Adaptive functional linear regression

Fabienne Comte, Jan Johannes

Adaptive functional linear regression

FABIENNE COMTE*
Université Paris Descartes,
Sorbonne Paris Cité

JAN JOHANNES*
Université catholique de Louvain

Abstract

We consider the estimation of the slope function in functional linear regression, where scalar responses are modeled in dependence of random functions. Cardot and Johannes [2010] have shown that a thresholded projection estimator can attain up to a constant minimax-rates of convergence in a general framework which allows to cover the prediction problem with respect to the mean squared prediction error as well as the estimation of the slope function and its derivatives. This estimation procedure, however, requires an optimal choice of a tuning parameter with regard to certain characteristics of the slope function and the covariance operator associated with the functional regressor. As this information is usually inaccessible in practice, we investigate a fully data-driven choice of the tuning parameter which combines model selection and Lepski's method. It is inspired by the recent work of Goldenshluger and Lepski [2011]. The tuning parameter is selected as minimizer of a stochastic penalized contrast function imitating Lepski's method among a random collection of admissible values. This choice of the tuning parameter depends only on the data and we show that within the general framework the resulting data-driven thresholded projection estimator can attain minimax-rates up to a constant over a variety of classes of slope functions and covariance operators. The results are illustrated considering different configurations which cover in particular the prediction problem as well as the estimation of the slope and its derivatives.

Keywords: Adaptation, Model selection, Lepski's method, linear Galerkin approach, prediction, derivative estimation, minimax theory.
AMS 2000 subject classifications: Primary 62GJ05; secondary 62G05, 62G20.

This work was supported by the IAP research network no. P6/03 of the Belgian Government (Belgian Science Policy).

---

*Laboratoire MAP5, UMR CNRS 8145, 45, rue des Saints-Pères, F-75270 Paris cedex 06, France, e-mail: `fabienne.comte@parisdescartes.fr`

*Institut de statistique, biostatistique et sciences actuarielles (ISBA), Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium, e-mail: `jan.johannes@uclouvain.be`

1 Introduction

In functional linear regression the dependence of a real-valued response Y on the variation of a random function X is studied. Typically the functional regressor X is assumed to be square-integrable or more generally to take its values in a separable Hilbert space $\mathbb{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ and norm $\|\cdot\|_{\mathbb{H}}$. Furthermore, we suppose that Y and X are centered, which simplifies the notations and that the dependence between Y and X is linear in the sense that

$$Y = \langle \beta, X \rangle_{\mathbb{H}} + \sigma\varepsilon, \quad \sigma > 0, \tag{1.1}$$

for some slope function $\beta \in \mathbb{H}$ and error term ε with mean zero and variance one. Assuming an independent and identically distributed (iid.) sample of (Y, X), the objective of this paper is the construction of a fully data driven estimation procedure of the slope function β which still can attain minimax-optimal rates of convergence.

Functional linear models have become very important in a diverse range of disciplines, including medicine, linguistics, chemometrics as well as econometrics (see for instance Ramsay and Silverman [2005] and Ferraty and Vieu [2006], for several case studies, or more specific, Forni and Reichlin [1998] and Preda and Saporta [2005] for applications in economics). The main class of estimation procedures of the slope function studied in the statistical literature are based on principal components regression (see *e.g.* Bosq [2000], Frank and Friedman [1993], Cardot et al. [1999], Cardot et al. [2007] or Müller and Stadtmüller [2005] in the context of generalized linear models). The second important class of estimators relies on minimizing a penalized least squares criterion which can be seen as generalization of the ridge regression (c.f. Marx and Eilers [1999] and Cardot et al. [2003]). More recently an estimator based on dimension reduction and threshold techniques has been proposed by Cardot and Johannes [2010] which borrows ideas from the inverse problems community (Efromovich and Koltchinskii [2001] and Hoffmann and Reiß [2008]). It is worth noting that all the proposed estimation procedures rely on the choice of at least one tuning parameter, which in turn, crucially influences the attainable accuracy of the constructed estimator.

It has been shown, for example in Cardot and Johannes [2010], that the attainable accuracy of an estimator of the slope β is essentially determined by a priori conditions imposed on both the slope function and the covariance operator Γ associated to the random function X (defined below). These conditions are usually captured by suitably chosen classes $\mathcal{F} \subset \mathbb{H}$ and $\mathcal{G}$ of slope functions and covariance operators respectively. Typically, the class $\mathcal{F}$ characterizes the level of smoothness of the slope function, while the class $\mathcal{G}$ specifies the decay of the sequence of eigenvalues of Γ. For example, Hall and Horowitz [2007] and Crambes et al. [2009] consider differentiable slope functions and a polynomial decay of the eigenvalues of Γ. Furthermore, given a weighted norm $\|\cdot\|_\omega$ and the completion $\mathcal{F}_\omega$ of $\mathbb{H}$ with respect to $\|\cdot\|_\omega$ we shall measure the performance of an estimator $\widehat{\beta}$ of β by its maximal $\mathcal{F}_\omega$-risk over a class $\mathcal{F} \subset \mathcal{F}_\omega$ of slope functions and a class $\mathcal{G}$ of covariance operators, that is

$$R_\omega[\widehat{\beta}; \mathcal{F}, \mathcal{G}] := \sup_{\beta \in \mathcal{F}} \sup_{\Gamma \in \mathcal{G}} \mathbb{E}\|\widehat{\beta} - \beta\|_\omega^2.$$

This general framework with appropriate choice of the weighted norm $\|\cdot\|_\omega$ allows us to cover the prediction problem with respect to the mean squared prediction error (see e.g. Cardot et al. [2003] or Crambes et al. [2009]) and the estimation not only of the slope function (see e.g. Hall and Horowitz [2007]) but also of its derivatives. For a detailed discussion, we refer to Cardot and Johannes [2010]. Having these applications in mind the additional condition $\mathcal{F} \subset \mathcal{F}_\omega$ only means that the estimation of a derivative of the slope function necessitates its existence. Assuming an iid. sample of (Y, X) of size n obeying model (1.1) Cardot and Johannes [2010] have derived a lower bound of the maximal weighted risk, that is

$$R_\omega^*[n; \mathcal{F}, \mathcal{G}] \leqslant C \ \inf_{\widehat{\beta}} R_\omega[\widehat{\beta}; \mathcal{F}, \mathcal{G}]$$

for some finite positive constant C where the infimum is taken over all possible estimators $\widehat{\beta}$. Moreover, they have shown that a thresholded projection estimator $\widehat{\beta}_{m_n^*}$ in dependence of an optimally chosen tuning parameter $m_n^* \in \mathbb{N}$ can attain this lower bound up to a constant $C > 0$,

$$R_\omega[\widehat{\beta}_{m_n^*}; \mathcal{F}, \mathcal{G}] \leqslant C \ R_\omega^*[n; \mathcal{F}, \mathcal{G}],$$

for a variety of classes $\mathcal{F}$ and $\mathcal{G}$. In other words, $R_\omega^*[n; \mathcal{F}, \mathcal{G}]$ is the minimax rate of convergence and $\widehat{\beta}_{m_n^*}$ is minimax-optimal. The optimal choice m_n^* of the tuning parameter, however, follows from a classical squared-bias-variance compromise and requires an a-priori knowledge about the classes $\mathcal{F}$ and $\mathcal{G}$, which is usually inaccessible in practice.

In this paper we propose a fully data driven method to select a tuning parameter $\widehat{m}$ in such a way that the resulting data-driven estimator $\widehat{\beta}_{\widehat{m}}$ can still attain the minimax-rate $R_\omega^*[n; \mathcal{F}, \mathcal{G}]$ up to a constant over a variety of classes $\mathcal{F}$ and $\mathcal{G}$. It is interesting to note that, considering a linear regression model with infinitely many regressors, Goldenshluger and Tsybakov [2001, 2003] propose an optimal data-driven prediction procedure allowing sharp oracle inequalities. However, a straightforward application of their results is not obvious to us since they assume a priori standardised regressors, which in turn, in functional linear regression necessitates the covariance operator Γ to be fully known in advance. In contrast, given a jointly normally distributed regressor and error term, Verzelen [2010] establishes sharp oracle inequalities for the prediction problem in case the covariance operator is not known in advance. Although, it is worth noting that considering the mean prediction error as risk eliminates the ill-posedness of the underlying problem, which in turn leads to faster minimax rates of convergences of the prediction error than, for examples, of the mean integrated squared error. On the other hand covering both of these two risks within the general framework discussed above Comte and Johannes [2010] consider functional linear regression with circular functional regressor which results in a partial knowledge of the associated covariance operator, i.e. its eigenfunctions are known in advance but the eigenvalues have to be estimated. In this situation Comte and Johannes [2010] have applied successfully a model selection approach which is inspired by the work of Barron et al. [1999] now extensively discussed in Massart [2007]. In the circular case, it is possible to develop the unknown slope function in the eigenbasis of the covariance operator, which in turn, allows to derive an orthogonal series estimator in dependence of a dimension parameter. This dimension parameter has been chosen fully data driven by a model selection approach and it is shown that the resulting data-driven orthogonal series estimator can attain minimax-optimal rates of convergence up to a constant. Although, the proof crucially relies on the possibility to write the orthogonal series estimator as a minimizer of a contrast.

In this paper we do not impose an a priori knowledge of the eigenbasis and, hence the orthogonal series estimator is no more accessible to us. Instead, we consider the thresholded projection estimator $\widehat{\beta}_m$ as presented in Cardot and Johannes [2010] which we did not succeed to write as a minimizer of a contrast. Therefore, our selection method combines model selection and Lepski's method (c.f. Lepski [1990] and its recent review in Mathé [2006]) which is inspired by a bandwidth selection method in kernel density estimation proposed recently by Goldenshluger and Lepski [2011]. Selecting the dimension parameter $\widehat{m}$ as minimizer of a stochastic penalized contrast function imitating Lepski's method among a random collection of admissible values we show that the fully data-driven estimator $\widehat{\beta}_{\widehat{m}}$ can attain the minimax-rate up to a constant $C > 0$, that is

$$R_\omega[\widehat{\beta}_{\widehat{m}}; \mathcal{F}, \mathcal{G}] \leqslant C \cdot R_\omega[n; \mathcal{F}, \mathcal{G}] \tag{1.2}$$

for a variety of classes $\mathcal{F}$ and $\mathcal{G}$. We shall emphasize that the proposed estimator can attain minimax-optimal rates without specifying in advance neither that the slope function belongs to a class of differentiable or analytic functions nor that the decay of the eigenvalues is polynomial or exponential. The only price for this flexibility is in term of the constant C which is asymptotically not equal to one, i.e. the oracle inequality (1.2) is not sharp.

The paper is organized as follows: in Section 2 we briefly introduce the thresholded projection estimator $\widehat{\beta}_m$ as proposed in Cardot and Johannes [2010]. We present the data driven method to select the tuning parameter and prove a first upper risk-bound for the fully data-driven estimator $\widehat{\beta}_{\widehat{m}}$ which emphasizes the key arguments. In section 3 we review the available minimax theory as presented in Cardot and Johannes [2010]. Within this general framework we derive upper risk-bounds for the fully-data driven estimator imposing additional assumptions on the distribution of the functional regressor X and the error term ε. Namely, we suppose first that X and ε are Gaussian random variables and second that they satisfy certain moment conditions. In both cases the proof of the upper risk-bound employs the key arguments given in Section 2, while more technical aspects are deferred to the appendix. The results in this paper are illustrated considering different configurations of classes $\mathcal{F}$ and $\mathcal{G}$. We recall the minimax-rates in this situations and show that up to a constant these rates are attained by the fully-data driven estimator.

2 Methodology.

Consider the functional linear model (1.1) where the random function X and the error term ε are independent. Let the centered random function X, i.e., $\mathbb{E}\langle X, h\rangle_{\mathbb{H}} = 0$ for all $h \in \mathbb{H}$, have a finite second moment, i.e., $\mathbb{E}\|X\|_{\mathbb{H}}^2 < \infty$. Multiplying both sides in (1.1) by $\langle X, h\rangle_{\mathbb{H}}$ and taking the expectation leads to the normal equation

$$\langle g, h\rangle_{\mathbb{H}} := \mathbb{E}[Y\langle X, h\rangle_{\mathbb{H}}] = \mathbb{E}[\langle \beta, X\rangle_{\mathbb{H}}\langle X, h\rangle_{\mathbb{H}}] =: \langle \Gamma\beta, h\rangle_{\mathbb{H}}, \quad \text{for all } h \in \mathbb{H}, \tag{2.1}$$

where g belongs to $\mathbb{H}$ and Γ denotes the covariance operator associated to the random function X. Throughout the paper we shall assume that there exists a solution $\beta \in \mathbb{H}$ of equation (2.1) and that the covariance operator Γ is strictly positive definite which ensures the identifiability of the slope function β (c.f. Cardot et al. [2003]). However, due to the finite second moment of X the associated covariance operator Γ has a finite trace, i.e. it is nuclear. Thereby, solving equation (2.1) to reconstruct the slope function β is an *ill-posed inverse problem* with the additional difficulty that Γ is unknown and has to be estimated (for a detailed discussion of ill-posed inverse problems in general we refer to Engl et al. [2000]).

2.1 Thresholded projection estimator

In this paper, we follow Cardot and Johannes [2010] and consider a linear Galerkin approach to derive an estimator of the slope function β. Here and subsequently, let $\{\psi_j, j \in \mathbb{N}\}$ be a pre-specified orthornormal basis in $\mathbb{H}$ which in general does not correspond to the eigenbasis of the operator Γ defined in (2.1). With respect to this basis, we consider for all $h \in \mathbb{H}$ the development $h = \sum_{j=1}^{\infty}[h]_j\psi_j$ where the sequence $([h]_j)_{j\geqslant 1}$ with generic elements $[h]_j := \langle h, \psi_j\rangle_{\mathbb{H}}$ is square-summable, i.e., $\|h\|_{\mathbb{H}}^2 = \sum_{j\geqslant 1}[h]_j^2 < \infty$. We will refer to any sequence $(a_n)_{n\in\mathbb{N}}$ as a whole by omitting its index as for example in «the sequence a». Furthermore, given $m \in \mathbb{N}$ denote $[h]_{\underline{m}} := ([h]_1, \dots, [h]_m)^t$ (where x^t denotes the transpose of x) and let $\mathbb{H}_m$ be the subspace of $\mathbb{H}$ spanned by the first m basis functions $\{\psi_1, \dots, \psi_m\}$. Obviously, if $h \in \mathbb{H}_m$ then the norm of h equals the euclidean norm of its coefficient vector $[h]_{\underline{m}}$, i.e., $\|h\|_{\mathbb{H}} = ([h]_{\underline{m}}^t[h]_{\underline{m}})^{1/2} =: \|[h]_{\underline{m}}\|$ with a slight abuse of notations. An element $\beta^m \in \mathbb{H}_m$ is called a Galerkin solution of equation (2.1), if

$$\|g - \Gamma\beta^m\|_{\mathbb{H}} \leqslant \|g - \Gamma\breve{\beta}\|_{\mathbb{H}}, \quad \forall\breve{\beta} \in \mathbb{H}_m. \tag{2.2}$$

Since the covariance operator Γ is strictly positive definite, it follows that the $(m \times m)$-dimensional covariance matrix $[\Gamma]_{\underline{m}} := \mathbb{E}([X]_{\underline{m}}[X]_{\underline{m}}^t)$ associated with the m-dimensional random vector $[X]_{\underline{m}}$ is strictly positive definite too. Consequently, the Galerkin solution $\beta^m \in \mathbb{H}_m$ is uniquely determined by $[\beta^m]_{\underline{m}} = [\Gamma]_{\underline{m}}^{-1}[g]_{\underline{m}}$ and $[\beta^m]_j = 0$ for all $j > m$. However, the Galerkin solution does generally not correspond to the orthogonal projection of the slope function onto the subspace $\mathbb{H}_m$. Moreover, let $(bias_m)_{m\geqslant 1}$ denote a sequence of approximation errors given by $bias_m := \sup_{k\geqslant m}\|\beta^k - \beta\|_\omega$, $m \geqslant 1$. It is important to note that in general without further assumptions the sequence $bias$ does not converge to zero. Here and subsequently, however, we restrict ourselves to classes $\mathcal{F}$ and $\mathcal{G}$ of slope functions and covariance operators respectively which ensure this convergence. Obviously, this is a minimal regularity condition for us since we aim to estimate the Galerkin solution. Assuming a sample $\{(Y_i, X_i)\}_{i=1}^n$ of (Y, X) of size n, it is natural to consider the estimators

$$\widehat{g} := \frac{1}{n}\sum_{i=1}^{n} Y_i\, X_i, \quad \text{and} \quad \widehat{\Gamma} := \frac{1}{n}\sum_{i=1}^{n} \langle \cdot, X_i\rangle_{\mathbb{H}} X_i$$

for g and Γ respectively. Moreover, let $[\widehat{\Gamma}]_{\underline{m}} := \frac{1}{n}\sum_{i=1}^{n}[X_i]_{\underline{m}}[X_i]_{\underline{m}}^t$ be the empirical $(m \times m)$-dimensional covariance matrix and note that $[\widehat{g}]_{\underline{m}} = \frac{1}{n}\sum_{i=1}^{n} Y_i[X_i]_{\underline{m}}$. Replacing in (2.2) the unknown quantities by their empirical counterparts let $\widetilde{\beta}^m \in \mathbb{H}_m$ be a Galerkin solution satisfying

$$\|\widehat{g} - \widehat{\Gamma}\widetilde{\beta}^m\|_{\mathbb{H}} \leqslant \|\widehat{g} - \widehat{\Gamma}\breve{\beta}\|_{\mathbb{H}}, \quad \forall \breve{\beta} \in \mathbb{H}_m.$$

Observe that there exists always a solution $\widetilde{\beta}^m$, but it might not be unique. Obviously, if $[\widehat{\Gamma}]_{\underline{m}}$ is non singular then $[\widetilde{\beta}^m]_{\underline{m}} = [\widehat{\Gamma}]_{\underline{m}}^{-1}[\widehat{g}]_{\underline{m}}$. We shall emphasize the multiplication with the inverse of the random matrix $[\widehat{\Gamma}]_{\underline{m}}$ which may result in an unstable estimator even in case $[\Gamma]_{\underline{m}}$ is well conditioned. Let $\mathbb{1}_{\{\|[\widehat{\Gamma}]_{\underline{m}}^{-1}\|_s \leqslant n\}}$ denote the indicator function which takes the value one if $[\widehat{\Gamma}]_{\underline{m}}$ is non-singular with spectral norm $\|[\widehat{\Gamma}]_{\underline{m}}^{-1}\|_s := \sup_{\|z\|=1}\|[\widehat{\Gamma}]_{\underline{m}}^{-1} z\|$ of its inverse bounded by n, and the value zero otherwise. The estimator $\widehat{\beta}_m$ of β proposed by Cardot and Johannes [2010] consists in thresholding the estimated Galerkin solution, that is,

$$\widehat{\beta}_m := \widetilde{\beta}^m\, \mathbb{1}_{\{\|[\widehat{\Gamma}]_{\underline{m}}^{-1}\|_s \leqslant n\}}\,. \tag{2.3}$$

In the next paragraph we introduce a data-driven method to select the dimension parameter $m \in \mathbb{N}$.

