\rho\rho_K) + r^2(c_\rho\rho_K) + \alpha_Q^2 + \alpha_M^2\right),\end{aligned}$$

where α_M^2, α_Q^2 are given in Lemma 3.D.7 and Lemma 3.D.8.

Proof of Lemma 3.D.9. We start by applying Lemma 3.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 (3.2.9) and the quantile properties in Lemma 3.D.2, we can rewrite

$$\begin{aligned}&\mathbb{E}[-2\zeta(f-f^*)(\mathbf{X})] \\ &\quad \leq Q_{1/4,K}[(f-f^*)^2 - 2\zeta(f-f^*)] + \alpha_M^2 \\ &\quad = \frac{\sigma+\sigma^*}{2c} Q_{1/4,K}\left[2c\frac{\ell_f - \ell_{f^*}}{\sigma+\sigma^*}\right] + \alpha_M^2 \\ &\quad = \frac{\sigma+\sigma^*}{2c} Q_{1/4,K}\left[R_c(\ell_{f^*}, \sigma^*, \ell_f, \sigma) - (\sigma-\sigma^*)\left(1 - 2\frac{\ell_f + \ell_{f^*}}{(\sigma+\sigma^*)^2}\right)\right] + \alpha_M^2 \\ &\quad \leq \frac{\sigma+\sigma^*}{2c}\left(Q_{1/2,K}\Big[R_c(\ell_{f^*}, \sigma^*, \ell_f, \sigma)\Big] - Q_{1/4,K}\left[(\sigma-\sigma^*)\left(1 - 2\frac{\ell_f + \ell_{f^*}}{(\sigma+\sigma^*)^2}\right)\right]\right) + \alpha_M^2 \\ &\quad \leq \frac{\sigma+\sigma^*}{2c}\left(Q_{1/2,K}\Big[R_c(\ell_{f^*}, \sigma^*, \ell_f, \sigma)\Big] + \mu(\|f\| - \|f^*\|)\right) + \frac{\sigma+\sigma^*}{2c}\mu\rho + \alpha_M^2 \\ &\qquad - \frac{\sigma+\sigma^*}{2c} Q_{1/4,K}\left[(\sigma-\sigma^*)\left(1 - 2\frac{\ell_f + \ell_{f^*}}{(\sigma+\sigma^*)^2}\right)\right] \\ &\quad = \frac{\sigma+\sigma^*}{2c} T_{K,\mu}(f^*, \sigma^*, f, \sigma) + \frac{\sigma+\sigma^*}{2c}\left(\mu\rho - Q_{1/4,K}\left[(\sigma-\sigma^*)\left(1 - 2\frac{\ell_f + \ell_{f^*}}{(\sigma+\sigma^*)^2}\right)\right]\right) + \alpha_M^2.\end{aligned}$$

Since $\sigma + \sigma^* \leq 2\sigma^* + \alpha_{K,c_\rho}$, it remains to show that

$$\begin{aligned}&-\frac{\sigma+\sigma^*}{2c} Q_{1/4,K}\left[(\sigma-\sigma^*)\left(1 - 2\frac{\ell_f + \ell_{f^*}}{(\sigma+\sigma^*)^2}\right)\right] \\ &\qquad \leq \frac{8(2\sigma^* + \alpha_{K,c_\rho})}{c\sigma^*(2\sigma^* - \alpha_{K,c_\rho})^2}\delta_{K,n}^2 + \frac{\alpha_{K,c_\rho}}{c(2\sigma^* - \alpha_{K,c_\rho})}\left(2\sigma^* r(c_\rho\rho_K) + r^2(c_\rho\rho_K) + \alpha_Q^2 + \alpha_M^2\right).\end{aligned} \tag{3.D.1}$$

