Nonparametric Linear Feature Learning in Regression Through Regularisation

Bertille Follain, Francis Bach

Nonparametric Linear Feature Learning in Regression Through Regularisation

Bertille Follain

Inria, PSL Research University
48 rue Barrault, 75013 Paris, France
e-mail: bertille.follain@inria.fr

Francis Bach

Inria, PSL Research University
48 rue Barrault, 75013 Paris, France
e-mail: francis.bach@inria.fr

Abstract: Representation learning plays a crucial role in automated feature selection, particularly in the context of high-dimensional data, where non-parametric methods often struggle. In this study, we focus on supervised learning scenarios where the pertinent information resides within a lower-dimensional linear subspace of the data, namely the multi-index model. If this subspace were known, it would greatly enhance prediction, computation, and interpretation. To address this challenge, we propose a novel method for joint linear feature learning and non-parametric function estimation, aimed at more effectively leveraging hidden features for learning. Our approach employs empirical risk minimisation, augmented with a penalty on function derivatives, ensuring versatility. Leveraging the orthogonality and rotation invariance properties of Hermite polynomials, we introduce our estimator, named **RegFeaL**. By using alternative minimisation, we iteratively rotate the data to improve alignment with leading directions. We establish that the expected risk of our method converges in high-probability to the minimal risk under minimal assumptions and with explicit rates. Additionally, we provide empirical results demonstrating the performance of **RegFeaL** in various experiments.

MSC2020 subject classifications: Primary 62G08, 62F10; secondary 65K10.
Keywords and phrases: multi-index model, sparsity, non-parametric regression, regularised empirical risk minimisation, alternating minimisation, Hermite polynomials.

Contents

arXiv: https://arxiv.org/abs/2307.12754

1. Introduction

The increasing availability of high-dimensional data has created a demand for effective feature selection methods that can handle complex datasets. Representation learning, which aims to automate the feature selection process, plays a crucial role in extracting meaningful information from such data. However, non-parametric methods often struggle in high-dimensional settings.

A sensible approach is to consider that there are a lower number of unknown relevant linear features, or linear transformations of the original data, that explain the relationship between the response and factors. A popular way to model this is to consider the multi-index model [33], where we assume that the prediction function is the composition of few linear features which form a linear subspace (the effective dimension reduction (e.d.r.) subspace) and a non-parametric function. The multi-index model has been used in practice in many fields, such as ecology [26] or bio-informatics [1]. If the features were known, learning would be much easier due to the lower dimensionality of the problem, and their low number allows for a simpler, more explainable model, as well as a lesser need for computational and storage resources. Although these relevant features are not known a priori, recognising their existence enables the development of methods that incorporate them, potentially resulting in better estimators for prediction.

Related work. A wide range of methods have been proposed to estimate the e.d.r. space in the context of multi-index models. Brillinger introduced the method of moments, initially designed for Gaussian data and an e.d.r. of dimension one [8]. This method uses specific moments to eliminate the unknown function and focuses solely on the influence of the e.d.r. space. Extensions of this approach for distributions with differentiable log-densities have been provided, resulting in the average derivative estimation (ADE) method [27].

To incorporate subspaces of any dimension, several methods have been proposed. Slicing methods, such as slice inverse regression (SIR) [20], use second-order moments to account for subspaces. Principal Hessian directions (PHD) [21] extend the approach to elliptically symmetric data. Combining these techniques, sliced average derivative estimation (SADE) [3] offers a comprehensive approach. However, these methods heavily rely on assumptions about the distribution shape and require prior knowledge of the distribution, limiting their applicability.

Iterative improvements have been suggested for both the one-dimensional latent subspace case [16] and the general case [11]. Other optimisation-based methods, such as local averaging, aim to minimise an objective function to estimate the subspace [13, 34]. Although these procedures exhibit favourable performance in practice, particularly the **MAVE** method [34], the theoretical guarantees provided by [34] show exponential dependency in the dimension of the original data. Nonetheless, the recent work by [19] has made significant contributions to sufficient dimension reduction (SDR) by providing robust theoretical results for high-dimensional data that do not exhibit exponential dependency. However, their method, designed primarily for dimension reduction and variable selection in the specific setting of the square loss, relies on the linearity condition, which holds for example under the assumption that the covariates follow an elliptically contoured distribution.

In our work, we consider regularising the empirical risk by incorporating derivatives, a technique employed in various contexts. Classical splines, such as Sobolev spaces regularisation [31], have used derivative-based regularisation. More recently, derivative regularisation has been employed in the context of semi-supervised learning [9], as well as in linear subspace estimation using SADE [3].

Contributions. We propose a novel approach for joint function estimation and effective dimension reduction space estimation in multi-index models.

We employ the empirical risk minimisation framework, compatible with a wide range of loss functions, which is regularised by a penalty on the derivatives of the prediction function. The proposed regularisation enforces dependence on a reduced set of projected dimensions. Our method addresses the discussed limitations of previous methods. Indeed the assumptions on the distribution of the covariates are minimal (typically subgaussianity of the norm), and does not require said distribution to be known a priori. We are also able to provide explicit rates for the high-probability convergence of the expected risk of our estimator to the minimal risk, again with limited assumptions.

To construct our estimator, which we coin **RegFeaL**, we exploit the advantageous properties of Hermite polynomials, which exhibit orthogonality and rotation invariance. By incorporating alternative minimisation on a variational formulation of the problem, we enable iterative rotation of the data to better align with the leading directions, as well as easy computation of the unknown relevant dimension of the e d.r. space. Furthermore, for the specific case of the variable selection problem, that is, when only a subset of the coordinates of the original data is relevant, we can simplify our proposed penalty term which yields a computationally more efficient algorithm.

While our primary objective is to leverage the existence of a dependency on only a few variables or features, we also offer principled ways to estimate the dimension of the feature space and select the relevant features.

We provide detailed explanations about the efficient computation of our estimator, ensuring its practical usability. Additionally, we present theoretical results that establish the high-probability convergence to the minimal risk of the expected risk of our estimator, with limited assumptions on the loss and data distribution. This allows for a deeper understanding of the performance of the method and the dependency on certain parameters such as the dimension of the original data and the number of samples.

To demonstrate the strengths of our approach, we conduct an extensive set of experiments focusing on training behaviour, dependency on sample size and dimension, and comparison to other methods.

Importantly, our regularisation strategy is applicable to a wide range of problems where empirical risk can be formulated, making it a versatile tool for feature learning and dimensionality reduction tasks, potentially extending beyond statistics to fields such as signal processing and control.

In summary, our contributions encompass the introduction of a novel empirical risk minimisation framework with derivative-based regularisation for prediction and e.d.r. space estimation in multi-index models. We provide efficient computational techniques, theoretical insights, and empirical evidence, highlighting the advantages of our proposed method.

Paper organisation. The paper is organised as follows: we begin by describing the problem, our penalties, and the use of Hermite polynomials in Section 2. Then, we address the question of effectively computing our estimator **RegFeaL** in Section 3. In Section 4, we discuss the convergence of the empirical risk of our estimator. In Section 5, we present numerical studies to illustrate the behaviour of **RegFeaL**. Finally, in Section 6, we summarise our findings, highlight the contributions of our research, and discuss potential future directions.

Notations. Let $\mathbb{N}$ denote the set of non-negative integers and $\mathbb{N}^*$ the set of positive integers. For $d \in \mathbb{N}$, let $[d] = 1, \dots, d$. Given $x \in \mathbb{R}^d$ and $a \in [d]$, x_a represents the a-th component of x. Similarly, for $S \subset [d]$, x_S denotes $(x_a)_{a \in S}$. Let $p, d \in \mathbb{N}^*$, and consider a matrix $A \in \mathbb{R}^{p \times d}$. The matrix A_S corresponds to the columns of A extracted using indices from S, while $A_{i,j}$ represents the element of A in the j-th position of row i. The cardinality of a set S is denoted

by $|S|$. I_d represents the $d \times d$ identity matrix, and O_d denotes the set of $d \times d$ orthogonal matrices. For any $d \times d$ matrix A, $\mathrm{tr}(A)$ denotes its trace, and $\mathrm{Diag}(A)$ represents the diagonal matrix of size $d \times d$ with the diagonal elements of A. The transpose of a matrix B is denoted by $B^\top$. For an invertible matrix Λ, Λ^{-1} represents its inverse. Given $\eta \in \mathbb{R}^d$, $\mathrm{Diag}(\eta)$ is the diagonal matrix of size $d \times d$ with η as its diagonal. For $r > 0$, $\|\eta\|_r = \left(\sum_{a=1}^d |\eta_a|^r\right)^{1/r}$. For any $\alpha \in \mathbb{N}^d$, $|\alpha| = \sum_{a=1}^d \alpha_a$.

2. Preliminaries

2.1. Problem description

We consider a standard regression problem, where we have access to a dataset $(x^{(i)}, y^{(i)})_{i\in[n]}, n \in \mathbb{N}^*$ consisting of independent and identically distributed (i.i.d.) realisations of a pair of random variables (X, Y) with probability measure ν on $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$. Our objective is to estimate the regression function $f^* := \arg\min_{f\in\mathcal{F}} \mathcal{R}(f)$, where $\mathcal{R}(f) := \mathbb{E}_\nu(\ell(Y, f(X)))$ is the risk, ℓ is a loss function and $\mathcal{F}$ a space of functions from $\mathbb{R}^d$ to $\mathbb{R}$. At this stage, we do not impose any assumptions regarding the choice of loss function or the data distribution.

We consider the multi-index model [33], i.e., a model where the regression function depends on a low-rank linear transformation of the original variables.

Assumption 2.1 (Feature learning)**.** *We assume that the regression function f^* can be expressed as the combination of a rank s linear transformation P and a function g^* from $\mathbb{R}^s \to \mathbb{R}$, i.e.,*

$$\exists s \in [d],\ \exists P \in \mathbb{R}^{d\times s},\ P^\top P = I_s,\ \exists g^* : \mathbb{R}^s \to \mathbb{R},\ \forall x \in \mathbb{R}^d,\ f^*(x) = g^*(P^\top x).$$

We do not assume any prior knowledge about the value of s. The model is nonparametric hence it remains broad. Our objective is to simultaneously estimate both f^* and the associated linear transformation P, as well as the dimension s, by means of regularised empirical risk minimisation. Recall the definition of the empirical risk $\widehat{\mathcal{R}}(f) := \frac{1}{n}\sum_{i=1}^n \ell(y^{(i)}, f(x^{(i)}))$. This approach offers versatility, allowing its application to various scenarios. Although our focus lies on the regression setting, we acknowledge the potential of the regularisation-based method for future work in any setting where a risk can be defined.

2.2. Penalising by derivatives

In this context, it is common to employ derivative-based regularisation techniques [3, 24]. Under mild regularity assumptions, if we express f as $f = g(Q^\top \cdot)$ with $Q \in \mathbb{R}^{d\times s}$, then for all $x \in \mathbb{R}^d$, $\nabla f(x) \cdot \nabla f(x)^\top = Q\nabla g(x) \cdot \nabla g(x)^\top Q^\top$, where $\nabla f(x) \in \mathbb{R}^d$ denotes the gradient of f at point x. Consequently, we observe that

$$\int_{\mathcal{X}} \nabla f \nabla f^\top \nu = \left(\int_{\mathcal{X}} \frac{\partial f}{\partial x_a}\frac{\partial f}{\partial x_b}\nu\right)_{a,b\in[d]}$$

has a rank of at most s. This observation motivates us to employ the rank of $\int_{\mathcal{X}} \nabla f \nabla f^\top \nu$ as a penalisation. However, the discontinuous nature of the rank makes this approach challenging for optimisation. To address this, we could penalise instead by $\operatorname{tr}\left(\int_{\mathcal{X}} \nabla f \nabla f^\top \nu\right)$ as a convex relaxation [23].

This strategy would extend the work of [24], which focuses on variable selection, a special case of feature learning. It corresponds to the constraint that P from Assumption 2.1 only contains 0 and 1 (with exactly a single one in each column), resulting in a model where the regression function depends on a limited number of the original variables.

Assumption 2.2 (Variable selection)**.** *We assume that f^*, the regression function, actually only depends on s of the d variables, i.e.,*

$$\exists s \in [d],\ \exists S \subset [d],\ |S| = s,\ \exists\ g^* : \mathbb{R}^s \to \mathbb{R},\ \forall x \in \mathbb{R}^d,\ f^*(x) = g^*(x_S).$$

In this variable selection setting, we can remark that it suffices to penalise by a simpler quantity. Specifically, under some mild regularity assumptions on the function f, f does not depend on variable x_a if and only if the partial derivative of f with respect to x_a, denoted by $\frac{\partial f}{\partial x_a}$, is null everywhere on $\mathcal{X}$. Hence, the task is to design a penalty that enforces sparsity in the dependence on different variables.

To address this, we can draw inspiration from the group Lasso [35], which extends the Lasso method to enable structured sparsity. The group Lasso encourages groups of related quantities to be selected or excluded together by penalising the sum over each group using an appropriate penalty. For example, the derivatives with regard to a variable x_a at data points $x^{(i)}$ should all be null if the function does not depend on variable x_a. Hence, they constitute a relevant group for group Lasso.

