Japanese Journal of Statistics and Data Science
https://doi.org/10.1007/s42081-025-00309-x

ORIGINAL PAPER

PAC-Bayes estimation for high-dimensional multi-index models with unknown active dimension

Maximilian F. Steffen[1]

Received: 22 March 2023 / Revised: 9 May 2025 / Accepted: 13 June 2025

Abstract

The multi-index model with sparse dimension reduction matrix is a popular approach to circumvent the curse of dimensionality in a high-dimensional regression setting. Building on the single-index analysis by Alquier, P. & Biau, G. (Journal of Machine Learning Research 14 (2013), 243–280), we develop a PAC-Bayesian estimation method for a possibly miss-specified multi-index model with unknown active dimension and an orthogonal dimension reduction matrix. Our main result is a non-asymptotic oracle inequality, which shows that the estimation method adapts to the active dimension of the model, the sparsity of the dimension reduction matrix and the regularity of the link function. Under a Sobolev regularity assumption on the link function the estimator achieves the minimax rate of convergence (up to a logarithmic factor) and no additional price is paid for the unknown active dimension. The method is illustrated with simulation examples.

Keywords Multi-index model · PAC-Bayes · Adaptive nonparametric estimation · Sparsity · Oracle inequality · Dimension reduction

Mathematics Subject Classification 62G08 · 62F15

1 Introduction

A standard task in supervised learning is to estimate or learn, respectively, the conditional expectation of a label $Y \in \mathbb{R}$ given a large vector $\mathrm{X} \in \mathbb{R}^p$ of explanatory random variables based on i.i.d. data $\mathcal{D}_n := (\mathrm{X}_i, Y_i)_{i=1,\dots,n}$ which are distributed as (X, Y). The corresponding nonparametric regression model reads as

✉ Maximilian F. Steffen
maximilian.steffen@kit.edu

[1] Department of Mathematics, Karlsruhe Institute of Technology, Karlsruhe, Germany

Published online: 02 September 2025

$$Y = f(\mathrm{X}) + \varepsilon \tag{1}$$

with an observation error ε satisfying $\mathbb{E}[\varepsilon|\mathrm{X}] = 0$ a.s. and the unknown regression function $f\colon \mathbb{R}^p \to \mathbb{R}$ given by $f = \mathbb{E}[Y|\mathrm{X} = \cdot]$. If the dimension p is large, the estimation problem suffers from the well-known curse of dimensionality. This is of particular importance in numerous recent applications where p may even exceed the sample size n.

A popular approach to reduce the effective dimension of the model is to impose a multi-index structure on f (Li, 1991). While we do not assume that the observations exactly follow a multi-index model, our method builds upon an approximation of the regression function of the form

$$f(\mathrm{x}) \approx g^*(W^*\mathrm{x}), \qquad \forall \mathrm{x} \in \mathbb{R}^p, \tag{2}$$

for some *active dimension* $d^* \ll p$, a sparse *dimension reduction matrix* $W^* \in \mathbb{R}^{d^* \times p}$ and a (measurable) *link function* $g^*\colon \mathbb{R}^{d^*} \to \mathbb{R}$. Under this model, the problem of estimating the high-dimensional regression function f boils down to estimating a lower-dimensional function g^* and a sparse $d^* \times p$ matrix. The sparsity assumption implies most of the covariates, approximately do not have an influence on the label. If we knew, which covariates are irrelevant, we could discard them entirely, further reducing the model complexity. However, even without this knowledge, we can benefit from a sparse dimension reduction matrix by constructing an estimator which adapts to this sparsity.

Note that the class of multi-index models covers various commonly studied regression models from the literature, see e.g. Hastie et al. (2009) for an overview. In particular, for $\mathrm{x} = (x_1, \dots, x_p)^\top, w, w_1, \dots, w_p \in \mathbb{R}^p, \mathrm{z} = (z_1, \dots, z_d)^\top \in \mathbb{R}^d$, $g, g_1, \dots, g_p\colon \mathbb{R} \to \mathbb{R}$ and $d \leqslant p$, we recover the following models:

(a) Linear regression: $f(\mathrm{x}) = w^\top \mathrm{x}$, which can be seen as a multi-index model with $d^* = 1$ and the identity as link function.

(b) Additive model: $f(\mathrm{x}) = \sum_{i=1}^{p} g_i(x_i)$, i.e. a multi-index model with $W = I_{p\times p}$ and $g(\mathrm{x}) = \sum_{i=1}^{d} g_i(x_i)$.

(c) Single-index model: $f(\mathrm{x}) = g(w^\top \mathrm{x})$, the special case of the multi-index model with $d^* = 1$.

(d) Projection pursuit regression: $f(\mathrm{x}) = \sum_{i=1}^{d} g_i(w_i^\top \mathrm{x})$, which is a multi-index model with $d^* = d$, dimension reduction matrix $W = (w_1, \dots, w_d)^\top$ and link function $g(\mathrm{z}) = \sum_{i=1}^{d} g_i(z_i)$.

In the high-dimensional linear regression, the curse of dimensionality can also be circumvented assuming sparsity of w by using the celebrated LASSO, see van de Geer et al. (2011). Alquier and Biau (2013) study the single-index model under a sparsity assumption on w.

Following the aforementioned Li (1991), the estimation of the space spanned by the rows of W^* has been studied extensively in the literature, see e.g. Hristache et al. (2001), Xia (2007) and Dalalyan et al. (2008), but under the assumption of a known active dimension d^*. While some research has been done on the estimation of d^* itself, see Xia et al. (2002) and Zhu et al. (2006), the estimation of the overall model has relied on estimating W^* and g^* separately to then analyze the propagation error, see Klock et al. (2021). The analysis of high-dimensional multi-index models, where $p \gg n$, is rather limited.

A common assumption in the theory of multi-index models is that the dimension reduction matrix is (semi-)orthogonal, i.e. $W^*(W^*)^\top = I_{d^* \times d^*}$ is the identity matrix. Indeed, this allows for the interpretation of $W^*\mathrm{X}$ as a rotation of the covariates, projected onto the first d^* coordinates followed by another rotation.

We use a PAC-Bayesian estimation approach based on the Gibbs-posterior (see Guedj (2019) and Alquier (2024) for an overview) which was originally developed by Catoni (2004, 2007) and has been adapted to the single-index model (i.e. $d^* = 1$) without miss-specification by Alquier and Biau (2013). Other applications of the Gibbs-posterior to nonparametric regression include additive models, see Guedj and Alquier (2013), and neural networks, see Bieringer et al. (2025) and Steffen and Trabs (2025).

In this paper we generalize the PAC-Bayes method for single index models to the more flexible class of multi-index models. In particular, we aim for a method which adapts to the unknown active dimension d^*, the sparsity of W^* and the regularity of g^* for a good approximation (2) based on the given data.

The PAC-Bayes approach relies on the following principle: With a prior Π for the parameters (W, g) we consider the Gibbs-posterior probability distribution $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$ whose Π-density is (up to normalization) given by

$$\frac{\mathrm{d}\Pi_\lambda(W, g \mid \mathcal{D}_n)}{\mathrm{d}\Pi} \propto \exp(-\lambda R_n(W, g)) \tag{3}$$

with a tuning parameter $\lambda > 0$ and empirical prediction risk

$$R_n(W^*, g^*) = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - g^*(W^*\mathrm{X}_i)\big)^2.$$

The estimator for f is obtained by simulating a random variable

$$(\widehat{W}_\lambda, \widehat{g}_\lambda) \sim \Pi_\lambda(\cdot \mid \mathcal{D}_n) \tag{4}$$

and setting $\widehat{f} := \widehat{g}_\lambda(\widehat{W}_\lambda \cdot)$. Other estimators based on the Gibbs-posterior are the posterior mean $\mathbb{E}[\widehat{f} \mid \mathcal{D}_n]$ and the maximum a posteriori (MAP) estimator. We focus on the estimator from (4) for clarity, but our results can easily be extended to the posterior mean.

While (3) coincides with the classical Bayesian posterior distribution only if $Y_i = g(W\mathrm{X}_i) + \varepsilon_i$ with i.i.d. $\varepsilon_i \sim \mathrm{N}(0, n/(2\lambda))$, the estimator $\widehat{f} := \widehat{g}_\lambda(\widehat{W}_\lambda \cdot)$ will achieve a small prediction error under quite mild model assumptions.

We choose a hierarchical prior that prefers models with a low active dimension, sparse dimension reduction matrices and regular link functions. Let Π be supported on $\bigcup_{d=1}^{p} \mathcal{S}_d \times \mathcal{G}_d$ for some classes $\mathcal{S}_d$ and $\mathcal{G}_d$ for W and g, respectively. For $\mathcal{S}_d$ we will study a class of sparse matrices while $\mathcal{G}_d$ will be given by finite wavelet approximations. The prior is uniform for a given sparsity and a wavelet projection level. The posterior weighs each pair of parameters (W, g) based on its empirical performance (with respect to the empirical loss function) on the data, where the tuning parameter λ controls the impact of $R_n(W, g)$.

We quantify the accuracy of the estimation procedure in terms of the excess risk

$$\mathcal{E}(W,g) := R(W,g) - \mathbb{E}[(Y - f(\mathrm{X}))^2] = \mathbb{E}[(g(W\mathrm{X}) - f(\mathrm{X}))^2], \tag{5}$$

where

$$R(W,g) := \mathbb{E}\big[(Y - g(W\mathrm{X}))^2\big] \tag{6}$$

is the prediction risk.

The paper is organized as follows: We explain our estimation method in Sect. 2 and state our main results in Sect. 3. In Sect. 4, we demonstrate the performance of our estimation method with simulation examples. The proofs have been postponed to Sect. 6.

Throughout, we denote the ℓ^q-norm of a vector $\mathrm{x} \in \mathbb{R}^p$ by $|\mathrm{x}|_q$ for $q \in [1, \infty]$ and, in particular, the Euclidean norm by $|\mathrm{x}| = |\mathrm{x}|_2$. Further, we set $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$ for $a, b \in \mathbb{R}$.

2 Construction of the prior

To construct the prior, we introduce for any dimension $d = 1, \dots, p$ classes $\mathcal{S}_d$ and $\mathcal{G}_d$ together with priors μ_d and ν_d for the dimension reduction matrix W and the link function g, respectively. Based on that we can then define the prior Π on $\bigcup_{d=1}^{p} \mathcal{S}_d \times \mathcal{G}_d$.

We start with a fixed active dimension $d \in \{1, \dots, p\}$. While it is common in the literature to assume that the "true" dimension reduction matrix W^* is (semi-)orthogonal, we do not need to impose this restriction on our estimation method. Instead, we simply generate an estimator with ℓ^2-standardized rows, i.e. for $W = (w_1, \dots, w_d)^\top \in \mathbb{R}^{d \times p}$ with row vectors $w_i = (w_{i,1}, \dots, w_{i,p}) \in \mathbb{R}^p$ we impose $|w_i| = 1$. To encode sparsity, let

$$\mathcal{I}_d := \big\{ I \,\big|\, \emptyset \neq I := I_1 \times \cdots \times I_d,\, I_1, \dots, I_d \subseteq \{1, \dots, p\} \big\}$$

contain all potential sets of *active coordinates,* that is I_i describes the active coordinates in the i-th argument of the link function. For $I = I_1 \times \cdots \times I_d \in \mathcal{I}_d$ the number of active coordinates is $\|I\| := \sum_{i=1}^{d} |I_i|$, where $|I_i|$ denotes the cardinality of I_i. Note that $\emptyset \neq I = I_1 \times \cdots \times I_d$ already implies $I_1, \dots, I_d \neq \emptyset$. The parameter set $\mathcal{S}_d(I)$ of sparse dimension reduction matrices is given by

$$\mathcal{S}_d(I) := \big\{W = (w_1, \dots, w_d)^\top \in \mathbb{R}^{d\times p} \mid w_i \in \mathcal{S}(I_i),\, i = 1, \dots, d\big\}, \qquad \text{where}$$
$$\mathcal{S}(I_i) := \big\{w_i = (w_{i,1}, \dots, w_{i,p}) \in \mathbb{R}^p \mid |w_i| = 1, \forall j \notin I_i : w_{i,j} = 0\big\}.$$

Finally, we define $\mathcal{S}_d = \bigcup_{I\in\mathcal{I}_d} \mathcal{S}_d(I)$. Note that $\mathcal{S}_d(I) \supseteq \widetilde{S}_d(I)$ for

$$\widetilde{S}_d(I) := \{W = (w_1, \dots, w_d)^\top \in \mathbb{R}^{d\times p} \mid |w_1| = \dots = |w_d| = 1, w_{i,j} \neq 0 \text{ iff } j \in I_i,\\ j = 1, \dots, p\}.$$

In $\widetilde{S}_d(I)$ the index set *I* exactly describes the sparsity of *W*. However, we consider the prior on the compact set $\mathcal{S}_d(I)$ to ensure the existence of solutions to minimization problems over $\mathcal{S}_d(I)$, which becomes relevant in Sect. 3.