First, by the quantile properties in Lemma 3.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

$$\begin{aligned}
&\frac{\sigma+\sigma^*}{2c} Q_{3/4,K}\left[(\sigma-\sigma^*)\left(2\frac{\ell_f+\ell_{f^*}}{(\sigma+\sigma^*)^2}-1\right)\right] \\
&\qquad = \frac{\sigma+\sigma^*}{2c} Q_{3/4,K}\left[(\sigma-\sigma^*)\left(\frac{4\ell_{f^*}}{(\sigma+\sigma^*)^2}-1\right)+(\sigma-\sigma^*)\frac{2(\ell_f-\ell_{f^*})}{(\sigma+\sigma^*)^2}\right] \\
&\qquad \le \frac{\sigma+\sigma^*}{2c} Q_{7/8,K}\left[(\sigma-\sigma^*)\left(\frac{4\ell_{f^*}}{(\sigma+\sigma^*)^2}-1\right)\right]+\frac{Q_{7/8,K}\left[(\sigma-\sigma^*)(\ell_f-\ell_{f^*})\right]}{c(\sigma+\sigma^*)} \\
&\qquad = \frac{\sigma+\sigma^*}{2c} T_1 + T_2.
\end{aligned} \tag{3.D.2}$$

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

$$\begin{aligned}
T_1 &= Q_{7/8,K}\left[(\sigma-\sigma^*)\left(\frac{4\ell_{f^*}}{(\sigma+\sigma^*)^2}-1\right)\right] \\
&\le \max\left\{\sup_{\sigma\in(\sigma^*,\sigma^*+\alpha_{K,c_\rho}]}(\sigma-\sigma^*)\left(\frac{4Q_{7/8,K}[\ell_{f^*}]}{(\sigma+\sigma^*)^2}-1\right), \sup_{\sigma\in[\sigma^*-\alpha_{K,c_\rho},\sigma^*)}(\sigma^*-\sigma)\left(1-\frac{4Q_{7/8,K}[\ell_{f^*}]}{(\sigma+\sigma^*)^2}\right)\right\}.
\end{aligned}$$

Thanks to Lemma 3.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

$$T_1 \le \max\left\{\sup_{\sigma\in(\sigma^*,\sigma^*+\alpha_{K,c_\rho}]}(\sigma-\sigma^*)\left(\frac{4(\sigma^{*2}+\delta_{K,n})}{(\sigma+\sigma^*)^2}-1\right), \sup_{\sigma\in[\sigma^*-\alpha_{K,c_\rho},\sigma^*)}(\sigma^*-\sigma)\left(1-\frac{4(\sigma^{*2}-\delta_{K,n})}{(\sigma+\sigma^*)^2}\right)\right\}. \tag{3.D.3}$$

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 $|\sigma_{a_+} - \sigma^*| \le 2a_+ - 2\sigma^* = 2\sigma^*(\sqrt{1+\delta_{K,n}/\sigma^{*2}}-1)$. By construction, the quantity $\delta_{K,n}/\sigma^{*2}$ is smaller than one, since

$$\frac{\delta_{K,n}^2}{\sigma^{*4}} = \frac{25\mu^{*4}K}{\sigma^{*4}n} = \frac{25\kappa^* K}{n} \le \frac{25\kappa_+ K}{n} \le 1$$

and $K \le n/(25\kappa_+)$. For all $x \in (0,1)$, the inequality $\sqrt{1+x} \le 1+x$ holds, so that

$$|\sigma_{a_+} - \sigma^*| \le 2\sigma^*\left(\sqrt{1+\frac{\delta_{K,n}}{\sigma^{*2}}}-1\right) \le 2\sigma^*\left(1+\frac{\delta_{K,n}}{\sigma^{*2}}-1\right) = \frac{2\delta_{K,n}}{\sigma^*}.$$