Combining these observations, [24] proposed a strategy using the fact that for all $a \in [d]$, f does not depend on x_a if and only if $\int_{\mathcal{X}} \left(\frac{\partial f}{\partial x_a}(x)\right)^2 \nu = 0$. They introduced penalties on each variable and summed them to obtain the penalty $\sum_{a=1}^d \left(\int_{\mathcal{X}} \left(\frac{\partial f}{\partial x_a}(x)\right)^2 \nu(x)\mathrm{d}x\right)^{1/2}$. However, since these quantities are intractable due to the unknown nature of ν, they use a data-dependent penalty instead

$$\sum_{a=1}^{d} \left(\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_a}(x^{(i)}) \right)^2 \right)^{1/2}.$$

By assuming that f belongs to some regular reproducing kernel Hilbert space (RKHS), the partial derivatives are easily computable, and so is the penalty [24] (for a good introduction to RKHS, see [2]). However, this regularisation by an estimate of the L^2 norms of derivatives in the context of RKHS is not suitable. Functions that depend on a single variable, such as x_1, do not belong to the RKHS, making it an inappropriate space for addressing this type of problem. Additionally, another regularisation by the norm in the RKHS is required, introducing an extra hyperparameter. Moreover, using derivatives only at the data points limits the exploitation of the power of regularity.

We are confronted with two challenges here. First, how can the penalisation scheme be improved for variable selection? Second, how can it be adapted for feature learning? While our primary goal is the latter, we consider the former as a by-product of our methodology.

To address both challenges, we employ Hermite polynomials [15], although it is worth noting that various other alternatives could have been considered for the first problem where rotation invariance is not needed.

2.3. Hermite polynomials for variable selection

To facilitate understanding, let us first consider the simpler case of variable selection. We employ multidimensional Hermite polynomials due to their suitability for both variable selection and feature learning. The normalised one-dimensional Hermite polynomials $(h_k(x))_{k\geq 0}$ form an orthonormal polynomial basis for the standard Gaussian measure on $\mathbb{R}$ with density $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. The first few polynomials are given by[1]

$$h_0(x) = 1,\ h_1(x) = x,\ h_2(x) = \frac{1}{\sqrt{2}}(x^2-1),\ h_3(x) = \frac{1}{\sqrt{6}}(x^3-3x).$$

These polynomials possess useful properties that allow their recursive computation and characterise their growth and their derivatives

$$h_{n+2}(x) = \frac{x}{\sqrt{n+2}}\cdot h_{n+1}(x) - \sqrt{\frac{n+1}{n+2}}\cdot h_n(x) \tag{2.1}$$

$$h_n'(x) = \sqrt{n}\cdot h_{n-1}(x) \tag{2.2}$$

$$|h_n(x)| \leq \exp(x^2/4). \tag{2.3}$$

The last property can be proved using Hermite functions and Cramer's inequality [28].

Next, we define the multivariate polynomials as follows

$$\big(H_\alpha\big)_{\alpha\in\mathbb{N}^d} \text{ where } \forall x\in\mathbb{R}^d,\ H_\alpha(x) = \prod_{a=1}^{d} h_{\alpha_a}(x_a). \tag{2.4}$$

This family forms an orthonormal basis of the space $L^2(q) := \big\{f : \mathbb{R}^d \to \mathbb{R},\ \int_{\mathbb{R}^d} f^2 q < +\infty\big\}$ where $q(x) = \frac{1}{(2\pi)^{d/2}}e^{-\|x\|^2/2}$ denotes the standard normal distribution on $\mathbb{R}^d$. We now present a Lemma which justifies the use of the multivariate Hermite polynomials in the variable selection setting.

Lemma 2.1 (Equivalence for dependency on variables)**.** *Let $f \in L^2(q)$ and express it as $f = \sum_{\alpha\in\mathbb{N}^d}\hat{f}(\alpha)H_\alpha$. Then for any $b\in[d]$,*

$$f \text{ does not depend on variable } x_b \iff \forall\alpha\in\big(\mathbb{N}^d\big)^*,\ \alpha_b\neq 0 \implies \hat{f}(\alpha) = 0.$$

[1] Given the regular "physicist" Hermite polynomials H_k (not to be confused with multivariate polynomials defined in Equation (2.4)), we have $h_k(x) = \frac{1}{\sqrt{2^k k!}}H_k(x/\sqrt{2})$ for any $k\in\mathbb{N}$ and for the "probabilist" Hermite polynomials He_k, we have $h_k(x) = \frac{1}{\sqrt{n!}}He_k(x)$.

Proof of Lemma 2.1. For $x \in \mathbb{R}^d$, we have $h_0(x) = 1$ and

$$f(x) = \underbrace{\hat{f}(0) + \sum_{\alpha \in (\mathbb{N}^d)^*,\ \alpha_b = 0} \hat{f}(\alpha) \prod_{a \in [d] \setminus \{b\}} h_{\alpha_a}(x_a)}_{\text{does not depend on } x_b} + \underbrace{\sum_{\alpha \in (\mathbb{N}^d)^*,\ \alpha_b > 0} \hat{f}(\alpha) \prod_{a \in [d]} h_{\alpha_a}(x_a)}_{\text{depends on } x_b},$$

i.e., f can be decomposed into two additive components, one of which does not depend on x_b. For the component that depends on x_b, it is the sum over $\alpha \in \mathbb{N}^d$ such that α_b is nonzero, yielding the result. □

We observe that when f does not depend on a variable, it corresponds to a specific sparsity pattern in the coefficients $\hat{f}(\alpha)$ with respect to the basis $(H_\alpha)_{\alpha \in \mathbb{N}^d}$. Indeed, if f does not depend on x_b, all coefficients $\hat{f}(\alpha)$ for α in the group $\{\alpha \in \left(\mathbb{N}^d\right)^*, \alpha_b > 0\}$ must be null. These groups overlap for different variables, and a similar argument holds for feature learning as we will see in Section 2.4. This specific sparsity pattern motivates the use of a penalty based on group Lasso [35], and more specifically overlapping group Lasso [17].

Hence, the Hermite polynomial basis is well-suited to this variable selection setting, while the space $L^2(q)$ is sufficiently large to describe a wide range of functions. However, it is worth noting that other spaces and well-adapted bases, such as any orthonormal basis of square-integrable functions, could also be used. Moreover, we use the Gaussian measure only to define the basis, and our method can be applied to all distributions.

To define a penalty relevant to variable selection, we examine the derivatives of H_α. Here, we decompose any $f \in \mathcal{F}$ as $f = \sum_{\alpha \in \mathbb{N}^d} \hat{f}(\alpha) H_\alpha$. Let e_a denote the a-th element of the canonical basis of $\mathbb{R}^d$, for $a \in [d]$. Using Equation (2.2), we obtain the following identities

$$\frac{\partial H_\alpha}{\partial x_a} = \sqrt{\alpha_a} H_{\alpha - e_a} \tag{2.5}$$

$$\frac{\partial f}{\partial x_a} = \sum_{\alpha \in (\mathbb{N}^d)^*} \sqrt{\alpha_a} \hat{f}(\alpha) H_{\alpha - e_a} \tag{2.6}$$

$$\int_{\mathbb{R}^d} \left(\frac{\partial f}{\partial x_a} \right)^2 q = \sum_{\alpha \in (\mathbb{N}^d)^*} \alpha_a \hat{f}(\alpha)^2. \tag{2.7}$$

However, we remark that Equation (2.7) corresponds to the expected version of the penalty proposed by [24] (when $\nu = q$), which we deemed not suitable for our problem: indeed, penalising the L^2-norm of derivatives does not impose enough regularity for statistically efficient non-parametric estimation and thus requires extra regularisation, as specified by [24].

We consider instead introducing a sequence $(c_k)_{k>0}$ of non-negative reals, to further regularise and avoid the need for additional regularisation. We consider

the space $\mathcal{F}$, spanned by the family composed of H_α for $\alpha = 0$ or $\alpha \in (\mathbb{N}^d)$ such that $c_{|\alpha|} > 0$, i.e., $\mathcal{F} := \mathrm{Span}\left(\{H_0\} \cup \{H_\alpha, \text{ for } \alpha \in (\mathbb{N}^d)^* \text{ such that } c_{|\alpha|>0}\}\right)$ and consider two penalties. First, we define a sparsity-inducing penalty, which depends on a hyper-parameter $r \in (0, +\infty)$

$$\Omega_{\mathrm{var}}(f) = \left(\sum_{a=1}^{d}\left(\sum_{\alpha\in(\mathbb{N}^d)^*} \alpha_a \frac{1}{c_{|\alpha|}}\hat{f}(\alpha)^2\right)^{r/2}\right)^{1/r}.$$

This penalty encourages sparsity in the dependence of f on individual variables, as it pushes quantities of the form $\left(\sum_{\alpha\in(\mathbb{N}^d)^*} \alpha_a \frac{1}{c_{|\alpha|}}\hat{f}(\alpha)^2\right)^{r/2}$ to be 0. When this is the case, we obtain that $\forall \alpha \in (\mathbb{N}^d)^*,\ \alpha_a \neq 0, \hat{f}(\alpha) = 0$, i.e., f does not depend on variable x_a (Lemma 2.1). When $r \geq 1$, Ω_{var} is a norm, which makes the problem easier to study from a theoretical point of view because if the loss is convex, this will yield a convex optimisation problem. However, estimators obtained through regularised empirical risk minimisation often suffer from bias due to the strong shrinkage associated with sparsity. Convex penalties can inadvertently reduce the significance of essential variables or features by excessive shrinkage to enforce sparsity. To address these issues, one can retrain on the set of selected variables or use concave penalties, which, despite presenting more analytical challenges, frequently deliver superior results by pushing the solution towards the boundary and enhancing sparsity [36, 5]. In this work, we adopt this strategy through the hyper-parameter r when $r < 1$, which is the choice used in practice, while $r = 1$ is used in the theoretical analysis.

The link with the nullity of the derivative can be seen using Equation (2.7)

$$\left(\sum_{\alpha\in(\mathbb{N}^d)^*} \alpha_a \frac{1}{c_{|\alpha|}}\hat{f}(\alpha)^2\right)^{r/2} = 0 \iff \int_{\mathbb{R}^d}\left(\frac{\partial f}{\partial x_a}\right)^2 q = 0.$$

Next, we introduce a smoothness-inducing norm, which penalises higher-order polynomials, i.e., those with large $|\alpha|$ (the dependence only on $|\alpha|$ is needed for future rotation invariance)

$$\Omega_0(f) = \left(\sum_{\alpha\in(\mathbb{N}^d)^*} \frac{1}{c_{|\alpha|}}\hat{f}(\alpha)^2\right)^{1/2}.$$

It is important to note that Ω_0 is not integrated into the theoretical analysis and will be used with a much smaller and fixed parameter compared to Ω_{var}. Its primary purpose is to enforce numerical stability during the optimisation procedure, as discussed in Section 3.

The choice of $(c_k)_{k\in\mathbb{N}^*}$ significantly influences the behaviour of the penalties. In this work, we will consider two specific choices: $c_k = \mathbb{1}_{k\leq M}$ for some $M \in \mathbb{N}$ and $c_k = \rho^k$ for some $\rho \in [0, 1)$. Both choices ensure that all three penalties are well-defined. Notably, when $M = 1$, Ω_{var} considered with the quadratic loss reduces to the basic Lasso problem with linear features [29].

It is worth mentioning that the coefficient $\hat{f}(0)$, which corresponds to the constant function $H_0 = 1$, is never penalised because it does not depend on any of the variables.

We then consider estimating f^* in the setting described in Assumption 2.2 by

$$f_{\text{var}}^{\lambda,\mu} := \arg\min_{f \in \mathcal{F}} \ \widehat{\mathcal{R}}(f) + \lambda \Omega_0^2(f) + \mu \Omega_{\text{var}}^r(f), \tag{2.8}$$

with λ a fixed parameter and μ a hyper-parameter to be estimated. When $r \geq 1$ and the loss is convex, we obtain a strongly-convex objective function, hence with a unique global minimiser. When $r < 1$, which we use in practice, only a local minimiser can be reached.

2.4. Hermite polynomials for feature learning

We now turn to the feature learning setting described in Assumption 2.1. The Hermite polynomials are particularly well-suited for feature learning, as they allow us to bridge the gap between variable selection and feature learning with only a minor modification of the previous penalties. This suitability is visible in some important properties which we now describe. First, the multivariate Hermite polynomials possess a rotation invariance property.

Lemma 2.2 (Rotational invariance property of Hermite polynomials)**.** *For any $x, x' \in \mathbb{R}^d$, any $k \in \mathbb{N}$ and any orthogonal matrix $R \in O_d$,*

$$\sum_{|\alpha|=k} H_\alpha(x) H_\alpha(x') = \sum_{|\alpha|=k} H_\alpha(Rx) H_\alpha(Rx').$$

The proof of this lemma is available in Appendix A.1. This property will be extremely useful to characterise the statistical behaviour of our methods, as discussed in Section 4. Another key property is that for any $R \in O_d$, the family $\left(H_\alpha(R\cdot)\right)_{\alpha \in \mathbb{N}^d}$ also forms a basis of $L^2(q)$. Consequently, we can express any $f \in \mathcal{F}$ in this basis.