To construct a prior measure μ_d on $\mathcal{S}_d$, we use the uniform distribution on the set of dimension reduction matrices with a given active dimension *d* and with sparsity $i = \|I\|$. These uniform distributions are then weighted geometrically such that sparse dimension reduction matrices are preferred by the prior. Denoting the uniform distribution on $\mathcal{S}_d(I)$ by $\mu_{d,I}$, the prior measure on $\mathcal{S}_d$ is thus given by the mixture

$$\mu_d := \sum_{i=d}^{dp} 2^{-i+d-1} \frac{1}{|\mathcal{I}_{d,i}|} \sum_{I\in\mathcal{I}_{d,i}} \mu_{d,I} \Big/ (1 - 2^{(1-p)d-1}) \qquad \text{where} \qquad \mathcal{I}_{d,i} := \{I \in \mathcal{I}_d \mid \|I\| = i\}.$$

Here and in the following two analogous constructions, the basis 2 of the geometrically decreasing weights can be replaced by an arbitrary fixed $a > 1$. The theoretical results remain unchanged up to constants.

To define a class $\mathcal{G}_d$ and a prior ν_d for the link function, we use a multivariate tensor product wavelet basis on $\mathbb{R}^d$, see e.g. Daubechies (1992); Giné and Nickl (2016). Let φ and ψ be a continuously differentiable scaling and wavelet function on $\mathbb{R}$, respectively, and write $\psi_0 := \varphi$, $\psi_1 := \psi$. We use compactly supported regular Daubechies wavelets. For $M \in \mathbb{N}_0, N \in \mathbb{N}$ we define the index set

$$\mathcal{Z}^d_{M,N} :=\{l = (0, l_2, 0) \mid l_2 \in \mathbb{Z}^d,\, |l_2|_\infty \leqslant N\}\\ \cup \big\{l = (l_1, l_2, l_3) \in \mathbb{N}_0 \times \mathbb{Z}^d \times \{0,1\}^d \mid l_1 \leqslant M, |l_2|_\infty \leqslant 2^{l_1}N, l_3 \neq 0\big\},$$

where l_1 is the approximation level, l_2 is a shift parameter and l_3 is due to the tensor structure. The system $(\Psi_l)_{l\in\mathcal{Z}^d_{\infty,\infty}}$ with

$$\Psi_l(\mathrm{x}) := 2^{l_1 d/2} \prod_{i=1}^d \psi_{l_{3,i}}(2^{l_1} x_i - l_{2,i}), \qquad \mathrm{x} \in \mathbb{R}^d, l = (l_1, l_2, l_3) \in \mathcal{Z}^d_{\infty,\infty},$$

is an orthonormal basis of $L^2(\mathbb{R}^d)$. In particular, each $g \in L^2(\mathbb{R}^d)$ admits a wavelet series representation $g = \sum_{l\in\mathcal{Z}^d_{\infty,\infty}} \langle g, \Psi_l\rangle \Psi_l$. Throughout, we fix a sufficiently large constant $N \in \mathbb{N}$ and abbreviate $\mathcal{Z}^d_M := \mathcal{Z}^d_{M,N}$. For $\xi > 0$ we define the compact wavelet coefficient ball

$$\begin{aligned}\mathcal{B}_{d,M}(\xi) &:= \{\beta \in \mathbb{R}^{\mathcal{Z}_M^d} \,|\, \|\beta\|_{\mathcal{B}} \leqslant \xi\}, &&\text{where}\\ \|\beta\|_{\mathcal{B}} &:= L^d \sum_{l \in \mathcal{Z}_M^d} 2^{l_1(d/2+1)}|\beta_l|, &&\text{with} \qquad L := \|\psi\|_\infty \vee \|\varphi\|_\infty \vee \|\psi'\|_\infty \vee \|\varphi'\|_\infty \vee 1,\end{aligned} \tag{7}$$

which determines the finite dimensional approximation space

$$\mathcal{G}_{d,M}(\xi) := \{g = \Phi_{d,M}(\beta) \,\big|\, \beta \in \mathcal{B}_{d,M}(\xi)\} \qquad \text{via} \qquad \Phi_{d,M}(\beta) := \sum_{l \in \mathcal{Z}_M^d} \beta_l \Psi_l, \beta \in \mathbb{R}^{\mathcal{Z}_M^d}.$$

For any $g = \Phi_{d,M}(\beta)$ we write $\|g\|_{\mathcal{B}} := \|\beta\|_{\mathcal{B}}$ which corresponds to the Besov norm with regularity $1 + d$ and integrability parameter 1 on $\operatorname{span}\{\Psi_l : l \in \mathcal{Z}_M^d\}$. In particular, we have for any $g \in \mathcal{G}_{d,M}(\xi)$

$$\|g\|_\infty \leqslant \|g\|_{\mathcal{B}} \leqslant \xi \qquad \text{and} \qquad \|(\nabla g)_i\|_\infty \leqslant \|g\|_{\mathcal{B}} \leqslant \xi, \qquad \forall i \in \{1, \dots, d\}. \tag{8}$$

For $C > 0$ we set $\mathcal{G}_d := \bigcup_{M=0}^n \mathcal{G}_{d,M}(C + 1)$.

The prior ν_d on $\mathcal{G}_d$ is defined as a random coefficient prior with uniformly distributed coefficients on $\mathcal{G}_{d,M}(C + 1)$ and geometrically decreasing weights in the approximation level *M*. To this end, let $\widetilde{\nu}_{d,M}$ be the uniform distribution on $\mathcal{B}_{d,M}(C + 1)$ and let $\nu_{d,M} := \widetilde{\nu}_{d,M}(\Phi_{d,M}^{-1}(\cdot))$ denote the push-forward measure of $\widetilde{\nu}_{d,M}$ under $\Phi_{d,M}$. Then, we set

$$\nu_d := \sum_{M=0}^{n} 2^{-M} \nu_{d,M} \Big/ (2 - 2^{-n}).$$

We can now define the prior for a fixed active dimension *d* as the product measure $\pi_d := \mu_d \otimes \nu_d$. Finally, we mix over all possible active dimensions to account for the fact that d^* is unknown. Encoding a preference for simple models, i.e. small active dimensions, via weights 2^{-d}, the final prior on $\bigcup_{d=1}^p \mathcal{S}_d \times \mathcal{G}_d$ is given by

$$\Pi = \sum_{d=1}^{p} 2^{-d} \pi_d \Big/ (1 - 2^{-p}). \tag{9}$$

Note that the structure of the prior ensures that drawing from Π yields a link function and a dimension reduction matrix with matching active dimension.

3 Oracle inequality

For an active dimension $d \in \{1, \dots, p\}$, an active index set $I \in \mathcal{I}_d$ of the dimension reduction matrix and an approximation level $M \in \{0, \dots, n\}$ of the link function, we define an *oracle choice* on $\mathcal{S}_d(I) \times \mathcal{G}_{d,M}(C)$ as

$$(W^*_{d,I}, g^*_{d,M}) \in \operatorname*{arg\,min}_{(W,g)\in\mathcal{S}_d(I)\times\mathcal{G}_{d,M}(C)} R(W,g). \tag{10}$$

Note that the minimization in g is over $\mathcal{G}_{d,M}(C)$, whereas the prior is defined on $\mathcal{G}_{d,M}(C+1)$ which ensures that a small neighborhood of $g^*_{d,M}$ is contained in the support of the prior. A solution to the minimization problem in (10) always exists since we have equivalently

$$(W^*_{d,I}, \beta^*_{d,M}) \in \operatorname*{arg\,min}_{(W,\beta)\in\mathcal{S}_d(I)\times\mathcal{B}_{d,M}(C)} \mathbb{E}\big[\big(Y-\Phi_{d,M}(\beta)(W\mathrm{X})\big)^2\big]$$

with compact $\mathcal{S}_d(I)\times\mathcal{B}_{d,M}(C)$ and continuous $(W,\beta)\mapsto\mathbb{E}\big[\big(Y-\Phi_{d,M}(\beta)(W\mathrm{X})\big)^2\big]$. Our main result gives a theoretical guarantee that the estimator $(\widehat{W}_\lambda, \widehat{g}_\lambda)$ from (4) is almost as good as the best oracle $(W^*_{d,I}, g^*_{d,M})$ for all possible active dimensions in terms of the excess risk. To this end, we need some mild assumptions on the regression model.

Assumption A

(a) **Bounded regression function:** For some constant $C \geqslant 1$ we have $\|f\|_\infty \leqslant C$.
(b) **Bounded inputs:** For some constant $K \geqslant 1$ we have $|\mathrm{X}|_\infty \leqslant K$ a.s.
(c) **Conditional sub-Gaussianity of observation noise:** There are constants $\sigma, \Gamma > 0$ such that

$$\mathbb{E}[|\varepsilon|^k|\mathrm{X}] \leqslant \frac{k!}{2}\sigma^2\Gamma^{k-2} \text{ a.s.,} \qquad \forall k \geqslant 2.$$

We obtain the following non-asymptotic oracle inequality. It generalizes Alquier and Biau (2013, Theorem 2) not only with respect to the multi-index approach with unknown active dimension, but also with respect to some technical but practically relevant aspects such as the ℓ^2-normalization of W and the wavelet basis.

Theorem 1 *(PAC-Bayes oracle inequality) Under Assumption A there are constants $Q_0, Q_1 > 0$ depending only on $C, \Gamma, \sigma > 0$ such that for $\lambda = n/Q_0$ and sufficiently large n we have for all $\delta \in (0,1)$ with a probability of at least $1-\delta$ that*

$$\mathcal{E}(\widehat{W}_\lambda, \widehat{g}_\lambda) \leqslant \min_{d,I,M}\Big(3\mathcal{E}(W^*_{d,I}, g^*_{d,M}) + \frac{Q_1}{n}\big(\|I\|\log(p\vee n) + 16^d N^d 2^{dM}\log(n) + \log(2/\delta)\big)\Big),$$

where the minimum is taken over all triplets (*d*, *I*, *M*) with $d \in \{1,\ldots,p\}$, $I \in \mathcal{I}_d$ and $M \in \{0,\ldots,n\}$.

Remark 2 Here and in the following, the $1-\delta$ probability in takes into account the randomness of the data and of the estimate. An explicit admissible choice for λ is $\lambda = n/((2C+1)(\Gamma\vee(2C+1)) + 4((2C+1)^2+4\sigma^2))$. The dependence of Q_1 on C, Γ, σ is at most quadratic and $n \geqslant n_0 = 5\vee(C+1)\vee K$ is sufficiently large.

The right-hand side of the oracle inequality can be interpreted similarly to the classical bias-variance decomposition in non-parametric statistics. The first term

$$\mathcal{E}(W^*_{d,I}, g^*_{d,M}) = \mathbb{E}[(g^*_{d,M}(W^*_{d,I}\mathrm{X}) - f(\mathrm{X}))^2]$$

quantifies the approximation error while second term is an upper bound for the stochastic error. In particular, we recover $\|I\| \log(p \vee n)/n$ (or $\|I\| \log(p)/n$ if $p \geqslant n$) as the typical error term for estimating sparse matrices with sparsity $\|I\|$, see van de Geer et al. (2011), while $16^d N^d 2^{dM} \log(n)/n$ is due to the estimation of $\mathcal{O}\big(16^d N^d 2^{dM}\big)$ many wavelet coefficients each with (squared) accuracy $\log(n)/n$ paying a logarithmic price for adaptivity.

The minimum over all (*d*, *I*, *M*) in the upper bound shows that the estimator adapts to the active dimension, the sparsity of the dimension reduction matrix and the regularity of the link function. One can show the same result in a multi-index model with a known active dimension d^* by using π_{d^*} as a prior instead of Π. The only difference (up to a different constant Q_1) in the result is that the minimum in the upper bound is only taken over all pairs $(I, M) \in \mathcal{I}_{d^*} \times \{0, \ldots, n\}$. Consequently, no additional price is paid for not knowing the true active dimension of the model.

In the well-specified setting and under assumptions on the distribution of $W^*\mathrm{X}$ as well as a Besov-type regularity assumption on the link function, we derive explicit convergence rates from Theorem 1.

Assumption B

(a) **Multi-index model:** There exist $d^* \in \{1, \ldots, p\}$, $W^* \in \mathcal{S}^{d^*}$ and $g^* : \mathbb{R}^{d^*} \to \mathbb{R}$ such that $f = g^*(W^* \cdot)$.
(b) **Bounded dimension reduced inputs:** For $B_1 \geqslant 1$, we have $|W^*\mathrm{X}|_\infty \leqslant B_1$.
(c) **Lebesgue density of dimension reduced inputs:** $W^*\mathrm{X}$ has a Lebesgue density ϱ on $\mathbb{R}^{d^*}$ bounded by a constant $B_2 \geqslant 1$.

For the true dimension reduction matrix W^* we write $\|W^*\|_0 := \|I^*\|$ for the minimal (with respect to $\| \cdot \|$) $I^* \in \mathcal{I}_{d^*}$ such that $W^* \in \mathcal{S}_{d^*}(I^*)$. The regularity of g^* will be measured in terms of its Besov norm. We recover Sobolev balls for $q = 2$, cf. Giné and Nickl (2016, (4.164)).

Definition 3 The Besov ellipsoid in $\mathbb{R}^{d^*}$ with regularity $\alpha > 0$ and integrability parameter $q \in [0, \infty)$ is given by

$$B^\alpha_{q,d^*}(\xi) := \Big\{ g \in L^2(\mathbb{R}^{d^*}) \,\Big|\, \sum_{l \in \mathcal{Z}^{d^*}_{\infty,\infty}} 2^{q l_1 \alpha} |\langle g, \Psi_l \rangle|^q \leqslant \xi^q \Big\} \tag{11}$$

for a radius $\xi > 0$.