2.2 Data-driven selection of the dimension parameter

Our selection method combines model selection (c.f. Barron et al. [1999] and its discussion in Massart [2007]) and Lepski's method (c.f. Lepski [1990]) borrowing ideas from Goldenshluger and Lepski [2011]. We select the dimension parameter as minimizer of a penalized contrast function depending on the weighted norm $\|\cdot\|_\omega$ which we formalize next. Let $(\omega_j)_{j\geqslant 1}$ be a strictly positive sequence of weights. We define for $h \in \mathbb{H}$ the weighted norm by $\|h\|_\omega^2 := \sum_{j=1}^{\infty}\omega_j[h]_j^2$. Furthermore, for $m \geqslant 1$, $[\nabla_\omega]_{\underline{m}}$ and $[\mathrm{Id}]_{\underline{m}}$ denotes respectively the m-dimensional diagonal matrix with diagonal entries $(\omega_j)_{1\leqslant j\leqslant m}$ and the identity matrix where for all $h \in \mathbb{H}_m$ we have $\|h\|_\omega^2 = [h]_{\underline{m}}^t[\nabla_\omega]_{\underline{m}}[h]_{\underline{m}} = \|[\nabla_\omega]_{\underline{m}}^{1/2}[h]_{\underline{m}}\|^2$. Given a sequence $K := ([K]_{\underline{k}})_{k\geqslant 1}$ of matrices, denote by

$$\Delta_m(K) := \max_{1\leqslant k\leqslant m}\|[\nabla_\omega]_{\underline{k}}^{1/2}[K]_{\underline{k}}^{-1}[\nabla_\omega]_{\underline{k}}^{1/2}\|_s \quad \text{and}$$
$$\delta_m(K) := m\,\Delta_m(K)\,\frac{\log(\Delta_m(K)\vee(m+2))}{\log(m+2)}. \tag{2.4}$$

Take as an example, $\Delta_m^\omega := \Delta_m(K)$ with $K = ([\mathrm{Id}]_{\underline{m}})_{m\geqslant 1}$ which satisfies $\Delta_m^\omega = \max_{1\leqslant k\leqslant m}\omega_k$. For $n \geqslant 1$, set $M_n^\omega := \max\left\{1 \leqslant m \leqslant \lfloor n^{1/4}\rfloor : \Delta_m^\omega \leqslant n\right\}$. The dimension parameter is selected among a collection of admissible values $\{1, \dots, \widehat{M}\}$ with random integer $\widehat{M}$ given by

$$\widehat{M} := \min\left\{2 \leqslant m \leqslant M_n^\omega : m\,\Delta_m^\omega\,\|[\widehat{\Gamma}]_{\underline{m}}^{-1}\|_s > \frac{n}{1+\log n}\right\} - 1, \tag{2.5}$$

where we set $\widehat{M}_n := M_n^\omega$ if the min runs over an empty set and $\lfloor a \rfloor$ denotes as usual the integer part of a. Furthermore we define a stochastic sequence of penalties $(\widehat{pen}_m)_{1\leqslant m\leqslant \widehat{M}_n}$ which takes its inspiration from Comte and Johannes [2010]. Let $\widehat{\delta}_m := \delta_m(K)$ with $K = ([\widehat{\Gamma}]_{\underline{m}})_{m\geqslant 1}$ and

$$\widehat{pen}_m := 14\,\kappa\,\widehat{\sigma}_m^2\,\widehat{\delta}_m\,n^{-1} \quad \text{with} \quad \widehat{\sigma}_m^2 := 2\Big(\frac{1}{n}\sum_{i=1}^n Y_i^2 + [\widehat{g}]_{\underline{m}}^t[\widehat{\Gamma}]_{\underline{m}}^{-1}[\widehat{g}]_{\underline{m}}\Big) \tag{2.6}$$

where κ is a positive constant to be chosen below. The random integer $\widehat{M}$ and the stochastic penalties $(\widehat{pen}_m)_{1\leqslant m\leqslant \widehat{M}_n}$ are used to define the sequence $(\widehat{\Psi}_m)_{1\leqslant m\leqslant \widehat{M}_n}$ of contrast by

$$\widehat{\Psi}_m := \max_{m\leqslant k\leqslant \widehat{M}}\left\{\|\widehat{\beta}_k - \widehat{\beta}_m\|_\omega^2 - \widehat{pen}_k\right\}.$$

Setting $\arg\min_{m\in A}\{a_m\} := \min\{m : a_m \leqslant a_{m'}, \forall m' \in A\}$ for a sequence $(a_m)_{m\geqslant 1}$ with minimal value in $A \subset \mathbb{N}$, we select the dimension parameter

$$\widehat{m} := \mathop{\arg\min}_{1\leqslant m\leqslant \widehat{M}}\left\{\widehat{\Psi}_m + \widehat{pen}_m\right\}. \tag{2.7}$$

The estimator of β is now given by $\widehat{\beta}_{\widehat{m}}$ and below we derive an upper bound for its risk. By construction the choice of the dimension parameter and hence the estimator $\widehat{\beta}_{\widehat{m}}$ do not rely on the regularity assumptions on the slope and the operator which we formalize in Section 3.

2.3 Upper risk bound for the data-driven thresholded projection estimator

The next assertion states the key argument in the proof of the upper risk-bound.

LEMMA 2.1. *Let $(bias_m)_{m\geqslant 1}$ be the sequence of approximation errors $bias_m = \sup_{m\leqslant k}\|\beta^k - \beta\|_\omega$. Consider an arbitrary sequence of penalties $(pen_m)_{m\geqslant 1}$, an upper bound $M \in \mathbb{N}$, and the sequence $(\Psi_m)_{m\geqslant 1}$ of contrasts given by $\Psi_m := \max_{m\leqslant k\leqslant M}\left\{\|\widehat{\beta}_k - \widehat{\beta}_m\|_\omega^2 - pen_k\right\}$. If the subsequence $(pen_1, \dots, pen_M)$ is non-decreasing, then we have for the selected model $\widetilde{m} := \arg\min_{1\leqslant m\leqslant M}\{\Psi_m + pen_m\}$ and for all $1 \leqslant m \leqslant M$ that*

$$\|\widehat{\beta}_{\widetilde{m}} - \beta\|_\omega^2 \leqslant 7\,pen_m + 78\,bias_m^2 + 42\max_{m\leqslant k\leqslant M}\left(\|\widehat{\beta}_k - \beta^k\|_\omega^2 - \frac{1}{6}\,pen_k\right)_+ \tag{2.8}$$

where $(a)_+ = \max(a, 0)$.

PROOF OF LEMMA 2.1. From the definition of $\widetilde{m}$ we deduce for all $1 \leqslant m \leqslant M$ that

$$\begin{aligned}\|\widehat{\beta}_{\widetilde{m}} - \beta\|_\omega^2 &\leqslant 3\Big\{\|\widehat{\beta}_{\widetilde{m}} - \widehat{\beta}_{\widetilde{m}\wedge m}\|_\omega^2 + \|\widehat{\beta}_{\widetilde{m}\wedge m} - \widehat{\beta}_m\|_\omega^2 + \|\widehat{\beta}_m - \beta\|_\omega^2\Big\} \\ &\leqslant 3\Big\{\Psi_m + pen_{\widetilde{m}} + \Psi_{\widetilde{m}} + pen_m + \|\widehat{\beta}_m - \beta\|_\omega^2\Big\} \\ &\leqslant 6\{\Psi_m + pen_m\} + 3\|\widehat{\beta}_m - \beta\|_\omega^2. \end{aligned} \tag{2.9}$$

Since $(pen_1, \dots, pen_M)$ is non-decreasing and $4\,bias_m^2 \geqslant \max_{m\leqslant k\leqslant M}\|\beta^k - \beta^m\|_\omega^2$, it is for all $1 \leqslant m \leqslant M$ easily verified that

$$\Psi_m \leqslant 6 \sup_{m\leqslant k\leqslant M}\Big(\|\widehat{\beta}_k - \beta^k\|_\omega^2 - \frac{1}{6}\,pen_k\Big)_+ + 12\,bias_m^2\,.$$

The last estimate allows us for all $1 \leqslant m \leqslant M$ to write

$$\|\widehat{\beta}_m - \beta\|_\omega^2 \leqslant \frac{1}{3}\,pen_m + 2\,bias_m^2 + 2\sup_{m\leqslant k\leqslant M}\Big(\|\widehat{\beta}_k - \beta^k\|_\omega^2 - \frac{1}{6}\,pen_k\Big)_+.$$

From the last inequality and (2.9), we obtain the assertion (2.8), which completes the proof. $\square$

In addition to the last assertion the proof of the upper risk bound requires two assumptions which we state next. For $n \geqslant 1$ and a positive sequence $a := (a_m)_{m\geqslant 1}$ denote

$$M_n(a) := \min\Big\{2 \leqslant m \leqslant M_n^\omega : m\,\Delta_m^\omega\, a_m > \frac{n}{1+\log n}\Big\} - 1 \tag{2.10}$$

where we set $M_n(a) := M_n^\omega$ if the set is empty. Observe that $\widehat{M}$ given in (2.5) satisfies $\widehat{M} = M_n(a)$ with $a = (\|[\widehat{\Gamma}]_{\underline{m}}^{-1}\|_s)_{m\geqslant 1}$. Consider for $m \geqslant 1$, $\delta_m^\Gamma := \delta_m(K)$ with $K = (\|[\Gamma]_{\underline{m}}^{-1}\|_s)_{m\geqslant 1}$ and

$$pen_m := \kappa\,\sigma_m^2\,\delta_m^\Gamma\,n^{-1} \quad \text{with} \quad \sigma_m^2 := 2\big(\mathbb{E}Y^2 + [g]_{\underline{m}}^t[\Gamma]_{\underline{m}}^{-1}[g]_{\underline{m}}\big) \tag{2.11}$$

which are obviously only theoretical counterparts of the random objects given in (2.6). In order to control the third right hand side term in the upper bound (2.8), the remainder term, we impose the following assumption, though we show in Section 3 under reasonable assumptions on the distribution of ε and X that it holds true for a wide range of classes $\mathcal{F}$ and $\mathcal{G}$.

ASSUMPTION 2.1. *There exist sequences $(m_n^\diamond)_{n\geqslant 1}$ and $(M_n^+)_{n\geqslant 1}$, and a constant K_1 such that*

$$\sup_{\beta\in\mathcal{F}}\sup_{\Gamma\in\mathcal{G}}\mathbb{E}\Big\{\sup_{m_n^\diamond\leqslant k\leqslant M_n^+}\Big(\|\widehat{\beta}_k - \beta^k\|_\omega^2 - \frac{1}{6}\,pen_k\Big)_+\Big\} \leqslant K_1\,n^{-1} \quad \textit{for all } n \geqslant 1.$$

In the following we decompose the risk with respect to an event $\mathcal{E}_n$ where $\widehat{pen}$ is comparable to its theoretical counterpart pen and $\widehat{M}$ lies between $m_n^\diamond$ and M_n^+ given by Assumption 2.1, and its complement $\mathcal{E}_n^c$. To be precise, we define the event

$$\mathcal{E}_n := \big\{pen_k \leqslant \widehat{pen}_k \leqslant 72\,pen_k,\ \forall\, m_n^\diamond \leqslant k \leqslant M_n^+\big\} \cap \Big\{m_n^\diamond \leqslant \widehat{M} \leqslant M_n^+\Big\} \tag{2.12}$$

and consider the elementary identity

$$\sup_{\beta\in\mathcal{F}}\sup_{\Gamma\in\mathcal{G}}\mathbb{E}\|\widehat{\beta}_{\widehat{m}} - \beta\|_\omega^2 = \sup_{\beta\in\mathcal{F}}\sup_{\Gamma\in\mathcal{G}}\mathbb{E}\big(\|\widehat{\beta}_{\widehat{m}} - \beta\|_\omega^2\,\mathbb{1}_{\mathcal{E}_n}\big) + \sup_{\beta\in\mathcal{F}}\sup_{\Gamma\in\mathcal{G}}\mathbb{E}\big(\|\widehat{\beta}_{\widehat{m}} - \beta\|_\omega^2\,\mathbb{1}_{\mathcal{E}_n^c}\big) \tag{2.13}$$

The conditions on the distribution of ε and X presented in the next section are also sufficient to show that the following assumption holds true.

Assumption 2.2. *There exists a constant $K_2 > 0$ such that*

$$\sup_{\beta\in\mathcal{F}}\sup_{\Gamma\in\mathcal{G}}\mathbb{E}\big(\|\widehat{\beta}_{\widehat{m}}-\beta\|_\omega^2\,\mathbb{1}_{\mathcal{E}_n^c}\big)\leqslant K_2\,n^{-1}\qquad \textit{for all } n\geqslant 1.$$

The next assertion provides an upper bound for the maximal $\mathcal{F}_\omega$-risk over the classes $\mathcal{F}$ and $\mathcal{G}$ of the thresholded projection estimator $\widehat{\beta}_{\widehat{m}}$ with data-driven choice $\widehat{m}$ given by (2.7).

Proposition 2.2. *If Assumption 2.1 and 2.2 hold true, then we have*

$$\mathcal{R}_\omega[\widehat{\beta}_{\widehat{m}};\mathcal{F},\mathcal{G}]\leqslant 504\sup_{\beta\in\mathcal{F}}\sup_{\Gamma\in\mathcal{G}}\{pen_{m_n^\diamond}+bias^2_{m_n^\diamond}\}+(504\,K_1+K_2)n^{-1}\qquad \textit{for all } n\geqslant 1.$$

Proof of Proposition 2.2. We make use of the elementary identity (2.13) and taking into account Assumption 2.2 we derive for all $n\geqslant 1$

$$\mathcal{R}_\omega[\widehat{\beta}_{\widehat{m}};\mathcal{F},\mathcal{G}]\leqslant \sup_{\beta\in\mathcal{F}}\sup_{\Gamma\in\mathcal{G}}\mathbb{E}\big(\|\widehat{\beta}_{\widehat{m}}-\beta\|_\omega^2\,\mathbb{1}_{\mathcal{E}_n}\big)+K_2n^{-1}.\tag{2.14}$$

We observe that the random subsequences $(\widehat{\sigma}_1^2,\dots,\widehat{\sigma}^2_{\widehat{M}})$ and hence $(\widehat{pen}_1,\dots,\widehat{pen}_{\widehat{M}})$ are by construction monotonically non-decreasing. Indeed, for all $1\leqslant m\leqslant k\leqslant\widehat{M}$ the identity $\langle\widehat{\Gamma}(\widehat{\beta}_k-\widehat{\beta}_m),(\widehat{\beta}_k-\widehat{\beta}_m)\rangle_{\mathbb{H}}=[\widehat{g}]_{\underline{k}}^t[\widehat{\Gamma}]_{\underline{k}}^{-1}[\widehat{g}]_{\underline{k}}-[\widehat{g}]_{\underline{m}}^t[\widehat{\Gamma}]_{\underline{m}}^{-1}[\widehat{g}]_{\underline{m}}$ holds true. Therefore, by using that $\widehat{\Gamma}$ is positive definite it follows that $[\widehat{g}]_{\underline{m}}^t[\widehat{\Gamma}]_{\underline{m}}^{-1}[\widehat{g}]_{\underline{m}}\leqslant[\widehat{g}]_{\underline{k}}^t[\widehat{\Gamma}]_{\underline{k}}^{-1}[\widehat{g}]_{\underline{k}}$, and hence $\widehat{\sigma}_m^2\leqslant\widehat{\sigma}_k^2$. Consequently, Lemma 2.1 is applicable for all $1\leqslant m\leqslant\widehat{M}$ and we obtain

$$\|\widehat{\beta}_{\widehat{m}}-\beta\|_\omega^2\leqslant 7\,\widehat{pen}_m+78\,bias_m^2+42\max_{m\leqslant k\leqslant\widehat{M}}\left(\|\widehat{\beta}_k-\beta^k\|_\omega^2-\frac{1}{6}\,\widehat{pen}_k\right)_+.$$

On the event $\mathcal{E}_n$ defined in (2.12) we deduce from the last bound that for all $n\geqslant 1$

$$\|\widehat{\beta}_{\widehat{m}}-\beta\|_\omega^2\,\mathbb{1}_{\mathcal{E}_n}\leqslant 504\,pen_{m_n^\diamond}+78\,bias^2_{m_n^\diamond}+42\sup_{m_n^\diamond\leqslant k\leqslant M_n^+}\left(\|\widehat{\beta}_k-\beta^k\|_\omega^2-\frac{1}{6}\,pen_k\right)_+$$

which by taking into account Assumption 2.1 implies that

$$\sup_{\beta\in\mathcal{F}}\sup_{\Gamma\in\mathcal{G}}\mathbb{E}\big(\|\widehat{\beta}_{\widehat{m}}-\beta\|_\omega^2\,\mathbb{1}_{\mathcal{E}_n}\big)\leqslant 504\sup_{\beta\in\mathcal{F}}\sup_{\Gamma\in\mathcal{G}}\{pen_{m_n^\diamond}+bias^2_{m_n^\diamond}\}+504\,K_1\,n^{-1}\qquad\text{for all } n\geqslant 1.$$

We obtain the claim of the proposition by combination of the last bound and (2.14). □

Remark 2.1. The upper risk-bound given in the last assertion is strongly reminiscent of a variance/squared-biased decomposition of the $\mathcal{F}_\omega$-risk associated with the estimator $\widehat{\beta}_{m_n^\diamond}$ employing the dimension parameter $m_n^\diamond$. Indeed, in many cases the penalty term pen_m is in the same order as the variance of the estimator $\widehat{\beta}_m$ (c.f. Illustration 3.1 **[P-P]** and **[E-P]** below). In this situation we obviously wish that the parameter $m_n^\diamond$ just realize the balance between both the variance and the squared-biased term which in many cases can lead to an optimal estimation procedure. However, the construction of the penalty term is more involved to ensure that Assumption 2.1 and 2.2 can be satisfied (c.f. Illustration 3.1 **[P-E]**). □

3 Minimax-optimality

In this section we recall first a general framework proposed by Cardot and Johannes [2010] which allows to derive minimax-optimal rates for the maximal $\mathcal{F}_\omega$-risk, $\sup_{\beta\in\mathcal{F}}\sup_{\Gamma\in\mathcal{G}}\mathbb{E}\|\widehat{\beta}-\beta\|_\omega^2$, over the classes $\mathcal{F}$ and $\mathcal{G}$. Placing us into this framework, we can derive the main results of this paper which state that the proposed data-driven procedure indeed can attain these minimax-rates.

3.1 Notations and basic assumptions

The additional regularity conditions $\beta \in \mathcal{F}$ and $\Gamma \in \mathcal{G}$ imposed on the slope function and the covariance operator, respectively, are characterized by different weighted norms in $\mathbb{H}$ with respect to the pre-specified orthonormal basis $\{\psi_j, j\in\mathbb{N}\}$ in $\mathbb{H}$, which we formalize now. Given a strictly positive sequence of weights $b=(b_j)_{j\geqslant 1}$ and a radius $r>0$, let $\mathcal{F}_b$ be the completion of $\mathbb{H}$ with respect to the weighted norm $\|\cdot\|_b$, then we consider in the following the ellipsoid $\mathcal{F}_b^r := \big\{h\in\mathcal{F}_b : \|h\|_b^2 \leqslant r\big\}$ as class of possible slope functions. Furthermore, as usual in the context of ill-posed inverse problems, we link the mapping properties of the covariance operator Γ and the regularity condition $\beta\in\mathcal{F}_b^r$. Therefore, consider the sequence $(\langle\Gamma\psi_j,\psi_j\rangle)_{j\geqslant 1}$ which sums up to $\mathbb{E}\|X\|_{\mathbb{H}}^2$, i.e. Γ is nuclear, and hence converges to zero. In what follows we impose restrictions on the decay of this sequence. Denote by $\mathcal{N}$ the set of all strictly positive nuclear operators defined on $\mathbb{H}$. Given a strictly positive sequence of weights γ and a constant $d\geqslant 1$ define the class $\mathcal{G}_\gamma^d\subset\mathcal{N}$ of covariance operators by

$$\mathcal{G}_\gamma^d := \Big\{T\in\mathcal{N}:\quad \|f\|_{\gamma^2}^2/d^2 \leqslant \|Tf\|^2 \leqslant d^2\,\|f\|_{\gamma^2}^2, \quad \forall f\in\mathbb{H}\Big\}$$

where arithmetic operations on sequences are defined element-wise, e.g. $\gamma^2=(\gamma_j^2)_{j\geqslant 1}$. Let us briefly discuss the last definition. If $T\in\mathcal{G}_\gamma^d$, then we have $d^{-1}\leqslant\langle T\psi_j,\psi_j\rangle/\gamma_j\leqslant d$, for all $j\geqslant 1$. Consequently, the sequence γ is necessarily summable, because T is nuclear. Moreover, if λ denotes the sequence of eigenvalues of T then $d^{-1}\leqslant\lambda_j/\gamma_j\leqslant d$, for all $j\geqslant 1$. In other words the sequence γ characterizes the decay of the eigenvalues of $T\in\mathcal{G}_\gamma^d$. We do not specify the sequences of weights ω, b and γ, but impose from now on the following minimal regularity conditions.

Assumption 3.1. *Let $(\omega_j)_{j\geqslant 1}$, $(b_j)_{j\geqslant 1}$, and $(\gamma_j)_{j\geqslant 1}$ be strictly positive sequences of weights with $b_1=1$, $\omega_1=1$, $\gamma_1=1$, and $\sum_{j=1}^\infty\gamma_j<\infty$ such that the sequences b^{-1}, ωb^{-1}, γ, and $\gamma^2\omega^{-1}$ are monotonically non-increasing and converging to zero.*

The last assumption is fairly mild. For example, assuming that ωb^{-1} is non-increasing, ensures that $\mathcal{F}_b^r\subset\mathcal{F}_\omega$. Furthermore, it is shown in Cardot and Johannes [2010] that the minimax rate $R_\omega^*[n;\mathcal{F}_b^r,\mathcal{G}_\gamma^d]$ is of order n^{-1} for all sequences γ and ω such that $\gamma^2\omega^{-1}$ is non-decreasing. We will illustrate all our results considering the following three configurations for the sequences ω, b and γ.