Now we repeat the same argument for the second function in (3.D.3), using $\sqrt{1-x} \ge 1-x$ for all $x \in (0,1)$, thus getting a point σ_{a_-} achieving the maximum such that $|\sigma_{a_-} - \sigma^*| \le 2\delta_{K,n}/\sigma^*$. By Lemma 3.A.1, we have $2\delta_{K,n}/\sigma^* < \alpha_{K,c_\rho} < \sigma^*$. With $\delta_a = 2\delta_{K,n}/\sigma^*$, this yields

$$\begin{aligned}
&Q_{7/8,K}\left[(\sigma-\sigma^*)\left(\frac{4\ell_{f^*}}{(\sigma+\sigma^*)^2}-1\right)\right]\\
&\quad\leq \max\left\{(\sigma^*-\sigma_{a_-})\left(1-\frac{4a_-^2}{(\sigma_{a_-}+\sigma^*)^2}\right),\ (\sigma_{a_+}-\sigma^*)\left(\frac{4a_+^2}{(\sigma_{a_+}+\sigma^*)^2}-1\right)\right\}\\
&\quad\leq \frac{2\delta_{K,n}}{\sigma^*}\max\left\{1-\frac{4\sigma^{*2}-4\delta_{K,n}}{(2\sigma^*-\delta_a)^2},\ \frac{4\sigma^{*2}+4\delta_{K,n}}{(2\sigma^*+\delta_a)^2}-1\right\}\\
&\quad= \frac{2\delta_{K,n}}{\sigma^*}\max\left\{\frac{4\sigma^*\delta_a+\delta_a^2+4\delta_{K,n}}{(2\sigma^*-\delta_a)^2},\ \frac{4\delta_{K,n}-4\sigma^*\delta_a-\delta_a^2}{(2\sigma^*+\delta_a)^2}\right\}\\
&\quad\leq \frac{16\delta_{K,n}^2}{\sigma^*(2\sigma^*-\delta_a)^2}\\
&\quad\leq \frac{16\delta_{K,n}^2}{\sigma^*(2\sigma^*-\alpha_{K,c_\rho})^2}.
\end{aligned}\tag{3.D.4}$$

It remains to bound T_2 in (3.D.2) in order to obtain (3.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 $\ell_{f^*}-\ell_f=2\zeta(f-f^*)-(f-f^*)^2$ and using Lemma 3.D.7 and Lemma 3.D.8 in the last inequality, we get

$$\begin{aligned}
T_2&=\frac{(\sigma-\sigma^*)}{c(\sigma+\sigma^*)}Q_{7/8,K}\left[(f-f^*)^2-2\zeta(f-f^*)\right]\\
&\leq\frac{\alpha_{K,c_\rho}}{c(2\sigma^*-\alpha_{K,c_\rho})}\left(Q_{15/16,K}\left[(f-f^*)^2\right]+Q_{15/16,K}\left[-2\zeta(f-f^*)\right]\right)\\
&\leq\frac{\alpha_{K,c_\rho}}{c(2\sigma^*-\alpha_{K,c_\rho})}\left(\|f-f^*\|_{2,\mathbf{X}}^2+\alpha_Q^2+\mathbb{E}[-2\zeta(f-f^*)(\mathbf{X})]+\alpha_M^2\right).
\end{aligned}$$

By the Cauchy-Schwarz inequality, $\mathbb{E}[-2\zeta(f-f^*)(\mathbf{X})]\leq 2\sigma^*\|f-f^*\|_{2,\mathbf{X}}\leq 2\sigma^* r(c_\rho\rho_K)$.

We now put together all the bounds found so far and conclude

$$\begin{aligned}
\mathbb{E}[-2\zeta(f-f^*)(\mathbf{X})]&\leq\frac{2\sigma^*+\alpha_{K,c_\rho}}{2c}T_{K,\mu}(f^*,\sigma^*,f,\sigma)+\frac{2\sigma^*+\alpha_{K,c_\rho}}{2c}\mu\rho+\alpha_M^2\\
&\quad+\frac{8(2\sigma^*+\alpha_{K,c_\rho})}{c\sigma^*(2\sigma^*-\alpha_{K,c_\rho})^2}\delta_{K,n}^2+\frac{\alpha_{K,c_\rho}}{c(2\sigma^*-\alpha_{K,c_\rho})}\left(2\sigma^* r(c_\rho\rho_K)+r^2(c_\rho\rho_K)+\alpha_Q^2+\alpha_M^2\right),
\end{aligned}$$

which gives the claim. □

Bibliography

[1] Adler, R. J., and Taylor, J. E. *Random fields and geometry*. Springer Monographs in Mathematics. Springer, New York, 2007.