Corollary 4 *(Convergence rate) Let the assumptions of Theorem 1 be fulfilled in addition to Assumption B. Take $\lambda = n/Q_0$ with Q_0 from Theorem 1. Suppose that*

$g^* \in B^{\alpha}_{2,d^*}(\xi)$ with $\xi = C(L^{d^*} 2N^{d^*/2} 16^{d^*/2})^{-1}$ for some $\alpha > 2 + d^*$. Then, for sufficiently large n and with a probability of at least $1 - \delta$, we have

$$\mathcal{E}(\widehat{W}_\lambda, \widehat{g}_\lambda) \leqslant Q_2\Big(\frac{\log n}{n}\Big)^{\frac{2\alpha}{2\alpha+d^*}} + Q_2\Big(\frac{\|W^*\|_0 \log(p \vee n)}{n} + \frac{\log(2/\delta)}{n}\Big),$$

where Q_2 is a constant only depending on $C, \Gamma, \sigma, N, B_1, B_2$ and d^*.

Remark 5 If W^* is sparse (i.e. $\|W^*\|_0$ is small), then the dominating term in the upper bound of the excess risk of the PAC-Bayesian estimator is of order

$$\Big(\frac{\log n}{n}\Big)^{\frac{2\alpha}{2\alpha+d^*}},$$

which is the usual minimax-optimal rate (up to a logarithmic factor) for such estimation problems, see e.g. Tsybakov (2009). Note that if d^* is substantially smaller than p, then we have successfully circumvented the curse of dimensionality, since the dimension which appears in the rate is now only d^*. As an alternative to the wavelet construction, one can use the multivariate trigonometric system on $[-1, 1]^{d^*}$, assume $\mathrm{X} \in [-1, 1]^p$ and ℓ^1-standardized rows of W^* (which ensures $W^*\mathrm{X} \in [-1, 1]^{d^*}$) leading to a more direct generalization of Alquier and Biau (2013). In particular, an analogous statement to Corollary 4 holds for an estimator based on the trigonometric system in the setting of W^* having ℓ^1-standardized rows and $\mathrm{X} \in [-1, 1]^p$. However, the ℓ^2-standardization seems more natural and is in line with the literature.

4 Simulation examples

In this section, we demonstrate the performance of our estimation method with simulation examples. To this end, we need to sample from the Gibbs posterior distribution. Since its normalizing constants are inaccessible, a common way to achieve this is through the Metropolis-Hastings algorithm where a Markov chain $(W^{(k)}, g^{(k)})_{k\in\mathbb{N}_0}$ is constructed which admits the Gibbs posterior as its invariant distribution. While the hierarchical prior allows for the construction of a method with desirable theoretical properties as demonstrated in Sect. 3, the implementation of such adaptive methods presents its own set of challenges, especially for our method with a multilevel hierarchical prior. The typical approach in the literature is to use a particular variant of the Metropolis-Hastings algorithm, namely the reversible-jump Markov chain Monte Carlo (RJMCMC) algorithm originally introduced by Green (1995). In a regression setting, it has successfully been applied to additive models (Guedj & Alquier, 2013) and single-index models (Alquier & Biau, 2013). To obtain a numerically feasible method, we restrict ourselves to the case of an active dimension $d^* = 2$ in a projection pursuit type regression which generalizes the implementation in the latter reference.

The Markov chain $(W^{(k)}, g^{(k)})_{k\in\mathbb{N}_0}$ is constructed as follows:

(a) Randomly initialize $(W^{(0)}, g^{(0)})$.

(b) For $k = 0, \dots, N$ with some fixed $N \in \mathbb{N}$:

(i) Given $(W^{(k)}, g^{(k)})$, draw $(\widetilde{W}, \widetilde{g})$ from some conditional proposal density $q_k(\cdot \mid W^{(k)}, g^{(k)})$.

(ii) Set

$$(W^{(k+1)}, g^{(k+1)}) = \begin{cases} (\widetilde{W}, \widetilde{g}), & \text{with probability } \alpha_k(\widetilde{W}, \widetilde{g} \mid W^{(k)}, g^{(k)}), \\ (W^{(k)}, g^{(k)}), & \text{with probability } 1 - \alpha_k(\widetilde{W}, \widetilde{g} \mid W^{(k)}, g^{(k)}), \end{cases}$$

where

$$\alpha_k(\widetilde{W}, \widetilde{g} \mid W^{(k)}, g^{(k)}) = \frac{\Pi_\lambda(\widetilde{W}, \widetilde{g} \mid \mathcal{D}_n)\Pi(\widetilde{W}, \widetilde{g}) q_k(W^{(k)}, g^{(k)} \mid \widetilde{W}, \widetilde{g})}{\Pi_\lambda(W^{(k)}, g^{(k)}) \mid \mathcal{D}_n)\Pi(W^{(k)}, g^{(k)}) q_k(\widetilde{W}, \widetilde{g} \mid W^{(k)}, g^{(k)})} \wedge 1. \quad (12)$$

As in Alquier and Biau (2013), we choose $q_k = q_1$ for odd k and $q_k = q_2$ for even k, where q_1 proposes to modify (W, g) and q_2 only proposes a modification to g. The main challenge is to construct q_1, q_2 such that the resulting algorithm is manageable. To explain this construction, we fix some notation. We write

$$g(x_1, x_2) = g_1(x_1) + g_2(x_2) = \sum_{j=1}^{m} \beta_{1,j}\varphi_j(x_1) + \sum_{j=1}^{m} \beta_{2,j}\varphi_j(x_2),$$

$$\widetilde{g}(x_1, x_2) = \widetilde{g}_1(x_1) + \widetilde{g}_2(x_2) = \sum_{j=1}^{\widetilde{m}} \widetilde{\beta}_{1,j}\varphi_j(x_1) + \sum_{j=1}^{\widetilde{m}} \widetilde{\beta}_{2,j}\varphi_j(x_2),$$

with $m, \widetilde{m} \in \mathbb{N}$, $\beta_{1,j}, \beta_{2,j}, \widetilde{\beta}_{1,j}, \widetilde{\beta}_{2,j} \in \mathbb{R}$ and a basis $(\varphi_j)_{j\in\mathbb{N}}$ consisting of univariate functions. For simplicity, we use the trigonometric system $\varphi_1(x) = 1, \varphi_{2j}(x) = \cos(\pi j x), \varphi_{2j+1}(x) = \sin(\pi j x)$ for $j \in \mathbb{N}$, but in theory other systems such as wavelets could be used. For the notation of the dimension reduction matrices, it is convenient to interpret them as a vector. In particular, we write

$$W = (w_1, \dots, w_{10}) \in \mathbb{R}^{10}, \qquad \text{with} \qquad |(w_1, \dots, w_5)| = |(w_6, \dots, w_{10})| = 1,$$
$$\widetilde{W} = (\widetilde{w}_1, \dots, \widetilde{w}_{10}) \in \mathbb{R}^{10}, \qquad \text{with} \qquad |(\widetilde{w}_1, \dots, \widetilde{w}_5)| = |(\widetilde{w}_6, \dots, \widetilde{w}_{10})| = 1$$

and denote by $I \subseteq \{1, \dots, 10\}$ the set of indices of the nonzero entries of W. Further, for some $s > 0$, we set

$$\rho(g \mid W, m) \propto \exp\Big(- \frac{1}{2\,s^2} \sum_{i=1}^{2} \sum_{j=1}^{m} \big(\beta_{i,j} - \beta^*_{i,j}(W, m)\big)^2 \Big),$$

where

$$\big(\beta^*_{i,j}(W,m)\big)_{i=1,2,j=0,\dots,m} \in \underset{(\beta_{i,j})_{i=1,2,j=1,\dots,m}\in\mathbb{R}^{2\times m}}{\arg\min}$$
$$\sum_{l=1}^{n}\Big(Y_l - \sum_{i=1}^{2}\sum_{j=1}^{m}\beta_{i,j}\varphi_j\big((w_{1+5(i-1)},\dots,w_{5+5(i-1)})\mathrm{X}_l\big)\Big)^2.$$

We can now define q_1, q_2. Starting with q_1, we set

$$q_1(\cdot \mid W,g) = \begin{cases} \frac{2}{3}q_{1,=}(\cdot \mid W,g) + \frac{1}{3}q_{1,+}(\cdot \mid W,g) & \text{if } |I| = 2,\\ \frac{1}{4}q_{1,+}(\cdot \mid W,g) + \frac{1}{2}q_{1,=}(\cdot \mid W,g) + \frac{1}{4}q_{1,-}(\cdot \mid W,g) & \text{if } 2 < |I| < p,\\ \frac{2}{3}q_{1,=}(\cdot \mid W,g) + \frac{1}{3}q_{1,-}(\cdot \mid W,g) & \text{if } |I| = p. \end{cases}$$

The idea is to achieve mixing over models with different sparsity by randomly proposing to add or remove an entry from *W* if this is possible. For instance, if $|I| = p$, no components can be added and if $|I| = 2$, no further entries can be set to zero as this would violate the standardization of $(w_1,\dots,w_5)$ and $(w_6,\dots,w_{10})$. In the case where we propose to leave the sparsity unchanged, we obtain $\widetilde{W}$ by adding noise to *W*, standardizing and, given the new $\widetilde{W}$, taking $\widetilde{g}$ as a noisy version of the least squares estimator. Specifically, we choose

$$q_{1,=}(\widetilde{W},\widetilde{g} \mid W,g) = q_{1,=}(\widetilde{W} \mid W)\rho(\widetilde{g} \mid \widetilde{W},m),$$

where $q_{1,=}(\widetilde{W} \mid W)$ denotes the density of $\widetilde{W}$ drawn from

$$\widetilde{W} \sim \Big(\frac{w_1 + U_1\mathbf{1}_{\{1\in I\}}}{|(w_1 + U_1\mathbf{1}_{\{1\in I\}},\dots,w_5 + U_5\mathbf{1}_{\{5\in I\}})|},\dots,\frac{w_5 + U_5\mathbf{1}_{\{5\in I\}}}{|(w_1 + U_1\mathbf{1}_{\{1\in I\}},\dots,w_5 + U_5\mathbf{1}_{\{5\in I\}})|},$$
$$\frac{w_6 + U_6\mathbf{1}_{\{6\in I\}}}{|(w_6 + U_6\mathbf{1}_{\{6\in I\}},\dots,w_{10} + U_{10}\mathbf{1}_{\{10\in I\}})|},\dots,\frac{w_{10} + U_{10}\mathbf{1}_{\{10\in I\}}}{|(w_6 + U_6\mathbf{1}_{\{6\in I\}},\dots,w_{10} + U_{10}\mathbf{1}_{\{10\in I\}})|}\Big)$$

with $(U_1,\dots,U_{10}) \sim \mathcal{U}([-\delta,\delta]^{10})$ for some $\delta > 0$. Adding or removing components is accomplished by $q_{1,+}$ and $q_{1,-}$, respectively. Denote by W_{-i} the vector *W* with the *i*-th entry set to zero and standardized in the first and second half of the entries, respectively. We set

$$q_{1,+}(\widetilde{W},\widetilde{g} \mid W,g) = \sum_{i\notin I} c_{+,i}\mathbf{1}_{\{\widetilde{W}_{-i}=W\}}\frac{\mathbf{1}_{\{|\widetilde{w}_i|<\delta\}}}{2\delta}\rho(\widetilde{g} \mid \widetilde{W},m)$$

with weights

$$c_{+,i} = \frac{\exp\Big(\Big|\sum_{l=1}^{n}\big(Y_l - g(W\mathrm{X}_l)\big)\mathrm{X}_{l,i}\Big|\Big)}{\sum_{j\notin I}\exp\Big(\Big|\sum_{l=1}^{n}\big(Y_l - g(W\mathrm{X}_l)\big)\mathrm{X}_{l,j}\Big|\Big)},$$

where $\mathrm{X}_{l,i}$ denotes the *i*-th entry of the *l*-th observation.

For the removal of components, we use

$$q_{1,-}(\widetilde{W}, \widetilde{g} \mid W, g) = \sum_{i \in I} c_{-,i} \mathbf{1}_{\{\widetilde{W} = W_{-i}\}} \rho(\widetilde{g} \mid \widetilde{W}, m),$$

where the weights are chosen as

$$c_{-,i} = \frac{\exp(-|w_i|^2) \mathbf{1}_{\{|w_i| < \delta\}}}{\sum_{j \in I} \exp(-|w_j|^2) \mathbf{1}_{\{|w_j| < \delta\}}}.$$

Choosing $\delta < 1$ ensures that the algorithm will not propose to remove the last remaining entry of $(w_1, \dots, w_5)$ or $(w_6, \dots, w_{10})$.