Moreover, we can characterise the derivatives of functions in $L^2(q)$ as in Equation (2.7). Let $f \in \mathcal{F}$ be written as $f = \sum_{\alpha \in \mathbb{R}^d} \hat{f}(\alpha) H_\alpha$, then using Equation (2.6), we have the following expressions for the derivatives

$$\int_{\mathbb{R}^d} \left(\frac{\partial f}{\partial x_a}\right)\left(\frac{\partial f}{\partial x_b}\right) q = \sum_{\alpha \in \mathbb{N}^d} \sqrt{(\alpha_a + 1)}\sqrt{(\alpha_b + 1)}\, \hat{f}(\alpha + e_a)\hat{f}(\alpha + e_b). \tag{2.9}$$

As before, we aim to enhance the regularisation using the sequence $(c_k)_{k>0}$ For $r \in (0, +\infty)$, we define

$$\begin{aligned} \Omega_{\text{feat}}(f) &= \left(\operatorname{tr}\left(M_f^{r/2}\right)\right)^{1/r} \\ \text{with } (M_f)_{a,b} &= \sum_{\alpha \in \mathbb{N}^d} \frac{1}{c_{|\alpha|+1}} \sqrt{\alpha_a + 1}\sqrt{\alpha_b + 1}\, \hat{f}(\alpha + e_a)\hat{f}(\alpha + e_b), \ a, b \in [d]. \end{aligned} \tag{2.10}$$

It is worth noting that M_f is a positive semi-definite matrix (see the proof of Lemma 2.3). The penalty Ω_{feat} pushes the eigenvalues of of M_f towards 0, and since the rank of M_f is equal to the number of its non-zero eigenvalues, the penalty encourages the rank of M_f to be low. It is crucial that $c_{|\alpha|}$ depends solely on $|\alpha|$ and not on any other quantities depending on α (e.g., $\max_{a\in[d]} \alpha_a$ for example). This property allows us to leverage the rotation invariance property described in Lemma 2.2, which is needed for our estimation algorithm in Section 3 and for obtaining statistical consistency results in Section 4.

Let us now examine some important properties of the proposed regularisation.

Lemma 2.3 (Properties of the regularisation)**.** *For any $f \in \mathcal{F}$, the following properties hold*

1. *Let $R \in O_d$, if we define $g = f(R\cdot)$, then $M_f = RM_gR^\top$ and $\Omega_{\text{feat}}(f) = \Omega_{\text{feat}}(g)$.*
2. *$\Omega_{\text{var}}(f) = \big(\operatorname{tr}\big(\operatorname{Diag}(M_f)^{r/2}\big)\big)^{1/r}$.*
3. *If M_f is diagonal, $\Omega_{\text{feat}}(f) = \Omega_{\text{var}}(f)$.*
4. *Let $M_f = UDU^\top$ be the eigendecomposition of M_f, where $U \in O_d$ and D is a diagonal matrix. If we define $g = f(U\cdot)$, then $M_g = D$ is diagonal and thus $\Omega_{\text{feat}}(f) = \Omega_{\text{var}}(g)$.*
5. *Let $M_f = UDU^\top$ be the eigendecomposition as above. If the rank of D is s, then $g = f(U\cdot)$ only depends on variables x_a where $D_a > 0$ and $f = g(U^\top\cdot)$ only depends on s linear transformations of the original coordinates, namely of $(U^\top x)_a$ for a such that $D_a > 0$.*
6. *If $r = 1$,*
$$\Omega_{\text{feat}}(f) \geq \inf_{R\in O_d} \Omega_{\text{var}}(f(R\cdot)).$$

Proof of Lemma 2.3. We proceed by proving each assertion separately.

1. We have for $z \in \mathbb{R}^d$
$$\begin{aligned} z^\top M_f z &= \sum_{a,b=1}^d \sum_{\alpha\in\mathbb{N}^d} \frac{1}{c_{|\alpha|+1}} z_a z_b \sqrt{\alpha_a+1}\sqrt{\alpha_b+1}\hat{f}(\alpha+e_a)\hat{f}(\alpha+e_b) \\ &= \sum_{a,b=1}^d \sum_{\alpha\in\mathbb{N}^d} \frac{1}{c_{|\alpha|+1}} z_a z_b \Big\langle \frac{\partial f}{\partial x_a}, H_\alpha \Big\rangle_{L^2(q)} \Big\langle \frac{\partial f}{\partial x_b}, H_\alpha \Big\rangle_{L^2(q)} \\ &= \sum_{\alpha\in\mathbb{N}^d} \frac{1}{c_{|\alpha|+1}} \big\langle z^\top \nabla f, H_\alpha \big\rangle^2_{L^2(q)}. \end{aligned}$$

This shows that M_f is positive semi-definite, writing $\mathcal{N}(0, I_d)$ for the stan-

dard normal distribution on $\mathbb{R}^d$, we then have

$$\begin{aligned}
z^\top M_g z &= \sum_{\alpha\in\mathbb{N}^d} \frac{1}{c_{|\alpha|+1}} \Big(\mathbb{E}_{X\sim\mathcal{N}(0,I_d)}\big(z^\top \nabla g(X) H_\alpha(X)\big)\Big)^2 \\
&= \sum_{\alpha\in\mathbb{N}^d} \frac{1}{c_{|\alpha|+1}} \Big(\mathbb{E}_{X\sim\mathcal{N}(0,I_d)}\big(z^\top R^\top \nabla f(RX) H_\alpha(X)\big)\Big)^2 \\
&\quad \text{as } \nabla g(X) = R^\top \nabla f(RX) \\
&= \sum_{\alpha\in\mathbb{N}^d} \frac{1}{c_{|\alpha|+1}} \Big(\mathbb{E}_{X\sim\mathcal{N}(0,I_d)}\big(z^\top R^\top \nabla f(RX) H_\alpha(RX)\big)\Big)^2 \\
&\quad \text{by Lemma 2.2,} \\
&= \sum_{\alpha\in\mathbb{N}^d} \frac{1}{c_{|\alpha|+1}} \Big(\mathbb{E}_{X\sim\mathcal{N}(0,I_d)}\big(z^\top R^\top \nabla f(X) H_\alpha(X)\big)\Big)^2 \\
&\quad \text{by rotation invariance of the standard Gaussian,} \\
&= z^\top R^\top M_f R z,
\end{aligned}$$

that is $M_g = R^\top M_f R$. The second assertion follows by the rotation invariance of the trace.

2. It suffices to see that for any $a \in [d]$

$$\mathrm{Diag}(M_f)_{a,a} = \sum_{\alpha\in\mathbb{N}^d} \frac{1}{c_{|\alpha|+1}} (\alpha_a+1)^2 \hat{f}(\alpha+e_a)^2 = \sum_{\alpha\in\mathbb{N}^d,\alpha_a>0} \frac{1}{c_{|\alpha|}} \alpha_a \hat{f}(\alpha)^2,$$

and therefore

$$\mathrm{tr}\left(\mathrm{Diag}(M_f)^{r/2}\right) = \sum_{a=1}^d \left(\sum_{\alpha\in\mathbb{N}^d} \alpha_a \frac{1}{c_{|\alpha|}} \hat{f}(\alpha)^2\right)^{r/2} = \Omega_{\mathrm{var}}(f)^r.$$

3. This is a direct consequence of the previous result, because of the definition of Ω_{feat}.
4. By applying the first result, we find that $\Omega_{\mathrm{feat}}(f) = \Omega_{\mathrm{feat}}(g)$ and $M_g = D$. Then, using the third result, we conclude that $\Omega_{\mathrm{var}}(g) = \Omega_{\mathrm{feat}}(g)$. This establishes the desired result.
5. Consider the function $g = f(U\cdot)$. From the previous result, we know that $M_g = D$ is diagonal. According to the definition of Ω_{var}, we have $D_a = 0$ if and only if g does not depend on variable x_a. Consequently, if the rank of D is s, then g only depends on s variables, specifically those for which $D_a > 0$. As a result, we can conclude that $f = g(U^\top \cdot)$ depends solely on $(U^\top x)_a$ for a such that $D_a > 0$.
6. Let us examine Ω_{feat} and Ω_{var} as follows

$$\Omega_{\mathrm{feat}}(f) = \left(\mathrm{tr}\left(M_f^{1/2}\right)\right), \qquad \Omega_{\mathrm{var}}(f) = \left(\mathrm{tr}\left(\mathrm{Diag}(M_f)^{1/2}\right)\right).$$

We can decompose M_f as $M_f = UDU^\top$ using its eigendecomposition. If we define $g = f(U\cdot)$, then $M_g = D$ is diagonal, and we have $\Omega_{\mathrm{feat}}(f) =$

$\Omega_{\text{feat}}(g) = \Omega_{\text{var}}(g)$. Consequently, we obtain the inequality

$$\Omega_{\text{feat}}(f) \geq \inf_{R \in O_d} \Omega_{\text{var}}(f(R\cdot)).$$

□

The rotation invariance of Ω_{feat} is crucial in the context of feature learning, as it ensures that the penalty is not biased towards specific directions. Similarly, Ω_0 is also rotation invariant, as can be seen using Lemma 2.2.

We observe that given a function f and its associated matrix M_f, we can construct a function g consisting of a rotation of the data and f in such a way that the feature penalty on f is equal to the variable selection penalty on g. This highlights that the feature learning setting extends the variable selection problem by allowing data rotation. Furthermore, we can easily determine if g depends only on a few variables, and therefore if f depends only on a few linear transformations of the data, which aligns with our assumption for f^*. The last assertion of Lemma 2.3 will be useful to show that the proof of the consistency for the variable penalty easily extends to the feature learning setting, see Section 4.

With these considerations, we proceed to estimate f^* in the setting described by Assumption 2.1 by solving

$$f_{\text{feat}}^{\lambda,\mu} := \arg\min_{f \in \mathcal{F}} \ \widehat{\mathcal{R}}(f) + \lambda \Omega_0^2(f) + \mu \Omega_{\text{feat}}^r(f), \tag{2.11}$$

with λ a fixed parameter and μ a hyper-parameter. We refer to this estimator as the **RegFeaL** (**R**egularised **F**eature **L**earning) estimator. As for the relevant features or variables and dimension, we discuss their computation in Section 3.1.

3. Estimator computation

The computation of the solution for the optimisation problems delineated by (2.8) and (2.11) requires the employment of several strategic methodologies, which we will now discuss.

3.1. Variational formulation

We first use the following quadratic variational formulation, similar to the approach presented in [5]. This formulation is necessary since it is not possible to directly optimise Equation (2.8) and Equation (2.11) due to the absence of closed-form solutions. Using other classical optimisation methods such as gradient-based methods would be less efficient as the overlapping group Lasso penalty we propose does not have efficient projection algorithms. Indeed, the variational formulation allows us to rewrite our optimisation problems as the minimisation over two variables of a specific quantity. Subsequently, we can alternate the minimisation with respect to each variable, leading to rapid convergence in practice.

We first give the following Lemma which is adapted from [18], which provides a variational formulation of sums of powers.

Lemma 3.1 (Variational formulation). *Let $r \in (0,2)$ and $u \in \mathbb{R}_+^d$, then*

$$\|u\|_{r/2}^{r/2} = \left(\sum_{a=1}^d u_a^{r/2}\right) = \min_{\eta \in \mathbb{R}_+^d,\ \|\eta\|_{r/(2-r)}=1} \sum_{a=1}^d \frac{u_a}{\eta_a},$$

with minimum attained at $\eta, \forall a \in [d],\ \eta_a = u_a^{(2-r)/2} / \left(\sum_{b=1}^d u_b^{r/2}\right)^{(2-r)/r}$.

Now, let us apply this approach to the penalty used for variable selection.

Lemma 3.2 (Variational formulation of variable selection penalty). *Let $f = \sum_{\alpha \in \mathbb{N}^d} \hat{f}(\alpha) H_\alpha \in \mathcal{F}$ and $r \in (0,2)$, then*

$$\begin{aligned}\Omega_{\mathrm{var}}^r(f) &= \min_{\eta \in \mathbb{R}_+^d,\ \|\eta\|_{r/(2-r)}=1} \sum_{a=1}^d \left(\sum_{\alpha \in (\mathbb{N}^d)^*} \alpha_a \frac{1}{c_{|\alpha|}} \hat{f}(\alpha)^2 \right) \eta_a^{-1} \\ &= \min_{\eta \in \mathbb{R}_+^d,\ \|\eta\|_{r/(2-r)}=1} \left(\sum_{\alpha \in (\mathbb{N}^d)^*} \alpha^\top \eta^{-1} \frac{1}{c_{|\alpha|}} \hat{f}(\alpha)^2 \right),\end{aligned}$$

where $\eta^{-1} = (1/\eta_1, \dots, 1/\eta_d)$ and where the minimum is reached for η such that

$$\forall a \in [d],\ \eta_a = \frac{\left(\sum_{\alpha \in (\mathbb{N}^d)^*} \alpha_a \frac{1}{c_{|\alpha|}} \hat{f}(\alpha)^2\right)^{(2-r)/2}}{\left(\sum_{b=1}^d \left(\sum_{\alpha \in (\mathbb{N}^d)^*} \alpha_b \frac{1}{c_{|\alpha|}} \hat{f}(\alpha)^2\right)^{r/2}\right)^{(2-r)/r}}. \tag{3.1}$$

Proof of Lemma 3.2. Recall $\Omega_{\mathrm{var}}(f) = \left(\sum_{a=1}^d \left(\sum_{\alpha \in (\mathbb{N}^d)^*} \frac{\alpha_a}{c_{|\alpha|}} \hat{f}(\alpha)^2\right)^{r/2}\right)^{1/r}$ and use Lemma 3.1 with $u_a = \sum_{\alpha \in (\mathbb{N}^d)^*} \alpha_a \frac{1}{c_{|\alpha|}} \hat{f}(\alpha)^2$. □