[2] Alon, N., Matias, Y., and Szegedy, M. The space complexity of approximating the frequency moments. *J. Comput. System Sci. 58*, 1, part 2 (1999), 137–147. Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996).

[3] Arendarczyk, M. On the asymptotics of supremum distribution for some iterated processes. *Extremes 20*, 2 (2017), 451–474.

[4] Aurzada, F., and Lifshits, M. On the small deviation problem for some iterated processes. *Electron. J. Probab. 14* (2009), no. 68, 1992–2010.

[5] Azzalini, A. Further results on a class of distributions which includes the normal ones. *Statistica (Bologna) 46*, 2 (1986), 199–208.

[6] Azzalini, A., and Dalla Valle, A. The multivariate skew-normal distribution. *Biometrika 83*, 4 (1996), 715–726.

[7] Bellec, P., Lecué, G., and Tsybakov, A. Towards the study of least squares estimators with convex penalty. In *Actes du 1er Congrès National de la SMF—Tours, 2016*, vol. 31 of *Sémin. Congr.* Soc. Math. France, Paris, 2017, pp. 109–136.

[8] Bellec, P., and Tsybakov, A. Bounds on the prediction error of penalized least squares estimators with convex penalty. In *Modern problems of stochastic analysis and statistics*, vol. 208 of *Springer Proc. Math. Stat.* Springer, Cham, 2017, pp. 315–333.

[9] Bellec, P. C., Lecué, G., and Tsybakov, A. B. Slope meets Lasso: improved oracle bounds and optimality. *Ann. Statist. 46*, 6B (2018), 3603–3642.

[10] Belloni, A., Chernozhukov, V., and Wang, L. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika 98*, 4 (2011), 791–806.

[11] Belloni, A., Chernozhukov, V., and Wang, L. Pivotal estimation via square-root Lasso in nonparametric regression. *Ann. Statist. 42*, 2 (2014), 757–788.

[12] Bickel, P. J., and Kleijn, B. J. K. The semiparametric Bernstein-von Mises theorem. *Ann. Statist. 40*, 1 (2012), 206–237.

[13] Bierkens, J., Grazzi, S., van der Meulen, F., and Schauer, M. Sticky PDMP samplers for sparse and local inference problems. *arXiv e-prints* (2021), arXiv:2103.08478.

[14] Bolley, F. Quantitative concentration inequalities on sample path space for mean field interaction. *ESAIM Probab. Stat. 14* (2010), 192–209.

[15] Burdzy, K. Some path properties of iterated Brownian motion. In *Seminar on Stochastic Processes, 1992 (Seattle, WA, 1992)*, vol. 33 of *Progr. Probab.* Birkhäuser Boston, Boston, MA, 1993, pp. 67–87.

[16] Casse, J., and Marckert, J.-F. Processes iterated *ad libitum*. *Stochastic Process. Appl. 126*, 11 (2016), 3353–3376.

[17] Castillo, I. Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat. 2* (2008), 1281–1299.

[18] Castillo, I. Semiparametric Bernstein–von Mises theorem and bias, illustrated with Gaussian process priors. *Sankhya A 74*, 2 (2012), 194–221.

[19] Castillo, I. A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields 152*, 1-2 (2012), 53–99.

[20] Castillo, I., and Mismer, R. Empirical Bayes analysis of spike and slab posterior distributions. *Electron. J. Stat. 12*, 2 (2018), 3953–4001.