Illustration 3.1. In all three cases, we take $\omega_j=j^{2s}$, $j\geqslant 1$. Moreover, let

[P-P] $b_j = j^{2p}$ and $\gamma_j = j^{-2a}$, $j \geqslant 1$, with $p > 0$, $a > 1/2$, and $p > s > -2a$;

[E-P] $b_j = \exp(j^{2p} - 1)$ and $\gamma_j = j^{-2a}$, $j \geqslant 1$, with $p > 0$, $a > 1/2$, and $s > -2a$;

[P-E] $b_j = j^{2p}$ and $\gamma_j = \exp(-j^{2a} + 1)$, $j \geqslant 1$, with $p > 0$, $a > 0$, and $p > s$;

then Assumption 3.1 is satisfied in all cases. □

Remark 3.1. In the configurations **[P-P]** and **[E-P]**, the case $s = -a$ can be interpreted as mean-prediction error (c.f. Cardot and Johannes [2010]). Moreover, if $\{\psi_j\}$ is the trigonometric basis and the value of s is an integer, then the weighted norm $\|h\|_\omega$ corresponds to the L^2-norm of the weak s-th derivative of h (c.f. Neubauer [1988]). In other words in this situation we consider as risk the mean integrated squared error when estimating the s-th derivative of β. Moreover, in the configurations **[P-P]** and **[P-E]**, the additional condition $p > s$ means that the slope function has at least $p \geqslant s+1$ weak derivatives, while for a value $p > 1$ in **[E-P]**, the slope function is assumed to be an analytic function (c.f. Kawata [1972]). □

3.2 Minimax optimal estimation reviewed

Let us first recall a lower bound of the maximal $\mathcal{F}_\omega$-risk over the classes $\mathcal{F}_b^r$ and $\mathcal{G}_\gamma^d$ due to Cardot and Johannes [2010]. Given an i.i.d. sample of (Y, X) of size n and sequences ω, b and γ satisfying Assumption 3.1 define

$$m_n^* := \arg\min_{m\geqslant 1}\left\{\max\left(\frac{\omega_m}{b_m}\ ,\ \sum_{j=1}^{m}\frac{\omega_j}{n\gamma_j}\right)\right\} \quad\text{and}\quad R_n^* := \max\left(\frac{\omega_{m_n^*}}{b_{m_n^*}}\ ,\ \sum_{j=1}^{m_n^*}\frac{\omega_j}{n\gamma_j}\right). \tag{3.1}$$

If in addition $\xi := \inf_{n\geqslant 1}\{(R_n^*)^{-1}\min(\omega_{m_n^*}b_{m_n^*}^{-1}\ ,\ \sum_{j=1}^{m_n^*}\omega_j(n\gamma_j)^{-1})\} > 0$, then there exists a constant $C := C(\sigma, r, d, \xi) > 0$ depending on σ, r, d and ξ only such that

$$\inf_{\widetilde{\beta}} \mathcal{R}_\omega^*[\widetilde{\beta}; \mathcal{F}_b^r, \mathcal{G}_\gamma^d] \geqslant C\, R_n^* \quad \text{for all } n \geqslant 1. \tag{3.2}$$

On the other hand considering the dimension parameter m_n^* given in (3.1) Cardot and Johannes [2010] have shown that the maximal risk $R_\omega^*[\widehat{\beta}_{m_n^*}; \mathcal{F}_b^r, \mathcal{G}_\gamma^d]$ of the estimator $\widehat{\beta}_{m_n^*}$ defined in (2.3) is bounded by R_n^* up to constant for a wide range of sequences ω, b, and γ, provided the random function X and the error ε satisfy certain additional moment conditions. In other words $R_n^* = R_\omega^*[n; \mathcal{F}_b^r, \mathcal{G}_\gamma^d]$ is the minimax-rate in this situation and the estimator $\widehat{\beta}_{m_n^*}$ is minimax optimal. Although, the definition of the dimension parameter m_n^* necessitates an a-priori knowledge of the sequences b and γ. In the remaining part of this paper we show that the data-driven choice of the dimension parameter constructed in Section 2 can automatically attain the minimax-rate R_n^* for a variety of sequences ω, b, and γ. Before, let us briefly illustrate the minimax result.

Illustration (continued) 3.2. Considering the three configurations (see Illustration 3.1), it has been shown in Cardot and Johannes [2010] that the estimator $\widehat{\beta}_{m_n^*}$ with m_n^* as given below attains the rate R_n^* up to a constant. We write for two strictly positive sequences $(a_n)_{n\geqslant 1}$ and $(b_n)_{n\geqslant 1}$ that $a_n \sim b_n$, if $(a_n/b_n)_{n\geqslant 1}$ is bounded away from 0 and infinity.

[P-P] It is easily seen that $m_n^* \sim n^{1/(2p+2a+1)}$ if $2s+2a+1>0$, $m_n^* \sim n^{1/[2(p-s)]}$ if $2s+2a+1<0$ and $m_n^* \sim (n/\log n)^{1/[2(p-s)]}$ if $2a+2s+1=0$, which in turn implies that $R_n^* \sim \max(n^{-(2p-2s)/(2a+2p+1)}, n^{-1})$, if $2s+2a+1\neq 0$ (and $R_n^*\sim \log(n)/n$ if $2s+2a+1=0$). Observe that an increasing value of a leads to a slower minimax-rate R_n^*. Therefore, the parameter a is called degree of ill-posedness (c.f. Natterer [1984]).

[E-P] If $2a+2s+1>0$, then $m_n^* \sim (\log n - \frac{2a+1}{2p}\log(\log n))^{1/(2p)}$ and $R_n^* \sim n^{-1}(\log n)^{(2a+1+2s)/(2p)}$. Furthermore, if $2a+2s+1<0$, then $m_n^*\sim(\log n + (s/p)\log(\log n))^{1/(2p)}$ and $R_n^*\sim n^{-1}$, while $R_n^* \sim \log(\log n)/n$ if $2a+2s+1=0$.

[P-E] We have $m_n^* \sim (\log n - \frac{2p+(2a-1)_+}{2a}\log(\log n))^{1/(2a)}$. Thereby, $R_n^* \sim (\log n)^{-(p-s)/a}$. The parameter a reflects again the degree of ill-posedness since an increasing value of a leads also here to a slower minimax-rate R_n^*. □

3.3 Minimax-optimality of the data-driven estimation procedure

Consider the thresholded projection estimator $\widehat{\beta}_{\widehat{m}}$ with data-driven choice $\widehat{m}$ of the dimension parameter. Supposing that the joint distribution of the random function X and the error term ε satisfies certain additional conditions, we will prove below that the Assumptions 2.1 and 2.2 formulated in Section 2 hold true. These assumptions rely on the existence of sequences $(m_n^\diamond)_{n\geqslant 1}$ and $(M_n^+)_{n\geqslant 1}$ which amongst others we define now referring only to the classes $\mathcal{F}_b^r$ and $\mathcal{G}_\gamma^d$. Keep in mind the notations given in (2.4) and (2.10). For $m\geqslant 1$ and $K=([\nabla_\gamma]_{\underline{m}})_{m\geqslant 1}$ define $\Delta_m^\gamma := \Delta_m(K)$ and $\delta_m^\gamma := \delta_m(K)$ where $\Delta_m^\gamma = \max_{1\leqslant k\leqslant m}\omega_j\gamma_j^{-1}$. Moreover, for $n\geqslant 1$ we set $M_n^- := M_n(a)$ with $a=(16d^3\gamma_m^{-1})_{m\geqslant 1}$ and $M_n^+ := M_n(a)$ with $a=([4d\gamma_m]^{-1})_{m\geqslant 1}$. Taking into account these notations we define for $n\geqslant 1$

$$m_n^\diamond := \mathop{\arg\min}_{1\leqslant m\leqslant M_n^-}\left\{\max\left(\frac{\omega_m}{b_m},\frac{\delta_m^\gamma}{n}\right)\right\} \quad\text{and}\quad R_n^\diamond := \max\left(\frac{\omega_{m_n^\diamond}}{b_{m_n^\diamond}},\frac{\delta_{m_n^\diamond}^\gamma}{n}\right)$$

satisfying $m_n^\diamond \leqslant M_n^- \leqslant M_n^+$. Furthermore, let $\Sigma := \Sigma(\mathcal{G}_\gamma^d)$ denote a finite constant such that

$$\Sigma \geqslant \sum_{j\geqslant 1}\gamma_j \quad\text{and}\quad \Sigma \geqslant \sum_{m\geqslant 1}\Delta_m^\gamma \exp\Big(-\frac{1}{16(1+\log(d))}\frac{m\log(\Delta_m^\gamma\vee(m+2))}{\log(m+2)}\Big) \tag{3.3}$$

which by construction always exists and depends on the class $\mathcal{G}_\gamma^d$ only. Let us illustrate the last definition by revisiting the three configurations for the sequences ω, b, and γ (Illustration 3.1).

Illustration (continued) 3.3. In the following we state the order of M_n^- and δ_m^γ which in turn are used to derive the order of $m_n^\diamond$ and $R_n^\diamond$.

[P-P] $M_n^- \sim \big(\frac{n}{1+\log n}\big)^{1+2a+(2s)_+}$, $\delta_m^\gamma \sim m^{1+(2a+2s)_+}$, and for all $p>(s)_+$ it follows $m_n^\diamond \sim m^{1/[1+2p-2s+(2a+2s)_+}$ and $R_n^\diamond \sim n^{-2(p-s)/[1+2p-2s+(2a+2s)_+]}$;

[E-P] $M_n^- \sim \big(\frac{n}{1+\log n}\big)^{1+2a+(2s)_+}$, $\delta_m^\gamma \sim m^{1+(2a+2s)_+}$, and for all $p>0$ it follows $m_n^\diamond \sim (\log n - \frac{1+2(a+s)_+-2s}{2p}\log(\log n))^{1/(2p)}$ and $R_n^\diamond \sim n^{-1}(\log n)^{[1+2(a+s)_+]/(2p)}$;

[P-E] $M_n^- \sim (\log n - \frac{1+2a+2(s)_+}{2a}\log(\log n))^{1/(2a)}$, $\delta_m^\gamma \sim m^{1+2s+2a}\exp(m^{2a})$, and for all $p>(s)_+$ it follows $m_n^\diamond \sim (\log n - \frac{1+2a+2p}{2a}\log(\log n))^{1/(2a)}$ and $R_n^\diamond \sim (\log n)^{-(p-s)/a}$. $\square$

We proceed by formalizing the additional conditions on the joint distribution of ε and X which in turn are used to prove that the Assumptions 2.1 and 2.2 hold true.

Imposing a joint normal distribution. Let us first assume that X is a centered Gaussian $\mathbb{H}$-valued random variable, that is, for all $k \geqslant 1$ and for all finite collections $\{h_1,\dots,h_k\} \subset \mathbb{H}$ the joint distribution of the real valued random variables $\langle X,h_1\rangle_{\mathbb{H}},\dots,\langle X,h_k\rangle_{\mathbb{H}}$ is Gaussian with zero mean vector and covariance matrix with generic elements $\mathbb{E}\langle h_j,X\rangle_{\mathbb{H}}\langle X,h_l\rangle_{\mathbb{H}}$, $1\leqslant j,l\leqslant k$. Moreover, suppose that the error term is standard normally distributed. The next assumption summarizes this situation.

ASSUMPTION 3.2. *The joint distribution of the random function X and the error ε is normal.*

The proof of the next assertion is more involved and hence deferred to Appendix C.

PROPOSITION 3.1. *Assume an iid. n-sample of (Y,X) obeying* (1.1) *and Assumption 3.2. Consider sequences ω, b and γ satisfying Assumption 3.1 and in the definition* (2.6) *and* (2.11) *of the penalty $\widehat{pen}$ and pen respectively set $\kappa = 96$. For the classes $\mathcal{F}_b^r$ and $\mathcal{G}_\gamma^d$, there exist finite constants $C_1 := C_1(d)$ and $C_2 := C_2(d)$ depending on d only such that the Assumptions 2.1 and 2.2 hold true, with $K_1 := C_1\,(\sigma^2+r)\,\Sigma$ and $K_2 := C_2\,(\sigma^2+r)\,\Sigma$ respectively.*

By taking the value $\kappa = 96$ the random penalty $\widehat{pen}$ and the random upper bound $\widehat{M}$ given in (2.6) and (2.5) respectively depend indeed only on the data, and hence the choice $\widehat{m}$ of the dimension parameter in (2.7) is fully data-driven. Moreover due to the last assertion we can apply Proposition 2.2 which in turn provides the key argument to prove the following upper risk-bound for the data-driven thresholded projection estimator $\widehat{\beta}_{\widehat{m}}$ with $\widehat{m}$ given by (2.7).

THEOREM 3.2. *Let the assumptions of Proposition 3.1 be satisfied. There exists a finite constant $K := K(d)$ depending on d only such that*

$$R_\omega[\widehat{\beta}_{\widehat{m}},\mathcal{F}_b^r,\mathcal{G}_\gamma^d] \leqslant K\,(\sigma^2+r)\big\{R_n^\diamond + \Sigma\, n^{-1}\big\} \quad \textit{for all } n \geqslant 1.$$

PROOF OF THEOREM 3.2. We shall provide in the appendix among others, the two technical Lemmas B.1 and B.2 which are used in the following. Moreover, denote by $K := K(d)$ a constant depending on d only which changes from line to line. Making use of Proposition 3.1, i.e., Assumptions 2.1 and 2.2 are satisfied, we can apply Proposition 2.2, and hence for all $n \geqslant 1$

$$R_\omega[\widehat{\beta}_{\widehat{m}},\mathcal{F}_b^r,\mathcal{G}_\gamma^d] \leqslant 504 \sup_{\beta\in\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\{pen_{m_n^\diamond} + bias^2_{m_n^\diamond}\} + K\,(\sigma^2+r)\,\Sigma\, n^{-1}. \tag{3.4}$$

Furthermore, if $\beta\in\mathcal{F}_b^r$ and $\Gamma\in\mathcal{G}_\gamma^d$ then firstly from (B.4) in Lemma B.1 follows that $bias^2_{m_n^\diamond} \leqslant 34\,d^8\,r\,\omega_{m_n^\diamond}\,b^{-1}_{m_n^\diamond}$ because $\gamma^2\omega^{-1}$ and ωb^{-1} are non increasing due to Assumption 3.1. Secondly, by combination of (i) and (iv) in Lemma B.2, it is easily verified that $pen_{m^\diamond} \leqslant K\,(\sigma^2{+}r)\delta^\gamma_{m^\diamond} n^{-1}$. Consequently, $\sup_{\beta\in\mathcal{F}^r_b}\sup_{\Gamma\in\mathcal{G}^d_\gamma}\{pen_{m^\diamond_n} + bias^2_{m^\diamond_n}\} \leqslant K\,(\sigma^2{+}r)\,R^\diamond_n$ for all $n \geqslant 1$ by combination of the last two estimates and the definition of $R^\diamond_n$ which in turn together with the upper bound (3.4) implies the assertion of the theorem. □

Imposing moment conditions. We dismiss now the Assumption 3.2 and formalize in place conditions on the moments of the random function X and the error term ε. In particular we use that for all $h \in \mathbb{H}$ with $\langle \Gamma h, h\rangle = 1$, the random variable $\langle h, X\rangle$ is standardized, i.e. has mean zero and variance one.

ASSUMPTION 3.3. *There exist a finite integer $k \geqslant 16$ and a finite constant $\eta \geqslant 1$ such that $\mathbb{E}|\varepsilon|^{4k} \leqslant \eta^{4k}$ and that for all $h \in \mathbb{H}$ with $\langle \Gamma h, h\rangle = 1$ the standardized random variable $\langle h, X\rangle$ satisfies $\mathbb{E}|\langle h, X\rangle|^{4k} \leqslant \eta^{4k}$.*

It is worth noting that for any Gaussian random function X with finite second moment Assumption 3.3 holds true, since for all $h \in \mathbb{H}$ with $\langle \Gamma h, h\rangle = 1$ the random variable $\langle h, X\rangle$ is standard normally distributed and hence $\mathbb{E}|\langle h, X\rangle|^{2k} = (2k{-}1)\cdot\ldots\cdot 5\cdot 3\cdot 1$. The proof of the next assertion is again rather involved and deferred to Appendix D. It follows, however, along the general lines of the proof of Proposition 2.2 though it is not a straightforward extension. Take as an example the concentration inequality for the random variable $\|[\Gamma]^{1/2}_{\underline{m}}([\widehat{g}]_{\underline{m}} - [\widehat{\Gamma}]_{\underline{m}}[\beta^m]_{\underline{m}})\|$ in Lemma C.3 in Appendix C which due to Assumption 3.2 is shown by employing elementary inequalities for Gaussian random variables. In contrast, the proof of an analogous result under Assumption 3.3 given in Lemma D.3 in Appendix D is based on an inequality due to Talagrand [1996] (Proposition D.1 in the appendix states a version as presented in Klein and Rio [2005]).

PROPOSITION 3.3. *Assume an iid. n-sample of (Y, X) obeying* (1.1) *and Assumption 3.3. Consider sequences ω, b and γ satisfying Assumption 3.1 and in the definition* (2.6) *and* (2.11) *of the penalty $\widehat{pen}$ and pen respectively, set $\kappa = 288$. For the classes $\mathcal{F}^r_b$ and $\mathcal{G}^d_\gamma$, there exist finite constants $C_1 := C_1(\sigma, \eta, \mathcal{F}^r_b, \mathcal{G}^d_\gamma)$ depending on σ, η and the classes $\mathcal{F}^r_b$ and $\mathcal{G}^d_\gamma$ only, and $C_2 := C_2(d)$ depending on d only, such that Assumptions 2.1 and 2.2 hold true with $K_1 := C_1\,\eta^{64}\,(\sigma^2 + r)\,\Sigma$ and $K_2 := C_2\,\eta^{64}\,(\sigma^2 + r)\,\Sigma$ respectively.*

We remark a change only in the constants when comparing the last proposition with Proposition 3.1. Note further that we need a larger value for the constant κ than in Proposition 3.1 although it is still a numerical constant and hence the choice $\widehat{m}$ given by (2.7) is again fully data-driven. Moreover, both values for the constant κ, though convenient for deriving the theory, may be far too large in practice and instead be determined by means of a simulation study as in Comte et al. [2006], for example. The next assertion provides an upper risk-bound for the data-driven thresholded projection estimator $\widehat{\beta}_{\widehat{m}}$ when imposing moment conditions.

THEOREM 3.4. *Let the assumptions of Proposition 3.3 be satisfied. There exist finite constants $K := K(d)$ depending on d only and $K' := K'(\sigma, \eta, \mathcal{F}^r_b, \mathcal{G}^d_\gamma)$ depending on σ, η and the classes $\mathcal{F}^r_b$ and $\mathcal{G}^d_\gamma$ only such that*

$$R_\omega[\widehat{\beta}_{\widehat{m}}, \mathcal{F}^r_b, \mathcal{G}^d_\gamma] \leqslant K\,(\sigma^2 + r)\left\{R^\diamond_n + K'\,\eta^{64}\,\Sigma\, n^{-1}\right\} \quad \text{for all } n \geqslant 1.$$

Proof of Theorem 3.4. Taking into account Proposition 3.3 rather than Proposition 3.1 we follow line by line the proof of Theorem 3.2 and hence we omit the details. □

Minimax-optimality. A comparison of the upper bounds in both Theorem 3.2 and Theorem 3.4 with the lower bound displayed in (3.2) shows that the data-driven estimator $\widehat{\beta}_{\widehat{m}}$ attains up to a constant the minimax-rate $R_n^* = \min_{1\leqslant m<\infty}\left\{\max\left(\frac{\omega_m}{b_m}, \sum_{j=1}^m \frac{\omega_j}{n\gamma_j}\right)\right\}$ only if $R_n^\diamond = \min_{1\leqslant m\leqslant M_n^-}\left\{\max\left(\frac{\omega_m}{b_m}, \frac{\delta_m^\gamma}{n}\right)\right\}$ has the same order as R_n^*. Note that, by construction, $\delta_m^\gamma \geqslant \sum_{j=1}^m \frac{\omega_j}{\gamma_j}$ for all $m\geqslant 1$. The next assertion is an immediate consequence of Theorem 3.2 and Theorem 3.4 and we omit its proof.