We can then rewrite (2.8) as

$$\begin{aligned} f_{\mathrm{var}}^{\lambda,\mu},\ \eta_{\mathrm{var}}^{\lambda,\mu} = \operatorname*{arg\,min}_{f \in \mathcal{F},\ \eta \in \mathbb{R}_+^d} \ & \widehat{\mathcal{R}}(f) + \sum_{\alpha \in (\mathbb{N}^d)^*} \frac{1}{c_{|\alpha|}} \hat{f}(\alpha)^2 (\lambda + \mu \alpha^\top \eta^{-1}) \\ \text{subject to} \quad & f = \sum_{\alpha \in \mathbb{N}^d} \hat{f}(\alpha) H_\alpha, \quad \|\eta\|_{r/(2-r)} = 1. \end{aligned} \tag{3.2}$$

Recall that $\Omega_{\mathrm{var}}(f) = \left(\sum_{a=1}^d \left(\sum_{\alpha \in (\mathbb{N}^d)^*} \alpha_a \frac{1}{c_{|\alpha|}} \hat{f}(\alpha)^2\right)^{r/2}\right)^{1/2}$. Each term $\left(\sum_{\alpha \in (\mathbb{N}^d)^*} \alpha_a \frac{1}{c_{|\alpha|}} \hat{f}(\alpha)^2\right)^{r/2}$ quantifies the dependency of f on the variable x_a. We then remark from the definition of $\eta_{\mathrm{var}}^{\lambda,\mu}$ in Equation (3.1), that

$$\forall a \in [d],\ \left(\eta_{\mathrm{var}}^{\lambda,\mu}\right)_a^{r/(2-r)} = \frac{\left(\sum_{\alpha \in (\mathbb{N}^d)^*} \alpha_a \frac{1}{c_{|\alpha|}} \hat{f}_{\mathrm{var}}^{\lambda,\mu}(\alpha)^2\right)^{r/2}}{\sum_{b=1}^d \left(\sum_{\alpha \in (\mathbb{N}^d)^*} \alpha_b \frac{1}{c_{|\alpha|}} \hat{f}_{\mathrm{var}}^{\lambda,\mu}(\alpha)^2\right)^{r/2}}.$$

Hence $\left(\eta_{\mathrm{var}}^{\lambda,\mu}\right)_a$ represents the variation of $f_{\mathrm{var}}^{\lambda,\mu}$ which is due to x_a. We can use $\eta_{\mathrm{var}}^{\lambda,\mu}$ to estimate the relevant underlying variables by using conventional techniques

such as thresholding. Specifically, we can consider a variable x_a to be relevant only if η_a is above some predetermined threshold, i.e. $\hat{S} := \{a \in [d],\ \left(\eta_{\text{var}}^{\lambda,\mu}\right)_a > t\}$ for some $t > 0$.

We can proceed in a similar manner for the feature learning setting.

Lemma 3.3 (Variational formulation of feature learning penalty)**.** *Let $f \in \mathcal{F}$, M_f from Equation (2.10), with $M_f = UDU^\top$ its eigendecomposition and $r \in (0,2)$, then*

$$\begin{aligned}\Omega_{\text{feat}}^r(f) = \min_{\Lambda \in \mathbb{R}^{d\times d}} &\operatorname{tr}\left(\Lambda^{-1} M_f\right)\\ \text{subject to } &\Lambda = R\operatorname{Diag}(\eta)R^\top\\ &R \in O_d,\ \eta \in \mathbb{R}_+^d,\ \|\eta\|_{r/(2-r)} = 1,\end{aligned}$$

where the minimum is attained for

$$\begin{aligned}\Lambda &= U\operatorname{Diag}(\eta)U^\top \\ \forall a \in [d],\ \eta_a &= \frac{D_a^{(2-r)/2}}{(\sum_{b=1}^d D_b^{r/2})^{(2-r)/r}}.\end{aligned} \tag{3.3}$$

This allows us to rewrite Equation (2.11) as

$$\begin{aligned}f_{\text{feat}}^{\lambda,\mu},\ \Lambda_{\text{feat}}^{\lambda,\mu} = \operatorname*{arg\,min}_{f\in\mathcal{F},\ \Lambda\in\mathbb{R}^{d\times d}} &\widehat{\mathcal{R}}(f) + \lambda\Omega_0^2(f) + \mu\operatorname{tr}\left(\Lambda^{-1}M_f\right)\\ \text{subject to } &\Lambda = R\operatorname{Diag}(\eta)R^\top\\ &R \in O_d,\ \eta \in \mathbb{R}_+^d,\ \|\eta\|_{r/(2-r)} = 1.\end{aligned}$$

Moreover, with $\Lambda = R\operatorname{Diag}(\eta)R^\top$ as above, if we write f in the rotated basis as $f = \sum_{\alpha\in\mathbb{N}^d} \hat{f}(\alpha)H_\alpha(R^\top\cdot)$, and $g = f(R\cdot) = \sum_{\alpha\in\mathbb{N}^d}\hat{f}(\alpha)H_\alpha$, we have $M_f = RM_gR^\top$ (Lemma 2.3). Therefore

$$\begin{aligned}\operatorname{tr}\left(\Lambda^{-1}M_f\right) = \operatorname{tr}\left(\operatorname{Diag}(\eta^{-1})M_g\right) &= \sum_{a=1}^d \eta_a^{-1}\sum_{\alpha\in(\mathbb{N}^d)^*}\frac{\alpha_a}{c_{|\alpha|}}\hat{f}(\alpha)^2\\ &= \sum_{\alpha\in(\mathbb{N}^d)^*}\frac{1}{c_{|\alpha|}}\hat{f}(\alpha)^2\alpha^\top\eta^{-1}.\end{aligned}$$

We can then rewrite Equation (3.4) as

$$\begin{aligned}f_{\text{feat}}^{\lambda,\mu},\ \Lambda_{\text{feat}}^{\lambda,\mu} = \operatorname*{arg\,min}_{f\in\mathcal{F},\ \Lambda\in\mathbb{R}^{d\times d}} &\widehat{\mathcal{R}}(f) + \sum_{\alpha\in(\mathbb{N}^d)^*}\frac{1}{c_{|\alpha|}}\hat{f}(\alpha)^2(\lambda + \mu\alpha^\top\eta^{-1})\\ \text{subject to } &\Lambda = R\operatorname{Diag}(\eta)R^\top\\ &R \in O_d,\ \eta \in \mathbb{R}_+^d,\ \|\eta\|_{r/(2-r)} = 1\\ &f = \sum_{\alpha\in\mathbb{N}^d}\hat{f}(\alpha)H_\alpha(R^\top\cdot).\end{aligned} \tag{3.4}$$

We see then that the feature learning problem can be viewed as an extension of the variable selection problem, where we additionally optimise over any possible data rotation. Conversely, the variable selection problem can be seen as a particular case of the feature learning problem, where the rotation matrix R is fixed to the identity matrix.

To estimate the dimension of the underlying feature space P and the features themselves, we use the eigendecomposition of $\Lambda_{\text{feat}}^{\lambda,\mu} = (R_{\text{feat}}^{\lambda,\mu})^\top \operatorname{Diag}(\eta_{\text{feat}}^{\lambda,\mu}) R_{\text{feat}}^{\lambda,\mu}$. By using the columns of $R_{\text{feat}}^{\lambda,\mu}$ corresponding to the selected features, denoted as $\hat{S} := \{a \in [d] \,|\, \left(\eta_{\text{feat}}^{\lambda,\mu}\right)_a > t\}$ for some threshold $t > 0$, we construct our feature estimator $\hat{P}$, i.e., $\hat{P} := \left(R_{\text{feat}}^{\lambda,\mu}\right)_{\hat{S}}$. We see that by employing alternating minimisation, we are able to simultaneously learn the regression function and the underlying features.

3.2. Optimisation procedure

We now discuss how to solve the optimisation problem using alternative minimisation, drawing on techniques described in [5]. In the following discussion, we will focus on the feature learning setting. However, it is important to note that by simply fixing $R = I_d$ in each equation, we can easily revert back to the variable selection case.

To solve Equation (3.4), we have observed that when the function f is fixed, the optimal Λ can be determined using Equation (3.3), which involves the matrix M_f.[2]

When Λ is fixed, we seek to solve the optimisation problem

$$\begin{aligned} &\arg\min_{f \in \mathcal{F}} \widehat{\mathcal{R}}(f) + \sum_{\alpha \in (\mathbb{N}^d)^*} \frac{1}{c_{|\alpha|}} \hat{f}(\alpha)^2 (\lambda + \mu \alpha^\top \eta^{-1}) \\ &\text{subject to } f = \sum_{\alpha \in \mathbb{N}^d} \hat{f}(\alpha) H_\alpha(R^\top \cdot), \end{aligned}$$

where $\Lambda = R \operatorname{Diag}(\eta) R^\top$. However, this can only be solved if $\widehat{\mathcal{R}}$ is known, i.e., for some chosen loss function ℓ. Until the end of Section 3, we consider the quadratic loss which is commonly used in regression problems and allows for closed-form solutions. Otherwise, iterative optimisation algorithms need to be employed. The problem is then

$$\begin{aligned} &\arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \left(y^{(i)} - f(x^{(i)})\right)^2 + \sum_{\alpha \in (\mathbb{N}^d)^*} \frac{1}{c_{|\alpha|}} \hat{f}(\alpha)^2 (\lambda + \mu \alpha^\top \eta^{-1}) \\ &\text{subject to } f = \sum_{\alpha \in \mathbb{N}^d} \hat{f}(\alpha) H_\alpha(R^\top \cdot). \end{aligned} \tag{3.5}$$

[2] If $f = \sum_\alpha \hat{f}(\alpha) H_\alpha(R^\top \cdot)$, to compute M_f, we can remark that with $g = f(R\cdot) = \sum_\alpha \hat{f}(\alpha) H_\alpha$, we have the usual formula for M_g from Equation (2.10) and $M_f = R M_g R^\top$.

If we write for any $x, x' \in \mathbb{R}^d$

$$k_\Lambda(x, x') = \sum_{\alpha \in (\mathbb{N}^d)^*} \frac{c_{|\alpha|} H_\alpha(R^\top x) H_\alpha(R^\top x')}{\lambda + \mu \alpha^\top \eta^{-1}}, \tag{3.6}$$

the function k_Λ verifies all properties required to be a reproducing kernel [2]. The condition for a function to be a reproducing kernel is that it is symmetric and that the associated kernel matrix is positive definite for any set of points. Specifically, for any $n \in \mathbb{N}$ and $x^{(1)}, \dots, x^{(n)}$, the matrix $K_\Lambda = (k_\Lambda(x^{(i)}, x^{(j)}))_{i,j \in [n]}$ must be positive definite (where $\lambda > 0$ is useful in this context). We can then apply the theory of reproducing kernel Hilbert spaces (RKHS). In this case, k_Λ serves as the reproducing kernel for the space $\mathcal{F}$, with associated norm $\| \cdot \|_\Lambda$, given by

$$\|f\|_\Lambda^2 = \sum_{\alpha \in (\mathbb{N}^d)^*} \frac{1}{c_{|\alpha|}} \hat{f}(\alpha)^2 (\lambda + \mu \alpha^\top \eta^{-1})$$

(note that $\hat{f}$ depends on Λ through R). We can interpret the problem as a standard kernel ridge regression, which we refer to as the "kernel point of view." By applying the representer theorem [2], we know that the solution to Equation (3.5) takes the form

$$f = \sum_{i=1}^n \delta_i^\Lambda k_\Lambda(x^{(i)}, \cdot) + \delta_0^\Lambda,$$

where δ^Λ and δ_0^Λ can be obtained in closed form using $Y = (y^{(1)}, \dots, y^{(n)})^\top$ and $K = (k_\Lambda(x^{(i)}, x^{(j)}))_{i,j \in [n]}$ as the minimisers of

$$\delta^\Lambda, \ \delta_0^\Lambda = \underset{\delta \in \mathbb{R}^n, \ \delta_0 \in \mathbb{R}}{\arg\min} \frac{1}{n} \|Y - K_\Lambda \delta - \delta_0 \mathbb{1}\|_2^2 + \delta^\top K_\Lambda \delta. \tag{3.7}$$

It is worth noting that the shape of the kernel defined in Equation (3.6) implies that features corresponding to $\alpha \in \mathbb{N}^d$ with large values of $\alpha^\top \eta^{-1}$ are penalised more. If η_a is small, indicating that it has been pushed down in the previous optimisation steps, it suggests that variable x_a or the direction $(R^\top x)_a$ may not be particularly useful for prediction. In such cases, for these variables/directions to be retained, they would need to contribute significantly more to the fit compared to others.

Furthermore, we observe that the parameter λ serves the purpose of ensuring numerical stability when solving linear systems, particularly when $\alpha^\top \eta^{-1}$ can be null. We recommend setting λ to a significantly smaller value than μ to achieve this desired stability (e.g, $\lambda = 10^{-8}/d^{(2-r)/r}$ in our experiments). In fact, it is possible to fix λ as a predetermined value, eliminating the need for it to be treated as a hyper-parameter.

3.3. Sampling approximation of the kernel

We remark that the kernel described in Equation (3.6) is defined as an infinite sum, which means it is not computable in practice. To overcome this challenge, we adopt an approximation approach using sampling.