[21] Castillo, I., and Nickl, R. Nonparametric Bernstein-von Mises theorems in Gaussian white noise. *Ann. Statist. 41*, 4 (2013), 1999–2028.

[22] Castillo, I., and Nickl, R. On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist. 42*, 5 (2014), 1941–1969.

[23] Castillo, I., and Rousseau, J. A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist. 43*, 6 (2015), 2353–2383.

[24] Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. Bayesian linear regression with sparse priors. *Ann. Statist. 43*, 5 (2015), 1986–2018.

[25] Castillo, I., and van der Vaart, A. Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *Ann. Statist. 40*, 4 (2012), 2069–2101.

[26] Chernozhukov, V., and Hong, H. Likelihood estimation and inference in a class of nonregular econometric models. *Econometrica 72*, 5 (2004), 1445–1480.

[27] Cohen, S., and Istas, J. *Fractional fields and applications*, vol. 73 of *Mathématiques & Applications (Berlin) [Mathematics & Applications].* Springer, Heidelberg, 2013. With a foreword by Stéphane Jaffard.

[28] Comminges, L., Collier, O., Ndaoud, M., and Tsybakov, A. B. Adaptive robust estimation in sparse vector model. *arXiv preprint arXiv:1802.04230* (2018).

[29] Cox, D. R., and Reid, N. Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B 49*, 1 (1987), 1–39. With a discussion.

[30] Cox, D. R., and Reid, N. A note on the calculation of adjusted profile likelihood. *J. Roy. Statist. Soc. Ser. B 55*, 2 (1993), 467–471.

[31] Curien, N., and Konstantopoulos, T. Iterating Brownian motions, ad libitum. *J. Theoret. Probab. 27*, 2 (2014), 433–448.

[32] Cutajar, K., Bonilla, E. V., Michiardi, P., and Filippone, M. Random feature expansions for deep gaussian processes. In *International Conference on Machine Learning* (2017), PMLR, pp. 884–893.

[33] Damianou, A., and Lawrence, N. Deep Gaussian processes. In *Artificial Intelligence and Statistics* (2013), pp. 207–215.

[34] de Jonge, R., and van Zanten, H. Semiparametric Bernstein–von Mises for the error standard deviation. *Electron. J. Stat. 7* (2013), 217–243.

[35] Derumigny, A. Improved bounds for square-root lasso and square-root slope. *Electron. J. Stat. 12*, 1 (2018), 741–766.

[36] Derumigny, A. *Some statistical results in high-dimensional dependence modeling.* PhD thesis, Université Paris-Saclay (ComUE), 2019.

[37] Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. Sub-Gaussian mean estimators. *Ann. Statist. 44*, 6 (2016), 2695–2725.

[38] Funaki, T. Probabilistic construction of the solution of some higher order parabolic differential equation. *Proc. Japan Acad. Ser. A Math. Sci. 55*, 5 (1979), 176–179.

[39] Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. Convergence rates of posterior distributions. *Ann. Statist. 28*, 2 (2000), 500–531.

[40] Ghosal, S., and Samanta, T. Asymptotic behaviour of Bayes estimates and posterior distributions in multiparameter nonregular cases. *Math. Methods Statist. 4*, 4 (1995), 361–388.

[41] Ghosal, S., and van der Vaart, A. *Fundamentals of nonparametric Bayesian inference*, vol. 44 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge, 2017.

[42] Giné, E., and Nickl, R. *Mathematical foundations of infinite-dimensional statistical models.* Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York, 2016.

[43] Giraud, C. *Introduction to high-dimensional statistics*, vol. 139 of *Monographs on Statistics and Applied Probability.* CRC Press, Boca Raton, FL, 2015.

[44] Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning.* Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2016.

[45] Gordon, R. D. Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *Ann. Math. Statistics 12* (1941), 364–366.

[46] Hayou, S., Doucet, A., and Rousseau, J. On the impact of the activation function on deep neural networks training. In *International Conference on Machine Learning* (2019), PMLR, pp. 2672–2680.