For q_2, we set

$$q_2(\cdot \mid W, g) = \begin{cases} \frac{2}{3} q_{2,=}(\cdot \mid W, g) + \frac{1}{3} q_{2,+}(\cdot \mid W, g) & \text{if } m = 1, \\ \frac{1}{4} q_{2,+}(\cdot \mid W, g) + \frac{1}{2} q_{2,=}(\cdot \mid W, g) + \frac{1}{4} q_{2,-}(\cdot \mid W, g) & \text{if } 1 < m < n, \\ \frac{2}{3} q_{2,=}(\cdot \mid W, g) + \frac{1}{3} q_{2,-}(\cdot \mid W, g) & \text{if } m = n, \end{cases}$$

where

$$\begin{aligned} q_{2,=}(\widetilde{W}, \widetilde{g} \mid W, g) &= \mathbf{1}_{\{W = \widetilde{W}\}} \rho(\widetilde{g} \mid W, m) \\ q_{2,+}(\widetilde{W}, \widetilde{g} \mid W, g) &= \mathbf{1}_{\{W = \widetilde{W}\}} \rho(\widetilde{g} \mid W, m + 1) \\ q_{2,-}(\widetilde{W}, \widetilde{g} \mid W, g) &= \mathbf{1}_{\{W = \widetilde{W}\}} \rho(\widetilde{g} \mid W, m - 1). \end{aligned}$$

This choice of q_2 allows for the mixing of models with a different number of basis elements representing the link function.

In theory, one could extend the algorithm to also accomplish mixing over different active dimensions. An idea would be to alternate between three steps, q_1, q_2 and q_3, where q_3 proposes to increase or decrease the active dimension of the model. However, constructing this proposal distribution q_3 such that the acceptance probability in (12) is large enough, on average, to allow for convergence of the algorithm in a reasonable number of steps is challenging, as a more complex proposal also leads to a more computationally costly acceptance probability. Also, evaluating the multivariate basis required for the full multi-index model in every step adds to this cost.

We now demonstrate the performance of this algorithm in simulation examples. First, we specify the experimental setup. For the true regression function *f*, we consider the following three models:

(a) Additive model: $f(\mathrm{x}) = g_1(x_1) + g_6(x_6)$ with $g_1(x_1) = -\sin(\pi x_1^2)$, $g_6(x_6) = -\cos(\pi x_6)$.
(b) Single-index model: $f(\mathrm{x}) = g(w^\top \mathrm{x})$ with $w = (1/\sqrt{2}, 1/\sqrt{2})^\top$ and $g = 2\cos(\pi \cdot)$.
(c) Projection pursuit regression: $f(\mathrm{x}) = g_1(w_1^\top \mathrm{x}) + g_2(w_2^\top \mathrm{x})$ with $w_1 = (3/5, 0, 4/5, 0, 0, 0, 0, 0, 0, 0)^\top$, $w_2 = (\sqrt{13}/4) \cdot (3/4, 1/2, 0, 0, 0, 0, 0, 0, 0, 0)^\top$ and $g_1 = 2\cos(\pi \cdot), g_2 = \sin(\pi \cdot)$.

We generate the i.i.d. data $\mathcal{D}_n := (\mathrm{X}_i, Y_i)_{i=1,\ldots,n} \subset \mathbb{R}^p \times \mathbb{R}$ with $n = 2000$ and $p = 10$ according to $f(\mathrm{X}) = Y + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 0.01)$. The data is then split into a training and a test sample with $n_{\text{train}} = 1600$ and $n_{\text{test}} = 400$, respectively. We run the RJMCMC algorithm with $\lambda = 5n_{\text{train}}$ on the training data for 5000 iterations to obtain our estimator $\widehat{f}$ for f.

The algorithm calculates the empirical prediction risk on the training sample in every step, as illustrated in the upper half of Fig. 1 for 25 Monte Carlo iterations in the projection pursuit model. From this figure, we note that most of the optimization occurs within the first 1000 steps, after which the chain stabilizes. The lower half of this figure illustrates how the algorithm tries to find the active coordinates of the model. The four active coordinates which it settles on in most runs are indeed correct in the projection pursuit regression specified for the experiment, namely the first and third entry of w_1 and the first two entries of w_2.

After training, the average of $\big(f(\mathrm{X}_i) - \widehat{f}(\mathrm{X}_i)\big)^2$ across the test sample is calculated and normalized by $n^{-1}\sum_{i=1}^{n} f(\mathrm{X}_i)^2$ for an approximation of the relative excess risk. For each of the models above, we repeat this procedure in a Monte Carlo simulation with 25 iterations. The results are summarized in the rows of Table 1 with $\widehat{f}$. Comparing the means with the much smaller trimmed means (trimmed at the 10% level on both ends) and the medians suggests that the larger means are due to outliers. This is underlined by the fairly large standard deviations and the upper half of Fig. 1. When checking the individual runs for all models, we observe that in about 1 out of 10 runs, the algorithm wrongly removes relevant coordinates from the model, fails to remove irrelevant ones or otherwise gets stuck at what is only a local optimum. In practice, this can be circumvented by running the algorithm multiple times on the same data.

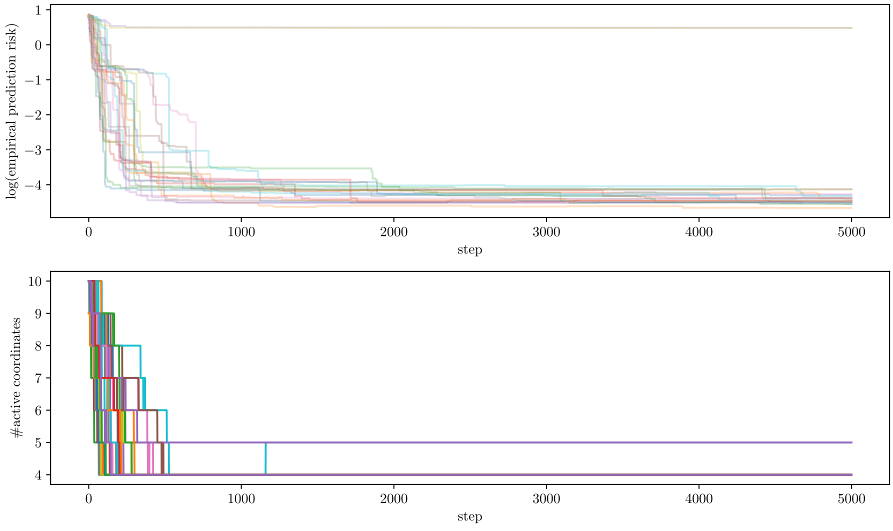

Fig. 1 Logarithm of the empirical prediction risk and number of active coordinates during training of 25 Monte Carlo iterations with 5000 steps each for the projection pursuit model

Table 1 Relative excess risk of the projection pursuit type estimator $\widehat{f}$ and the single-index estimator $\widetilde{f}$ in a Monte Carlo simulation with 25 iterations

Model	Estimator	Mean (standard deviation)	Trimmed mean	Median (inter-quartile range)
additive	$\widehat{f}$	0.0850 (0.1210)	0.0629	0.0289 (0.1193)
	$\widetilde{f}$	0.1429 (0.0090)	0.1428	0.1426 (0.0157)
single-index	$\widehat{f}$	0.0304 (0.1477)	0.0003	0.0002 (0.0004)
	$\widetilde{f}$	0.0310 (0.1386)	0.0027	0.0019 (0.0032)
projection pursuit	$\widehat{f}$	0.0528 (0.1760)	0.0009	0.0007 (0.0004)
	$\widetilde{f}$	0.2724 (0.2024)	0.2279	0.2011 (0.0232)

In order to compare the above numerical performance with a reference method, we implement the algorithm by Alquier and Biau (2013) in the same settings. The results are shown in the rows of Table 1 with $\widetilde{f}$. The performance on data generated from a single-index model is comparable to that of our method. The other settings showcase that our generalized method is more widely applicable.

5 Discussion

Overall, our results showcase the flexibility of the PAC-Bayesian estimation approach based on the Gibbs-posterior to construct adaptive estimators in a wide spectrum of models. Our theoretical results demonstrate that the estimation procedure generalizes that of Alquier and Biau (2013) from the single-index model to the multi-index model with unknown active dimension. As illustrated in Sect. 4, it is applicable to more general settings than single-index data, which is in line with the theory.

A remaining issue is that our implementation does not access the full Gibbs-posterior with a mixing prior in the active dimension, but only a simplified version in a projection pursuit type model with fixed active dimension. However, theoretical guarantees for this simplified estimator could be studied with the same techniques. For this, one would consider the Gibbs-posterior with respect to a simpler prior. Following the construction in Sect. 2, the prior would be chosen as $\widetilde{\pi}_d := \widetilde{\mu}_d \otimes \widetilde{\nu}_d$, where $\widetilde{\mu}_d$ draws a level of sparsity from a geometric distribution and given this level of sparsity draws a $d \times p$-matrix with appropriately normalized row-vectors $w_1, \dots, w_d$. $\widetilde{\nu}_d$ draws a link function by drawing a basis projection level *m* from a geometric distribution, given *m* uniformly draws coefficients $(\beta_{i,j})_{i=1,\dots,m,j=1,\dots,d}$ in basis representation and then returns the resulting function. In particular, for a given draw $(\widehat{w}_1, \dots, \widehat{w}_d)$, $(\widehat{\beta}_{i,j})_{i=1,\dots,m,j=1,\dots,d}$ from the Gibbs-posterior with prior $\widetilde{\pi}_d$, the estimate for *f* is

$$\widehat{f} = \sum_{i=1}^{d}\sum_{j=1}^{m}\widehat{\beta}_{i,j}\varphi_j(\widehat{w}_i\cdot),$$

where $(\varphi_j)_{j\in\mathbb{N}}$ is a sufficiently regular orthonormal basis of $L^2(\mathbb{R})$ such as the trigonometric system or a wavelet basis. Meanwhile, our theory is applicable to the much more general and adaptive Gibbs-posterior with the prior constructed in Sect. 2. Since the mixing prior introduces computational complexity as explained in Sect. 4, an implementation of the full method remains challenging.

6 Proofs

We begin with a general PAC-Bayes bound to then prove the main results. The proof strategy is line with the PAC-Bayes literature, see e.g. (Alquier & Biau, 2013). A similar result can be found in Steffen and Trabs (2025). The main difference is the change of the prior and, consequently, the change of the integration variables. This results in slightly different constants. The proofs of the auxiliary results are postponed to Sect. 6.3.

For probability measures μ,ν on a measurable space $(E,\mathscr{A})$, the *Kullback–Leibler divergence* of μ with respect to ν is defined via

$$\mathrm{KL}(\mu\mid\nu) := \begin{cases} \int\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu, & \text{if } \mu\ll\nu \\ \infty, & \text{otherwise}\end{cases}. \tag{13}$$

Lemma 6 *(PAC-Bayes bound) Grant Assumption A and set $V := 8(2C+1)(\Gamma\vee(2C+1))$. Then, we have for any $\lambda\in(0,n/V)$ and any $\mathcal{D}_n$ -dependent (in a measurable way) probability measure $\varrho\ll\Pi$ that*

$$\mathcal{E}(\widehat{W}_\lambda,\widehat{g}_\lambda) \leqslant 3\int\mathcal{E}\,\mathrm{d}\varrho + \frac{4}{\lambda}\big(\mathrm{KL}(\varrho\mid\Pi)+\log(2/\delta)\big) \tag{14}$$

with probability of at least $1-\delta$.

Proof For $(W,g)\in\bigcup_{d=1}^{p}\mathcal{S}_d\times\mathcal{G}_d$, we set $\mathcal{E}_n(W,g) := \frac{1}{n}\sum_{i=1}^{n}Z_i$ with centered and independent random variables

$$Z_i := (Y_i-g(W\mathrm{X}_i))^2-(Y_i-f(\mathrm{X}_i))^2 = -\big(2\varepsilon_i+f(\mathrm{X}_i)-g(W\mathrm{X}_i)\big)\big(g(W\mathrm{X}_i)-f(\mathrm{X}_i)\big).$$

Owing to (8), g is bounded $C+1$. By Assumption A, f is bounded by C and ε_i is sub-Gaussian. Hence,

$$\mathbb{E}[Z_i^2] = \mathbb{E}\big[\big(2\varepsilon_i+f(\mathrm{X}_i)-g(W\mathrm{X}_i)\big)^2\big(g(W\mathrm{X}_i)-f(\mathrm{X}_i)\big)^2\big] \leqslant 2(4\sigma^2+(2C+1)^2)\mathcal{E}(W,g) =: U$$

and for $k\geqslant 3$ we have

$$\begin{aligned}\mathbb{E}[(Z_i)_+^k] &\leqslant \mathbb{E}\big[|2\varepsilon_i + f(\mathrm{X}_i) - g(W\mathrm{X}_i)|^k |g(W\mathrm{X}_i) - f(\mathrm{X}_i)|^{k-2}(g(W\mathrm{X}_i) - f(\mathrm{X}_i))^2\big]\\ &\leqslant (2C+1)^{k-2}\mathbb{E}\big[|2\varepsilon_i + f(\mathrm{X}_i) - g(W\mathrm{X}_i)|^k (g(W\mathrm{X}) - f(\mathrm{X}))^2\big]\\ &\leqslant (2C+1)^{k-2}2^{k-1}\big(k!2^{k-1}\sigma^2\Gamma^{k-2} + (2C+1)^k\big)\mathcal{E}(W,g)\\ &\leqslant (2C+1)^{k-2}k!8^{k-2}\big(\Gamma^{k-2}\vee(2C+1)^{k-2}\big)U\\ &= k!UV^{k-2}.\end{aligned}$$