Let us define $C(\eta) = \sum_{\alpha\in(\mathbb{N}^d)^*} \frac{c_{|\alpha|}}{\lambda+\mu\alpha^\top\eta^{-1}}$. By defining $h(\alpha) = \frac{1}{C(\eta)}\frac{c_{|\alpha|}}{\lambda+\mu\alpha^\top\eta^{-1}}$, for all $\alpha \in (\mathbb{N}^d)^*$ we obtain a probability distribution on $(\mathbb{N}^d)^*$. Consequently, we can express the kernel $k_\Lambda(x,x')$ as $C(\eta)\mathbb{E}_{\alpha\sim h}\big(H_\alpha(R^\top x)H_\alpha(R^\top x')\big)$.

Sampling from the distribution h can be challenging, particularly in high-dimensional settings. Therefore, we employ an importance sampling technique. For the first choice $c_{|\alpha|} = \mathbb{1}_{|\alpha|\le M}$, the kernel $k_\Lambda(x,x')$ can be expressed as

$$\begin{aligned} k_\Lambda(x,x') &= \sum_{\alpha\in(\mathbb{N}^d)^*,|\alpha|\le M} \frac{H_\alpha(R^\top x)H_\alpha(R^\top x')}{\lambda+\mu\alpha^\top\eta^{-1}} \\ &= \binom{M+d}{d}\mathbb{E}_{\alpha\sim\mathcal{U}\{|\alpha|\le M\}}\left(\frac{H_\alpha(R^\top x)H_\alpha(R^\top x')}{\lambda+\mu\alpha^\top\eta^{-1}}\right), \end{aligned}$$

where $\mathcal{U}\{|\alpha| \le M\}$ is the uniform distribution over $\{\alpha \in (\mathbb{N}^d)^*,\ |\alpha| \le M\}$. Sampling from this uniform distribution can be achieved by selecting a subset B of size d uniformly from the set $[M+d]$, sorting the subset into $B_1 < \cdots < B_d$, setting $B_0 = 0$, and using the differences between consecutive values to construct α. Specifically, for each $a \in [d]$, we set $\alpha_a = B_a - B_{a-1} - 1$. If the resulting α is the null tuple, it is rejected, and the sampling process is repeated.

For the choice $c_{|\alpha|} = \rho^{|\alpha|}$ the kernel is

$$k_\Lambda(x,x') = \sum_{\alpha\in(\mathbb{N}^d)^*} \frac{\rho^{|\alpha|}}{\lambda+\mu\alpha^\top\eta^{-1}} H_\alpha(R^\top x)H_\alpha(R^\top x').$$

We have developed a methodology called "group sampling" that addresses the challenges of sampling from the distribution h. To initialise the sampling, we set all components of η to be equal. This choice ensures unbiasedness among the possible directions while satisfying the constraint $\|\eta\|_{r/(2-r)} = 1$. As a result, the kernel takes the form

$$k_\Lambda(x,x') = \sum_{\alpha\in(\mathbb{N}^d)^*} \frac{\rho^{|\alpha|}}{\lambda+\mu|\alpha|d^{(2-r)/r}} H_\alpha(R^\top x)H_\alpha(R^\top x').$$

We can directly sample from the distribution proportional to $\frac{\rho^{|\alpha|}}{\lambda+\mu|\alpha|d^{(2-r)/r}}$. The sampling process involves two steps. First, we sample an integer k from the distribution

$$k \sim \binom{k+d-1}{d-1}\frac{\rho^k}{\lambda+\mu d^{(2-r)/r}k}.$$

To perform this sampling, we can precompute a table of probabilities for different values of k up to a chosen maximum value (e.g., 40). We then normalise these

probabilities and use them to sample the value of k. Once we have obtained k, it represents the cardinality of α. In the second step, we sample α uniformly from the set $\alpha \in (\mathbb{N}^d)^*,\ |\alpha| = k$. This sampling procedure is exact, except for the controlled approximation introduced by the choice of the maximum value.

We can develop an importance sampling scheme for the other optimisation steps when the components of η are not equal. Here are the steps.

1. Sort the components of η in ascending order and find the largest gap between consecutive values. Divide the set $[d]$ into two groups: Group 1, containing the components above the top of the gap, with size d_1, and Group 2, containing the remaining components, with size d_2.
2. Define $\tilde{\eta}_1$ as the minimum value among the components in Group 1, and $\tilde{\eta}_2$ as the maximum value among the components in Group 2.
3. Sample k_1 and k_2 from the distribution

$$k_1, k_2 \sim \binom{k_1 + d_1 - 1}{d_2 - 1}\binom{k_2 + d_2 - 1}{d_2 - 1}\frac{\rho^{k_1+k_2}}{\lambda + \mu\left(\frac{k_1}{\tilde{\eta}_1} + \frac{k_2}{\tilde{\eta}_2}\right)},$$

   where k_1 and k_2 represent $|\alpha^{(1)}|$ and $|\alpha^{(2)}|$ respectively, and $\alpha^{(1)}$ corresponds to the components in Group 1.
4. Sample $\alpha^{(1)}$ uniformly from the set $\alpha \in (\mathbb{N}^{d_1})^,\ |\alpha| = k_1$, and sample $\alpha^{(2)}$ uniformly from the set $\alpha \in (\mathbb{N}^{d_2})^,\ |\alpha| = k_2$.
5. This yields

$$\begin{aligned} k_\Lambda(x, x') &= \sum_{\alpha \in (\mathbb{N}^d)^*} \frac{C(\tilde{\eta})}{C(\tilde{\eta})} \frac{\rho^{|\alpha|}}{\lambda + \mu\alpha^\top \tilde{\eta}^{-1}} \frac{\lambda + \mu\alpha^\top \tilde{\eta}^{-1}}{\lambda + \mu\alpha^\top \eta^{-1}} H_\alpha(R^\top x) H_\alpha(R^\top x') \\ &= \mathbb{E}_{\alpha \sim \text{Group sampling}}\left(C(\tilde{\eta}) \frac{\lambda + \mu\alpha^\top \tilde{\eta}^{-1}}{\lambda + \mu\alpha^\top \eta^{-1}} H_\alpha(R^\top x) H_\alpha(R^\top x')\right), \end{aligned}$$

   with $C(\tilde{\eta})$ a normalising constant.

By using this importance sampling scheme, we can approximate the desired distribution accurately, even when the components of η are not equal.

We observe that with the group sampling approach, the distribution of α is influenced by η through the grouping process, as well as through the values of $\tilde{\eta}_1$ and $\tilde{\eta}_2$. As the optimisation progresses, the sampled tuples exhibit specific patterns: in directions that are deemed unimportant (corresponding to small values of η_a), α_a tends to be close to zero, while in directions that are considered important (corresponding to large η_a), α_a is more widely distributed.[3]

[3]It is worth noting that using the geometric distribution independently on each dimension of α would have been a simpler approach. However, this method becomes highly inefficient as the dimensionality increases, since it would involve sampling numerous α tuples with low importance weights (as determined by $\frac{H_\alpha(R^\top x) H_\alpha(R^\top x')}{\lambda + \mu\alpha^\top \eta^{-1}}$) due to their alignment with directions where η is very small (i.e., $\alpha^\top \eta^{-1}$ is small).

No matter the sampling scheme, we sample $\alpha^{(1)}, \dots, \alpha^{(m)}$ from some distribution with importance weight $w(\alpha)$, yielding

$$k_\Lambda(x, x') \approx \sum_{j=1}^{m} w(\alpha^{(i)}) H_{\alpha^{(j)}}(R^\top x) H_{\alpha^{(j)}}(R^\top x').$$

We use this formula to compute the kernel matrix $K_\Lambda = \left(k_\Lambda(x^{(i)}, x^{(j)})\right)_{i,j\in[n]}$. Instead of approximating the matrix K_Λ to use in Equation (3.7), we can also consider the equivalent explicit "feature point of view" by writing f in the form

$$f = \sum_{j=1}^{m} \theta_j w(\alpha^{(j)}) H_{\alpha^{(j)}}(R^\top \cdot) + \theta_0 H_0,$$

where

$$\theta^\Lambda,\ \theta_0^\Lambda = \underset{\theta\in\mathbb{R}^m,\ \theta_0\in\mathbb{R}}{\arg\min} \frac{1}{n}\|Y - \Phi\theta - \theta_0 \mathbb{1}\|_2^2 + \|\theta\|_2^2, \tag{3.8}$$

with $\Phi \in \mathbb{R}^{n\times m}$ the matrix filled with $w(\alpha^{(j)}) H_{\alpha^{(j)}}(R^\top x^{(i)})$. This is computationally advantageous when $n > m$. Otherwise, we use the kernel point of view. In both cases, we can use (θ, θ_0) or (δ, δ_0) to rewrite f as $\sum_{\alpha\in\mathbb{N}^d} \hat{f}(\alpha) H_\alpha(R^\top \cdot)$. We remark that $\hat{f}(\alpha) = 0$ when α has not been sampled.

The pseudo-code for the **RegFeaL** method is provided in Algorithm 1.

Algorithm 1: RegFeaL pseudocode

```
for i ∈ [n_iter] do
    if i = 0 then
        η ← 1/d^{(2−r)/r};
        R ← I_d;
    else
        if feature learning then
            Update R and η as in Equation (3.3);
        else
            Update η as in Equation (3.1);
        end
    end
    Sample α^(1), ..., α^(m) using group sampling as in Section 3.3 with η ;
    Compute importance weights w(α^(1)), ..., w(α^(m));
    Compute Hermite features Φ ∈ R^{n×m}, Φ_{i,j} = w(α^(j)) H_{α^(j)}(R^⊤ x^(i)) ;
    if n > m then
        Update θ as in Equation (3.8);
    else
        Update δ as in Equation (3.7);
    end
end
```

In terms of numerical complexity, each iteration has a cost of

$$\mathcal{O}\Bigg(\underbrace{nm'd + nd^2}_{\text{Hermite features}} + \underbrace{d^2(m')^2 + d^3}_{M_f \text{ and its eigendecomposition}} + \underbrace{md}_{\text{Sampling}} + \underbrace{nm'\max(n, m')}_{\text{Computing } \theta \text{ or } \delta}\Bigg),$$

where m' is the number of unique tuples sampled (which is necessarily smaller than m, and can be much smaller when η is sparse). The parameter m can be chosen to achieve a balance between computational cost and performance, but selecting an excessively small value for m may adversely affect performance. In practice, the number of iterations required for convergence is typically very small (less than 10), as demonstrated in Section 5. Additionally, it is worth noting that the computation cost of δ in the feature point of view could be reduced through the use of the Nyström approximation [25].

4. Statistical properties

We now consider the statistical properties of **RegFeaL**. We always take $r = 1$ and we do not consider the approximation due to the computation of the estimators in this section. Our goal is to provide a high-probability bound on the expected risk of **RegFeaL** to gain insights into its generalisation properties under minimal assumptions to obtain a very general result. We do not consider the consistency of the e.d.r. space estimation, as this usually requires much stronger assumptions, such as the linearity condition, the gradient along the relevant directions to be large enough in norm, or constraint on the loss to be the square loss, for example [10, 19].

We leverage the results presented in [4], which provide bounds on the maximum difference between empirical and expected risk, in terms of the expectation over the class of functions with bounded norm. These bounds are expressed in terms of the Rademacher complexity of the set $\{f \in \mathcal{F},\ \Omega(f) \leq D\}$, where $D > 0$ is a fixed bound. By employing these results, we can obtain a probabilistic bound on the constrained estimator and apply McDiarmid's inequality [7] to establish a result in probability. Ultimately, Theorem 4.1 provides a probabilistic bound for the **RegFeaL** estimator, leveraging the aforementioned results as well as the optimality conditions satisfied by the estimator.

4.1. Setup

We start by making assumptions about the data used to train the model.

Assumption 4.1 (Data)**.** *$\mathcal{D} = (x^{(i)}, y^{(i)})_{i\in[n]}$ is a set of i.i.d data, with (X, Y) a pair of random variables such that $\forall i \in [n],\ (x^{(i)}, y^{(i)}) \sim (X, Y)$.*

Notice that we do not make strong assumptions on the distribution of the data, such as independence of the covariates or constraint to be elliptically contoured, nor do we require is to be known a priori.

Let us introduce some definitions. Let $(c_k)_{k>0}$ be a non-null sequence of positive reals. We define the function space $\mathcal{F}$ as Span $\big(\{H_0\} \cup \{H_\alpha,\ \text{for } \alpha \in (\mathbb{N}^d)^* \text{such that } c_{|\alpha|>0}\}\big)$. Let ℓ be a loss function on $\mathbb{R} \times \mathbb{R}$, and let the expected

risk $\mathcal{R}$ and the empirical risk $\widehat{\mathcal{R}}$ be

$$\mathcal{R}(f) = \mathbb{E}_{X,Y}\big(\ell(Y, f(X))\big) \quad \text{and} \quad \widehat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}, f(x^{(i)})).$$

We define the functional norm $\Omega(f)$ for any $f \in \mathcal{F}$ as $\Omega(f) := \Omega_{\text{feat}}(f) + |\hat{f}(0)|$ or $\Omega(f) := \Omega_{\text{var}}(f) + |\hat{f}(0)|$, where $\hat{f}(0)$ represents the constant coefficient of f. It is important to note that the constraint on the constant coefficient is not necessary in practice, but we include it for the purpose of theoretical analysis (we could also add a small weight on $|\hat{f}(0)|$ to this effect). We define the regularised empirical risk $\widehat{\mathcal{R}}\mu(f)$ for $\mu > 0$ as follows

$$\widehat{\mathcal{R}}\mu(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}, f(x^{(i)})) + \mu\Omega(f).$$

We denote our estimator as $f^\mu := \arg\min_{f\in\mathcal{F}} \widehat{\mathcal{R}}_\mu$. In order to establish theoretical results, we will rely on the following assumptions.