[47] Hewitt, E., and Savage, L. J. Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc. 80* (1955), 470–501.

[48] Hochberg, K. J., and Orsingher, E. Composition of stochastic processes governed by higher-order parabolic and hyperbolic equations. *J. Theoret. Probab. 9*, 2 (1996), 511–532.

[49] Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. On adaptive posterior concentration rates. *Ann. Statist. 43*, 5 (2015), 2259–2295.

[50] Jerrum, M. R., Valiant, L. G., and Vazirani, V. V. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci. 43*, 2-3 (1986), 169–188.

[51] Kiefer, J., and Wolfowitz, J. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist. 27* (1956), 887–906.

[52] Kobayashi, K. Small ball probabilities for a class of time-changed self-similar processes. *Statist. Probab. Lett. 110* (2016), 155–161.

[53] Le Cam, L. *Asymptotic methods in statistical decision theory.* Springer Series in Statistics. Springer-Verlag, New York, 1986.

[54] Le Cam, L. Maximum likelihood: an introduction. *International Statistical Review/Revue Internationale de Statistique* (1990), 153–171.

[55] Lecué, G., and Lerasle, M. Robust machine learning by median-of-means: theory and practice. *Ann. Statist. 48*, 2 (2020), 906–931.

[56] Lecué, G., and Mendelson, S. Learning subgaussian classes: upper and minimax bounds (2013). *Topics in Learning Theory-Societe Mathematique de France,(S. Boucheron and N. Vayatis Eds.)* (2013).

[57] Lecué, G., and Mendelson, S. Regularization and the small-ball method I: Sparse recovery. *Ann. Statist. 46*, 2 (2018), 611–641.

[58] Levin, L. A. Notes for Miscellaneous Lectures. *arXiv e-prints* (mar 2005), cs/0503039.

[59] Li, W. V., and Shao, Q.-M. Gaussian processes: inequalities, small ball probabilities and applications. In *Stochastic processes: theory and methods*, vol. 19 of *Handbook of Statist.* North-Holland, Amsterdam, 2001, pp. 533–597.

[60] Lugosi, G., and Mendelson, S. Regularization, sparse recovery, and median-of-means tournaments. *Bernoulli 25*, 3 (2019), 2075–2106.

[61] Lugosi, G., and Mendelson, S. Risk minimization by median-of-means tournaments. *J. Eur. Math. Soc. (JEMS) 22*, 3 (2020), 925–965.

[62] Massart, P. *Concentration inequalities and model selection*, vol. 1896 of *Lecture Notes in Mathematics.* Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[63] Matthews, A., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. *arXiv e-prints* (Apr. 2018), arXiv:1804.11271.

[64] McCullagh, P., and Tibshirani, R. A simple method for the adjustment of profile likelihoods. *J. Roy. Statist. Soc. Ser. B 52*, 2 (1990), 325–344.

[65] Meijer, E., Rohwedder, S., and Wansbeek, T. Measurement error in earnings data: using a mixture model approach to combine survey and register data. *J. Bus. Econom. Statist. 30*, 2 (2012), 191–201.

[66] Mendelson, S. Upper bounds on product and multiplier empirical processes. *Stochastic Process. Appl. 126*, 12 (2016), 3652–3680.

[67] Mendelson, S. On multiplier processes under weak moment assumptions. In *Geometric aspects of functional analysis*, vol. 2169 of *Lecture Notes in Math.* Springer, Cham, 2017, pp. 301–318.

[68] Neal, R. M. *Bayesian learning for neural networks*, vol. 118 of *Lecture Notes in Statistics.* Springer, New York, NY, 1996.

[69] Nemirovsky, A. S., and Yudin, D. B. a. *Problem complexity and method efficiency in optimization.* A Wiley-Interscience Publication. John Wiley & Sons, Inc., New

York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[70] NEYMAN, J., AND SCOTT, E. L. Consistent estimates based on partially consistent observations. *Econometrica 16* (1948), 1–32.