Corollary 3.5. *Let the assumptions of either Theorem 3.2 or Theorem 3.4 be satisfied. If in addition* $\xi^\diamond := \sup_{n\geqslant 1}\{R_n^\diamond/R_n^*\}<\infty$ *holds true, then we have for all* $n\geqslant 1$

$$R_\omega[\widehat{\beta}_{\widehat{m}};\mathcal{F}_b^r,\mathcal{G}_\gamma^d] \leqslant C\cdot \inf_{\widetilde{\beta}} R_\omega[\widetilde{\beta};\mathcal{F}_b^r,\mathcal{G}_\gamma^d]$$

where the infimum is taken over all possible estimators $\widetilde{\beta}$ *and* C *is a finite positive constant.*

Remark 3.2. In the last assertion $\xi^\diamond = \sup_{n\geqslant 1}\{R_n^\diamond/R_n^*\}<\infty$ is for example satisfied if the following two conditions hold simultaneously true: (i) $m_n^*\leqslant M_n^-$ for all $n\geqslant 1$ and (ii) $\Delta_m^\gamma = \max_{1\leqslant j\leqslant m}\omega_j\gamma_j^{-1}\leqslant Cm^{-1}\sum_{j=1}^m\omega_j\gamma_j^{-1}$ and $\log(\Delta_m^\gamma\vee(m+2))\leqslant C\log(m+2)$ for all $m\geqslant 1$. Observe that (ii) which implies $\delta_m^\gamma\leqslant C\sum_{j=1}^m\frac{\omega_j}{\gamma_j}$ is satisfied in case Δ_m^γ is in the order of a power of m (e.g. Illustration 3.2 **[P-P]** and **[E-P]**). If this term has an exponential order with respect to m (e.g. Illustration 3.2 **[P-E]**), then a deterioration of the term δ_m^γ compared to the variance term $\sum_{j=1}^m\frac{\omega_j}{\gamma_j}$ is possible. However, no loss in terms of the rate may occur, i.e., $\xi^\diamond<\infty$, when the squared-bias term $\omega_{m_n^\diamond}b_{m_n^\diamond}^{-1}$ dominates the variance term $n^{-1}\delta_{m_n^\diamond}^\gamma$ (for a detailed discussion in a deconvolution context we refer to Butucea and Tsybakov [2007a,b]).□

Let us illustrate the performance of the data-driven thresholded projection estimator $\widehat{\beta}_{\widehat{m}}$ considering the three configurations for the sequences ω, b, and γ (see Illustration 3.1 above).

Proposition 3.6. *Assume an iid. n-sample of* (Y,X) *satisfying* (1.1) *and let either Assumption 3.2 or Assumption 3.3 hold true where we set respectively* $\kappa=96$ *or* $\kappa=288$ *in* (2.6). *The fully data-driven estimator* $\widehat{\beta}_{\widehat{m}}$ *attains the minimax-rates* R_n^* *(see Illustration 3.2), up to a constant, in the three cases introduced in the Illustration 3.1, if we additionally assume* $a+s\geqslant 0$ *in the cases* **[P-P]** *and* **[E-P]**.

Proof of Proposition 3.6. Under the stated conditions it is easily verified that the assumptions of either Theorem 3.2 or Theorem 3.4 are satisfied. Moreover, the rates R_n^* (Illustration 3.2) and $R_n^\diamond$ (Illustration 3.3) are of the same order if we additionally assume $a+s\geqslant 0$ in the cases **[P-P]** and **[E-P]**. Therefore we can apply Corollary 3.5 which implies the assertion. □

Appendix

This section gathers preliminary technical results and the proofs of Proposition 3.1 and 3.3.

A Notations

We begin by defining and recalling notations to be used in all proofs. Given $m \geqslant 1$, $\mathbb{H}_m$ denotes the subspace of $\mathbb{H}$ spanned by the functions $\{\psi_1, \dots, \psi_m\}$. Π_m and $\Pi_m^\perp$ denote the orthogonal projections on $\mathbb{H}_m$ and its orthogonal complement $\mathbb{H}_m^\perp$, respectively. If K is an operator mapping $\mathbb{H}$ to itself and if we restrict $\Pi_m K \Pi_m$ to an operator from $\mathbb{H}_m$ to itself, then it can be represented by a matrix $[K]_{\underline{m}}$ with generic entries $\langle \psi_j, K\psi_l \rangle_{\mathbb{H}} =: [K]_{j,l}$ for $1 \leqslant j,l \leqslant m$. The spectral norm of $[K]_{\underline{m}}$ is denoted by $\|[K]_{\underline{m}}\|_s$ and the inverse matrix of $[K]_{\underline{m}}$ by $[K]_{\underline{m}}^{-1}$. Furthermore, keeping in mind the notations given in (2.4) and (2.10) we use for all $m \geqslant 1$ and $n \geqslant 1$

$$
\begin{aligned}
&\Delta_m^\omega = \Delta_m([\nabla_\omega]),\ \Delta_m^\Gamma = \Delta_m([\Gamma]),\ \Lambda_m^\Gamma := \frac{\log(\Delta_m^\Gamma \vee (m+2))}{\log(m+2)},\ \delta_m^\Gamma = m\Delta_m^\Gamma\Lambda_m^\Gamma = \delta_m([\Gamma]),\\
&\Delta_m^\gamma = \max_{1\leqslant k\leqslant m} \omega_j\gamma_j^{-1} = \Delta_m([\nabla_\gamma]),\ \Lambda_m^\gamma := \frac{\log(\Delta_m^\gamma \vee (m+2))}{\log(m+2)},\ \delta_m^\gamma = m\Delta_m^\gamma\Lambda_m^\gamma = \delta_m([\nabla_\gamma]),\\
&\widehat{\Delta}_m = \Delta_m([\widehat{\Gamma}]),\quad \widehat{\Lambda}_m := \frac{\log(\widehat{\Delta}_m \vee (m+2))}{\log(m+2)},\quad \widehat{\delta}_m := m\widehat{\Delta}_m\widehat{\Lambda}_m = \delta_m([\widehat{\Gamma}]),\\
&\widehat{M} = M_n\big((\|[\widehat{\Gamma}]_{\underline{m}}^{-1}\|_s)_{m\geqslant 1}\big),\quad M_n^- = M_n(16d^3\gamma^{-1}),\quad M_n^+ = M_n([4d\gamma]^{-1}),\\
&pen_m = \kappa\sigma_m^2 m\Delta_m^\Gamma\Lambda_m^\Gamma n^{-1} \quad \text{and} \quad \widehat{pen}_m = 14\kappa\widehat{\sigma}_m^2 m\widehat{\Delta}_m\widehat{\Lambda}_m n^{-1}.
\end{aligned}
\tag{A.1}
$$

Recall that $[\widehat{\Gamma}]_{\underline{m}} = \frac{1}{n}\sum_{i=1}^n [X_i]_{\underline{m}}[X_i]_{\underline{m}}^t$ and $[\widehat{g}]_{\underline{m}} = \frac{1}{n}\sum_{i=1}^n Y_i[X_i]_{\underline{m}}$ where $[\Gamma]_{\underline{m}} = \mathbb{E}[X]_{\underline{m}}[X]_{\underline{m}}^t$ and $[g]_{\underline{m}} = \mathbb{E}Y[X]_{\underline{m}}$. Given a Galerkin solution $\beta^m \in \mathbb{H}_m$, $m \geqslant 1$, of equation (1.2), let $Z_m := Y - \langle \beta^m, X\rangle_{\mathbb{H}} = \sigma\varepsilon + \langle \beta - \beta^m, X\rangle_{\mathbb{H}}$, and denote $\rho_m^2 := \mathbb{E}Z_m^2 = \sigma^2 + \langle \Gamma(\beta-\beta^m), (\beta-\beta^m)\rangle_{\mathbb{H}}$, $\sigma_Y^2 := \mathbb{E}Y^2 = \sigma^2 + \langle \Gamma\beta, \beta\rangle_{\mathbb{H}}$ and $\sigma_m^2 = 2\big(\sigma_Y^2 + [g]_{\underline{m}}^t[\Gamma]_{\underline{m}}^{-1}[g]_{\underline{m}}\big)$ where we used that ε and X are uncorrelated. Define the random matrix $[\Xi]_{\underline{m}}$ and random vector $[W]_{\underline{m}}$ respectively by

$$
[\Xi]_{\underline{m}} := [\Gamma]_{\underline{m}}^{-1/2}[\widehat{\Gamma}]_{\underline{m}}[\Gamma]_{\underline{m}}^{-1/2} - [\mathrm{Id}]_{\underline{m}}, \quad \text{and} \quad [W]_{\underline{m}} := [\widehat{g}]_{\underline{m}} - [\widehat{\Gamma}]_{\underline{m}}[\beta^m]_{\underline{m}},
$$

where $\mathbb{E}[\Xi]_{\underline{m}} = 0$, because $\mathbb{E}[\widehat{\Gamma}]_{\underline{m}} = [\Gamma]_{\underline{m}}$, and $\mathbb{E}[W]_{\underline{m}} = [\Gamma(\beta - \beta^m)]_{\underline{m}} = 0$. Define further $\widehat{\sigma}_Y^2 := n^{-1}\sum_{i=1}^n Y_i^2$, the events

$$
\begin{aligned}
&\Omega_{m,n} := \{\|[\widehat{\Gamma}]_{\underline{m}}^{-1}\|_s \leqslant n\}, \quad \mho_{m,n} := \{8\|[\Xi]_{\underline{m}}\|_s \leqslant 1\},\\
&\mathcal{A}_n := \{1/2 \leqslant \widehat{\sigma}_Y^2/\sigma_Y^2 \leqslant 3/2\}, \quad \mathcal{B}_n := \{\|[\Xi]_{\underline{k}}\|_s \leqslant 1/8, \forall 1 \leqslant k \leqslant M_n^\omega\},\\
&\mathcal{C}_n := \{[W]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[W]_{\underline{k}} \leqslant \frac{1}{8}([g]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[g]_{\underline{k}} + \sigma_Y^2), \forall 1 \leqslant k \leqslant M_n^\omega\},
\end{aligned}
\tag{A.2}
$$

and their complements $\Omega_{m,n}^c$, $\mho_{m,n}^c$, $\mathcal{A}_n^c$, $\mathcal{B}_n^c$, and $\mathcal{C}_n^c$, respectively. Furthermore, we will denote by C universal numerical constants and by $C(\cdot)$ constants depending only on the arguments. In both cases, the values of the constants may change from line to line.

B Preliminary results

This section gathers preliminary results where we only exploit that the sequences ω, b and γ satisfy Assumption 3.1. The proof of the next lemma can be found in Johannes and Schenk [2010].

LEMMA B.1. *Let $\Gamma \in \mathcal{G}_\gamma^d$ where the sequence γ satisfies Assumption 3.1, then we have*

$$\sup_{m\in\mathbb{N}}\Big\{\gamma_m \|[\Gamma]_{\underline{m}}^{-1}\|_s\Big\} \leqslant 4d^3, \tag{B.1}$$

$$\sup_{m\in\mathbb{N}}\|[\nabla_\gamma]_{\underline{m}}^{1/2}[\Gamma]_{\underline{m}}^{-1}[\nabla_\gamma]_{\underline{m}}^{1/2}\|_s \leqslant 4d^3, \tag{B.2}$$

$$\sup_{m\in\mathbb{N}}\|[\nabla_\gamma]_{\underline{m}}^{-1/2}[\Gamma]_{\underline{m}}[\nabla_\gamma]_{\underline{m}}^{-1/2}\|_s \leqslant d. \tag{B.3}$$

Let in addition $\beta \in \mathcal{F}_b^r$ with sequence b satisfying Assumption 3.1. If β^m denotes a Galerkin solution of $g = \Gamma\beta$ then for each strictly positive sequence $w := (w_j)_{j\geqslant 1}$ such that w/b is non increasing we obtain for all $m \in \mathbb{N}$

$$\|\beta - \beta^m\|_w^2 \leqslant 34\, d^8\, r\, \frac{w_m}{b_m} \max\left(1, \frac{\gamma_m^2}{w_m}\max_{1\leqslant j\leqslant m}\frac{w_j}{\gamma_j^2}\right), \tag{B.4}$$

$$\|\beta^m\|_b^2 \leqslant 34\, d^8\, r \quad \textit{and} \quad \|\Gamma^{1/2}(\beta-\beta^m)\|_{\mathbb{H}}^2 \leqslant 34\, d^9\, r\, \gamma_m b_m^{-1}. \tag{B.5}$$

LEMMA B.2. *Let $D := (4d^3)$ and let Assumption 3.1 be satisfied. If $\Gamma \in \mathcal{G}_\gamma^d$ then it holds*

(i) $d^{-1} \leqslant \gamma_m\|[\Gamma]_{\underline{m}}^{-1}\|_s \leqslant D$, $d^{-1} \leqslant \Delta_m^\Gamma/\Delta_m^\gamma \leqslant D$, $(1+\log d)^{-1} \leqslant \Lambda_m^\Gamma/\Lambda_m^\gamma \leqslant (1+\log D)$, *and* $d^{-1}(1+\log d)^{-1} \leqslant \delta_m^\Gamma/\delta_m^\gamma \leqslant D(1+\log D)$ *for all* $m \geqslant 1$,

(ii) $\delta_{M_n^+}^\gamma \leqslant n4D(1+\log D)$ *and* $\delta_{M_n^+}^\Gamma \leqslant n4D^2(1+2\log D)$ *for all* $n \geqslant 1$,

(iii) $n \geqslant 2\max_{1\leqslant m\leqslant M_n^+}\|[\Gamma]_{\underline{m}}^{-1}\|$ *if* $n \geqslant 2D$ *and* $\Delta_{M_n^+}^\omega M_n^+(1+\log n) \geqslant 8D^2$.

If in addition $\beta \in \mathcal{F}_b^r$ then we have for all $m \geqslant 1$

(iv) $\rho_m^2 \leqslant \sigma_m^2 \leqslant 2(\sigma^2 + 35d^9 r)$

PROOF OF LEMMA B.2. Proof of (i). Due to (B.1) and (B.3) in Lemma B.1, we have for all $\Gamma \in \mathcal{G}_\gamma^d$ and for all $m \geqslant 1$ that $\|[\Gamma]_{\underline{m}}^{-1}\|_s \leqslant 4d^3\gamma_m^{-1}$ and $\gamma_m^{-1} \leqslant d\|[\Gamma]_{\underline{m}}^{-1}\|_s$. Thus, given $D = (4d^3)$ for all $m \geqslant 1$ we have $d^{-1} \leqslant \|[\Gamma]_{\underline{m}}^{-1}\|_s\gamma_m \leqslant D$. Moreover, the monotonicity of γ implies $d^{-1} \leqslant \gamma_M \max_{1\leqslant m\leqslant M}\|[\Gamma]_{\underline{m}}^{-1}\|_s \leqslant D$. From these estimates we obtain (i).
Proof of (ii). Observe that $\Delta_{M_n^+}^\gamma \leqslant \Delta_{M_n^+}^\omega\gamma_{M_n^+}^{-1}$. In case $M_n^+ = 1$ the assertion is trivial, since $\Delta_1^\omega\gamma_1^{-1} = 1$ due to Assumption 3.1. Thus, consider $M_n^\omega \geqslant M_n^+ > 1$, which implies $\min_{1\leqslant j\leqslant M_n^+}\{\gamma_j(j\Delta_j^\omega)^{-1}\} \geqslant (1+\log n)(4Dn)^{-1}$, and hence $M_n^+\Delta_{M_n^+}^\gamma \leqslant 4Dn(1+\log n)^{-1}$, $\Lambda_{M_n^+}^\gamma \leqslant (1+\log D)(1+\log n)$, $M_n^+\Delta_{M_n^+}^\Gamma \leqslant 4D^2n(1+\log n)^{-1}$ and $\Lambda_{M_n^+}^\Gamma \leqslant (1+2\log D)(1+\log n)$. The assertion (ii) follows now by combination of these estimates.
Proof of (iii). By employing that $D\gamma_{M_n^+}^{-1} \geqslant \max_{1\leqslant m\leqslant M_n^+}\|[\Gamma]_{\underline{m}}^{-1}\|$, the assertion (iii) follows in case $M_n^+ = 1$ from $\gamma_1 = 1$, while in case $M_n^+ > 1$, we use $M_n^+\Delta_{M_n^+}^\omega\gamma_{M_n^+}^{-1} \leqslant 4Dn(1+\log n)^{-1}$.
Proof of (iv). Since ε and X are centered it follows from $[\beta^m]_{\underline{m}} = [\Gamma]_{\underline{m}}^{-1}[g]_{\underline{m}}$ that $\rho_m^2 \leqslant 2\big(\mathbb{E}Y^2 + \mathbb{E}|\langle\beta^m, X\rangle_{\mathbb{H}}|^2\big) = 2\big(\sigma_Y^2 + [g]_{\underline{m}}^t[\Gamma]_{\underline{m}}^{-1}[g]_{\underline{m}}\big) = \sigma_m^2$. Moreover, by employing successively the inequality of Heinz [1951], i.e. $\|\Gamma^{1/2}\beta\|^2 \leqslant d\|\beta\|_\gamma^2$, and Assumption 3.1, i.e., γ and b^{-1} are non-increasing, the identity $\sigma_Y^2 = \sigma^2 + \langle\Gamma\beta,\beta\rangle_{\mathbb{H}}$ implies

$$\sigma_Y^2 \leqslant \sigma^2 + d\|\beta\|_\gamma^2 \leqslant \sigma^2 + dr. \tag{B.6}$$

Furthermore, from (B.3) and (B.5) in Lemma B.1 we obtain

$$[g]_{\underline{m}}^t[\Gamma]_{\underline{m}}^{-1}[g]_{\underline{m}} \leqslant d\|\beta^m\|_\gamma^2 \leqslant 34d^9 r. \tag{B.7}$$

The assertion (iv) follows now from (B.6) and (B.7), which completes the proof. □

LEMMA B.3. *For all $n, m \geqslant 1$ we have*

$$\left\{\frac{1}{4} < \frac{\|[\widehat{\Gamma}]_{\underline{m}}^{-1}\|_s}{\|[\Gamma]_{\underline{m}}^{-1}\|_s} \leqslant 4, \forall\, 1 \leqslant m \leqslant M_n^\omega\right\} \subset \left\{M_n^- \leqslant \widehat{M} \leqslant M_n^+\right\}.$$

PROOF OF LEMMA B.3. Let $\widehat{\tau}_m = \|[\widehat{\Gamma}]_{\underline{m}}^{-1}\|_s^{-1}$ and recall that $1 \leqslant \widehat{M} \leqslant M_n^\omega$ with

$$\left\{\widehat{M} = M\right\} = \begin{cases} & \left\{\frac{\widehat{\tau}_{M+1}}{(M+1)\Delta_{M+1}^\omega} < \frac{1+\log n}{n}\right\}, \quad M = 1, \\ \left\{\min\limits_{2\leqslant m\leqslant M} \frac{\widehat{\tau}_m}{m\Delta_m^\omega} \geqslant \frac{1+\log n}{n}\right\} \bigcap & \left\{\frac{\widehat{\tau}_{M+1}}{(M+1)\Delta_{M+1}^\omega} < \frac{1+\log n}{n}\right\}, \quad 1 < M < M_n^\omega, \\ \left\{\min\limits_{2\leqslant m\leqslant M} \frac{\widehat{\tau}_m}{m\Delta_m^\omega} \geqslant \frac{1+\log n}{n}\right\}, & \qquad M = M_n^\omega. \end{cases}$$

Given $\tau_m := \|[\Gamma]_{\underline{m}}^{-1}\|_s^{-1}$ we have $D^{-1} \leqslant \tau_m/\gamma_m \leqslant d$, $m \geqslant 1$ due to (i) in Lemma B.2 which we use to prove the following two assertions

$$\left\{\widehat{M} < M_n^-\right\} \subset \left\{\min_{1\leqslant m\leqslant M_n^\omega} : \frac{\widehat{\tau}_m}{\tau_m} < \frac{1}{4}\right\}, \tag{B.8}$$

$$\left\{\widehat{M} > M_n^+\right\} \subset \left\{\max_{1\leqslant m\leqslant M_n^\omega} \frac{\widehat{\tau}_m}{\tau_m} \geqslant 4\right\}. \tag{B.9}$$