Assumption 4.2 (Problem assumptions)**.**
1. *The true regression function* $f^* := \arg\min_{f\in\mathcal{F}} \mathcal{R}(f)$ *exists.*
2. *For some* $D > 0$*, the loss function* ℓ *is* G*-Lipschitz continuous in its second argument for any value of its first argument, i.e.,* $\forall y \in \mathcal{Y},\ \forall x, x' \in \mathcal{X},\ \forall f \in \mathcal{F}$ *such that* $\Omega(f) \le D,\ |\ell(y, f(x)) - \ell(y, f(x'))| \le G \cdot |f(x) - f(x')|$.
3. *For some* $D > 0$, $\ell_\infty := \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}, f\in\mathcal{F}, \Omega(f)\le D} \ell(y, f(x))$ *is finite.*
4. *The loss* ℓ *is convex on* $\mathbb{R}\times\mathbb{R}$ *.*

For our main result, we will use $D = 2\Omega(f^*)$. These assumptions are commonly used in the analysis of nonparametric regression [14]. Many commonly used loss functions in regression problems, such as the quadratic loss, absolute mean error, Huber loss, or logistic loss, are convex. The Lipschitz continuity condition holds for all of these losses, except for the quadratic loss, which we handle separately, for example by exploiting the boundedness of the data. If the data is bounded (i.e., $\mathcal{X}\times\mathcal{Y}$ is bounded in $\mathbb{R}^d\times\mathbb{R}$), then $\sup_{x\in\mathcal{X}, f\in\mathcal{F}, \Omega(f)\le D} |f(x)|$ is bounded for any $D > 0$.[4] We can then use the convexity of the loss ℓ and boundedness of $\mathcal{Y}$ to justify that ℓ_∞ is well-defined. For the quadratic loss, in this setting, it satisfies Assumption 4.2.2 because $(y - f(x))^2 - (y - f(x'))^2 = (f(x') - f(x))(y - f(x) + y - f(x'))$, and we can then take$G := \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}, f\in\mathcal{F}, \Omega(f)\le D} |y - f(x) + y - f(x')|$.

[4]This can be seen for Ω_{var} by noticing that $\Omega(f)$ can be written as $|\hat{f}(0)| + \sum_{a=1}^d \Theta_a(f) \ge \big(|\hat{f}(0)|^2 + \sum_{a=1}^d \Theta_a(f)^2\big)^{1/2}$, with the latter being an RKHS norm with reproducing kernel $k(x,x') = 1 + \sum_{\alpha\in(\mathbb{N}^d)^*} \frac{c_{|\alpha|}}{|\alpha|} H_\alpha(x) H_\alpha(x')$. It follows that $f(x) = \langle f, k(X,\cdot)\rangle \le \hat{f}(0) + \Omega(f)\sqrt{k(x,x)}$ which is bounded if x is bounded.

4.2. Rademacher complexity

First, we apply the Lipschitz continuity assumption to bound the supremum over a set of functions of the difference between the empirical risk and expected risk, in expectation over the dataset.

Lemma 4.1 (Use of Gaussian complexity)**.** *Let $\mathcal{G}$ be any set of functions, then under Assumption 4.1, and Assumption 4.2.2,*

$$\mathbb{E}_{\mathcal{D}}\left(\sup_{f\in\mathcal{G}}\left(\mathcal{R}(f)-\widehat{\mathcal{R}}(f)\right)+\sup_{f\in\mathcal{G}}\left(\widehat{\mathcal{R}}(f)-\mathcal{R}(f)\right)\right)\leq 4\sqrt{\frac{\pi}{2}}G\cdot G_n(\mathcal{G}),$$

where

$$G_n(\mathcal{G}):=\mathbb{E}_{\mathcal{D},\varepsilon\sim\mathcal{N}(0,I_n)}\left(\sup_{f\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x^{(i)})\right)$$

is the Gaussian complexity of the set $\mathcal{G}$ [6].

See Appendix A.2 for the proof, which we include for the sake of completeness. This is a close variation of the work presented in [4]. We now need to bound the Gaussian complexity, when we consider subsets of the working space $\mathcal{F}$ with bounded norm.

Lemma 4.2 (Bound on Gaussian complexity)**.** *Let $D>0$, with $\mathcal{G}:=\{f\in\mathcal{F},\ \Omega(f)\leq D\}$ with Ω defined as in Section 4.1, under Assumption 4.1, we have*

$$G_n(\mathcal{G})\leq D\cdot\sqrt{\frac{1}{n}\left(1+\sum_{\alpha\in(\mathbb{N}^d)^*}\frac{c_{|\alpha|}}{|\alpha|}\mathbb{E}_X\left(H_\alpha(X)^2\right)\right)}.$$

We remark that the result depends heavily on the data distribution through the expectations $\mathbb{E}_X(H_\alpha(X)^2)$ and the design of the norm through $(c_k)_{k>0}$. We discuss these in more details in Section 4.4.

Proof of Lemma 4.2. We first consider the norm Ω_{var}. Let $f\in\mathcal{G}$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x^{(i)})=\sum_{\alpha\in\mathbb{N}^d}\hat{f}(\alpha)\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i H_\alpha(x^{(i)})\right)=\sum_{\alpha\in\mathbb{N}^d}\hat{f}(\alpha)\hat{\xi}(\alpha),$$

with ξ an infinite vector indexed by $(\mathbb{N}^d)$, $\hat{\xi}(\alpha)=\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i H_\alpha(x^{(i)})$. Therefore

$$\sup_{f\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x^{(i)})=\sup_{f\in\mathcal{G}}\sum_{\alpha\in\mathbb{N}^d}\hat{f}(\alpha)\hat{\xi}(\alpha)=D\cdot\Omega^*_{\text{var}}(\xi).$$

Now since Ω_{var} is the sum of $d+1$ semi-norms $\Theta_0,\Theta_1,\dots,\Theta_d$, with

$$\Theta_0(f)=|\hat{f}(0)|$$
$$\Theta_a(f)=\left(\sum_{\alpha\in(\mathbb{N}^d)^*}\frac{\alpha_a}{c_{|\alpha|}}\hat{f}(\alpha)^2\right)^{1/2},\forall a\in[d],$$

we have

$$\Omega^*_{\mathrm{var}}(\xi) = \inf_{\xi=\sum_{a=0}^d \xi_a} \sup_{a\in\{0,\dots,d\}} \Theta^*_a(\xi_a).$$

This is an extension of the fact that the set $\Omega^*_{\mathrm{var}}(\xi) \leq 1$ is the subdifferential of Ω_{var} at $f = 0$, and thus the sum of the d subdifferentials of $\Omega_0, \dots, \Omega_d$ at $f = 0$. We consider $a \in [d]$, we have

$$\Omega^*_a(\xi_a)^2 = \sum_{\alpha\in\mathbb{N}^d, \alpha_a>0} \hat{\xi}_a(\alpha)^2 \frac{c_{|\alpha|}}{\alpha_a},$$

and $\Omega^*_0(\xi)^2 = \hat{\xi}(0)^2$.

If we choose $\forall \alpha \in (\mathbb{N}^d)^*$, $\hat{\xi}_a(\alpha) = \frac{\sqrt{\alpha_a}}{\sum_b \sqrt{\alpha_b}}\hat{\xi}(\alpha)$, $\hat{\xi}_0(\alpha) = 0$, $\hat{\xi}_a(0) = 0$ and $\hat{\xi}_0(0) = \hat{\xi}(0)$, we have:

$$\begin{aligned}
\Omega^*_{\mathrm{var}}(\xi)^2 &\leq \sup\left(\sup_{a\in[d]} \sum_{\alpha\in\mathbb{N}^d, \alpha_a>0} \hat{\xi}_a(\alpha)^2 \frac{c_{|\alpha|}}{\alpha_a}, \hat{\xi}_0(0)^2\right) \\
&\leq \sup\left(\sup_{a\in[d]} \sum_{\alpha\in\mathbb{N}^d, \alpha_a>0} \hat{\xi}(\alpha)^2 \frac{c_{|\alpha|}}{\left(\sum_b \sqrt{\alpha_b}\right)^2}, \hat{\xi}(0)^2\right) \\
&\leq \sum_{\alpha\in\mathbb{N}^d} \hat{\xi}(\alpha)^2 \left(\frac{c_{|\alpha|}}{|\alpha|}\mathbb{1}_{|\alpha|>0} + \mathbb{1}_{|\alpha|=0}\right).
\end{aligned}$$

Let $W^2 = \mathrm{Diag}\left(\frac{c_{|\alpha|}}{|\alpha|}\mathbb{1}_{|\alpha|>0} + \mathbb{1}_{|\alpha|=0}\right)$ and Φ the design matrix of all $H_\alpha(x^{(i)})$ (with n rows and infinitely many columns indexed with $\alpha \in \mathbb{N}^d$). We have $\hat{\xi} = \frac{1}{n}\Phi^\top \varepsilon$, and

$$\Omega^*_{\mathrm{var}}(\xi)^2 \leq \hat{\xi}^\top W^2 \hat{\xi} = \frac{1}{n^2}\varepsilon^\top \Phi W^2 \Phi^\top \varepsilon.$$

We compute the expectation of $\Omega^*_{\mathrm{var}}(\xi)^2$ for $\varepsilon \sim \mathcal{N}(0, I_n)$, and get

$$\begin{aligned}
\mathbb{E}_\varepsilon\left(\Omega^*_{\mathrm{var}}(\xi)^2\right) \leq \mathbb{E}_\varepsilon\left(\frac{1}{n^2}\varepsilon^\top \Phi W^2 \Phi^\top \varepsilon\right) &= \frac{1}{n^2}\,\mathrm{tr}\left(\Phi W^2 \Phi^\top\right) \\
&= \frac{1}{n} + \frac{1}{n^2} \sum_{\alpha\in(\mathbb{N}^d)^*} \sum_{i=1}^n \frac{c_{|\alpha|}}{|\alpha|} H_\alpha(x^{(i)})^2.
\end{aligned}$$

We now take expectations with regards to the data $\mathcal{D}$ and get

$$\mathbb{E}_{\mathcal{D},\varepsilon}\left(\Omega^*_{\mathrm{var}}(\xi)^2\right) \leq \frac{1}{n} \sum_{\alpha\in(\mathbb{N}^d)^*} \frac{c_{|\alpha|}}{|\alpha|} \mathbb{E}_X\left(H_\alpha(X)^2\right) + \frac{1}{n}.$$

Using Cauchy-Schwartz, $\mathbb{E}_{\mathcal{D},\varepsilon}\left(\Omega^*_{\mathrm{var}}(\xi)\right) \leq \sqrt{\frac{1}{n}\left(1 + \sum_{\alpha\in(\mathbb{N}^d)^*} \frac{c_{|\alpha|}}{|\alpha|}\mathbb{E}_X\left(H_\alpha(X)^2\right)\right)}$.

From Lemma 2.3, we have

$$\Omega_{\mathrm{feat}}(f) \geq \inf_{R\in O_d} \Omega_{\mathrm{var}}(f(R\cdot)).$$

Then, for an infinite vector ξ indexed by $\mathbb{N}^d$, with $\hat{\xi}(\alpha) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i H_\alpha(x^{(i)})$ and ξ_R an infinite vector indexed by $\mathbb{N}^d$ with $\hat{\xi}_R(\alpha) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i H_\alpha(Rx^{(i)})$, we have $\Omega^*_{\text{feat}}(\xi) \leq \sup_{R\in O_d} \Omega^*_{\text{var}}(\xi_R)$.

Therefore

$$\sup_{R\in O_d} \Omega^*_{\text{var}}(\xi_R) \leq \sup_{R\in O_d} \frac{1}{n^2}\varepsilon^\top \Phi_R W^2 \Phi_R^\top \varepsilon,$$

with Φ_R the design matrix of all $H_\alpha(Rx^{(i)})$ (with n rows and infinitely many columns indexed with $\alpha \in \mathbb{N}^d$). Therefore using Lemma 2.2,

$$\begin{aligned}
\varepsilon^\top \Phi_R W^2 \Phi_R^\top \varepsilon &= \sum_{i,j} \varepsilon_i \varepsilon_j \left(1 + \sum_{\alpha\in(\mathbb{N}^d)^*} \frac{c_{|\alpha|}}{|\alpha|} H_\alpha(Rx^{(i)}) H_\alpha(Rx^{(j)})\right) \\
&= \sum_{i,j} \varepsilon_i \varepsilon_j \left(1 + \sum_{k=1}^{+\infty} \frac{c_k}{k} \sum_{\alpha\in\mathbb{N}^d,\ |\alpha|=k} H_\alpha(Rx^{(i)}) H_\alpha(Rx^{(j)})\right) \\
&= \sum_{i,j} \varepsilon_i \varepsilon_j \left(1 + \sum_{k=1}^{+\infty} \frac{c_k}{k} \sum_{|\alpha|=k} H_\alpha(x^{(i)}) H_\alpha(x^{(j)})\right) = \varepsilon^\top \Phi W^2 \Phi^\top \varepsilon,
\end{aligned}$$

which is independent of R, therefore yielding exactly the same result as for Ω_{var} once expectation with regards to ε and the data is taken. □

4.3. Statistical convergence

To gain insight into the proof technique, we initially establish an expectation-based result for the constrained estimator instead of the regularised estimator. We bound the expected risk of the function that minimises the empirical risk over the set of functions with a bounded norm, in expectation over the dataset. To accomplish this, we use Lemma 4.1 and Lemma 4.2.