As $\mathcal{E}_n(W,g)$ is centered around $\mathcal{E}(W,g)$, a variant of Bernstein's inequality, see Massart (2007, inequality (2.21)), yields for $\lambda\in(0,n/V)$ and $C_{n,\lambda} := \frac{\lambda}{n}\frac{2((2C+1)^2+4\sigma^2)}{1-V\lambda/n}$ that

$$\mathbb{E}\big[\exp\big(\lambda\big(\mathcal{E}_n(W,g) - \mathcal{E}(W,g)\big)\big)\big] \leqslant \exp\Big(\frac{U\lambda^2}{n(1-V\lambda/n)}\Big) = \exp\big(C_{n,\lambda}\lambda\mathcal{E}(W,g)\big),$$

which we can rewrite as

$$\mathbb{E}\big[\exp\big(\lambda\mathcal{E}_n(W,g) - \lambda(1+C_{n,\lambda})\mathcal{E}(W,g) - \log(\delta^{-1})\big)\big] \leqslant \delta. \tag{15}$$

Note that the same arguments can be applied if we replace Z_i by $-Z_i$ leading to

$$\mathbb{E}\big[\exp\big(\lambda(1-C_{n,\lambda})\mathcal{E}(W,g) - \lambda\mathcal{E}_n(W,g) - \log(\delta^{-1})\big)\big] \leqslant \delta. \tag{16}$$

Integrating both sides of (15) and (16) in (*W*, *g*) with respect to Π and applying Fubini's theorem, we conclude

$$\mathbb{E}\Big[\int\exp\big(\lambda(1-C_{n,\lambda})\mathcal{E}(W,g) - \lambda\mathcal{E}_n(W,g) - \log(\delta^{-1})\big)\,\mathrm{d}\Pi(W,g)\Big] \leqslant \delta \qquad \text{and} \tag{17}$$

$$\mathbb{E}\Big[\int\exp\big(\lambda\mathcal{E}_n(W,g) - \lambda(1+C_{n,\lambda})\mathcal{E}(W,g) - \log(\delta^{-1})\big)\,\mathrm{d}\Pi(W,g)\Big] \leqslant \delta. \tag{18}$$

The Radon-Nikodym density of the posterior distribution $\Pi_\lambda(\cdot\mid\mathcal{D}_n)\ll\Pi$ with respect to Π is given by

$$\frac{\mathrm{d}\Pi_\lambda(W,g\mid\mathcal{D}_n)}{\mathrm{d}\Pi} = D_\lambda^{-1}\exp(-\lambda R_n(W,g)), \qquad D_\lambda := \int\exp(-\lambda R_n(W,g))\,\mathrm{d}\Pi(W,g). \tag{19}$$

Therefore, (17) gives

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n,(\widehat{W},\widehat{g})\sim\Pi_\lambda(\cdot|\mathcal{D}_n)}\big[\exp\big(\lambda(1-C_{n,\lambda})\mathcal{E}(\widehat{W},\widehat{g})-\lambda\mathcal{E}_n(\widehat{W},\widehat{g})-\log(\delta^{-1})\\
&\quad+\lambda R_n(\widehat{W},\widehat{g})+\log D_\lambda\big)\big]\\
&\quad=\mathbb{E}_{\mathcal{D}_n,(\widehat{W},\widehat{g})\sim\Pi_\lambda(\cdot|\mathcal{D}_n)}\Big[\exp\Big(\lambda(1-C_{n,\lambda})\mathcal{E}(\widehat{W},\widehat{g})-\lambda\mathcal{E}_n(\widehat{W},\widehat{g})\\
&\quad-\log(\delta^{-1})-\log\Big(\frac{\mathrm{d}\Pi_\lambda(\widehat{W},\widehat{g}\mid\mathcal{D}_n)}{\mathrm{d}\Pi}\Big)\Big)\Big]\\
&\quad=\mathbb{E}_{\mathcal{D}_n}\Big[\int\exp\big(\lambda(1-C_{n,\lambda})\mathcal{E}(W,g)-\lambda\mathcal{E}_n(W,g)-\log(\delta^{-1})\big)\,\mathrm{d}\Pi(W,g)\Big]\leqslant\delta.
\end{aligned}$$

Using the inequality $\mathbf{1}_{[0,\infty)}(x)\leqslant \mathrm{e}^{\lambda x}$ for all $x\in\mathbb{R}$, we deduce for $(\widehat{W},\widehat{g})\sim\Pi_\lambda(\cdot\mid\mathcal{D}_n)$ with a probability not larger than δ that

$$\big(1-C_{n,\lambda}\big)\mathcal{E}(\widehat{W},\widehat{g})-\mathcal{E}_n(\widehat{W},\widehat{g})+R_n(\widehat{W},\widehat{g})-\lambda^{-1}\big(\log(\delta^{-1})-\log D_\lambda\big)\geqslant 0.$$

As $1-C_{n,\lambda}>0$, we have with a probability of at least $1-\delta$ that

$$\mathcal{E}(\widehat{W},\widehat{g})\leqslant(1-C_{n,\lambda})^{-1}\big(-R_n(I_{p\times p},f)+\lambda^{-1}\big(\log(\delta^{-1})-\log D_\lambda\big)\big).$$

Equation (5.2.1) in Catoni (2004) yields

$$\begin{aligned}
-\log D_\lambda&=-\log\Big(\int\exp(-\lambda R_n(W,g))\,\mathrm{d}\Pi(W,g)\Big)\\
&=\inf_{\varrho\ll\Pi}\Big(\mathrm{KL}(\varrho\mid\Pi)+\int\lambda R_n(W,g)\,\mathrm{d}\varrho(W,g)\Big).
\end{aligned}\tag{20}$$

Therefore, for any $\varrho\ll\Pi$, it holds with probability of at least $1-\delta$ that

$$\mathcal{E}(\widehat{W},\widehat{g})\leqslant(1-C_{n,\lambda})^{-1}\Big(\int\mathcal{E}_n(W,g)\,\mathrm{d}\varrho(W,g)+\lambda^{-1}\big(\log(\delta^{-1})+\mathrm{KL}(\varrho\mid\Pi)\big)\Big).\tag{21}$$

To control the leading integral term in (21), we use Jensen's inequality and (18) to obtain

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\Big[\exp\Big(\int\lambda\mathcal{E}_n(W,g)-\lambda(1+C_{n,\lambda})\mathcal{E}(W,g)\,\mathrm{d}\varrho(W,g)-\mathrm{KL}(\varrho\mid\Pi)-\log(\delta^{-1})\Big)\Big]\\
&\qquad=\mathbb{E}_{\mathcal{D}_n}\Big[\exp\Big(\int\lambda\mathcal{E}_n(W,g)-\lambda(1+C_{n,\lambda})\mathcal{E}(W,g)-\log\Big(\frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}(W,g)\Big)\\
&\qquad-\log(\delta^{-1})\,\mathrm{d}\varrho(W,g)\Big)\Big]\\
&\qquad\leqslant\mathbb{E}_{\mathcal{D}_n,(W,g)\sim\varrho}\Big[\exp\Big(\lambda\mathcal{E}_n(W,g)-\lambda(1+C_{n,\lambda})\mathcal{E}(W,g)\\
&\qquad-\log\Big(\frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}(W,g)\Big)-\log(\delta^{-1})\Big)\Big]\\
&\qquad=\mathbb{E}_{\mathcal{D}_n}\Big[\int\exp\big(\lambda\mathcal{E}_n(W,g)-\lambda(1+C_{n,\lambda})\mathcal{E}(W,g)-\log(\delta^{-1})\big)\,\mathrm{d}\Pi(W,g)\Big]\leqslant\delta.
\end{aligned}$$

Again by $\mathbf{1}_{[0,\infty)}(x) \leqslant \mathrm{e}^{\lambda x}$, we deduce that with probability of at least $1-\delta$

$$\int \mathcal{E}_n(W,g)\,\mathrm{d}\varrho(W,g) \leqslant (1+C_{n,\lambda})\int \mathcal{E}(W,g)\,\mathrm{d}\varrho(W,g) + \lambda^{-1}\big(\mathrm{KL}(\varrho \mid \Pi) + \log(\delta^{-1})\big).$$

Combined with (21), we conclude that with probability of at least $1-2\delta$

$$\mathcal{E}(\widehat{W},\widehat{g}) \leqslant (1-C_{n,\lambda})^{-1}\Big((1+C_{n,\lambda})\int \mathcal{E}(W,g)\,\mathrm{d}\varrho(W,g) + \frac{2}{\lambda}\big(\mathrm{KL}(\varrho \mid \Pi) + \log(\delta^{-1})\big)\Big),$$

which yields the claimed bound since $C_{n,\lambda} \leqslant 1/2$. □

6.1 Proof of Theorem 1

We extend the proof strategy by Alquier and Biau (2013) to the multi-index setting with unknown active dimension.

Lemma 6 with $\lambda = n/(V+4((2C+1)^2+4\sigma^2))$ and $\varrho \ll \Pi$ ensures that

$$\mathcal{E}(\widehat{W}_\lambda,\widehat{g}_\lambda) \leqslant 3\int \mathcal{E}(W,g)\,\mathrm{d}\varrho(W,g) + \frac{Q_3}{n}\big(\mathrm{KL}(\varrho \mid \Pi) + \log(2/\delta)\big) \tag{22}$$

with a probability of at least $1-\delta$, where the constant Q_3 only depends on C,Γ and σ.

To choose ϱ, we fix some triplet (d,I,M) with $d \in \{1,\dots,p\}$, $I = I_1 \times \cdots \times I_d \in \mathcal{I}_d$, $M \in \{0,\dots,n\}$ as well as $\eta,\gamma \in (0,1]$ and set

$$\varrho := \varrho_{d,I,M,\eta,\gamma} := \varrho^1_{d,I,\eta} \otimes \varrho^2_{d,M,\gamma}, \tag{23}$$

where $\varrho^1_{d,I,\eta}$ and $\varrho^2_{d,M,\gamma}$ are the uniform distribution with respect to $\mu_{d,I}$ and $\nu_{d,M}$ on a ball of radius η and γ around the oracle $W^*_{d,I} = (w^*_{d,I,1},\dots,w^*_{d,I,d})^\top \in \mathbb{R}^{d\times p}$ and $g^*_{d,M}$, respectively. Specifically, we set

$$\begin{aligned}
&\frac{\mathrm{d}\varrho^1_{d,I,\eta}}{\mathrm{d}\mu_{d,I}}(W) := \prod_{i=1}^d \frac{\mathrm{d}\varrho^{1,i}_{d,I,\eta}}{\mathrm{d}\mu_{I_i}}(w_i),\ \forall W = (w_1,\dots,w_d)^\top, \quad \text{where}\\
&\frac{\mathrm{d}\varrho^{1,i}_{d,I,\eta}}{\mathrm{d}\mu_{I_i}}(w_i) \propto \mathbf{1}_{\{|w_i - w^*_{d,I,i}| \leqslant \eta\}} \quad \text{and} \quad \frac{\mathrm{d}\varrho^2_{d,M,\gamma}}{\mathrm{d}\nu_{d,M}}(g) \propto \mathbf{1}_{\{\|g-g^*_{d,M}\|_{\mathcal{B}} \leqslant \gamma\}},
\end{aligned} \tag{24}$$

where μ_{I_i} denotes the uniform distribution on $\mathcal{S}(I_i)$. To complete the proof, we need to bound the terms on the right hand side of (22) for this choice of ϱ.

First, we deal with the Kullback–Leibler divergence term using the following two lemmas:

Lemma 7 *For $\varrho = \varrho_{d,I,M,\eta,\gamma}$ from (23) and with $\pi_{d,I,M} = \mu_{d,I} \otimes \nu_{d,M}$, we have*

$\mathrm{KL}(\varrho \mid \Pi) \leqslant \|I\| \log(\mathrm{e}p) + (\|I\| + M + 2)\log(2) + \mathrm{KL}(\varrho \mid \pi_{d,I,M}) =: T_1 + \mathrm{KL}(\varrho \mid \pi_{d,I,M})$.

Lemma 8 *For $\varrho = \varrho_{d,I,M,\eta,\gamma}$ from (23) and with $\pi_{d,I,M} = \mu_{d,I} \otimes \nu_{d,M}$, we have*

$$\mathrm{KL}(\varrho \mid \pi_{d,I,M}) \leqslant \|I\| \log(5/\eta) + 16^d N^d 2^{dM} \log((C+1)/\gamma) =: T_2.$$

Thus,

$$\mathcal{E}(\widehat{W}_\lambda, \widehat{g}_\lambda) \leqslant 3 \int \mathcal{E}(W,g)\,\mathrm{d}\varrho(W,g) + \frac{Q_3}{n}\big(T_1 + T_2 + \log(2/\delta)\big) \tag{25}$$

with a probability of at least $1-\delta$.