Obviously, the assertion of Lemma B.3 follows now by combination of (B.8) and (B.9). Consider (B.8) which is trivial in case $M_n^- = 1$. If $M_n^- > 1$ we have $\min\limits_{1\leqslant m\leqslant M_n^-} \frac{\gamma_m}{m\omega_m^+} \geqslant \frac{4D(1+\log n)}{n}$ and, hence $\min\limits_{1\leqslant m\leqslant M_n^-} \frac{\tau_m}{m\Delta_m^\omega} \geqslant \frac{4(1+\log n)}{n}$. By exploiting the last estimate we obtain

$$\begin{aligned}\left\{\widehat{M} < M_n^\omega\right\} \cap \left\{\widehat{M} < M_n^-\right\} &= \bigcup_{M=1}^{M_n^- -1} \left\{\widehat{M} = M\right\} \\ &\subset \bigcup_{M=1}^{M_n^- -1} \left\{\frac{\widehat{\tau}_{M+1}}{(M+1)\Delta_{M+1}^\omega} < \frac{1+\log n}{n}\right\} = \left\{\min_{2\leqslant m\leqslant M_n^-} \frac{\widehat{\tau}_m}{m\Delta_m^\omega} < \frac{1+\log n}{n}\right\} \\ &\subset \left\{\min_{1\leqslant m\leqslant M_n^-} \frac{\widehat{\tau}_m}{\tau_m} < 1/4\right\}\end{aligned}$$

while trivially $\left\{\widehat{M} = M_n^\omega\right\} \cap \left\{\widehat{M} < M_n^-\right\} = \emptyset$, which proves (B.8) because $M_n^- \leqslant M_n^\omega$. Consider (B.9) which is trivial in case $M_n^+ = M_n^\omega$. If $M_n^+ < M_n^\omega$, then $\frac{\tau_{M_n^+ +1}}{(M_n^+ +1)\Delta_{M_n^+ +1}^\omega} < \frac{(1+\log n)}{4n}$, and hence

$$\left\{\widehat{M}>1\right\}\cap\left\{\widehat{M}>M_n^+\right\}=\bigcup_{M=M_n^++1}^{M_n^\omega}\left\{\widehat{M}=M\right\}$$
$$\subset\bigcup_{M=M_n^++1}^{M_n^\omega}\left\{\min_{2\leqslant m\leqslant M}\frac{\widehat{\tau}_m}{m\Delta_m^\omega}\geqslant\frac{1+\log n}{n}\right\}=\left\{\min_{2\leqslant m\leqslant (M_n^++1)}\frac{\widehat{\tau}_m}{m\Delta_m^\omega}\geqslant\frac{1+\log n}{n}\right\}$$
$$\subset\left\{\frac{\widehat{\tau}_{M_n^++1}}{\tau_{M_n^++1}}\geqslant 4\right\}$$

while trivially $\{\widehat{M}=1\}\cap\{\widehat{M}>M_n^+\}=\emptyset$ which shows (B.9) and completes the proof. □

Lemma B.4. *Let $\mathcal{A}_n$, $\mathcal{B}_n$ and $\mathcal{C}_n$ as in* (A.2). *For all $n\geqslant 1$ it holds true that*
$$\mathcal{A}_n\cap\mathcal{B}_n\cap\mathcal{C}_n\subset\{pen_k\leqslant\widehat{pen}_k\leqslant 72\,pen_k, 1\leqslant k\leqslant M_n^\omega\}\cap\{M_n^-\leqslant\widehat{M}\leqslant M_n^+\}.$$

Proof of Lemma B.4. Let $M_n^\omega\geqslant k\geqslant 1$. If $\|[\Xi]_{\underline{k}}\|_s\leqslant 1/8$, i.e. on the event $\mathcal{B}_n$, it is easily verified that $\|([\mathrm{Id}]_{\underline{k}}+[\Xi]_{\underline{k}})^{-1}-[\mathrm{Id}]_{\underline{k}}\|_s\leqslant 1/7$ which we exploit to conclude

$$\begin{aligned}6/7\|[\nabla_\omega]_{\underline{k}}^{1/2}[\Gamma]_{\underline{k}}^{-1}[\nabla_\omega]_{\underline{k}}^{1/2}\|_s\leqslant\|[\nabla_\omega]_{\underline{k}}^{1/2}[\widehat{\Gamma}]_{\underline{k}}^{-1}[\nabla_\omega]_{\underline{k}}^{1/2}\|_s\leqslant 8/7\|[\nabla_\omega]_{\underline{k}}^{1/2}[\Gamma]_{\underline{k}}^{-1}[\nabla_\omega]_{\underline{k}}^{1/2}\|_s,\\ 6/7\|[\Gamma]_{\underline{k}}^{-1}\|_s\leqslant\|[\widehat{\Gamma}]_{\underline{k}}^{-1}\|_s\leqslant 8/7\|[\Gamma]_{\underline{k}}^{-1}\|_s\quad\text{and}\\ 6/7x^t[\Gamma]_{\underline{k}}^{-1}x\leqslant x^t[\widehat{\Gamma}]_{\underline{k}}^{-1}x\leqslant 8/7x^t[\Gamma]_{\underline{k}}^{-1}x,\quad\text{for all } x\in\mathbb{R}^k,\end{aligned}\tag{B.10}$$

and, consequently

$$(6/7)[\widehat{g}]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[\widehat{g}]_{\underline{k}}\leqslant[\widehat{g}]_{\underline{k}}^t[\widehat{\Gamma}]_{\underline{k}}^{-1}[\widehat{g}]_{\underline{k}}\leqslant(8/7)[\widehat{g}]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[\widehat{g}]_{\underline{k}}.\tag{B.11}$$

Moreover, from $\|[\Xi]_{\underline{k}}\|_s\leqslant 1/8$ we obtain after some algebra,

$$\begin{aligned}&[g]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[g]_{\underline{k}}\leqslant\frac{1}{16}[g]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[g]_{\underline{k}}+4[W]_{\underline{k}}[\Gamma]_{\underline{k}}^{-1}[W]_{\underline{k}}+2[\widehat{g}]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[\widehat{g}]_{\underline{k}},\\ &[\widehat{g}]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[\widehat{g}]_{\underline{k}}\leqslant\frac{33}{16}[g]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[g]_{\underline{k}}+4[W]_{\underline{k}}[\Gamma]_{\underline{k}}^{-1}[W]_{\underline{k}}.\end{aligned}$$

Combining each of these estimates with (B.11) yields

$$\begin{aligned}&(15/16)[g]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[g]_{\underline{k}}\leqslant 4[W]_{\underline{k}}[\Gamma]_{\underline{k}}^{-1}[W]_{\underline{k}}+(7/3)[\widehat{g}]_{\underline{k}}^t[\widehat{\Gamma}]_{\underline{k}}^{-1}[\widehat{g}]_{\underline{k}},\\ &(7/8)[\widehat{g}]_{\underline{k}}^t[\widehat{\Gamma}]_{\underline{k}}^{-1}[\widehat{g}]_{\underline{k}}\leqslant(33/16)[g]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[g]_{\underline{k}}+4[W]_{\underline{k}}[\Gamma]_{\underline{k}}^{-1}[W]_{\underline{k}}.\end{aligned}$$

If in addition $[W]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[W]_{\underline{k}}\leqslant\frac{1}{8}([g]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[g]_{\underline{k}}+\sigma_Y^2)$, i.e., on the event $\mathcal{C}_n$, then the last two estimates imply respectively

$$\begin{aligned}&(7/16)([g]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[g]_{\underline{k}}+\sigma_Y^2)\leqslant(15/16)\sigma_Y^2+(7/3)[\widehat{g}]_{\underline{k}}^t[\widehat{\Gamma}]_{\underline{k}}^{-1}[\widehat{g}]_{\underline{k}},\\ &(7/8)[\widehat{g}]_{\underline{k}}^t[\widehat{\Gamma}]_{\underline{k}}^{-1}[\widehat{g}]_{\underline{k}}\leqslant(41/16)[g]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[g]_{\underline{k}}+(1/2)\sigma_Y^2,\end{aligned}$$

and hence in case $1/2 \leqslant \widehat{\sigma}_Y^2/\sigma_Y^2 \leqslant 3/2$, i.e., on the event $\mathcal{A}_n$, we obtain

$$(7/16)([g]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[g]_{\underline{k}} + \sigma_Y^2) \leqslant (15/8)\widehat{\sigma}_Y^2 + (7/3)[\widehat{g}]_{\underline{k}}^t[\widehat{\Gamma}]_{\underline{k}}^{-1}[\widehat{g}]_{\underline{k}},$$
$$(7/8)([\widehat{g}]_{\underline{k}}^t[\widehat{\Gamma}]_{\underline{k}}^{-1}[\widehat{g}]_{\underline{k}} + \widehat{\sigma}_Y^2) \leqslant (41/16)[g]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[g]_{\underline{k}} + (29/16)\sigma_Y^2.$$

Combining the last two estimates we have

$$\frac{1}{6}(2[g]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[g]_{\underline{k}} + 2\sigma_Y^2) \leqslant (2[\widehat{g}]_{\underline{k}}^t[\widehat{\Gamma}]_{\underline{k}}^{-1}[\widehat{g}]_{\underline{k}} + 2\widehat{\sigma}_Y^2) \leqslant 3(2[g]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[g]_{\underline{k}} + 2\sigma_Y^2).$$

Since on the event $\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n$ the last estimate and (B.10) hold for all $1 \leqslant k \leqslant M_n^\omega$ it follows

$$\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n \subset \Big\{(1/6)\sigma_m^2 \leqslant \widehat{\sigma}_m^2 \leqslant 3\sigma_m^2 \text{ and } (6/7)\Delta_m^\Gamma \leqslant \widehat{\Delta}_m \leqslant (8/7)\Delta_m^\Gamma,\ \forall 1 \leqslant m \leqslant M_n^\omega\Big\}.$$

From $\widehat{\Lambda}_m = \frac{\log(\widehat{\Delta}_m \vee (m+2))}{\log(m+2)}$ it is easily seen that $(6/7) \leqslant \widehat{\Delta}_m/\Delta_m^\Gamma \leqslant (8/7)$ implies

$$1/2 \leqslant (1 + \log(7/6))^{-1} \leqslant \widehat{\Lambda}_m/\Lambda_m^\Gamma \leqslant (1 + \log(8/7)) \leqslant 3/2.$$

Taking into account the last estimates and the definitions $pen_m = \kappa\sigma_m^2 m\Delta_m^\Gamma\Lambda_m^\Gamma n^{-1}$ and $\widehat{pen}_m = 14\kappa\widehat{\sigma}_m^2 m\widehat{\Delta}_m\widehat{\Lambda}_m n^{-1}$ we obtain

$$\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n \subset \{pen_m \leqslant \widehat{pen}_m \leqslant 72\, pen_m,\ \forall 1 \leqslant m \leqslant M_n^\omega\}. \tag{B.12}$$

On the other hand, by exploiting successively (B.10) and Lemma B.3 we have

$$\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n \subset \left\{\frac{6}{7} \leqslant \frac{\|[\widehat{\Gamma}]_{\underline{m}}^{-1}\|_s}{\|[\Gamma]_{\underline{m}}^{-1}\|_s} \leqslant \frac{8}{7},\ \forall 1 \leqslant m \leqslant M_n^\omega\right\} \subset \Big\{M_n^- \leqslant \widehat{M} \leqslant M_n^+\Big\}. \tag{B.13}$$

From (B.12) and (B.13) follows the assertion of the lemma, which completes the proof. □

Lemma B.5. *For all $m, n \geqslant 1$ with $n \geqslant (8/7)\|[\Gamma]_{\underline{m}}^{-1}\|_s$ we have $\mho_{m,n} \subset \Omega_{m,n}$.*

Proof of Lemma B.5. Taking into account the identity $[\widehat{\Gamma}]_{\underline{m}} = [\Gamma]_{\underline{m}}^{1/2}\{[\mathrm{Id}]_{\underline{m}} + [\Xi]_{\underline{m}}\}[\Gamma]_{\underline{m}}^{1/2}$ we observe that $\|[\Xi]_{\underline{m}}\|_s \leqslant 1/8$ implies $\|[\widehat{\Gamma}]_{\underline{m}}^{-1}\|_s \leqslant (8/7)\|[\Gamma]_{\underline{m}}^{-1}\|_s$ due to the usual Neumann series argument. If $n \geqslant (8/7)\|[\Gamma]_{\underline{m}}^{-1}\|_s$, then the last assertion implies $\mho_{m,n} \subset \Omega_{m,n}$, which proves the lemma. □

C Proof of Proposition 3.1

We will suppose throughout this section that the conditions of Proposition 3.1 are satisfied and thus Assumption 3.1 particularly holds true which allows us to employ the Lemmas B.1-B.5 stated in Section B. Moreover, we show first technical assertions (Lemma C.1- C.5) where we exploit Assumption 3.2, i.e. X and ε are jointly normally distributed. They are used below to prove that the Assumptions 2.1 and 2.2 are satisfied (Proposition C.6 and C.7 respectively), which is the claim of Proposition 3.1.

We begin by recalling elementary properties due to the Assumption 3.2 which are frequently used in this section. Given $f \in \mathbb{H}$ the random variable $\langle f, X\rangle_{\mathbb{H}}$ is normally distributed with mean zero and variance $\langle \Gamma f, f\rangle_{\mathbb{H}}$. Consider the Galerkin solution β^m and $h \in \mathbb{H}_m$ then the random variables $\langle \beta - \beta^m, X\rangle_{\mathbb{H}}$ and $\langle h, X\rangle_{\mathbb{H}}$ are independent. Thereby, $Z_m = Y - \langle \beta^m, X\rangle_{\mathbb{H}} = \sigma\varepsilon + \langle \beta - \beta^m, X\rangle_{\mathbb{H}}$ and $[X]_{\underline{m}}$ are independent, normally distributed with mean zero, and, respectively, variance ρ_m^2 and covariance matrix $[\Gamma]_{\underline{m}}$. Consequently, $(\rho_m^{-1} Z_m, [X]_m^t [\Gamma]_m^{-1/2})$ is a $(m+1)$-dimensional vector of iid. standard normally distributed random variables. Let us further state elementary inequalities for Gaussian random variables.

Lemma C.1. *Let $\{U_i, V_{ij}, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m\}$ be independent and standard normally distributed random variables. Then we have for all $\eta > 0$ and $\zeta \geqslant 4m/n$*

$$P\left(n^{-1/2}\sum_{i=1}^{n}(U_i^2 - 1) \geqslant \eta\right) \leqslant \exp\left(-\frac{1}{8}\frac{\eta^2}{1+\eta\, n^{-1/2}}\right); \tag{C.1}$$

$$P\left(n^{-1}\left|\sum_{i=1}^{n} U_i V_{i1}\right| \geqslant \eta\right) \leqslant \frac{\eta n^{1/2}+1}{\eta n^{1/2}}\exp\left(-\frac{n}{4}\min\left\{\eta^2, 1/4\right\}\right); \tag{C.2}$$

$$P\left(n^{-2}\sum_{j=1}^{m}\left|\sum_{i=1}^{n} U_i V_{ij}\right|^2 \geqslant \zeta\right) \leqslant \exp\big(\frac{-n}{16}\big) + \exp\big(\frac{-\zeta n}{64}\big); \tag{C.3}$$

and for all $c \geqslant 1$ and $a_1, \dots, a_m \geqslant 0$ that

$$\mathbb{E}\left(\sum_{i=1}^{n} U_i^2 - 2\,c\,n\right)_+ \leqslant 16\exp\big(\frac{-c\,n}{16}\big) \tag{C.4}$$

$$\mathbb{E}\left(\sum_{j=1}^{m}\left|n^{-1/2}\sum_{i=1}^{n} U_i V_{ij}\right|^2 - 4\,c\,m\right)_+ \leqslant 16\exp\big(\frac{-c\,m}{16}\big) + 32\frac{c\,m}{n}\exp\big(\frac{-n}{16}\big) \tag{C.5}$$

$$\mathbb{E}\left(\sum_{j=1}^{m} a_j\left|\sum_{i=1}^{n} U_i V_{ij}\right|^2\right)^2 = n(n+2)\Big(\sum_{j=1}^{m} a_j^2 + (\sum_{j=1}^{m} a_j)^2\Big) \tag{C.6}$$

Proof of Lemma C.1. Define $W := \sum_{i=1}^n U_i^2$ and $Z_j := (\sum_{i=1}^n U_i^2)^{-1/2}\sum_{i=1}^n U_i V_{ij}$. Obviously, W has χ_n^2 distribution with n degrees of freedom and $Z_1, \dots, Z_m$ given $U_1, \dots, U_n$ are independent and standard normally distributed, which we use below without further reference. From the estimate (C.1) given in Dahlhaus and Polonik [2006] (Proposition A.1) follows

$$\begin{aligned} P\left(n^{-2}\left|\sum_{i=1}^{n} U_i V_{i1}\right|^2 \geqslant \eta^2\right) &\leqslant P(n^{-1}W \geqslant 2) + \mathbb{E}\big[P\big(2n^{-1}|Z_1|^2 \geqslant \eta^2 \big| U_1, \dots, U_n\big)\big] \\ &\leqslant \exp\left(-\frac{n}{16}\right) + \frac{1}{\sqrt{\pi\eta^2 n}}\exp\left(-\frac{\eta^2 n}{4}\right), \end{aligned}$$

which implies (C.2). The estimate (C.3) follows analogously and we omit the details. By employing (C.1), $2c-1\geqslant c$ and $n^{-1}(cn+t)\geqslant 1$ we obtain (C.4). Indeed,

$$\begin{aligned}\mathbb{E}\bigg(\sum_{i=1}^n U_i^2-2\,c\,n\bigg)_+ &= \int_0^\infty P\bigg(n^{-1/2}\sum_{i=1}^n(U_i^2-1)\geqslant n^{-1/2}(cn+t)\bigg)dt\\ &\leqslant \int_0^\infty \exp\bigg(-\frac{1}{8}\frac{n^{-1}(cn+t)^2}{1+n^{-1}(cn+t)}\bigg)dt\leqslant \int_0^\infty \exp\bigg(-\frac{1}{16}(cn+t)\bigg)dt\\ &= \exp\bigg(-\frac{cn}{16}\bigg)\int_0^\infty \exp\bigg(-\frac{t}{16}\bigg)dt = 16\exp\bigg(-\frac{cn}{16}\bigg)\end{aligned}$$

From the last estimate and (C.1) follows (C.5), because

$$\begin{aligned}&\mathbb{E}\bigg(\sum_{j=1}^m\bigg|n^{-1/2}\sum_{i=1}^n U_iV_{ij}\bigg|^2-4\,c\,m\bigg)_+\\ &\qquad\leqslant \mathbb{E}\bigg[n^{-1}W\mathbb{E}\bigg[\bigg(\sum_{j=1}^m|Z_j|^2-2\,c\,m\bigg)_+\bigg|U_1,\dots,U_n\bigg]+2\,c\,m\,n^{-1}(W-2n)_+\bigg]\\ &\qquad\leqslant 16\exp\bigg(-\frac{c\,m}{16}\bigg)\mathbb{E}[n^{-1}W]+32\,\frac{c\,m}{n}\exp\bigg(-\frac{n}{16}\bigg).\end{aligned}$$

It remains to prove (C.6) which can be realized as follows (keep in mind that $\mathbb{E}[W^2]=n(n+2)$)

$$\begin{aligned}\mathbb{E}\bigg(\sum_{j=1}^m a_j\bigg|\sum_{i=1}^n U_iV_{ij}\bigg|^2\bigg)^2 &= \mathbb{E}\bigg[W^2\mathbb{E}\bigg[\big(\sum_{j=1}^m a_j|Z_j|^2\big)^2\bigg|U_1,\dots,U_n\bigg]\bigg]\\ &= \mathbb{E}[W^2]\bigg(\sum_{j=1}^m a_j^2+\big(\sum_{j=1}^m a_j\big)^2\bigg).\end{aligned}$$

□

Lemma C.2. *For all $n,m\geqslant 1$ we have*

$$n^2\rho_m^{-4}\mathbb{E}\|[W]_{\underline{m}}\|^4\leqslant 6\,(\mathbb{E}\|X\|^2)^2. \tag{C.7}$$

Furthermore, there exist a numerical constant $C>0$ such that for all $n\geqslant 1$

$$n^8\max_{1\leqslant m\leqslant\lfloor n^{1/4}\rfloor}P\bigg(\frac{[W]_{\underline{m}}^t[\Gamma]_{\underline{m}}^{-1}[W]_{\underline{m}}}{\rho_m^2}>\frac{1}{16}\bigg)\leqslant C; \tag{C.8}$$

$$n^8\max_{1\leqslant m\leqslant\lfloor n^{1/4}\rfloor}P\big(\|[\Xi]_{\underline{m}}\|_s>1/8\big)\leqslant C; \tag{C.9}$$

$$n^7P\big(\{1/2\leqslant\widehat{\sigma}_Y^2/\sigma_Y^2\leqslant 3/2\}^c\big)\leqslant C. \tag{C.10}$$