Lemma 4.3 (Expected risk of constrained estimator)**.** *Let $D > \Omega(f^*)$, let $f^D = \arg\min_{f\in\mathcal{F},\ \Omega(f)\leq D} \widehat{\mathcal{R}}(f)$, under Assumptions 4.1, 4.2.1 and 4.2.2,*

$$\mathbb{E}_{\mathcal{D}}\big(\mathcal{R}(f^D)\big) \leq \mathcal{R}(f^*) + \frac{4GD}{\sqrt{n}}\sqrt{\frac{\pi}{2}}\sqrt{1 + \sum_{\alpha\in(\mathbb{N}^d)^*} \frac{c_{|\alpha|}}{|\alpha|}\mathbb{E}_X\big(H_\alpha(X)^2\big)}.$$

Proof of Lemma 4.3. With $\mathcal{G} := \{f \in \mathcal{F},\ \Omega(f) \leq D\}$, we have the classical decomposition of the excess risk

$$\begin{aligned}
\mathcal{R}(f^D) - \mathcal{R}(f^*) &= \mathcal{R}(f^D) - \widehat{\mathcal{R}}(f^D) + \widehat{\mathcal{R}}(f^D) - \widehat{\mathcal{R}}(f^*) + \widehat{\mathcal{R}}(f^*) - \mathcal{R}(f^*) \\
&\leq \mathcal{R}(f^D) - \widehat{\mathcal{R}}(f^D) + \widehat{\mathcal{R}}(f^*) - \mathcal{R}(f^*) \\
&\leq \sup_{f\in\mathcal{G}} \mathcal{R}(f) - \widehat{\mathcal{R}}(f) + \sup_{f\in\mathcal{G}} \widehat{\mathcal{R}}(f) - \mathcal{R}(f).
\end{aligned}$$

We then take the expectation over the data on both sides and use Lemma 4.1 and Lemma 4.2

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\big(\mathcal{R}(f^D)\big) - \mathcal{R}(f^*) &\leq \mathbb{E}_{\mathcal{D}}\big(\sup_{f\in\mathcal{F}} \mathcal{R}(f) - \widehat{\mathcal{R}}(f) + \sup_{f\in\mathcal{F}} \widehat{\mathcal{R}}(f) - \mathcal{R}(f)\big) \\ &\leq 4\sqrt{\frac{\pi}{2}} G \cdot G_n(\mathcal{G}) \\ &\leq \frac{4GD}{\sqrt{n}}\sqrt{\frac{\pi}{2}}\sqrt{1 + \sum_{\alpha\in(\mathbb{N}^d)^*} \frac{c_{|\alpha|}}{|\alpha|}\mathbb{E}_X\big(H_\alpha(X)^2\big)},\end{aligned}$$

hence the desired result. □

In addition to the expectation-based result, obtaining a result with high probability for the constrained estimator is also of interest. This is achieved in Lemma A.1, presented in Appendix A.3, by using McDiarmid's inequality [7]. To apply this inequality, an additional assumption is required: the boundedness of the loss (Assumption 4.2.3). However, the most significant and relevant result is the one obtained for the estimator that minimises the regularised empirical risk. This result is more realistic and imposes the additional requirement of convexity of the loss function.

Theorem 4.1 (High-probability bound on expected risk of regularised estimator)**.** *Under Assumption 4.1 and Assumptions 4.2.1, 4.2.2, 4.2.3 with* $D = 2\Omega(f^*)$*, 4.2.4, then for any* $\delta \in (0,1)$*, with the choice of regularising parameter*

$$\mu = \frac{8G}{\sqrt{n}}\sqrt{\frac{\pi}{2}}\sqrt{1 + \sum_{\alpha\in(\mathbb{N}^d)^*} \frac{c_{|\alpha|}}{|\alpha|}\mathbb{E}_X\big(H_\alpha(X)^2\big)} + \frac{\ell_\infty 2\sqrt{2}}{\Omega(f^*)\sqrt{n}}\sqrt{\log\frac{2}{\delta}},$$

with probability larger than $1-\delta$

$$\begin{aligned}&\mathcal{R}(f^\mu) \leq \mathcal{R}(f^*) \\ &+ \Omega(f^*)\left(\frac{16G}{\sqrt{n}}\sqrt{\frac{\pi}{2}}\sqrt{1 + \sum_{\alpha\in(\mathbb{N}^d)^*} \frac{c_{|\alpha|}}{|\alpha|}\mathbb{E}_X\big(H_\alpha(X)^2\big)}\right) + \frac{\ell_\infty 4\sqrt{2}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}\end{aligned}$$

and $\Omega(f^\mu) \leq 2\Omega(f^*)$.

We now discuss the meaning of Theorem 4.1. The theorem states that with high probability, under the appropriate choice of the regularisation parameter, the norm of the estimator f^μ, $\Omega(f^\mu)$, is bounded by twice the norm of the true regression function f^*, $\Omega(f^*)$. We remark that the choice of regularisation parameter depends on $\Omega(f^*)$, however, this is not the case in the bounded setting, see the discussion in Section 4.4. Under Assumption 2.1 (feature learning setting) or Assumption 2.2 (variable selection setting), we know that $\Omega(f^*)$ does not depend explicitly on d but only on s, the underlying number of variables or dimension of the linear subspace.

The norm $\Omega(f^*)$ also helps us bound the difference between the expected risk of the estimator $\mathcal{R}(f^\mu)$ and the expected risk of the true regression function

$\mathcal{R}(f^*)$. This difference, denoted as $\mathcal{R}(f^\mu) - \mathcal{R}(f^*)$, has a dependency on the number of samples n, with a convergence rate of $n^{-1/2}$, as expected for a Lipschitz loss and a well-specified model. However, the dependency on the dimension d of the original data is somewhat concealed in $\sqrt{1 + \sum_{\alpha \in (\mathbb{N}^d)^*} \frac{c_{|\alpha|}}{|\alpha|} \mathbb{E}_X\big(H_\alpha(X)^2\big)}$. We provide a detailed analysis of this dependency for specific choices of the data distribution X and the sequence $(c_k)_{k>0}$ in Section 4.4.

Proof of Theorem 4.1. The proof is adapted from [4]. Define $f^{\mu *}$ as the minimiser of $\mathcal{R}_\mu := \mathcal{R} + \mu\Omega$ over $\mathcal{F}$. Now, for $D > 0, \tau > 0$ define the following convex set

$$\mathcal{C}_{D,\tau} = \{f \in \mathcal{F}, \Omega(f) \leq D, \mathcal{R}_\mu(f) - \mathcal{R}_\mu(f^{\mu *}) \leq \tau\}.$$

It has boundary

$$\partial\mathcal{C}_{D,\tau} = \{f \in \mathcal{F}, \Omega(f) \leq D, \mathcal{R}_\mu(f) - \mathcal{R}_\mu(f^{\mu *}) = \tau\},$$

i.e., the second constraint is the saturated one, for well chosen D and τ. This is because, if we consider a f such that $\Omega(f) = D$, since the optimality conditions for $f^{\mu *}$ give that $\Omega^*(\mathcal{R}'(f^{\mu *})) \leq \mu$, (with $\mathcal{R}'$ any subgradient of $\mathcal{R}$ which necessarily exists because $\mathcal{R}$ is convex since ℓ is convex) we have

$$\begin{aligned}
\mathcal{R}_\mu(f) - \mathcal{R}_\mu(f^{\mu *}) &= \mathcal{R}(f) + \mu\Omega(f) - \mathcal{R}(f^{\mu *}) - \mu\Omega(f^{\mu *}) \\
&\geq \langle \mathcal{R}'(f^{\mu *}), (f - f^{\mu *}) \rangle + \mu\Omega(f) - \mu\Omega(f^{\mu *}) \\
&\quad \text{by convexity with } \langle \cdot, \cdot \rangle \text{ associated to } \Omega \\
&\geq -\Omega^*\big(\mathcal{R}'(f^{\mu *})\big)\Omega(f - f^{\mu *}) + \mu\Omega(f) - \mu\Omega(f^{\mu *}) \\
&\quad \text{by Holder's inequality} \\
&\geq -\mu\Omega(f - f^{\mu *}) + \mu\Omega(f) - \mu\Omega(f^{\mu *}) \text{ by optimality of } f^{\mu *} \\
&\geq 2\mu\Omega(f) - 2\mu\Omega(f^{\mu *}) \text{ by the triangular inequality} \\
&\geq 2\mu D - 2\mu\Omega(f^{\mu *}) \text{ since } \Omega(f) = D, \\
&\geq 2\mu\Omega(f^*) \text{ by choosing } D = 2\Omega(f^*), \text{ since } \Omega(f^*) \geq \Omega(f^{\mu *}) \\
&\geq \tau, \text{ by choosing } \tau = \mu\Omega(f^*),
\end{aligned}$$

hence the desired result on the active constraint of the boundary. We now fix $\tau = \mu\Omega(f^*)$ and $D = 2\Omega(f^*)$.

Now if f^μ does not belong to $\mathcal{C}_{D,\tau}$, since $f^{\mu *}$ does, there is an element f in the segment $[f^\mu, f^{\mu *}]$ that belongs to $\partial\mathcal{C}_{D,\tau}$, i.e, $\Omega(f) \leq D$ and $\mathcal{R}_\mu(f) - \mathcal{R}_\mu(f^{\mu *}) = \tau$. Because the loss is convex, we have that $\widehat{\mathcal{R}}_\mu(f) \leq \max\{\widehat{\mathcal{R}}_\mu(f^\mu), \widehat{\mathcal{R}}_\mu(f^{\mu *})\} = \widehat{\mathcal{R}}_\mu(f^{\mu *})$. Therefore

$$\begin{aligned}
\tau = \mathcal{R}_\mu(f) - \mathcal{R}_\mu(f^{\mu *}) &\leq \mathcal{R}_\mu(f) - \mathcal{R}_\mu(f^{\mu *}) + \widehat{\mathcal{R}}_\mu(f^{\mu *}) - \widehat{\mathcal{R}}_\mu(f) \\
&\leq \mathcal{R}(f) - \widehat{\mathcal{R}}(f) + \widehat{\mathcal{R}}(f^{\mu *}) - \mathcal{R}(f^{\mu *}). \qquad (4.1)
\end{aligned}$$

From the proof of Lemma A.1, for all $\delta \in (0,1)$

$$\sup_{f\in\mathcal{F},\ \Omega(f)\le D} \mathcal{R}(f) - \widehat{\mathcal{R}}(f) + \sup_{f\in\mathcal{F},\ \Omega(f)\le D} \widehat{\mathcal{R}}(f) - \mathcal{R}(f) \le \frac{4GD}{\sqrt{n}}\sqrt{\frac{\pi}{2}}\sqrt{1+\sum_{\alpha\in(\mathbb{N}^d)^*}\frac{c_{|\alpha|}}{|\alpha|}\mathbb{E}_X\left(H_\alpha(X)^2\right)} + \frac{\ell_\infty 2\sqrt{2}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}$$

with probability larger than $1-\delta$.

We apply this to the RHS of Equation (4.1) (as $\Omega(f)\le D$ and $\Omega(f^{\mu*})\le D$), which is smaller than $\frac{4GD}{\sqrt{n}}\sqrt{\frac{\pi}{2}}\sqrt{1+\sum_{\alpha\in(\mathbb{N}^d)^*}\frac{c_{|\alpha|}}{|\alpha|}\mathbb{E}_X\left(H_\alpha(X)^2\right)} + \frac{\ell_\infty 2\sqrt{2}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}$ with probability larger than $1-\delta$.