[71] PELUCHETTI, S., AND FAVARO, S. Infinitely deep neural networks as diffusion processes. In *International Conference on Artificial Intelligence and Statistics* (2020), PMLR, pp. 1126–1136.

[72] POLSON, N. G., AND ROČKOVÁ, V. Posterior concentration for sparse deep learning. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 930–941.

[73] RASMUSSEN, C. E. *Gaussian Processes in Machine Learning.* Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 63–71.

[74] RAY, K., AND SZABÓ, B. Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association* (2021). to appear.

[75] REISS, M., AND SCHMIDT-HIEBER, J. Nonparametric Bayesian analysis of the compound Poisson prior for support boundary recovery. *Ann. Statist. 48*, 3 (2020), 1432–1451.

[76] SCHMIDT-HIEBER, J. Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist. 48*, 4 (2020), 1875–1897.

[77] SPOKOINY, V., AND PANOV, M. Accuracy of Gaussian approximation in nonparametric Bernstein – von Mises Theorem. *arXiv e-prints* (Oct. 2019), arXiv:1910.06028.

[78] SWEETING, T. Discussion of "Parameter orthogonality and approximate conditional inference". *J. Roy. Statist. Soc. Ser. B 49*, 1 (1987), 20–21.

[79] TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B 58*, 1 (1996), 267–288.

[80] VAN DER PAS, S. L., SALOMOND, J.-B., AND SCHMIDT-HIEBER, J. Conditions for posterior contraction in the sparse normal means problem. *Electron. J. Stat. 10*, 1 (2016), 976–1000.

[81] VAN DER VAART, A. W. *Asymptotic statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge, 1998.

[82] VAN DER VAART, A. W., AND VAN ZANTEN, J. H. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist. 36*, 3 (2008), 1435–1463.

[83] van der Vaart, A. W., and van Zanten, J. H. Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, vol. 3 of *Inst. Math. Stat. (IMS) Collect.* Inst. Math. Statist., Beachwood, OH, 2008, pp. 200–222.

[84] van der Vaart, A. W., and Wellner, J. A. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

Summary

This thesis investigates Bayesian and frequentist procedures for challenging high-dimensional estimation problems.

In a Gaussian sequence model, we study the Bayesian approach to estimate the common variance of the observations. A fraction of the means is known to be zero, whereas the non-zero means are treated as nuisance parameters. This model is non-standard in the sense that it induces inconsistent maximum likelihood. We show a general inconsistency result: the posterior distribution does not contract around the true variance as long as the nuisance parameters are drawn from an i.i.d. proper distribution. We also show that consistency is retained by a hierarchical Gaussian mixture prior. For the latter, we recover the asymptotic shape of the posterior in the Bernstein-von Mises sense and show it is non-Gaussian in the case of small means.

In the nonparametric regression model, we study the Bayesian approach to the estimation of a regression function that is characterized by some underlying composition structure, parametrized by a graph and a smoothness index. This model is inspired by deep learning methods, which work well when complex objects have to be built from simpler features. In previous work, a frequentist estimator based on deep neural networks has been shown to be adaptive with respect to the underlying structure and achieve minimax estimation rates. We characterize the contraction rates of the posterior distribution arising from priors induced by the composition of Gaussian processes. With a suitable model selection prior, we show that the posterior achieves the minimax rates of estimation.

In the nonparametric least-squares regression model, we study a frequentist approach to estimate the regression function and the standard deviation of the residuals. The dataset consists of i.i.d. observations contaminated by a small number of outliers, and heavy-tailed residuals. For the case of known standard deviation, robust median-of-means procedures are available, and we extend them to the case of unknown standard deviation. In the sparse linear regression case, the median-of-means estimator yields a robust version of the Lasso, whereas our method yields a robust version of the square-root Lasso thanks to a scale-invariance argument. We also provide an aggregated estimator achieving minimax convergence rates while being adaptive to the unknown sparsity level.