Proof of Lemma C.2. Let $n,m\geqslant 1$ be fixed, denote by $(\lambda_j,e_j)_{1\leqslant j\leqslant m}$ an eigenvalue decomposition of $[\Gamma]_{\underline{m}}$. Define $U_i:=(\sigma\varepsilon_i+\langle\beta-\beta^m,X_i\rangle_{\mathbb{H}})/\rho_m$ and $V_{ij}:=(\lambda_j^{-1/2}e_j^t[X_i]_{\underline{m}})$, $1\leqslant i\leqslant n$, $1\leqslant j\leqslant m$, where $U_1,\dots,U_n,V_{11},\dots,V_{nm}$ are independent and standard normally distributed

random variables.
Proof of (C.7) and (C.8). Taking into account $\sum_{j=1}^m \lambda_j \leqslant \mathbb{E}\|X\|_{\mathbb{H}}^2$ and the identities $n^4\rho_m^{-4}\|[W]_{\underline{m}}\|^4 = (\sum_{j=1}^m \lambda_j(\sum_{i=1}^n U_iV_{ij})^2)^2$ and $([W]_{\underline{m}}^t[\Gamma]_{\underline{m}}^{-1}[W]_{\underline{m}})/\rho_m^2 = n^{-2}\sum_{j=1}^m(\sum_{i=1}^n U_iV_{ij})^2$ the assertions (C.7) and (C.8) follow, respectively, from (C.6) and (C.3) in Lemma C.1 (with $a_j = \lambda_j$).
Proof of (C.9). Since $n\|[\Xi]_{\underline{m}}\|_s \leqslant m\max_{1\leqslant j,l\leqslant m}|\sum_{i=1}^n(V_{ij}V_{il}-\delta_{jl})|$ we obtain due to (C.1) and (C.2) in Lemma C.1 that for all $\eta>0$

$$
\begin{aligned}
P(\|[\Xi]_{\underline{m}}\|_s \geqslant \eta) &\leqslant \sum_{1\leqslant j,l\leqslant m} P(|n^{-1}\sum_{i=1}^n(V_{ij}V_{il}-\delta_{jl})| \geqslant \eta/m)\\
&\leqslant m^2\max\left\{P(|n^{-1}\sum_{i=1}^n V_{i1}V_{i2}| \geqslant \eta/m), P(|n^{-1/2}\sum_{i=1}^n(V_{i1}^2-1)| \geqslant n^{1/2}\eta/m)\right\}\\
&\leqslant m^2\max\left\{(1+\frac{m}{\eta n^{1/2}})\exp\Big(-\frac{n}{4}\min\{\eta^2/m^2, 1/4\}\Big), 2\exp\Big(-\frac{1}{8}\frac{n\eta^2/m^2}{1+\eta/m}\Big)\right\}.
\end{aligned}
$$

Moreover, for all $\eta \leqslant m/2$ the last bound simplifies to

$$
P(\|[\Xi]_{\underline{m}}\|_s \geqslant \eta) \leqslant m^2\max\left\{1+\frac{2m}{\eta n^{1/2}}, 2\right\}\exp\Big(-\frac{1}{12}\frac{n\eta^2}{m^2}\Big).
$$

and it is easily seen that the last bound implies (C.9).
Proof of (C.10). Since $Y_1/\sigma_Y, \dots, Y_n/\sigma_Y$ are independent and standard normally distributed, by exploiting that $\{1/2 \leqslant \widehat{\sigma}_Y^2/\sigma_Y^2 \leqslant 3/2\}^c \subset \{|n^{-1}\sum_{i=1}^n Y_i^2/\sigma_Y^2 - 1| > 1/2\}$, (C.10) follows from (C.1) in Lemma C.1, which completes the proof. □

Lemma C.3. *We have for all $c \geqslant 1$ and $n, m \geqslant 1$*

$$
\mathbb{E}\left(\frac{n[W]_{\underline{m}}^t[\Gamma]_{\underline{m}}^{-1}[W]_{\underline{m}}}{\rho_m^2} - 4\,c\,m\right)_+ \leqslant 16\exp\big(\frac{-c\,m}{16}\big) + 32\frac{c\,m}{n}\exp\big(\frac{-n}{16}\big).
$$

Proof of Lemma C.3. The assertion follows from (C.5) in Lemma C.1 and the identity $n\|[\Gamma]_{\underline{m}}^{-1/2}[W]_{\underline{m}}\|^2\rho_m^{-2} = \sum_{j=1}^m(n^{-1/2}\sum_{i=1}^n U_iV_{ij})^2$ derived in the proof of Lemma C.2. □

Lemma C.4. *There exists a constant $C(d)$ only depending on d such that for all $n \geqslant 1$*

$$
\sup_{\beta\in\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\sum_{k=m_n^\diamond}^{M_n^+}\Delta_k^\Gamma\,\mathbb{E}\Big([W]_{\underline{k}}^t[\Gamma]_{\underline{k}}^{-1}[W]_{\underline{k}} - 4\,\sigma_k^2\frac{k\,\Lambda_k^\Gamma}{n}\Big)_+ \leqslant C(d)(\sigma^2+r)\Sigma n^{-1}.
$$

Proof of Lemma C.4. The key argument of the proof is the estimate given in Lemma C.3 with $c = \Lambda_k^\Gamma$. Taking into account this upper bound and that for all $\beta \in \mathcal{F}_\beta^r$ and $\Gamma \in \mathcal{G}_\gamma^d$ the estimates $\Delta_k^\Gamma \leqslant 4d^3\Delta_k^\gamma$, $(1+\log d)^{-1}\Lambda_k^\gamma \leqslant \Lambda_k^\Gamma$, $\delta_{M_n^+}^\Gamma \leqslant nCd^6(1+\log d)$ (recall that $\delta_m^\Gamma = m\Delta_m^\Gamma\Lambda_m^\Gamma$) and $\rho_k^2 \leqslant \sigma_k^2 \leqslant 2(\sigma^2+35d^6r)$ (Lemma B.2 (i), (ii) and (iv) respectively) hold true, we obtain

$$\sum_{k=m_n^\diamond}^{M_n^+} \Delta_k^\Gamma \, \mathbb{E}\bigg([W]_{\underline{k}}^t [\Gamma]_{\underline{k}}^{-1} [W]_{\underline{k}} - 4\,\sigma_k^2 \,\frac{k\,\Lambda_k^\Gamma}{n} \bigg)_+$$
$$\leqslant \sum_{k=1}^{M_n^+} \frac{\sigma_k^2 \Delta_k^\Gamma}{n} \mathbb{E}\bigg(\frac{n[W]_{\underline{k}}^t [\Gamma]_{\underline{k}}^{-1} [W]_{\underline{k}}}{\rho_k^2} - 4\,k\,\Lambda_k^\Gamma \bigg)_+$$
$$\leqslant C(d)(\sigma^2 + r)\, n^{-1} \left\{ \sum_{k=1}^{M_n^+} \Delta_k^\gamma \exp\Big(- \frac{k\Lambda_k^\gamma}{16(1+\log d)} \Big) + M_n^+ \exp\big(-n/16\big) \right\}.$$

Finally, exploiting that the constant Σ satisfies (3.3) and that $M_n^+ \exp\big(-n/16\big) \leqslant C$ for all $n \geqslant 1$ we obtain the assertion of the lemma, which completes the proof. □

Lemma C.5. *There exist a numerical constant C and a constant $C(d)$ only depending on d such that for all $n \geqslant 1$ we have*

$$\sup_{\beta\in\mathcal{F}_b^r} \sup_{\Gamma\in\mathcal{G}_\gamma^d} \Big\{ n^6 (M_n^+)^2 \max_{1\leqslant m\leqslant M_n^+} P\big(\mho_{m,n}^c\big) \Big\} \leqslant C; \tag{C.11}$$

$$\sup_{\beta\in\mathcal{F}_b^r} \sup_{\Gamma\in\mathcal{G}_\gamma^d} \Big\{ n\, M_n^+ \max_{1\leqslant m\leqslant M_n^+} P\big(\Omega_{m,n}^c\big) \Big\} \leqslant C(d); \tag{C.12}$$

$$\sup_{\beta\in\mathcal{F}_b^r} \sup_{\Gamma\in\mathcal{G}_\gamma^d} \big\{ n^7 P(\mathcal{E}_n^c) \big\} \leqslant C. \tag{C.13}$$

Proof of Lemma C.5. Since $M_n^+ \leqslant \lfloor n^{1/4} \rfloor$ and $\mho_{m,n}^c = \big\{ \|[\Xi]_{\underline{m}}\| > 1/8 \big\}$ the assertion (C.11) follows from (C.8) in Lemma C.2. Consider (C.12). Let $n_o := n_o(d) := \exp(128 d^6) \geqslant 8d^3$, and consequently $\Delta_{M_n^+}^\omega (M_n^+ \log n) \geqslant 128 d^6$ for all $n \geqslant n_o$. We distinguish in the following the cases $n < n_o$ and $n \geqslant n_o$. First, consider $1 \leqslant n \leqslant n_o$. Obviously, we have $M_n^+ \max_{1\leqslant m\leqslant M_n^+} P(\Omega_{m,n}^c) \leqslant M_n^+ \leqslant n^{-1} n_o^{5/4} \leqslant C(d) n^{-1}$ since $M_n^+ \leqslant n^{1/4}$ and n_o depends on d only. On the other hand, if $n \geqslant n_o$ then from Lemma B.2 (iii) follows $n \geqslant 2 \max_{1\leqslant m\leqslant M_n^+} \|[\Gamma]_{\underline{m}}^{-1}\|$, and hence $\mho_{m,n} \subset \Omega_{m,n}$ for all $1 \leqslant m \leqslant M_n^+$ by employing Lemma B.5. From (C.11) we conclude $M_n^+ \max_{1\leqslant m\leqslant M_n^+} P(\Omega_{m,n}^c) \leqslant M_n^+ \max_{1\leqslant m\leqslant M_n^+} P(\mho_{m,n}^c) \leqslant C n^{-3}$. By combination of the two cases we obtain (C.12). It remains to show (C.13). Consider the events $\mathcal{A}_n$, $\mathcal{B}_n$ and $\mathcal{C}_n$ defined in (A.2), where $\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n \subset \mathcal{E}_n$ due to Lemma B.4. Moreover we have $n^7 P(\mathcal{A}_n^c) \leqslant C$, $n^7 P(\mathcal{B}_n^c) \leqslant C$, and $n^7 P(\mathcal{C}_n^c) \leqslant C$, due to (C.10), (C.9) and (C.8) in Lemma C.2 respectively (keep in mind that $\lfloor n^{1/4} \rfloor \geqslant M_n^\omega$ and $2(\sigma_Y^2 + [g]_{\underline{k}}^t [\Gamma]_{\underline{k}}^{-1} [g]_{\underline{k}}) = \sigma_k^2 \geqslant \rho_k^2$). Combining these estimates we obtain (C.13), which completes the proof. □

Proposition C.6. *Let $\kappa = 96$ in the definition of the penalty pen given in* (2.11)*. There exists a constant $C(d)$ such that for all $n \geqslant 1$ we have*

$$\sup_{\beta\in\mathcal{F}_b^r} \sup_{\Gamma\in\mathcal{G}_\gamma^d} \mathbb{E}\bigg\{ \sup_{m_n^\diamond \leqslant k \leqslant M_n^+} \Big(\|\widehat{\beta}_k - \beta^k\|_\omega^2 - \frac{1}{6}\, pen_k \Big)_+ \bigg\} \leqslant C(d)(\sigma^2 + r)\Sigma\, n^{-1}.$$

PROOF OF PROPOSITION C.6. We use the identity $[\widehat{\beta}_k - \beta^k]_{\underline{k}} = [\widehat{\Gamma}]_{\underline{k}}^{-1}[W]_{\underline{k}}\,\mathbb{1}_{\Omega_{k,n}} - [\beta^k]_{\underline{k}}\,\mathbb{1}_{\Omega_{k,n}^c}$, and obtain

$$\|\widehat{\beta}_k - \beta^k\|_\omega^2 = \|[\nabla_\omega]_{\underline{k}}^{1/2}[\widehat{\Gamma}]_{\underline{k}}^{-1}[W]_{\underline{k}}\|^2\,\mathbb{1}_{\Omega_{k,n}} + \|\beta^k\|_\omega^2\,\mathbb{1}_{\Omega_{k,n}^c}. \tag{C.14}$$

Exploiting further $\|([\mathrm{Id}]_{\underline{k}} + [\Xi]_{\underline{k}})^{-1}\|_s\,\mathbb{1}_{\mho_{k,n}} \leqslant 2$, the identity $[\widehat{\Gamma}]_{\underline{k}} = [\Gamma]_{\underline{k}}^{1/2}\{[\mathrm{Id}]_{\underline{k}} + [\Xi]_{\underline{k}}\}[\Gamma]_{\underline{k}}^{1/2}$ and the definition of Δ_k^Γ it follows that $\|[\nabla_\omega]_{\underline{k}}^{1/2}[\widehat{\Gamma}]_{\underline{k}}^{-1}[W]_{\underline{k}}\|^2\,\mathbb{1}_{\mho_{k,n}} \leqslant 4\Delta_k^\Gamma\|[\Gamma]_{\underline{k}}^{-1/2}[W]_{\underline{k}}\|^2$. On the other hand, we have $\|[\nabla_\omega]_{\underline{k}}^{1/2}[\widehat{\Gamma}]_{\underline{k}}^{-1}[W]_{\underline{k}}\|^2\,\mathbb{1}_{\Omega_{k,n}} \leqslant \Delta_k^\omega n^2\|[W]_{\underline{k}}\|^2$. From these estimates and $\|\beta^k\|_\omega \leqslant \|\beta^k\|_b$ (ωb^{-1} is non-increasing due to Assumption 3.1) we deduce for all $k \geqslant 1$

$$\|\widehat{\beta}_k - \beta^k\|_\omega^2 \leqslant 4\Delta_k^\Gamma\|[\Gamma]_{\underline{k}}^{-1/2}[W]_{\underline{k}}\|^2 + \Delta_k^\omega\, n^2\|[W]_{\underline{k}}\|^2\,\mathbb{1}_{\mho_{k,n}^c} + \|\beta^k\|_b^2\,\mathbb{1}_{\Omega_{k,n}^c}.$$

Taking into account this upper bound, the notations Δ_k^Γ and Λ_k^Γ given in (A.1), and the definition $pen_k = 96\sigma_k^2 k\Delta_k^\Gamma\Lambda_k^\Gamma n^{-1}$ we obtain for all $\beta \in \mathcal{F}_b^r$ and $\Gamma \in \mathcal{G}_\gamma^d$ that

$$\mathbb{E}\bigg\{\sup_{m_n^\diamond \leqslant k \leqslant M_n^+}\Big(\|\widehat{\beta}_k - \beta^k\|_\omega^2 - \frac{1}{6}\,pen_k\Big)_+\bigg\} \leqslant 4\sum_{k=m_n^\diamond}^{M_n^+}\Delta_k^\Gamma\,\mathbb{E}\bigg(\|[\Gamma]_{\underline{k}}^{-1/2}[W]_{\underline{k}}\|^2 - 4\,\sigma_k^2\,\frac{k\,\Lambda_k^\Gamma}{n}\bigg)_+$$
$$+ \sum_{k=m_n^\diamond}^{M_n^+} n^3\big(\mathbb{E}\|[W]_{\underline{k}}\|^4\big)^{1/2}\big(P(\mho_{k,n}^c)\big)^{1/2} + \sum_{k=m_n^\diamond}^{M_n^+}\|\beta^k\|_b^2 P(\Omega_{k,n}^c)$$

Consider the second and third right hand side term. By exploiting, respectively, (C.7) in Lemma C.2 and (B.5) in Lemma B.1 together with $\rho_m^2 \leqslant 2(\sigma^2 + 35d^6r)$ (Lemma B.2 (iv)) these two terms are bounded by

$$6(\sigma^2 + 35d^6r)\mathbb{E}\|X\|^2 n^2 M_n^+ \max_{m_n^\diamond \leqslant k \leqslant M_n^+}\big(P(\mho_{k,n}^c)\big)^{1/2} + 34d^8rM_n^+ \max_{m_n^\diamond \leqslant k \leqslant M_n^+} P(\Omega_{k,n}^c).$$

Combining this upper bound, the property $\mathbb{E}\|X\|^2 \leqslant d\sum_{j\geqslant 1}\gamma_j \leqslant d\Sigma$ and the estimates given in Lemma C.5 we deduce for all $\beta \in \mathcal{F}_b^r$ and $\Gamma \in \mathcal{G}_\gamma^d$ that

$$\sup_{\beta\in\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\mathbb{E}\bigg\{\sup_{m_n^\diamond \leqslant k \leqslant M_n^+}\Big(\|\widehat{\beta}_k - \beta^k\|_\omega^2 - \frac{1}{6}\,pen_k\Big)_+\bigg\} \leqslant C(d)(\sigma^2 + r)\Sigma\, n^{-1} +$$
$$4\sup_{\beta\in\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\sum_{k=m_n^\diamond}^{M_n^+}\Delta_k^\Gamma\,\mathbb{E}\bigg(\|[\Gamma]_{\underline{k}}^{-1/2}[W]_{\underline{k}}\|^2 - 4\,\sigma_k^2\,\frac{k\,\Lambda_k^\Gamma}{n}\bigg)_+$$

The result of the proposition follows now by replacing the last right hand side term by its upper bound given in Lemma C.4, which completes the proof. □

PROPOSITION C.7. *Let $\kappa = 96$ in the definition of pen and $\widehat{pen}$ given in* (2.11) *and* (2.6) *respectively. There exists a constant $C(d)$ such that for all $n \geqslant 1$ we have*

$$\sup_{\beta\in\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\mathbb{E}(\|\widehat{\beta}_{\widehat{m}} - \beta\|_\omega^2\,\mathbb{1}_{\mathcal{E}_n^c}) \leqslant C(d)(\sigma^2 + r)\Sigma\, n^{-1}.$$

Proof of Proposition C.7. Taking into account the decomposition (C.14) and the estimate $\|[\nabla_\omega]_{\underline{k}}^{1/2}[\widehat{\Gamma}]_{\underline{k}}^{-1}[W]_{\underline{k}}\|^2 \mathbb{1}_{\Omega_{k,n}} \leqslant \Delta_k^\omega n^2 \|[W]_{\underline{k}}\|^2$ given in the proof of Proposition C.6 we conclude

$$\|\widehat{\beta}_k - \beta\|_\omega^2 \leqslant 2\Delta_k^\omega n^2 \|[W]_{\underline{k}}\|^2 + 2\|\beta^k\|_\omega^2 + 2\|\beta\|_\omega^2, \quad \text{for all } k \geqslant 1.$$

By exploiting (B.5) in Lemma B.1 together with $\|\beta^k\|_\omega \leqslant \|\beta^k\|_b$ (ωb^{-1} is non-increasing due to Assumption 3.1) we obtain for all $\beta \in \mathcal{F}_b^r$ and $\Gamma \in \mathcal{G}_\gamma^d$ and for all $k \geqslant 1$ that

$$\|\widehat{\beta}_k - \beta\|_\omega^2 \leqslant 2\Delta_k^\omega n^2 \|[W]_{\underline{k}}\|^2 + 2(34d^8 r + r).$$

Since $1 \leqslant \widehat{m} \leqslant M_n^\omega$ and $\max_{1\leqslant k\leqslant M_n^\omega} \Delta_k^\omega \leqslant n$ it follows for all $\beta \in \mathcal{F}_b^r$ and $\Gamma \in \mathcal{G}_\gamma^d$ that

$$\mathbb{E}(\|\widehat{\beta}_{\widehat{m}} - \beta\|_\omega^2 \mathbb{1}_{\mathcal{E}_n^c}) \leqslant 2n^3 M_n^\omega \max_{1\leqslant k\leqslant M_n^\omega} \big(\mathbb{E}\|[W]_{\underline{k}}\|^4\big)^{1/2} |P(\mathcal{E}_n^c)|^{1/2} + 2(34d^8 r + r) M_n^\omega P(\mathcal{E}_n^c).$$

From (C.7) in Lemma C.2 together with $\rho_m^2 \leqslant 2(\sigma^2 + 35d^6 r)$ (Lemma B.2) and $\mathbb{E}\|X\|^2 \leqslant d\Sigma$ we conclude for all $\beta \in \mathcal{F}_b^r$ and $\Gamma \in \mathcal{G}_\gamma^d$ that

$$\mathbb{E}(\|\widehat{\beta}_{\widehat{m}} - \beta\|_\omega^2 \mathbb{1}_{\mathcal{E}_n^c}) \leqslant 12(\sigma^2 + 35d^6 r) d\Sigma n^2 M_n^\omega |P(\mathcal{E}_n^c)|^{1/2} + 2(34d^8 r + r) M_n^\omega P(\mathcal{E}_n^c).$$

The result of the proposition follows now from $M_n^\omega \leqslant \lfloor n^{1/4} \rfloor$ and by replacing the probability $P(\mathcal{E}_n^c)$ by its upper bound Cn^{-7} given in Lemma C.5, which completes the proof. □

Proof of Proposition 3.1. The assertion follows from Proposition C.6 and Proposition C.7 and we omit the details. □