Second, we control the integral term in (25) by splitting it into

$$\begin{aligned}\int \mathcal{E}(W,g)\,\mathrm{d}\varrho(W,g) &= \mathcal{E}(W^*_{d,I}, g^*_{d,M}) \\ &\quad+ \int \mathbb{E}[(g^*_{d,M}(W^*_{d,I}\mathrm{X}) - g(W^*_{d,I}\mathrm{X}))^2]\,\mathrm{d}\varrho(W,g) \\ &\quad+ \int \mathbb{E}[(g(W^*_{d,I}\mathrm{X}) - g(W\mathrm{X}))^2]\,\mathrm{d}\varrho(W,g) \\ &\quad+ \int \mathbb{E}[2(Y - g^*_{d,M}(W^*_{d,I}\mathrm{X}))(g^*_{d,M}(W^*_{d,I}\mathrm{X}) - g(W^*_{d,I}\mathrm{X}))]\,\mathrm{d}\varrho(W,g) \\ &\quad+ \int \mathbb{E}[2(Y - g^*_{d,M}(W^*_{d,I}\mathrm{X}))(g(W^*_{d,I}\mathrm{X}) - g(W\mathrm{X}))]\,\mathrm{d}\varrho(W,g) \\ &\quad+ \int \mathbb{E}[2(g^*_{d,M}(W^*_{d,I}\mathrm{X}) - g(W^*_{d,I}\mathrm{X}))(g(W^*_{d,I}\mathrm{X}) \\ &\qquad - g(W\mathrm{X}))]\,\mathrm{d}\varrho(W,g) \\ &=: \mathcal{E}(W^*_{d,I}, g^*_{d,M}) + U_1 + U_2 + U_3 + U_4 + U_5\end{aligned} \tag{26}$$

and treating the terms $U_1, \ldots, U_5$ sequentially.

Recall the definition of $\|\cdot\|_{\mathcal{B}}$ from (7) and the paragraph thereafter. As in (8), $g = \Phi_{d,M}(\beta) \in \mathcal{G}_{d,M}(C+1)$ with $\|\beta - \beta^*_{d,M}\|_{\mathcal{B}} \leqslant \gamma$ implies

$$\|g - g^*_{d,M}\|_\infty \leqslant \|g - g^*_{d,M}\|_{\mathcal{B}} = \|\beta - \beta^*_{d,M}\|_{\mathcal{B}} \leqslant \gamma.$$

As a consequence, we obtain

$$\begin{aligned} U_1 &= \int \mathbb{E}[(g^*_{d,M}(W^*_{d,I}\mathrm{X}) - g(W^*_{d,I}\mathrm{X}))^2]\,\mathrm{d}\varrho^2_{d,M,\gamma}(g) \\ &\leqslant \int \sup_{\mathrm{x}\in\mathbb{R}^d} (g^*_{d,M}(\mathrm{x}) - g(\mathrm{x}))^2\,\mathrm{d}\varrho^2_{d,M,\gamma}(g) \leqslant \gamma^2. \end{aligned} \tag{27}$$

Any $g \in \mathcal{G}_{d,M}(C+1)$ is differentiable as a linear combination of only finitely many basis elements. Therefore, applying the fundamental theorem of calculus to the mapping

$$h : [-1,1] \to \mathbb{R},\ s \mapsto g((W + s(W^*_{d,I} - W))\mathrm{X}(\omega))$$

with a fixed $\omega \in \Omega$ (which we omit from here on) yields

$$g(W^*_{d,I}\mathrm{X}) - g(W\mathrm{X}) = \int_0^1 h'(s)\,\mathrm{d}s = \Big\langle (W^*_{d,I} - W)\mathrm{X}, \int_0^1 \nabla g((W + s(W^*_{d,I} - W))\mathrm{X})\,\mathrm{d}s \Big\rangle.$$

Combined with (8) and applying the Cauchy–Schwarz inequality twice, we obtain

$$\begin{aligned} |g(W^*_{d,I}\mathrm{X}) - g(W\mathrm{X})| &= \Big| \Big\langle (W^*_{d,I} - W)\mathrm{X}, \int_0^1 \nabla g((W + s(W^*_{d,I} - W))\mathrm{X})\,\mathrm{d}s \Big\rangle \Big| \\ &\leqslant \big|(W^*_{d,I} - W)\mathrm{X}\big| \Big| \int_0^1 \nabla g((W + s(W^*_{d,I} - W))\mathrm{X})\,\mathrm{d}s \Big| \\ &\leqslant |\mathrm{X}| \Big(\sum_{i=1}^d |w^*_{d,I,i} - w_i|^2 \Big)^{1/2} \sqrt{d}C \\ &\leqslant pK \Big(\sum_{i=1}^d |w^*_{d,I,i} - w_i|^2 \Big)^{1/2} \sqrt{d}C \ \ \mathbb{P}\text{-a.s.} \end{aligned} \tag{28}$$

Using (28), we deduce

$$\begin{aligned} U_2 &= \int \mathbb{E}[(g(W^*_{d,I}\mathrm{X}) - g(W\mathrm{X}))^2]\,\mathrm{d}\varrho^1_{d,I,\eta} \otimes \varrho^2_{d,M,\gamma}(W, g) \\ &\leqslant d(pKC)^2 \int \cdots \int \sum_{i=1}^d |w^*_{d,I,i} - w_i|^2\,\mathrm{d}\varrho^{1,1}_{d,I,\eta}(w_1) \ldots \mathrm{d}\varrho^{1,d}_{d,I,\eta}(w_d) \\ &\leqslant (dpKC\eta)^2. \end{aligned} \tag{29}$$

By construction, $\varrho^2_{d,M,\gamma}$ is centered around $g^*_{d,M}$ and thus

$$\int g(\mathrm{x})\,\mathrm{d}\varrho^2_{d,M,\gamma}(g) = g^*_{d,M}(\mathrm{x}),\ \forall \mathrm{x} \in \mathbb{R}^p. \tag{30}$$

In particular, we have

$$U_3 = 0. \tag{31}$$

Using Fubini's theorem together with (30), and applying the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} |U_4| &= 2\Big|\mathbb{E}\Big[(Y - g^*_{d,M}(W^*_{d,I}\mathrm{X}))\int\int g(W^*_{d,I}\mathrm{X}) - g(W\mathrm{X})\,\mathrm{d}\varrho^2_{d,M,\gamma}(g)\,\mathrm{d}\varrho^1_{d,I,\eta}(W)\Big]\Big| \\ &= 2\Big|\mathbb{E}\Big[(Y - g^*_{d,M}(W^*_{d,I}\mathrm{X}))\int g(W^*_{d,I}\mathrm{X}) - g^*_{d,M}(W\mathrm{X})\,\mathrm{d}\varrho^1_{d,I,\eta}(W)\Big]\Big| \\ &\leqslant 2\Big(R(W^*_{d,I}, g^*_{d,M})\mathbb{E}\Big[\Big(\int g^*_{d,M}(W^*_{d,I}\mathrm{X}) - g^*_{d,M}(W\mathrm{X})\,\mathrm{d}\varrho^1_{d,I,\eta}(W)\Big)^2\Big]\Big)^{1/2}. \end{aligned} \tag{32}$$

Repeating the argument from treating U_2, but now with $g = g^*_{d,M}$, we have

$$\mathbb{E}\Big[\Big(\int g^*_{d,M}(W^*_{d,I}\mathrm{X}) - g^*_{d,M}(W\mathrm{X})\,\mathrm{d}\varrho^1_{d,I,\eta}(W)\Big)^2\Big] \leqslant (dpKC\eta)^2. \tag{33}$$

Clearly, $g \equiv 0 \in \mathcal{G}_{d,M}(C)$ and thus we get by definition of $(W^*_{d,I}, g^*_{d,M})$ that

$$R(W^*_{d,I}, g^*_{d,M}) \leqslant R(W^*_{d,I}, g) = \mathbb{E}[Y^2] = \mathbb{E}[f(\mathrm{X})^2] + 2\mathbb{E}\big[f(\mathrm{X})\mathbb{E}[\varepsilon|\mathrm{X}]\big] + \mathbb{E}[\varepsilon^2] \leqslant C^2 + \sigma^2. \tag{34}$$

Plugging (33) and (34) into (32), we have

$$|U_4| \leqslant 2dpKC\eta\sqrt{C^2 + \sigma^2}. \tag{35}$$

Finally, applying (28) again yields

$$\begin{aligned} |U_5| &\leqslant 2\int \mathbb{E}[|g^*_{d,M}(W^*_{d,I}\mathrm{X}) - g(W^*_{d,I}\mathrm{X})||g(W^*_{d,I}\mathrm{X}) - g(W\mathrm{X})|]\,\mathrm{d}\varrho(W, g) \\ &\leqslant 2\sqrt{d}pKC\mathbb{E}\Big[\int |g^*_{d,M}(W^*_{d,I}\mathrm{X}) - g(W^*_{d,I}\mathrm{X})|\Big(\sum_{i=1}^d |w^*_{d,I,i} - w_i|^2\Big)^{1/2} \\ &\quad \mathrm{d}\varrho((w_1, \ldots, w_d)^\top, g)\Big] \\ &= 2\sqrt{d}pKC\int |g^*_{d,M}(W^*_{d,I}\mathrm{X}(\omega)) - g(W^*_{d,I}\mathrm{X}(\omega))| \\ &\quad \cdot \Big(\sum_{i=1}^d |w^*_{d,I,i} - w_i|^2\Big)^{1/2}\,\mathrm{d}\mathbb{P}\otimes\varrho(\omega, (w_1, \ldots, w_d)^\top, g) \\ &\leqslant 2\sqrt{d}pKC\Big(\int \big(g^*_{d,M}(W^*_{d,I}\mathrm{X}(\omega)) - g(W^*_{d,I}\mathrm{X}(\omega))\big)^2\,\mathrm{d}\mathbb{P}\otimes\varrho(\omega, (w_1, \ldots, w_d)^\top, g)\Big)^{1/2} \\ &\quad \cdot \Big(\int\sum_{i=1}^d |w^*_{d,I,i} - w_i|^2\,\mathrm{d}\mathbb{P}\otimes\varrho(\omega, (w_1, \ldots, w_d)^\top, g)\Big)^{1/2} \end{aligned} \tag{36}$$

$$\begin{aligned} &\leqslant 2\sqrt{d}pKC\Big(\int \mathbb{E}[(g(W^*_{d,I}\mathrm{X}) - g(W^*_{d,I}\mathrm{X}))^2]\,\mathrm{d}\varrho^2_{d,M,\gamma}(g)\Big)^{1/2} \\ &\quad \cdot \Big(\int\sum_{i=1}^d |w^*_{d,I,i} - w_i|^2\,\mathrm{d}\varrho^1_{d,I,\eta}((w_1, \ldots, w_d)^\top)\Big)^{1/2} \\ &\leqslant 2dpKC\eta\gamma, \end{aligned} \tag{37}$$

where (36) follows from the Cauchy-Schwarz inequality for integration with respect to the product measure $\mathbb{P} \otimes \varrho^1_{d,I,\eta} \otimes \varrho^2_{d,M,\gamma}$.

Choosing $\eta = (2dpKC\sqrt{C^2+\sigma^2 n})^{-1}, \gamma = n^{-1}$ when summarizing (26), (27), (29), (31), (35) and (37), we have

$$\begin{aligned}\int \mathcal{E}(W,g)\,\mathrm{d}\varrho(W,g) &\leqslant \mathcal{E}(W^*_{d,I}, g^*_{d,M}) + \gamma^2 + (dpKC\eta)^2 \\ &\quad + 2dpKC\eta\sqrt{C^2+\sigma^2} + 2dpKC\eta\gamma \\ &\leqslant \mathcal{E}(W^*_{d,I}, g^*_{d,M}) + \frac{4}{n}.\end{aligned}$$

With these choices for η and γ, we get for sufficiently large *n* that

$$\begin{aligned}T_1 + T_2 &= \|I\|\log(\mathrm{e}p) + (\|I\| + M + 2)\log(2) + \|I\|\log(5dpn) + 16^d N^d 2^{dM}\log((C+1)n) \\ &\leqslant Q_4\big(\|I\|\log(p \vee n) + 16^d N^d 2^{dM}\log(n)\big)\end{aligned}$$

with a constant Q_4 independent of the parameters involved.