Samenvatting

Deze thesis onderzoekt Bayesiaanse en frequentistische procedures voor uitdagende hoogdimensionale schattingsproblemen.

We bestuderen de Bayesiaanse benadering om de gemeenschappelijke variantie van waarnemingen in een Gaussiaans sequentiemodel te schatten. Een deel van de gemiddelden is bekend en gelijk aan nul. De gemiddelden die niet gelijk zijn aan nul worden behandeld als hinderlijke parameters. Dit model is niet-standaard aangezien het inconsistente maximale waarschijnlijkheid veroorzaakt. We tonen een algemeen inconsistentieresultaat aan: de posterieure verdeling vertoont geen contractie rond de werkelijke variantie zolang de hinderlijke parameters worden gehaald uit een identieke en onderling onafhankelijke 'proper' verdeling. We tonen ook aan dat de consistentie wordt behouden door een hiërarchisch Gaussiaans mengsel prior. Voor dit laatste vinden we de asymptotische vorm van de posterior in de zin van Bernstein-Van Mises en we tonen aan dat dit niet-Gaussiaans is in het geval van kleine gemiddelden.

In het niet-parametrische regressiemodel bestuderen we de Bayesiaanse benadering van de schatting van een regressiefunctie die wordt gekenmerkt door een onderliggende compositiestructuur, geparametriseerd door een grafiek en een gladheidsindex. Dit model is geïnspireerd op *deep learning*-methoden, die goed werken wanneer complexe objecten worden gebouwd met eenvoudigere functies. In eerder onderzoek is aangetoond dat een frequentistische schatter op basis van neurale netwerken zich aanpast aan de onderliggende structuur en dat deze schatter de minimax-schattingssnelheden bereikt. We karakteriseren de contractiesnelheden van de posterior-verdeling die voortkomt uit de priors geïnduceerd door de samenstelling van Gauss-processen. Met een geschikte prior modelselectie tonen we aan dat de posterior de minimax schattingssnelheid bereikt.

In het niet-parametrische kleinste-kwadratenregressiemodel bestuderen we een frequentistische benadering van de schatting van de regressiefunctie en de standaarddeviatie van de residuen. De dataset bestaat uit onafhankelijke en identiek verdeelde waarnemingen die worden vervuild met een paar uitschieters en residuen met een zware staart. In het geval van een bekende standaarddeviatie zijn robuuste *median-of-means* procedures beschikbaar, en wij breiden deze uit naar het geval van een onbekende standaarddeviatie. In het geval van schaarse lineaire regressie levert de *median-of-means*-schatter een robuuste versie van de Lasso, terwijl onze methode een robuuste versie van de wortel Lasso oplevert dankzij een schaalinvariantie-argument. We geven ook een samengevoegde schatter die minimax-convergentiesnelheden bereikt en zich ook aanpast aan het onbekende schaarsteniveau.

Curriculum Vitæ

Gianluca Finocchio is an Italian mathematician born in Tychy (Poland) on December 1st, 1991. He completed secondary school in summer 2010 at Istituto Superiore “S. Spaventa”, Atessa. In the autumn of the same year, he began his studies in Mathematics at the University of Pisa. There, he graduated with a bachelor’s degree in Mathematics in 2014 and a master’s degree in Mathematics (*cum laude*) in 2017. His master thesis focused on the mean field theory of systems of interacting particles and was written under the supervision of Prof. Franco Flandoli and Dr. Dario Trevisan.

In spring 2017 he started his PhD career at Leiden University, working on high-dimensional Bayesian statistics under the supervision of Prof. Johannes Schmidt-Hieber. In spring 2019 he moved to the University of Twente and continued his research by studying the Bayesian counterpart of deep neural networks. Under the additional supervision of Dr. Katharina Proksch and Dr. Alexis Derumigny, he worked on robust estimation via median-of-means. His four-years research project was financed by the Dutch Research Council (NWO).