D Proof of Proposition 3.3

We assume throughout this section that the conditions of Proposition 3.3 are satisfied which allows us to employ the Lemma B.1-B.5 stated in Section B. We formulate first preliminary results (Proposition D.1 and Lemma D.2- D.5) which rely on the moment conditions imposed through Assumption 3.3. They are used below to prove that the Assumptions 2.1 and 2.2 are satisfied (Proposition D.6 and D.7 respectively), which is the claim of Proposition 3.3. We begin by gathering elementary bounds due to Assumption 3.3. Let k be given by Assumption 3.3 then for all $m \geqslant 1$ we have

$$\mathbb{E}|Z_m|^{4k} \leqslant \rho_m^2 \eta^{4k}, \quad \mathbb{E}|Y|^{4k} \leqslant \sigma_Y^{4k} \eta^{4k}, \quad \max_{1\leqslant j\leqslant m} \mathbb{E}|([\Gamma]_{\underline{m}}^{-1/2}[X]_{\underline{m}})_j|^{4k} \leqslant \eta^{4k},$$
$$\mathbb{E}\big|\langle \beta - \beta^m, X\rangle_{\mathbb{H}}\big|^{4k} \leqslant \|\Gamma^{1/2}(\beta^m - \beta)\|_{\mathbb{H}}^{4k} \eta^{4k}, \quad \mathbb{E}\big|[X]_{\underline{m}}^t [\Gamma]_{\underline{m}}^{-1} [X]_{\underline{m}}\big|^{2k} \leqslant m^{2k} \eta^{4k}.$$

Moreover, if V is a non negative random variable with $\mathbb{E}V^k < \infty$ then the elementary inequality $\mathbb{E}V \mathbb{1}_{\{V\geqslant t\}} \leqslant t^{-k+1}\mathbb{E}V^k$ holds true for all $t > 0$. Taking into account this estimate we obtain under Assumption 3.3, that for all $m, n \geqslant 1$

$$\mathbb{E}\varepsilon^2 \mathbb{1}_{\{|\varepsilon|>n^{1/6}\}} \leqslant \eta^{32} n^{-5},$$
$$\mathbb{E}\big|\langle \beta - \beta^m, X\rangle_{\mathbb{H}}\big|^2 \mathbb{1}_{\{|\langle \beta-\beta^m, X\rangle_{\mathbb{H}}| > \|\Gamma^{1/2}(\beta^m-\beta)\|_{\mathbb{H}} n^{1/6}\}} \leqslant \eta^{32} \|\Gamma^{1/2}(\beta^m - \beta)\|_{\mathbb{H}}^2 n^{-5},$$
$$\mathbb{E}|[X]_{\underline{m}}^t [\Gamma]_{\underline{m}}^{-1} [X]_{\underline{m}}|^2 \mathbb{1}_{\{[X]_{\underline{m}}^t [\Gamma]_{\underline{m}}^{-1} [X]_{\underline{m}} > mn^{1/3}\}} \leqslant \eta^{32} m^2 n^{-14/3} \quad \text{(D.1)}$$

and by employing Markov's inequality

$$P(|\varepsilon| > n^{1/6}) \leqslant n^{-16/3}\eta^{32}, \quad P(|\langle \beta - \beta^m, X\rangle_{\mathbb{H}}| > \|\Gamma^{1/2}(\beta^m - \beta)\|_{\mathbb{H}} n^{1/6}) \leqslant n^{-16/3}\eta^{32}. \quad \text{(D.2)}$$

We exploit these bounds in the following proofs. Moreover, the key argument used in the proof of Lemma D.3 is the following inequality due to Talagrand [1996] (see e.g. Klein and Rio [2005]).

Proposition D.1 (Talagrand's Inequality). *Let $T_1, \dots, T_n$ be independent $\mathcal{T}$-valued random variables and $\nu_s^* = (1/n)\sum_{i=1}^n \big[\nu_s(T_i) - \mathbb{E}[\nu_s(T_i)]\big]$, for ν_s belonging to a countable class $\{\nu_s : s \in \mathbb{S}\}$ of measurable functions. Then, for $\varepsilon > 0$,*

$$\mathbb{E}\Big(\sup_{s\in\mathbb{S}} |\nu_s^*|^2 - 2(1+2\varepsilon)H^2\Big)_+ \leqslant C\left(\frac{v}{n}\exp(-K_1\varepsilon\frac{nH^2}{v}) + \frac{h^2}{n^2C^2(\varepsilon)}\exp(-K_2C(\varepsilon)\sqrt{\varepsilon}\frac{nH}{h})\right)$$

with $K_1 = 1/6$, $K_2 = 1/(21\sqrt{2})$, $C(\varepsilon) = \sqrt{1+\varepsilon} - 1$ and C a universal constant and where

$$\sup_{s\in\mathbb{S}}\sup_{t\in\mathcal{T}} |\nu_s(t)| \leqslant h, \quad \mathbb{E}\left[\sup_{s\in\mathbb{S}} |\nu_s^*|\right] \leqslant H, \quad \sup_{s\in\mathbb{S}} \frac{1}{n}\sum_{i=1}^n \mathbb{V}\mathrm{ar}(\nu_s(T_i)) \leqslant v.$$

Lemma D.2. *There exist a numerical constant $C > 0$ such that for all $n \geqslant 1$*

$$n^2 \sup_{m\geqslant 1} \rho_m^{-4}\mathbb{E}\|[W]_{\underline{m}}\|^4 \leqslant C\eta^8(\mathbb{E}\|X\|_{\mathbb{H}}^2)^2; \tag{D.3}$$

$$n^8 \max_{1\leqslant m\leqslant \lfloor n^{1/4}\rfloor} P\left(\frac{[W]_{\underline{m}}^t[\Gamma]_{\underline{m}}^{-1}[W]_{\underline{m}}}{\rho_m^2} > \frac{1}{16}\right) \leqslant C\eta^{64}; \tag{D.4}$$

$$n^8 \max_{1\leqslant m\leqslant \lfloor n^{1/4}\rfloor} P\big(\|[\Xi]_{\underline{m}}\|_s > 1/8\big) \leqslant C(\eta); \tag{D.5}$$

$$n^7 P\big(\{1/2 \leqslant \widehat{\sigma}_Y^2/\sigma_Y^2 \leqslant 3/2\}^c\big) \leqslant C\eta^{64}. \tag{D.6}$$

Proof of Lemma D.2. Let $n, m \geqslant 1$ be fixed, denote by $(\lambda_j, e_j)_{1\leqslant j\leqslant m}$ an eigenvalue decomposition of $[\Gamma]_{\underline{m}}$. Define $U_i := (\sigma\varepsilon_i + \langle \beta - \beta^m, X_i\rangle_{\mathbb{H}})/\rho_m$ and $V_{ij} := (\lambda_j^{-1/2} e_j^t [X_i]_{\underline{m}})$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant m$. Keep in mind that $\mathbb{E}|U_i|^{4k} \leqslant \eta^{4k}$, $\mathbb{E}|V_{ij}|^{4k} \leqslant \eta^{4k}$ and $\mathbb{E}|U_iV_{ij}|^{2k} \leqslant \eta^{4k}$ for some $k \geqslant 16$ due to Assumption 3.3 and $U_1V_{1j}, \dots, U_nV_{nj}$ are independent and centered random variables for all $1 \leqslant j \leqslant m$.

Proof of (D.3) and (D.4). Consider the identities $n^4\rho_m^{-4}\|[W]_{\underline{m}}\|^4 = (\sum_{j=1}^m \lambda_j(\sum_{i=1}^n U_iV_{ij})^2)^2$ and $([W]_{\underline{m}}^t[\Gamma]_{\underline{m}}^{-1}[W]_{\underline{m}})/\rho_m^2 = n^{-2}\sum_{j=1}^m(\sum_{i=1}^n U_iV_{ij})^2$. We apply successively Minkowski's (respectively Jensen's) inequality and Theorem 2.10 in Petrov [1995], which leads to

$$n^2\rho_m^{-4}\mathbb{E}\|[W]_{\underline{m}}\|^4 = n^{-2}\mathbb{E}\Big|\sum_{j=1}^m \lambda_j(\sum_{i=1}^n U_iV_{ij})^2\Big|^2 \leqslant n^{-2}\Big[\sum_{j=1}^m \lambda_j\Big(\mathbb{E}\big|\sum_{i=1}^n U_iV_{ij}\big|^4\Big)^{1/2}\Big]^2$$

$$\leqslant C\Big[\sum_{j=1}^m \lambda_j \max_{1\leqslant i\leqslant n}\Big(\mathbb{E}\big|U_iV_{ij}\big|^4\Big)^{1/2}\Big]^2 \leqslant C\eta^8\Big[\sum_{j=1}^m \lambda_j\Big]^2;$$

$$m^{-k}n^k\rho_m^{-2k}\mathbb{E}\|[\Gamma]_{\underline{m}}^{-1/2}[W]_{\underline{m}}\|^{2k} = n^{-k}\mathbb{E}\Big|m^{-1}\sum_{j=1}^m(\sum_{i=1}^n U_iV_{ij})^2\Big|^k \leqslant n^{-k}m^{-1}\sum_{j=1}^m\mathbb{E}\big|\sum_{i=1}^n U_iV_{ij}\big|^{2k}$$

$$\leqslant C(k)m^{-1}\sum_{j=1}^m \max_{1\leqslant i\leqslant n}\mathbb{E}\big|U_iV_{ij}\big|^{2k} \leqslant C(k)\eta^{4k}.$$

The first estimate implies (D.3) since $\sum_{j=1}^m\lambda_j \leqslant \mathbb{E}\|X\|_{\mathbb{H}}^2$. By employing Markov's inequality the second estimate with $k=16$ implies (D.4), that is

$$\max_{1\leqslant m\leqslant\lfloor n^{1/4}\rfloor} P\left(\frac{[W]_{\underline{m}}^t[\Gamma]_{\underline{m}}^{-1}[W]_{\underline{m}}}{\rho_m^2} > \frac{1}{16}\right) \leqslant Cn^{-16}\eta^{64}\max_{1\leqslant m\leqslant\lfloor n^{1/4}\rfloor} m^{16} \leqslant Cn^{-12}\eta^{64}.$$

Proof of (D.5). Since $V_{1j}V_{1l}-\delta_{jl},\dots,V_{nj}V_{nl}-\delta_{jl}$ are independent and centered random variables with $\mathbb{E}|V_{ij}V_{il}-\delta_{jl}|^{2k} \leqslant C\eta^{4k}$ for all $1\leqslant j,l\leqslant m$ it follows from Theorem 2.10 in Petrov [1995] that $n^k\mathbb{E}\big|n^{-1}\sum_{i=1}^n(V_{ij}V_{il}-\delta_{jl})\big|^{2k} \leqslant C(k)\eta^{4k}$. By employing the elementary inequality $\|[\Xi]_m\|_s^2 \leqslant \sum_{1\leqslant j,l\leqslant m}|V_{ij}V_{il}-\delta_{jl}|^2$, Jensen's inequality and the last bound we obtain $m^{-2k}n^k\mathbb{E}\|[\Xi]_{\underline{m}}\|_s^{2k} \leqslant C(k)\eta^{4k}$. Applying Markov's inequality and the last bound with $k=16$ we conclude

$$\max_{1\leqslant m\leqslant\lfloor n^{1/4}\rfloor} P\left(\|[\Xi]_{\underline{m}}\|_s > \frac{1}{8}\right) \leqslant Cn^{-16}\eta^{64}\max_{1\leqslant m\leqslant\lfloor n^{1/4}\rfloor} m^{32} \leqslant Cn^{-8}\eta^{64}$$

which proves the assertion (D.5).
Proof of (D.6). Since $Y_1^2/\sigma_Y^2-1,\dots,Y_n^2/\sigma_Y-1$ are independent and and centered random variables with $\mathbb{E}\big|Y_i^2/\sigma_Y^2-1\big|^{2k} \leqslant C(k)\eta^{4k}$ it follows from Theorem 2.10 in Petrov [1995] that $\mathbb{E}\big|n^{-1}\sum_{i=1}^n Y_i^2/\sigma_Y^2-1\big|^{2k} \leqslant C(k)n^{-k}\eta^{4k}$. Employing Markov's inequality and the last bound with $k=16$ we deduce $P\big(|n^{-1}\sum_{i=1}^n Y_i^2/\sigma_Y^2-1|>1/2\big) \leqslant Cn^{-16}\eta^{64}$. Thereby, the assertion (C.10) follows from the last bound by exploiting that $\{1/2\leqslant\widehat{\sigma}_Y^2/\sigma_Y^2\leqslant 3/2\}^c \subset \{|n^{-1}\sum_{i=1}^n Y_i^2/\sigma_Y^2-1|>1/2\}$, which completes the proof. $\square$

Lemma D.3. *Let $\varsigma_m := \sigma+\eta^2\|\Gamma^{1/2}(\beta^m-\beta)\|_{\mathbb{H}}$, $m\geqslant 1$. There exists a numerical constant C such that for all $\lfloor n^{1/4}\rfloor \geqslant m\geqslant 1$ we have*

$$\mathbb{E}\left(\|[\Gamma]_{\underline{m}}^{-1/2}[W_n]_{\underline{m}}\|^2 - 12\varsigma_m^2\frac{m\Lambda_m^\Gamma}{n}\right)_+ \leqslant C\frac{\varsigma_m^2}{n}\Big\{\exp\Big(-\frac{m\Lambda_m^\Gamma}{6}\Big)+\exp\Big(-\frac{n^{1/6}}{100}\Big)+\frac{\eta^{32}}{n^2}\Big\}.$$

Proof. Let $1\leqslant m\leqslant n$ be fixed and $\mathbb{S}^m := \{z\in\mathbb{R}^m : z^tz\leqslant 1\}$. Define the subsets $\mathcal{E}_n := \{e\in\mathbb{R} : |e|\leqslant n^{1/6}\}$, $\mathcal{X}_{1n} := \{x\in\mathbb{H} : |\langle\beta-\beta^m,x\rangle_{\mathbb{H}}| \leqslant \|\Gamma^{1/2}(\beta-\beta^m)\|_{\mathbb{H}}n^{1/6}\}$,

$\mathcal{X}_{2n} := \{x \in \mathbb{H} : [x]^t_{\underline{m}}[\Gamma]^{-1}_{\underline{m}}[x]_{\underline{m}} \leqslant mn^{1/3}\}$ and $\mathcal{X}_n := \mathcal{X}_{1n} \cap \mathcal{X}_{2n}$. Given $e \in \mathbb{R}$, $x \in \mathbb{H}$ and $s \in \mathbb{S}^m$ we set

$$\begin{aligned}\nu_s(e,x) &:= (\sigma e + \langle \beta - \beta^m, x\rangle_{\mathbb{H}}) s^t [\Gamma]^{-1/2}_{\underline{m}}[x]_{\underline{m}} \, \mathbb{1}_{\{e\in\mathcal{E}_n, x\in\mathcal{X}_n\}},\\ R_s(e,x) &:= (\sigma e + \langle \beta - \beta^m, x\rangle_{\mathbb{H}}) s^t [\Gamma]^{-1/2}_{\underline{m}}[x]_{\underline{m}}(1 - \mathbb{1}_{\{e\in\mathcal{E}_n, x\in\mathcal{X}_n\}}).\end{aligned}$$

Let $\nu_s^* := (1/n)\sum_{i=1}^n\{\nu_s(\varepsilon_i, X_i) - \mathbb{E}\nu_s(\varepsilon_i, X_i)\}$ and $R_s^* := (1/n)\sum_{i=1}^n\{R_s(\varepsilon_i, X_i) - \mathbb{E}R_s(\varepsilon_i, X_i)\}$, then it is easily seen that $\|[\Gamma]^{-1/2}_{\underline{m}}[W_n]_{\underline{m}}\|^2 = \sup_{s\in\mathbb{S}^m}|\nu_s^* + R_s^*|^2$ and hence

$$\begin{aligned}\mathbb{E}\Big(\|[\Gamma]^{-1/2}_{\underline{m}}[W_n]_{\underline{m}}\|^2 - 12\varsigma_m^2\frac{m\Lambda_m^\Gamma}{n}\Big)_+ &\leqslant 2\mathbb{E}\Big(\sup_{s\in\mathbb{S}^m}|\nu_s^*|^2 - 6\varsigma_m^2\frac{m\Lambda_m^\Gamma}{n}\Big)_+\\ &\quad + 2\mathbb{E}\sup_{s\in\mathbb{S}^m}|R_s^*|^2 =: 2\{T_1+T_2\}, \qquad \text{(D.7)}\end{aligned}$$

where we bound the terms T_1 and T_2 on the right hand side separately.
Consider first T_1 which we estimate by employing Talagrand's inequality. Obviously, we have

$$\begin{aligned}\sup_{e\in\mathbb{R},x\in\mathbb{H}}\sup_{s\in\mathbb{S}^m}|\nu_s(e,x)|^2 &= \sup_{e\in\mathbb{R},x\in\mathbb{H}}(\sigma e + \langle\beta-\beta^m, x\rangle_{\mathbb{H}})^2[x]^t_{\underline{m}}[\Gamma]^{-1}_{\underline{m}}[x]_{\underline{m}}\,\mathbb{1}_{\{e\in\mathcal{E}_n,x\in\mathcal{X}_n\}}\\ &\leqslant (\sigma + \|\Gamma^{1/2}(\beta^m-\beta)\|_{\mathbb{H}})^2 n^{2/3}m \leqslant \varsigma_m^2 n^{2/3} m =: h^2 \qquad \text{(D.8)}\end{aligned}$$

By employing the independence of ε and X it is easily seen that

$$\begin{aligned}&n\,\mathbb{E}\sup_{s\in\mathbb{S}^m}|\nu_s^*|^2 \leqslant \sigma^2 m + \mathbb{E}|\langle\beta-\beta^m, X\rangle_{\mathbb{H}}|^2[X]^t_{\underline{m}}[\Gamma]^{-1}_{\underline{m}}[X]_{\underline{m}},\\ &\sup_{s\in\mathbb{S}^m}\frac{1}{n}\sum_{i=1}^n\mathbb{V}\mathrm{ar}(\nu_s(\varepsilon_i,X_i)) \leqslant \sigma^2 + \sup_{s\in\mathbb{S}^m}\mathbb{E}|\langle\beta-\beta^m,X\rangle_{\mathbb{H}}|^2|s^t[\Gamma]^{-1/2}_{\underline{m}}[X]_{\underline{m}}|^2.\end{aligned}$$

By applying the Cauchy-Schwarz inequality together with $\mathbb{E}\|[\Gamma]^{-1/2}_{\underline{m}}[X]_{\underline{m}}\|^4 \leqslant m^2\eta^4$ and $\mathbb{E}\big|\langle\beta-\beta^m, X\rangle_{\mathbb{H}}\big|^4 \leqslant \|\Gamma^{1/2}(\beta^m-\beta)\|^4_{\mathbb{H}}\eta^4$ we obtain

$$\mathbb{E}\sup_{s\in\mathbb{S}^m}|\nu_s^*|^2 \leqslant \frac{m}{n}(\sigma^2 + \|\Gamma^{1/2}(\beta-\beta^m)\|^2_{\mathbb{H}}\eta^4) \leqslant \varsigma_m^2\frac{m\Lambda_m^\Gamma}{n} =: H^2, \qquad \text{(D.9)}$$

and taking in addition into account that $\mathbb{E}|s^t[\Gamma]^{-1/2}_{\underline{m}}[X]_{\underline{m}}|^4 \leqslant \eta^4$ for all $s \in \mathbb{S}^m$ we obtain

$$\sup_{s\in\mathbb{S}^m}\frac{1}{n}\sum_{i=1}^n\mathbb{V}\mathrm{ar}(\nu_s(\varepsilon_i,X_i)) \leqslant \sigma^2 + \|\Gamma^{1/2}(\beta^m-\beta)\|^2_{\mathbb{H}}\eta^4 \leqslant \varsigma_m^2 =: v. \qquad \text{(D.10)}$$