Summarizing the above, we arrive at

$$\mathcal{E}(\widehat{W}_\lambda, \widehat{g}_\lambda) \leqslant 3\mathcal{E}(W^*_{d,I}, g^*_{d,M}) + \frac{Q_5}{n}\big(\|I\|\log(p \vee n) + 16^d N^d 2^{dM}\log(n) + \log(2/\delta)\big) \tag{38}$$

with a probability of at least $1-\delta$, where Q_5 is a constant only depending on C, Γ and σ. Note that the upper bound in (38) is deterministic. Choosing a triplet (*d*, *I*, *M*) such that this upper bound is minimized (which is always possible, since there are only finitely many choices for (*d*, *I*, *M*)) completes the proof of Theorem 1. □

6.2 Proof of Corollary 4

Plugging in d^* and I^* in the minimum in Theorem 1, we obtain that

$$\begin{aligned}\mathcal{E}(\widehat{W}_\lambda, \widehat{g}_\lambda) &\leqslant \min_{d,I,M}\Big(3\mathcal{E}(W^*_{d,I}, f^*_{d,M}) + \frac{Q_1}{n}\big(\|I\|\log(p \vee n) + 16^d N^d 2^{dM}\log(n) + \log(2/\delta)\big)\Big) \\ &\leqslant \min_{\substack{0 \leqslant M \leqslant n,\\ g \in \mathcal{G}_{d^*,M}(C)}}\Big(3\mathcal{E}(W^*, g) + \frac{Q_1}{n}\big(\|W^*\|_0\log(p \vee n) + 16^d N^d 2^{dM}\log(n) + \log(2/\delta)\big)\Big)\end{aligned} \tag{39}$$

with a probability of at least $1-\delta$. The rest of the proof consists of choosing *M* to balance the terms on the right hand side of (39) by using an approximation of g^*, namely

$$g_M = \sum_{l \in \mathcal{Z}^{d^*}_M} \langle g^*, \Psi_l\rangle \Psi_l \tag{40}$$

and then determining the projection level *M*. To do this, we have to verify that g_M is a valid choice for *g* in the sense that $g_M \in \mathcal{G}_{d^*,M}(C)$. Indeed, the Cauchy-Schwarz inequality ensures that

$$\begin{aligned} L^{d^*} \sum_{l\in\mathcal{Z}_M^{d^*}} 2^{l_1(d/2+1)}|\langle g^*,\Psi_l\rangle| &\leqslant L^{d^*}\Big(\sum_{l\in\mathcal{Z}_M^{d^*}} 2^{2l_1(1-\alpha+d^*/2)}\Big)^{1/2}\Big(\sum_{l\in\mathcal{Z}_M^{d^*}} 2^{2l_1\alpha}|\langle g^*,\Psi_l\rangle|^2\Big)^{1/2} \\ &\leqslant L^{d^*}2N^{d^*/2}16^{d^*/2}\Big(\sum_{l\in\mathcal{Z}_{\infty,\infty}^{d^*}} 2^{2l_1\alpha}|\langle g^*,\Psi_l\rangle|^2\Big)^{1/2} \\ &\leqslant C, \end{aligned} \tag{41}$$

since $g^* \in B^{\alpha}_{2,d^*}(\xi)$ with $\xi = C(L^{d^*}2N^{d^*/2}16^{d^*/2})^{-1}$. Denote the Lebesgue measure on $\mathbb{R}^{d^*}$ by λ^{d^*}. Using that $W^*\mathrm{X}$ admits a Lebesgue-density $\varrho \leqslant B_2$ by Assumption B, we see that g_M admits an excess risk of

$$\begin{aligned} \mathcal{E}(W^*, g_M) &= \mathbb{E}[(g_M(W^*\mathrm{X}) - g^*(W^*\mathrm{X}))^2] \\ &= \int_{[-B_1,B_1]^{d^*}} \varrho(\mathrm{x})(g_M(\mathrm{x}) - g^*(\mathrm{x}))^2 \lambda^{d^*}(\mathrm{dx}) \\ &\leqslant B_2 \int_{[-B_1,B_1]^{d^*}} (g_M(\mathrm{x}) - g^*(\mathrm{x}))^2 \lambda^{d^*}(\mathrm{dx}) \\ &\leqslant B_2 \sum_{l\in\mathcal{Z}_{\infty,N}^{d^*}\setminus\mathcal{Z}_M^{d^*}} |\langle g^*,\Psi_l\rangle|^2 \\ &\leqslant B_2 2^{-2\alpha M} \sum_{l\in\mathcal{Z}_{\infty,N}^{d^*}\setminus\mathcal{Z}_M^{d^*}} 2^{2l_1\alpha}|\langle g^*,\Psi_l\rangle|^2 \\ &\leqslant B_2 2^{-2\alpha M}(2N^{d^*/2}16^{d^*/2})^{-1}CL^{-d^*}. \end{aligned} \tag{42}$$

Applying (42) to (39), we see for sufficiently large *n* and with a probability of at least $1-\delta$ that

$$\mathcal{E}(\widehat{W}_\lambda, \widehat{g}_\lambda) \leqslant Q_6 \min_{0\leqslant M\leqslant n} \Big(2^{-2\alpha M} + 2^{d^*M}\frac{\log n}{n} + \frac{\|W^*\|_0 \log(p\vee n)}{n} + \frac{\log(2/\delta)}{n}\Big)$$

with a constant Q_6 only depending on $C, \Gamma, \sigma, N, B_1, B_2$ and d^*. To balance the order of the terms depending on *M*, we choose

$$M = \Big\lceil \log\Big(\frac{n}{\log n}\Big)\Big/\big((2\alpha + d^*)\log(2)\big)\Big\rceil,$$

which completes the proof. □

6.3 Proofs of auxiliary lemmas

6.3.1 Proof of Lemma 7

We employ another auxiliary lemma:

Lemma 9 *It holds that*

$$\begin{aligned}\mathrm{KL}(\varrho_{d,I,M,\eta,\gamma} \mid \Pi) =&\log(G(d,I,M)) + \mathrm{KL}(\varrho_{d,I,M,\eta,\gamma} \mid \pi_{d,I,M}), \qquad \textit{where} \\ G(d,I,M) :=&(1-2^{-p})(1-2^{(1-p)d-1})(2-2^{-n})2^{\|I\|+M+1}|\mathcal{I}_{d,\|I\|}|.\end{aligned}$$

Now, we can combine $|\mathcal{I}_{d,\|I\|}| \leqslant \binom{dp}{\|I\|}$ with the basic inequality $\binom{dp}{\|I\|} \leqslant \left(\frac{dp\mathrm{e}}{\|I\|}\right)^{\|I\|}$ and $\|I\| \geqslant d$ to obtain

$$\begin{aligned}\mathrm{KL}(\varrho_{d,I,M,\eta,\gamma} \mid \Pi) &= \log(G(d,I,M)) + \mathrm{KL}(\varrho_{d,I,M,\eta,\gamma} \mid \pi_{d,I,M}) \\ &\leqslant \log\left(2^{\|I\|+M+2}\binom{dp}{\|I\|}\right) + \mathrm{KL}(\varrho_{d,I,M,\eta,\gamma} \mid \pi_{d,I,M}) \\ &\leqslant \|I\|\log(\mathrm{e}p) + (\|I\|+M+2)\log(2) + \mathrm{KL}(\varrho_{d,I,M,\eta,\gamma} \mid \pi_{d,I,M}).\end{aligned}$$

□

6.3.2 Proof of Lemma 8

We split the proof into two further auxiliary lemmas:

Lemma 10 *For $\varrho^1_{d,I,\eta}$ from (24), we have*

$$\mathrm{KL}(\varrho^1_{d,I,\eta} \mid \mu_{d,I}) \leqslant \|I\|\log(5/\eta).$$

Lemma 11 *For $\varrho^2_{d,M,\gamma}$ from (24), we have*

$$\mathrm{KL}(\varrho^2_{d,M,\gamma} \mid \nu_{d,M}) \leqslant 16^d N^d 2^{dM}\log((C+1)/\gamma).$$

The assertion then follows directly via

$$\begin{aligned}\mathrm{KL}(\varrho \mid \pi_{d,I,M}) &= \mathrm{KL}(\varrho^1_{d,I,\eta} \otimes \varrho^2_{d,M,\gamma} \mid \mu_{d,I} \otimes \nu_{d,M}) \\ &= \mathrm{KL}(\varrho^1_{d,I,\eta} \mid \mu_{d,I}) + \mathrm{KL}(\varrho^2_{d,M,\gamma} \mid \nu_{d,M}) \\ &\leqslant \|I\|\log(5/\eta) + 16^d N^d 2^{dM}\log((C+1)/\gamma).\end{aligned}$$

□

6.3.3 Proof of Lemma 9

To simplify the notation we write $\varrho = \varrho_{d,I,M,\eta,\gamma}$ and $\pi_{d,I,M} = \mu_{d,I} \otimes \nu_{d,M}$. We will show that

$$\frac{\mathrm{d}\varrho}{\mathrm{d}\Pi} = G(d,I,M)\frac{\mathrm{d}\varrho}{\mathrm{d}\pi_{d,I,M}}, \tag{43}$$

from which we can deduce

$$\begin{aligned}\mathrm{KL}(\varrho \mid \Pi) = \int \log\left(\frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}\right) \mathrm{d}\varrho &= \log(G(d,I,M)) + \int \log\left(\frac{\mathrm{d}\varrho}{\mathrm{d}\pi_{d,I,M}}\right) \mathrm{d}\varrho \\ &= \log(G(d,I,M)) + \mathrm{KL}(\varrho \mid \pi_{d,I,M}).\end{aligned}$$

For (43), we need to show that

$$\varrho(A) = \int_A G(d,I,M)^{-1} \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}\, \mathrm{d}\pi_{d,I,M} \tag{44}$$

holds for all Borel-measurable sets $A = A_1 \times A_2 \subseteq \bigcup_{d=1}^{p} \mathcal{S}_d \times \mathcal{G}_d$. Observe that for

$$\begin{aligned}\mathcal{S}_{d,\Leftrightarrow}(J) &:= \{W = (w_1, \ldots, w_d)^\top \in \mathcal{S}_d \mid (w_{i,j} \neq 0 \Leftrightarrow j \in J_i)\, \forall i \in \{1, \ldots, d\}, \\ &\qquad j \in \{1, \ldots, p\}\} \\ \mathcal{G}_{d,\widetilde{M},\neq}(C+1) &:= \left\{g = \Phi_{d,\widetilde{M}}(\beta) \in \mathcal{G}_{d,\widetilde{M}}(C+1) \,\middle|\, \exists l \in \mathcal{Z}^d_{\widetilde{M}} : |l_2|_\infty = 2^{l_1}\widetilde{M},\ \beta_l \neq 0\right\}\end{aligned}$$

with $J = J_1 \times \cdots \times J_d \in \mathcal{I}_d$ and $\widetilde{M} \in \{0, \ldots, n\}$, we have

$$\mu_{d,J}(\mathcal{S}_{d,\Leftrightarrow}(J)) = \nu_{d,\widetilde{M}}(\mathcal{G}_{d,\widetilde{M},\neq}(C+1)) = 1. \tag{45}$$

In particular, (45) holds for $J = I$ and $\widetilde{M} = M$. Since also $\varrho(\mathcal{S}_{d,\Leftrightarrow}(I) \times \mathcal{G}_{d,M,\neq}(C+1)) = 1$, no generality is lost in additionally assuming that

$$A_1 \subseteq \mathcal{S}_{d,\Leftrightarrow}(I) \qquad \text{and} \qquad A_2 \subseteq \mathcal{G}_{d,M,\neq}(C+1).$$

Note that

$$\mathcal{S}_{d,\Leftrightarrow}(J) \cap \mathcal{S}_{d,\Leftrightarrow}(I) = \emptyset\, \forall J \neq I \quad \text{and} \quad \mathcal{G}_{d,\widetilde{M},\neq}(C+1) \cap \mathcal{G}_{d,M,\neq}(C+1) = \emptyset\, \forall \widetilde{M} \neq M. \tag{46}$$

Combining (45) with (46), we find

$$\int_{A_1} \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}(W,g)\, \mathrm{d}\mu_{d,J}(W) = 0$$

for any $g \in A_2$ and $J \neq I$. Similarly, we have for any $W \in A_1$ and $\widetilde{M} \neq M$ that

$$\int_{A_2} \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}(W,g)\, \mathrm{d}\nu_{d,\widetilde{M}}(g) = 0.$$

Therefore, repeated application of Fubini's theorem yields

$$\begin{aligned}\varrho(A) &= \int_A \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}\,\mathrm{d}\Pi = \sum_{c=1}^{p} 2^{-c} \int_A \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}\,\mathrm{d}\pi_c \Big/ (1-2^{-p}) \\ &= \int_A \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}\,\mathrm{d}\mu_d \otimes \nu_d \Big/ \big(2^d(1-2^{-p})\big) \\ &= \int_{A_1\times A_2} \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}(W,f)\,\mathrm{d}\mu_d \otimes \nu_d(W,g) \Big/ \big(2^d(1-2^{-p})\big) \\ &= G(d,I,M)^{-1} \int_{A_1\times A_2} \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}(W,g)\,\mathrm{d}\mu_{d,I} \otimes \nu_{d,M}(W,f) \\ &= G(d,I,M)^{-1} \int_A \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}\,\mathrm{d}\pi_{d,I,M}.\end{aligned}$$

Thus, we have shown (44). □

6.3.4 Proof of Lemma 10

We will show that

$$\mathrm{KL}(\varrho^{1,i}_{d,I,\eta} \mid \mu_{I_i}) \leqslant |I_i| \log(5/\eta), \qquad \forall i = 1,\dots,d. \tag{47}$$

The assertion follows immediately via

$$\mathrm{KL}(\varrho \mid \mu_{d,I}) = \mathrm{KL}\Big(\bigotimes_{i=1}^{d} \varrho^{1,i}_{d,I,\eta} \,\Big|\, \bigotimes_{i=1}^{d} \mu_{I_i}\Big) = \sum_{i=1}^{d} \mathrm{KL}(\varrho^{1,i}_{d,I,\eta} \mid \mu_{I_i}) \leqslant \|I\| \log(5/\eta),$$

where the first equality holds barring a slight breach of conventions for product measures. To show (47), we fix $i \in \{1,\dots,d\}$ and for simplicity of the notation, we set $\varrho := \varrho^{1,i}_{d,I,\eta}$ and $J = I_i$. Plugging the μ_J-density of ϱ into the definition of the Kullback–Leibler divergence, we easily obtain

$$\mathrm{KL}(\varrho \mid \mu_J) = -\log\Big(\int \mathbf{1}_{\{|w-w^*_{d,I,i}|\leqslant\eta\}}\mu_J(\mathrm{d}w)\Big) = -\log(\widetilde{\mu}(\{w \in \mathbb{R}^{|J|} \mid |w-\widetilde{w}^*| \leqslant \eta\})), \tag{48}$$

where $\widetilde{w}^*$ is the projection of $w^*_{d,I,i}$ onto the coordinates whose indices are elements of J and $\widetilde{\mu}_J$ denotes the uniform distribution on the unit sphere in $\mathbb{R}^{|J|}$.

We want to show a lower bound for $\widetilde{\mu}(\{w \in \mathbb{R}^{|J|} \mid |w-\widetilde{w}^*| \leqslant \eta\})$, which is the proportion of the surface of the unit sphere in $\mathbb{R}^{|J|}$ covered by the η-ball around $\widetilde{w}^*$ to the surface of the entire unit sphere. By rotational symmetry of the uniform distribution on this sphere, no generality is lost by assuming $\widetilde{w}^* = (1,0,\dots,0)^\top \in \mathbb{R}^{|J|}$.

For any $\widetilde{w} = (\widetilde{w}_1,\dots,w_{|J|})^\top \neq 0$ with $|\widetilde{w}| \leqslant 1$, the η-ball around $\widetilde{w}^*$ covers the same part of the surface of the unit sphere as the $\widetilde{\eta}$-ball around $\widetilde{w}$, where $\widetilde{\eta} = \sqrt{\eta^2|\widetilde{w}| - 2|\widetilde{w}| + |\widetilde{w}|^2 + 1}$. Using this dependence between $\widetilde{w}$, η and $\widetilde{\eta}$, it is easily checked that $|\widetilde{w}| = \sqrt{1-\widetilde{\eta}^2}$ is solved for $0 < \widetilde{\eta} = \sqrt{\eta^2 - \eta^4/4} < 1$. Henceforth, we fix this $\widetilde{\eta}$. Suppose that $N_{\widetilde{\eta}}$ is the smallest number of $\widetilde{\eta}$-balls with centers in the unit ball that suffice to cover the entire unit ball. Denote their centers by $t_1, \ldots, t_{N_{\widetilde{\eta}}}$. In particular, these balls cover the entire unit sphere and therefore, at least one of them covers at least $N_{\widetilde{\eta}}^{-1}$ of the surface of the unit sphere. This can only be the case for $t_j \neq 0$, because otherwise $\widetilde{\eta} < 1$ implies $\{w \in \mathbb{R}^{|J|} \mid |w - t_j| \leqslant \widetilde{\eta}\} \cap \{w \in \mathbb{R}^{|J|} \mid |w| = 1\} = \emptyset$. If we now change the length of such a $w := t_j \neq 0$ (without changing its orientation and without changing $\widetilde{\eta}$), the coverage of the unit sphere provided by the corresponding $\widetilde{\eta}$-ball also changes. In particular, we will show that if $|w| \neq \sqrt{1 - \widetilde{\eta}^2}$, then decreasing (or increasing) $|w|$ towards $\sqrt{1 - \widetilde{\eta}^2}$, enlarges the coverage of the corresponding ball on the unit sphere. Thus, the proportional coverage of the $\widetilde{\eta}$-ball around $|w|^{-1}\sqrt{1 - \widetilde{\eta}^2}w$ (which has, as we showed above, the same proportional coverage of the unit sphere as the η-ball around $\widetilde{w}^*$ that we are actually trying to control) is bounded from below by the proportional coverage of the $\widetilde{\eta}$-ball around $\widetilde{w}$, which in turn is bounded from below by $N_{\widetilde{\eta}}^{-1}$. Using the fact that $N_{\widetilde{\eta}} \leqslant (3/\widetilde{\eta})^{|J|}$ combined with $\widetilde{\eta} \geqslant \eta/\sqrt{2}$, we have

$$N_{\widetilde{\eta}}^{-1} \geqslant (3/\widetilde{\eta})^{-|J|} \geqslant (3\sqrt{2}/\eta)^{-|J|} \geqslant (5/\eta)^{-|J|}$$

and therefore (47) follows from (48) with

$$\mathrm{KL}(\varrho \mid \mu_J) = -\log(\widetilde{\mu}(\{w \in \mathbb{R}^{|J|} \mid |w - \widetilde{w}^*| \leqslant \eta\})) \leqslant -\log(N_{\widetilde{\eta}}^{-1}) \leqslant |J| \log(5/\eta).$$

It remains to show that changing the length of $w \neq 0$ towards $\sqrt{1 - \widetilde{\eta}^2}$ increases the proportional coverage of the corresponding $\widetilde{\eta}$-ball. By rotational symmetry, we can assume $w = (w_1, 0, \ldots, 0)^\top \in \mathbb{R}^{|J|}$ with some $0 < w_1 \leqslant 1$ at no loss of generality. Now, it is sufficient to show that

$$\{y \in \mathbb{R}^{|J|} \mid |y| = 1,\ |y - w| \leqslant \widetilde{\eta}\} \subseteq \{y \in \mathbb{R}^{|J|} \mid |y| = 1,\ |y - \widetilde{w}| \leqslant \widetilde{\eta}\},$$

where $\widetilde{w} = (\sqrt{1 - \widetilde{\eta}^2}, 0, \ldots, 0)^\top \in \mathbb{R}^{|J|}$. In this setting, and as $\eta, \widetilde{\eta} > 0$, the relationship

$$\widetilde{\eta}^2 = \eta^2 |\widetilde{w}| - 2|\widetilde{w}| + |\widetilde{w}|^2 + 1 \tag{49}$$

is equivalent to

$$\eta^2 = (2\widetilde{w}_1 - \widetilde{w}_1^2 - 1 + \widetilde{\eta}^2)/\widetilde{w}_1. \tag{50}$$

Using elementary calculus together with the fact that $\widetilde{\eta} < 1$, it is straightforward to see

$$(2w_1 - w_1^2 - 1 + \widetilde{\eta}^2)/w_1 \leqslant 2(1 - \sqrt{1 - \widetilde{\eta}^2}).$$

In combination with the relationship between η and $\widetilde{\eta}$, we obtain

$$\begin{aligned}\{y\in\mathbb{R}^{|J|}\mid |y-w|\leqslant\widetilde{\eta},\,|y|=1\} &= \{y\in\mathbb{R}^{|J|}\mid |y|=1,\,|y-\widetilde{w}^*|^2\\ &\qquad\leqslant(2w_1-w_1^2-1+\widetilde{\eta}^2)/w_1\}\\ &\subseteq\{y\in\mathbb{R}^{|J|}\mid |y|=1,\,|y-\widetilde{w}^*|^2\leqslant 2(1-\sqrt{1-\widetilde{\eta}^2})\}\\ &=\{y\in\mathbb{R}^{|J|}\mid |y|=1,\,|y-\widetilde{w}^*|^2\leqslant\eta^2\}\\ &=\{y\in\mathbb{R}^{|J|}\mid |y|=1,\,|y-\widetilde{w}|\leqslant\widetilde{\eta}\}.\end{aligned}$$

□

6.3.5 Proof of Lemma 11

To simplify the notation, we write $\widetilde{g}^*=g^*_{d,M}$ and $\widetilde{\beta}^*=\beta^*_{d,M}$. We will show that

$$\int\mathbf{1}_{\{\|g-\widetilde{g}^*\|_{\mathcal{B}}\leqslant\gamma\}}\mathrm{d}\nu_{d,M}(g)=((C+1)/\gamma)^{-|\mathcal{Z}^d_M|}. \tag{51}$$

The assertion follows directly with

$$\begin{aligned}\mathrm{KL}(\varrho^2_{d,M,\gamma}\mid\nu_{d,M})=-\log\Big(\int\mathbf{1}_{\{\|g-g^*\|_{\mathcal{B}}\leqslant\gamma\}}\mathrm{d}\nu_{d,M}(g)\Big)&=|\mathcal{Z}^d_M|\log((C+1)/\gamma)\\ &\leqslant 16^dN^d2^{dM}\log((C+1)/\gamma).\end{aligned}$$

We now verify (51) using the definition of $\nu_{d,M}$. If we let $\lambda^{\mathcal{Z}^d_M}$ denote the Lebesgue measure on $\mathbb{R}^{\mathcal{Z}^d_M}$, it holds that

$$\begin{aligned}\int\mathbf{1}_{\{\|g-\widetilde{g}^*\|_{\mathcal{B}}\leqslant\gamma\}}\,\mathrm{d}\nu_{d,M}(g) &= \int\mathbf{1}_{\{\|\Phi_{d,M}(\beta)-\widetilde{g}^*\|_{\mathcal{B}}\leqslant\gamma\}}\,\mathrm{d}\widetilde{\nu}_{d,M}(\beta)\\ &=\int\mathbf{1}_{\{\|\beta-\widetilde{\beta}^*\|_{\mathcal{B}}\leqslant\gamma\}}\,\mathrm{d}\widetilde{\nu}_{d,M}(\beta)\\ &=\frac{\int\mathbf{1}_{\{\|\beta\|_{\mathcal{B}}\leqslant C+1\}}\mathbf{1}_{\{\|\beta-\widetilde{\beta}^*\|_{\mathcal{B}}\leqslant\gamma\}}\,\mathrm{d}\lambda^{\mathcal{Z}^d_M}(\beta)}{\int\mathbf{1}_{\{\|\beta-\widetilde{\beta}^*\|_{\mathcal{B}}\leqslant C+1\}}\,\mathrm{d}\lambda^{\mathcal{Z}^d_M}(\beta)}\\ &=\frac{\int\mathbf{1}_{\{\|\beta-\widetilde{\beta}^*\|_{\mathcal{B}}\leqslant\gamma\}}\,\mathrm{d}\lambda^{\mathcal{Z}^d_M}(\beta)}{\int\mathbf{1}_{\{\|\beta-\widetilde{\beta}^*\|_{\mathcal{B}}\leqslant C+1\}}\,\mathrm{d}\lambda^{\mathcal{Z}^d_M}(\beta)}\\ &=\Big(\frac{\gamma}{C+1}\Big)^{|\mathcal{Z}^d_M|}\frac{\int\mathbf{1}_{\{\|\beta-\widetilde{\beta}^*\|_{\mathcal{B}}\leqslant 1\}}\,\mathrm{d}\lambda^{\mathcal{Z}^d_M}(\beta)}{\int\mathbf{1}_{\{\|\beta-\widetilde{\beta}^*\|_{\mathcal{B}}\leqslant 1\}}\,\mathrm{d}\lambda^{\mathcal{Z}^d_M}(\beta)}\\ &=\Big(\frac{C+1}{\gamma}\Big)^{-|\mathcal{Z}^d_M|},\end{aligned}$$

where we have used that

$$\|\beta\|_{\mathcal{B}}\leqslant\|\widetilde{\beta}^*\|_{\mathcal{B}}+\|\beta-\widetilde{\beta}^*\|_{\mathcal{B}}\leqslant C+\gamma\leqslant C+1$$

on $\{\|\beta - \widetilde{\beta}^*\|_{\mathcal{B}} \leqslant \gamma\}$ in the fourth equality. This implies that if the second indicator in the integral in the numerator is 1, so is the first. □

Acknowledgements The author would like to thank Pierre Alquier for initial remarks on the implementation and Mathias Trabs for helpful comments.

Funding Open Access funding enabled and organized by Projekt DEAL. This study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through project TR 1349/3-1. The author acknowledges support by the state of Baden-Württemberg through bwHPC. Large parts of this research were carried out while the author was affiliated to Universität Hamburg.

Declarations

Conflict of interest The author has no relevant financial interests to declare.

References

Alquier, P. (2024). User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends in Machine Learning, 17*(2), 174–303.

Alquier, P., & Biau, G. (2013). Sparse single-index model. *Journal of Machine Learning Research, 14*, 243–280.

Bieringer, S., Kasieczka, G., Steffen, M. F., & Trabs, M. (2025). The surrogate Gibbs-posterior of a corrected stochastic MALA: Towards uncertainty quantification for neural networks, *arXiv preprint* arXiv:2310.09335.

Catoni, O. (2004). *Statistical learning theory and stochastic optimization*. Springer.

Catoni, O. (2007). *PAC-Bayesian supervised classification: The thermodynamics of statistical learning*, volume 56 of *Lecture Notes-Monograph Series*. Institute of Mathematical Statistics.

Dalalyan, A. S., Juditsky, A., & Spokoiny, V. (2008). A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research, 9*, 1647–1678.

Daubechies, I. (1992). *Ten lectures on wavelets*, volume 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Giné, E. & Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models*. Number 40 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika, 82*(4), 711–732.

Guedj, B. (2019). A primer on PAC-Bayesian learning. In *Proceedings of the 2nd Congress of the French Mathematical Society* (Vol. 33, pp. 391–414).

Guedj, B., & Alquier, P. (2013). PAC-Bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics, 7*, 264–291.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction,* (Vol. 2). Springer.

Hristache, M., Juditsky, A., Polzehl, J., & Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *The Annals of Statistics, 29*(6), 1537–1566.

Klock, T., Lanteri, A., & Vigogna, S. (2021). Estimating multi-index models with response-conditional least squares. *Electronic Journal of Statistics, 15*(1), 589–629.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association, 86*(414), 316–327.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer.

Steffen, M. F. & Trabs, M. (2025). A PAC-Bayes oracle inequality for sparse neural networks. In *Springer Proceedings in Mathematics & Statistics*, to appear.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

van de Geer, S., Bühlmann, P., & Zhou, S. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electronic Journal of Statistics, 5*, 688–749.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics, 35*(6), 2654–2690.

Xia, Y., Tong, H., Li, W. K., & Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society. Series B. Statistical Methodology, 64*(3), 363–410.

Zhu, L., Miao, B., & Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association, 101*(474), 630–643.