Combining (D.8), (D.9), (D.10) due to Talagrand's inequality (Lemma D.1 with $\varepsilon = 1$) follows

$$\begin{aligned}\mathbb{E}\Big(\sup_{s\in\mathbb{S}^m}|\nu_s^*|^2 - 6\varsigma_m^2\frac{m\Lambda_m^\Gamma}{n}\Big)_+ &\leqslant C\Big\{\frac{\varsigma_m^2}{n}\exp\Big(-\frac{m\Lambda_m^\Gamma}{6}\Big) + \frac{\varsigma_m^2 m}{n^{4/3}}\exp\Big(-\frac{n^{1/6}}{100}\Big)\Big\}\\ &\leqslant C\frac{\varsigma_m^2}{n}\Big\{\exp\Big(-\frac{m\Lambda_m^\Gamma}{6}\Big) + \exp\Big(-\frac{n^{1/6}}{100}\Big)\Big\} \qquad \text{(D.11)}\end{aligned}$$

where we used that $m \leqslant \lfloor n^{1/4} \rfloor$.
Consider T_2 on the right hand side of (D.7). By employing $[X]_{\underline{m}}[\Gamma]_{\underline{m}}^{-1}[X]_{\underline{m}}\, \mathbb{1}_{\{X\in\mathcal{X}_{2,n}\}} \leqslant mn^{1/3}$ and $\mathcal{X}_n = \mathcal{X}_{1n} \cap \mathcal{X}_{2n}$ we have

$$\begin{aligned}
n\mathbb{E} \sup_{s\in\mathbb{S}^m} |R_s^*|^2 &\leqslant \mathbb{E}(\sigma\varepsilon + \langle \beta - \beta^m, X\rangle_{\mathbb{H}})^2 [X]_{\underline{m}}[\Gamma]_{\underline{m}}^{-1}[X]_{\underline{m}}(1 - \mathbb{1}_{\{\varepsilon\in\mathcal{E}_n, X\in\mathcal{X}_n\}}) \\
&\leqslant \mathbb{E}(\sigma\varepsilon + \langle \beta - \beta^m, X\rangle_{\mathbb{H}})^2 [X]_{\underline{m}}[\Gamma]_{\underline{m}}^{-1}[X]_{\underline{m}}\, \mathbb{1}_{\{X\notin\mathcal{X}_{2,n}\}} \\
&\qquad + mn^{1/3}\mathbb{E}(\sigma\varepsilon + \langle \beta - \beta^m, X\rangle_{\mathbb{H}})^2(\mathbb{1}_{\{\varepsilon\notin\mathcal{E}_n\}} + \mathbb{1}_{\{X\notin\mathcal{X}_{1n}\}}).
\end{aligned}$$

Taking into account that $\mathbb{E}(\sigma\varepsilon + \langle \beta - \beta^m, X\rangle_{\mathbb{H}})^4 \leqslant (\sigma^2 + \|\Gamma^{1/2}(\beta-\beta^m)\|_{\mathbb{H}}^2)^2\eta^4$, $\mathbb{E}\varepsilon^2 = 1$ and $\mathbb{E}|\langle \beta - \beta^m, X\rangle_{\mathbb{H}}|^2 = \|\Gamma^{1/2}(\beta-\beta^m)\|_{\mathbb{H}}^2$ from the independence between ε and X follows

$$\begin{aligned}
n\mathbb{E} \sup_{s\in\mathbb{S}^m} |R_s^*|^2 &\leqslant (\sigma^2 + \|\Gamma^{1/2}(\beta-\beta^m)\|_{\mathbb{H}}^2)\eta^2 \Big(\mathbb{E}|[X]_{\underline{m}}[\Gamma]_{\underline{m}}^{-1}[X]_{\underline{m}}|^2\, \mathbb{1}_{\{X\notin\mathcal{X}_{2,n}\}}\Big)^{1/2} \\
&\qquad + mn^{1/3}\Big\{\sigma^2\mathbb{E}\varepsilon^2\, \mathbb{1}_{\{\varepsilon\notin\mathcal{E}_n\}} + \|\Gamma^{1/2}(\beta-\beta^m)\|_{\mathbb{H}}^2 P(\varepsilon\notin\mathcal{E}_n) \\
&\qquad\qquad + \sigma^2 P(X\notin\mathcal{X}_{1n}) + \mathbb{E}|\langle \beta - \beta^m, X\rangle_{\mathbb{H}}|^2\, \mathbb{1}_{\{X\notin\mathcal{X}_{1n}\}}\Big\}.
\end{aligned}$$

We exploit now the estimates given in (D.1) and (D.2). Thereby, we obtain

$$n\mathbb{E} \sup_{s\in\mathbb{S}^m} |R_s^*|^2 \leqslant C(\sigma^2 + \|\Gamma^{1/2}(\beta-\beta^m)\|_{\mathbb{H}}^2)\eta^{32} m n^{-7/3} \leqslant C\varsigma_m^2 \eta^{32} n^{-2}$$

where we used that $m \leqslant \lfloor n^{1/4} \rfloor$. Keeping in mind the decomposition (D.7) the last bound and (D.11) imply together the claim of Lemma D.3 which completes the proof. □

LEMMA D.4. *There exists a constant $K := K(\sigma, \eta, \mathcal{F}_b^r, \mathcal{G}_\gamma^d)$ depending on σ, η and the classes $\mathcal{F}_b^r$ and $\mathcal{G}_\gamma^d$ only such that for all $n \geqslant 1$ we have*

$$\sup_{\beta\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d} \sum_{m=m_n^\diamond}^{M_n^+} \Delta_m^\Gamma \mathbb{E}\Big(\|[\Gamma]_{\underline{m}}^{-1/2}[W_n]_{\underline{m}}\|^2 - 12\sigma_m^2 \frac{m\Lambda_m^\Gamma}{n}\Big)_+ \leqslant K\,\eta^{32}\,(\sigma^2 + r)\,\Sigma\, n^{-1}.$$

PROOF. We begin our proof with the observation that there exists an integer $n_o := n_o(\sigma, \eta, \mathcal{F}_b^r, \mathcal{G}_\gamma^d)$ depending on σ, η and the classes $\mathcal{F}_b^r$ and $\mathcal{G}_\gamma^d$ only such that for all $n \geqslant n_o$ and for all $m \geqslant m_n^\diamond$ we have $\varsigma_m^2 \leqslant 2(\sigma^2 + \|\Gamma^{1/2}\beta\|_{\mathbb{H}}^2 + [g]_{\underline{m}}^t[\Gamma]_{\underline{m}}^{-1}[g]_{\underline{m}}) = 2(\sigma_Y^2 + [g]_{\underline{m}}^t[\Gamma]_{\underline{m}}^{-1}[g]_{\underline{m}}) = \sigma_m^2$. Indeed, we have $1/m_n^\diamond = o(1)$ as $n\to\infty$ and $|\varsigma_m^2 - \sigma^2| = o(1)$ as $m\to\infty$ because $\varsigma_m = \sigma + \eta^2\|\Gamma^{1/2}(\beta^m - \beta)\|_{\mathbb{H}}$ and $\|\Gamma^{1/2}(\beta^m-\beta)\|_{\mathbb{H}}^2 \leqslant 34\, d^9\, r\, \gamma_m b_m^{-1}$ due to (B.5) in Lemma B.1. We distinguish in the following the cases $n < n_o$ and $n \geqslant n_o$. First, consider $n < n_o$. Due to (D.3) in Lemma D.2 and $\rho_m^2 \leqslant 2(\sigma^2 + 35d^6 r)$ (Lemma B.2 (iv)) we have for all $m \geqslant 1$

$$\mathbb{E}\Big(\|[\Gamma]_{\underline{m}}^{-1/2}[W_n]_{\underline{m}}\|^2 - 12\sigma_m^2\frac{m\Lambda_m^\Gamma}{n}\Big)_+ \leqslant \mathbb{E}\|[\Gamma]_{\underline{m}}^{-1/2}[W_n]_{\underline{m}}\|^2 \leqslant C\frac{m}{n}\eta^4(\sigma^2 + d^6 r).$$

Hence, $M_n^+ \leqslant \lfloor n^{1/4}\rfloor$ and $m\Delta_m^\Gamma \leqslant \delta_{M_n^+}^\Gamma \leqslant nC(d)$ for all $1 \leqslant m \leqslant M_n^+$ (Lemma B.2 (ii)) imply

$$\sup_{\beta\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d} \sum_{m=m_n^\diamond}^{M_n^+} \Delta_m^\Gamma \mathbb{E}\Big(\|[\Gamma]_{\underline{m}}^{-1/2}[W_n]_{\underline{m}}\|^2 - 12\sigma_m^2 \frac{m\Lambda_m^\Gamma}{n}\Big)_+ \leqslant n^{-1}C(d)\, n_o^{5/4}\,\eta^4(\sigma^2 + r).$$

The last bound implies the assertion of the lemma for all $1 \leqslant n < n_o$ because n_o depends on σ, η and the classes $\mathcal{F}_b^r$ and $\mathcal{G}_\gamma^d$ only. Consider now $n \geqslant n_o$ where we have $\varsigma_m^2 \leqslant \sigma_m^2$ for all $m \geqslant m_n^\diamond$. Thereby, we can apply Lemma D.3, which gives

$$\sup_{\beta\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\sum_{m=m_n^\diamond}^{M_n^+}\Delta_m^\Gamma\mathbb{E}\Big(\|[\Gamma]_{\underline{m}}^{-1/2}[W_n]_{\underline{m}}\|^2-12\sigma_m^2\frac{m\Lambda_m^\Gamma}{n}\Big)_+$$
$$\leqslant C\sup_{\beta\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\sum_{m=m_n^\diamond}^{M_n^+}\frac{\varsigma_m^2\Delta_m^\Gamma}{n}\Big\{\exp\Big(-\frac{m\Lambda_m^\Gamma}{6}\Big)+\exp\Big(-\frac{n^{1/6}}{100}\Big)+\frac{\eta^{32}}{n^2}\Big\}.$$

Taking into account the estimates $\Delta_k^\Gamma \leqslant 4d^3\Delta_k^\gamma$, $\Lambda_k^\Gamma \geqslant (1+\log d)^{-1}\Lambda_k^\gamma$, $M_n^+\Delta_{M_n^+}^\Gamma \leqslant \delta_{M_n^+}^\Gamma \leqslant nCd^6(1+\log d)$ and $\varsigma_k^2 \leqslant \sigma_k^2 \leqslant 2(\sigma^2+35d^6r)$ (Lemma B.2 (i), (ii) and (iv) respectively) follows

$$\sup_{\beta\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\sum_{m=m_n^\diamond}^{M_n^+}\Delta_m^\Gamma\mathbb{E}\Big(\|[\Gamma]_{\underline{m}}^{-1/2}[W_n]_{\underline{m}}\|^2-12\sigma_m^2\frac{m\Lambda_m^\Gamma}{n}\Big)_+\leqslant C(d)(\sigma^2+r)n^{-1}$$
$$\times\sup_{\beta\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\left\{\sum_{m=m_n^\diamond}^{M_n^+}\Delta_m^\gamma\exp\Big(-\frac{m\Lambda_m^\gamma}{6(1+\log d)}\Big)+n\exp\big(-\frac{n^{1/6}}{100}\big)+\frac{\eta^{32}}{n}\right\}.$$

Finally, exploit that $\Sigma = \Sigma(\mathcal{G}_\gamma^d)$ satisfies (3.3) and $n\exp\big(-n^{1/6}/100\big) \leqslant C$ which in turn implies the claim of the lemma for all $n \geqslant n_o$, i.e.,

$$\sup_{\beta\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\sum_{m=m_n^\diamond}^{M_n^+}\Delta_m^\Gamma\mathbb{E}\Big(\|[\Gamma]_{\underline{m}}^{-1/2}[W_n]_{\underline{m}}\|^2-12\sigma_m^2\frac{m\Lambda_m^\Gamma}{n}\Big)_+\leqslant C(d)\eta^{32}(\sigma^2+r)\,\Sigma\,n^{-1}.$$

Combining the cases $n < n_o$ and $n \geqslant n_o$ completes the proof. □

Lemma D.5. *There exist a numerical constant C and a constant $C(d)$ only depending on d such that for all $n \geqslant 1$ we have*

$$\sup_{\beta\in\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\Big\{n^6(M_n^+)^2\max_{1\leqslant m\leqslant M_n^+}P\big(\mho_{m,n}^c\big)\Big\}\leqslant C\eta^{64};$$
$$\sup_{\beta\in\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\Big\{n\,M_n^+\max_{1\leqslant m\leqslant M_n^+}P\big(\Omega_{m,n}^c\big)\Big\}\leqslant C(d)\eta^{64};$$
$$\sup_{\beta\in\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\big\{n^7P(\mathcal{E}_n^c)\big\}\leqslant C\eta^{64}.$$

Proof of Lemma D.5. By employing Lemma D.2 rather than Lemma C.2 the proof of the lemma follows along the lines of the proof of Lemma C.5, and we omit the details. □

Proposition D.6. *Let $\kappa = 288$ in the definition of the penalty pen given in* (2.11). *There exists a constant $K := K(\sigma,\eta,\mathcal{F}_b^r,\mathcal{G}_\gamma^d)$ depending on σ, η and the classes $\mathcal{F}_b^r$ and $\mathcal{G}_\gamma^d$ only such that for all $n \geqslant 1$ we have*

$$\sup_{\beta\in\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\mathbb{E}\Big\{\sup_{m_n^\diamond\leqslant k\leqslant M_n^+}\Big(\|\widehat{\beta}_k-\beta^k\|_\omega^2-\frac{1}{6}\,pen_k\Big)_+\Big\}\leqslant K\,\eta^{64}\,(\sigma^2+r)\,\Sigma\,n^{-1}.$$

Proof of Proposition D.6. We follow line by line the proof of Proposition C.6 . Keeping in mind that $pen_k = 288\sigma_k^2 k\Delta_k^\Gamma \Lambda_k^\Gamma n^{-1}$ we obtain

$$\mathbb{E}\Big\{\sup_{m_n^\diamond \leqslant k \leqslant M_n^+}\Big(\|\widehat{\beta}_k - \beta^k\|_\omega^2 - \frac{1}{6}\,pen_k\Big)_+\Big\} \leqslant 4\sum_{k=m_n^\diamond}^{M_n^+}\Delta_k^\Gamma \mathbb{E}\Big(\|[\Gamma]_{\underline{k}}^{-1/2}[W]_{\underline{k}}\|^2 - 12\sigma_k^2\frac{k\Lambda_k^\Gamma}{n}\Big)_+$$
$$+ \sum_{k=m_n^\diamond}^{M_n^+} n^3\big(\mathbb{E}\|[W]_{\underline{k}}\|^4\big)^{1/2}\big(P(\mho_{k,n}^c)\big)^{1/2} + \sum_{k=m_n^\diamond}^{M_n^+}\|\beta^k\|_\omega^2 P(\Omega_{k,n}^c).$$

The second and third right hand side term we bound due to Lemma D.2 and D.5, i.e.,

$$\sup_{\beta\in\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\mathbb{E}\Big\{\sup_{m_n^\diamond \leqslant k \leqslant M_n^+}\Big(\|\widehat{\beta}_k - \beta^k\|_\omega^2 - \frac{1}{6}\,pen_k\Big)_+\Big\} \leqslant C(d)\eta^{64}\,(\sigma^2+r)\Sigma\, n^{-1}$$
$$+ 4\sup_{\beta\in\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\sum_{k=m_n^\diamond}^{M_n^+}\Delta_k^\Gamma \mathbb{E}\Big(\|[\Gamma]_{\underline{k}}^{-1/2}[W]_{\underline{k}}\|^2 - 12\sigma_k^2\frac{k\Lambda_k^\Gamma}{n}\Big)_+,$$

and hence by employing the bound given in Lemma D.4 we complete the proof. □

Proposition D.7. *Let $\kappa = 288$ in the definition of pen and $\widehat{pen}$ given in* (2.11) *and* (2.6) *respectively. There exists a constant $C(d)$ such that for all $n \geqslant 1$ we have*

$$\sup_{\beta\in\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\mathbb{E}(\|\widehat{\beta}_{\widehat{m}} - \beta\|_\omega^2\,\mathbb{1}_{\mathcal{E}_n^c}) \leqslant C(d)\,\eta^{64}\,(\sigma^2+r)\,\Sigma\,n^{-1}.$$

Proof of Proposition D.7. Taking into account (D.3) in Lemma D.2 rather than (C.7) in Lemma C.2 we follow line by line the proof of Proposition C.7 and conclude that

$$\sup_{\beta\in\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}\mathbb{E}(\|\widehat{\beta}_{\widehat{m}} - \beta\|_\omega^2\,\mathbb{1}_{\mathcal{E}^c}) \leqslant C(d)(\sigma^2+r)\eta^8\Sigma n^{5/2}\sup_{\beta\in\mathcal{F}_b^r}\sup_{\Gamma\in\mathcal{G}_\gamma^d}|P(\mathcal{E}_n^c)|^{1/2}.$$

The assertion follows now with help of Lemma D.5, which completes the proof. □

Proof of Proposition 3.3. The assertion follows from Proposition D.6 and Proposition D.7 and we omit the details. □

References

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3):301–413, 1999.

D. Bosq. *Linear Processes in Function Spaces.*, volume 149 of *Lecture Notes in Statistics.* Springer-Verlag, 2000.

C. Butucea and A. B. Tsybakov. Sharp optimality in density deconvolution with dominating bias. I. *Teor. Veroyatn. Primen.*, 52(1):111–128, 2007a.

C. Butucea and A. B. Tsybakov. Sharp optimality in density deconvolution with dominating bias. II. *Teor. Veroyatn. Primen.*, 52(2):336–349, 2007b.

H. Cardot and J. Johannes. Thresholding projection estimators in functional linear models. *Journal of Multivariate Analysis*, 101(2):395–408, 2010.

H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statistics & Probability Letters*, 45(11-22), 1999.

H. Cardot, F. Ferraty, and P. Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, 13:571–591, 2003.

H. Cardot, A. Mas, and P. Sarda. CLT in functional linear regression models. *Probability Theory and Related Fields*, 138(3-4):325–361, 2007.

F. Comte and J. Johannes. Adaptive estimation in circular functional linear models. *Mathematical Methods of Statistics*, 19:42–63, 2010.

F. Comte, Y. Rozenholc, and M.-L. Taupin. Penalized contrast estimator for density deconvolution. *The Canadian Journal of Statistics*, 37(3):431–452, 2006.

C. Crambes, A. Kneip, and P. Sarda. Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, 37(1):35–72, 2009.

R. Dahlhaus and W. Polonik. Nonparametric quasi-maximum likelihood estimation for Gaussian locally stationary processes. *The Annals of Statistics*, 34(6):2790–2824, 2006.

S. Efromovich and V. Koltchinskii. On inverse problems with unknown operators. *IEEE Transactions on Information Theory*, 47(7):2876–2894, 2001.

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems.* Kluwer Academic, Dordrecht, 2000.

F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Methods, Theory, Applications and Implementations.* Springer-Verlag, London, 2006.

M. Forni and L. Reichlin. Let's get real: A factor analytical approach to disaggregated business cycle dynamics. *Review of Economic Studies*, 65:453–473, 1998.

I. Frank and J. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–148, 1993.

A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39:1608–1632, 2011.

A. Goldenshluger and A. Tsybakov. Adaptive prediction and estimation in linear regression with infinitely many parameters. *The Annals of Statistics*, 29(6):1601–1619, 2001.

A. Goldenshluger and A. Tsybakov. Optimal prediction for linear regression with infinitely many parameters. *Journal of Multivariate Analysis*, 84(1):40–60, 2003.

P. Hall and J. L. Horowitz. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91, 2007.

E. Heinz. Beiträge zur störungstheorie der spektralzerlegung. *Mathematische Annalen*, 123: 415–438, 1951.

M. Hoffmann and M. Reiß. Nonlinear estimation for linear inverse problems with error in the operator. *The Annals of Statistics*, 36(1):310–336, 2008.

J. Johannes and R. Schenk. On rate optimal local estimation in functional linear model. http://arxiv.org/abs/0902.0645, Université catholique de Louvain, 2010.

T. Kawata. *Fourier analysis in probability theory.* Academic Press, New York, 1972.

T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

O. V. Lepski. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability and its Applications*, 35:454–466, 1990.

B. D. Marx and P. H. Eilers. Generalized linear regression on sampled signals and curves : a p-spline approach. *Technometrics*, 41:1–13, 1999.

P. Massart. *Concentration inequalities and model selection*, volume 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.

P. Mathé. The Lepskiĭ principle revisited. *Inverse Problems*, 22(3):11–15, 2006.

H.-G. Müller and U. Stadtmüller. Generalized functional linear models. *The Annals of Statistics*, 33:774–805, 2005.

F. Natterer. Error bounds for Tikhonov regularization in Hilbert scales. *Applicable Analysis*, 18:29–37, 1984.

A. Neubauer. When do Sobolev spaces form a Hilbert scale? *Proceedings of the American Mathematical Society*, 103(2):557–562, 1988.

V. V. Petrov. *Limit theorems of probability theory. Sequences of independent random variables.* Oxford Studies in Probability. Clarendon Press., Oxford, 4. edition, 1995.

C. Preda and G. Saporta. Pls regression on a stochastic process. *Computational Statistics & Data Analysis*, 48:149 –158, 2005.

J. Ramsay and B. Silverman. *Functional Data Analysis.* Springer, New York, second ed. edition, 2005.

M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126(3):505–563, 1996.

N. Verzelen. High-dimensional Gaussian model selection on a Gaussian design. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 46(2):480–524, 2010.