Now if τ is such that

$$\frac{4GD}{\sqrt{n}}\sqrt{\frac{\pi}{2}}\sqrt{1+\sum_{\alpha\in(\mathbb{N}^d)^*}\frac{c_{|\alpha|}}{|\alpha|}\mathbb{E}_X\left(H_\alpha(X)^2\right)} + \frac{\ell_\infty 2\sqrt{2}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}} \le \tau, \text{ i.e.,}$$

$$\Omega(f^*)\frac{8G}{\sqrt{n}}\sqrt{\frac{\pi}{2}}\sqrt{1+\sum_{\alpha\in(\mathbb{N}^d)^*}\frac{c_{|\alpha|}}{|\alpha|}\mathbb{E}_X\left(H_\alpha(X)^2\right)} + \frac{\ell_\infty 2\sqrt{2}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}} \le \mu\Omega(f^*)$$

$$\frac{8G}{\sqrt{n}}\sqrt{\frac{\pi}{2}}\sqrt{1+\sum_{\alpha\in(\mathbb{N}^d)^*}\frac{c_{|\alpha|}}{|\alpha|}\mathbb{E}_X\left(H_\alpha(X)^2\right)} + \frac{\ell_\infty 2\sqrt{2}}{\sqrt{n}\Omega(f^*)}\sqrt{\log\frac{1}{\delta}} \le \mu$$

then f^μ belongs to $\mathcal{C}_{D,\tau}$ with probability larger than $1-\delta$. If we choose $\mu = \frac{8GD}{\sqrt{n}}\sqrt{\frac{\pi}{2}}\sqrt{1+\sum_{\alpha\in(\mathbb{N}^d)^*}\frac{c_{|\alpha|}}{|\alpha|}\mathbb{E}_X\left(H_\alpha(X)^2\right)} + \frac{\ell_\infty 2\sqrt{2}}{\Omega(f^*)\sqrt{n}}\sqrt{\log\frac{1}{\delta}}$, then

$$\begin{aligned}
\mathcal{R}_\mu(f^\mu) &\le \mathcal{R}_\mu(f^{\mu*}) + \tau \\
&\le \mathcal{R}_\mu(f^{\mu*}) + \tau \\
&\le \mathcal{R}_\mu(f^*) + \tau \\
&\le \mathcal{R}(f^*) + \mu\Omega(f^*) + \tau \\
&\le \mathcal{R}(f^*) + 2\mu\Omega(f^*) \\
&\le \mathcal{R}(f^*) \\
&\quad + \Omega(f^*)\left(\frac{16G}{\sqrt{n}}\sqrt{\frac{\pi}{2}}\sqrt{1+\sum_{\alpha\in(\mathbb{N}^d)^*}\frac{c_{|\alpha|}}{|\alpha|}\mathbb{E}_X\left(H_\alpha(X)^2\right)}\right) + \frac{\ell_\infty 4\sqrt{2}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}
\end{aligned}$$

and $\Omega(f^\mu)\le D = 2\Omega(f^*)$ with probability larger than $1-\delta$. □

4.4. Dependence on problem parameters

As we have seen, Theorem 4.1 depends on some quantities we detail now. First, we provide a definition of subgaussian real variables, as given by [30].

Definition 4.1 (Subgaussian Variables)**.** *Let Z be a real-valued (not necessarily centred) random variable. Z is subgaussian with variance proxy σ^2 if and only if*

$$\forall t > 0, \max\left(\mathbb{P}(Z \geq t), \mathbb{P}(Z \leq -t)\right) \leq e^{-\frac{t^2}{2\sigma^2}}.$$

Data distribution. To begin, we aim to establish an upper bound for the expectation of the squared Hermite polynomials over the covariates.

Lemma 4.4 (Analysis of data-dependent terms in Theorem 4.1)**.** *Let $\alpha \in \mathbb{N}^d$.*

1. *If $X \sim \mathcal{N}(0, I_d)$, then*
$$\mathbb{E}_X(H_\alpha(X)^2) = 1.$$
2. *If X is such that $\|X\|_2 \leq R$ a.s., then*
$$\mathbb{E}_X(H_\alpha(X)^2) \leq e^{\frac{R^2}{2}}.$$
3. *If X is such that $\|X\|_2$ is a subgaussian variable with variance proxy bounded by $\sigma^2 < 1/(36e)$, then*
$$\mathbb{E}_X(H_\alpha(X)^2) \leq e^{36e\sigma^2} \leq e.$$

The proof of this lemma is provided in Appendix A.4. Note that independence between the coordinates is not required, except in the first case, which is an illustration of the definition of the Hermite polynomials. It is worth noting that except in the Gaussian case, the bounds may not be ideal with respect to their dependency on d. However, these bounds rely heavily on the bound for Hermite polynomials in Equation (2.3), which is valid for all points on the real line and for all one-dimensional Hermite polynomials. Thus, it is expected that better bounds in expectation are possible.

Choice of $(c_k)_{k>0}$. The quantities in Theorem 4.1 are influenced by the design of the penalty, which is determined by the choice of the sequence $(c_k)_{k>0}$. This dependency is observed in $\Omega(f^*)$, ℓ_∞, and $\sqrt{1 + \sum_{\alpha \in (\mathbb{N}^d)^*} \frac{c_{|\alpha|}}{|\alpha|} \mathbb{E}_X\left(H_\alpha(X)^2\right)}$. It is worth noting that the bounds provided in Lemma 4.4 do not rely on the specific value of α. Therefore, our focus is now on bounding the summation term $\sum_{\alpha \in (\mathbb{N}^d)^*} \frac{c_{|\alpha|}}{|\alpha|}$.

Lemma 4.5 (Analysis of terms depending on $(c_k)_{k>0}$ in Theorem 4.1)**.** *If $c_{|\alpha|} = \rho^{|\alpha|}$, with $\rho \in (0,1)$*

$$\sum_{\alpha \in (\mathbb{N}^d)^*} \frac{c_{|\alpha|}}{|\alpha|} \leq \frac{1}{(1-\rho)^d}$$

and if $c_{|\alpha|} = \mathbb{1}_{|\alpha| \leq M}$,

$$\sum_{\alpha \in (\mathbb{N}^d)^*} \frac{c_{|\alpha|}}{|\alpha|} \leq \frac{M+1}{d} \binom{M+d}{M+1}.$$

The proof of this result can be found in Appendix A.5. By combining the different results, in the case of bounded data, for example, we can derive a corollary of Theorem 4.1 as follows

Corollary 4.1 (High-probability bound on expected risk of regularised estimator for bounded data)**.** *Under Assumption 4.1 and Assumptions 4.2.1, 4.2.2, 4.2.3 with* $D = 2\Omega(f^*)$*, 4.2.4, if* $\|X\|_2 \leq R$ *a.s.,* $(c_k)_{k>0} = (\rho^k)_{k>0}$*, then for any* $\delta \in (0,1)$*, with the choice of regularising parameter*

$$\mu = \frac{G}{\sqrt{n}}\sqrt{1 + \frac{e^{R^2/2}}{(1-\rho)^d}}\left(8\sqrt{\frac{\pi}{2}} + 2\sqrt{2}\sqrt{\log\frac{2}{\delta}}\right),$$

with probability larger than $1-\delta$

$$\mathcal{R}(f^\mu) \leq \mathcal{R}(f^*) + \Omega(f^*)\frac{G}{\sqrt{n}}\sqrt{1 + \frac{e^{R^2/2}}{(1-\rho)^d}}\left(16\sqrt{\frac{\pi}{2}} + 4\sqrt{2}\sqrt{\log\frac{2}{\delta}}\right)$$

and $\Omega(f^\mu) \leq 2\Omega(f^*)$.

The proof is provided in Appendix A.6. We note that the choice of the regularisation parameter is independent of the unknown norm $\Omega(f^*)$ or the distribution of X, as long as R is known. In the derived bound, the value of G can be independent of d for certain loss functions such as the logistic loss. We observe that $\Omega(f^*)$ does not depend on the dimension d, but solely on the number of variables or the dimension of the linear subspace s. It is important to note that the method exhibits a strong dependence on the dimension, which does not overcome the curse of dimensionality. However, this is merely the first step towards solving the multi-index model through regularised empirical risk minimisation, leaving room for future work and improvements.

5. Numerical study

In this section, we present the numerical results that demonstrate the behaviour and performance of **RegFeaL**. The implementation of the estimator, as well as the code to run the experiments, can be accessed online at `https://github.com/BertilleFollain/RegFeaL`. The **RegFeaL** estimator class is designed to be compatible with the Scikit-learn API [22], ensuring seamless integration with existing machine learning workflows.

5.1. Setup

We describe the experiment setup, which includes data simulation, training procedure and metrics for evaluation.

Data. In each generated dataset, depending on whether we consider feature learning or variable selection, we construct the linear subspace P differently. In the feature learning case, we sample a matrix from the set of $d \times d$ orthogonal matrices O_d and select its first s columns to form P. For variable selection, we simply consider the first s variables to be the relevant ones. Note that while our experiments were conducted with independently generated covariates, our method is invariant to rotations (in the feature learning case) and sign changes of the data (in both feature learning and variable selection). As such, it is robust to potential correlation between the covariates. The i.i.d dataset $(x^{(i)}, y^{(i)})_{i\in[n]}$ is then generated as follows

$$
\begin{aligned}
&X \sim \mathcal{U}\{[-\sqrt{3}, \sqrt{3}]\}^d \\
&f^*(x) = \sin(2(P^\top x)_1) + \sin(2(P^\top x)_2), \forall x \in \mathbb{R}^d \text{ (sinus dataset)} \\
&f^*(x) = (P^\top x)_1 + (P^\top x)_2 - (P^\top x)_1^2 - (P^\top x)_2^2 + 2(P^\top x)_1 (P^\top x)_2^3 - 4, \\
&\forall x \in \mathbb{R}^d \text{ (polynomial dataset)} \\
&Y = f^*(X) + \sigma\varepsilon,\ \varepsilon \sim \mathcal{N}(0,1).
\end{aligned}
$$

Each component of X has mean 0 and variance 1. Notably, in both datasets, the true regression function f^* depends on $s = 2$ linear combinations of the original variables. The importance of the noise can be controlled through the parameter σ. The test set $(x^{(i)}_{\text{test}}, y^{(i)}_{\text{test}})_{i\in[n_{\text{test}}]}$ is generated in a similar manner as the training set.

Training. The loss that we consider is the quadratic loss. We train **RegFeaL** on the training set with fixed values of λ and r, and we cross-validate on μ and ρ. The number of iterations n_{iter} depends on the experiment. Some of the parameters are the same in all experiments, such as $n_{\text{test}} = 5000, s = 2, \lambda = 10^{-8}/d^{(2-r)/r}, r = 0.33$.

The values of the grid used for cross-validation can be found in Appendix B. The training pipeline differs between Experiment 1 and Experiments 2 and 3.

In Experiment 1, for each parameter tuple (ρ, μ), we estimate the number of relevant dimensions $\hat{s}$ using $\hat{s} := |\{a \in [d], (\eta^{\lambda,\mu}_{\text{feat}})_a^{r/(2-r)} \geq 1/d\}|$. Recall that $\eta_a^{r/(2-r)}$, represents the importance of feature a, and at initialisation, it is set to $1/d$ for all $a \in [d]$. We then select $\hat{P}$ as the set of $\hat{s}$ eigenvectors of $\Lambda^{\lambda,\mu}_{\text{feat}}$ corresponding to the $\hat{s}$ largest eigenvalues. Finally, we train a final regressor using Multivariate Adaptive Regression Splines (MARS) [12] on the dataset $(\hat{P}^\top x^{(i)}, y^{(i)})_{i\in[n]}$.

In Experiments 2 and 3, we simply use the output $f^{\lambda,\mu}_{\text{feat}}$ of Algorithm 1 as the prediction function. In both cases, the R^2 score is used as the evaluation metric, which is described in Equation (5.1).

Metrics. We evaluate the performance of **RegFeaL** using two metrics: the R^2 score [32] for regression performance and an adapted Grassmannian distance for feature learning performance.

The R^2 score is computed as

$$1 - \frac{\sum_{i=1}^{n_{\text{test}}} (y_{\text{test}}^{(i)} - y_{\text{pred}}^{(i)})^2}{\sum_{i=1}^{n_{\text{test}}} (y_{\text{test}}^{(i)} - \bar{y}_{\text{test}})^2}, \tag{5.1}$$

where $\bar{y}_{\text{test}}$ is the mean of the test response values. The R^2 score can be computed on both the training and test sets. A score of 1 indicates the best possible performance, while a score of $-\infty$ indicates the worst performance. A constant estimator that predicts the average response value corresponds to a score of 0.

For the feature learning score, we compute the Grassmannian distance between the true subspace P and the estimated subspace $\hat{P}$, which corresponds to the s largest eigenvalue for the score computation. Note that the knowledge of s is only necessary to compute this score and not necessary for training. Note also that this is not the same $\hat{P}$ that was used to retrain **MARS** in Experiment 1, as the dimension of that one is estimated. The score is defined as

$$\|P(P^\top P)^{-1}P^\top - \hat{P}(\hat{P}^\top \hat{P})^{-1}\hat{P}^\top\|^2/(2s) \text{ if } s \leq d/2$$
$$\|P(P^\top P)^{-1}P^\top - \hat{P}(\hat{P}^\top \hat{P})^{-1}\hat{P}^\top\|^2/(2(d-s)) \text{ if } s > d/2,$$

where s is the number of relevant dimensions. The best possible score is 1, indicating a perfect match between the true and estimated subspaces, while a score of 0 indicates no correspondence between the subspaces.

In the setting of variable selection, this discussion can be adapted as discussed in Section 3. The omitted details of the experiments can be found in Appendix B.

5.2. Results

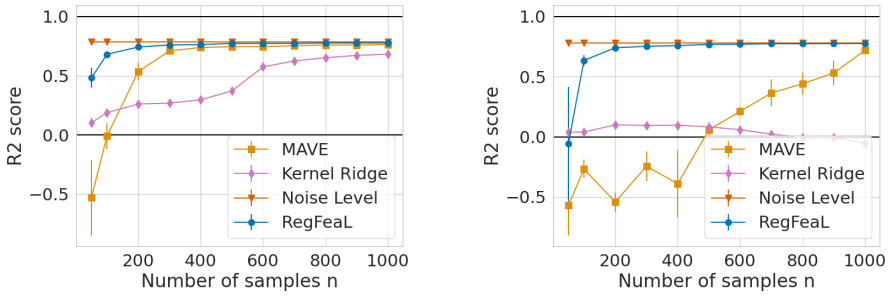
We now provide the results of the experiments.

Experiment 1. In this experiment, we investigate the dependence on the dimension of the variables d and the number of samples n. We perform the training procedure described earlier, including the retraining step using **MARS** [12] on the projected data. We evaluate the performance on both the sinus dataset and the polynomial dataset with noise levels $\sigma = 0.5$ and $\sigma = 2.5$ respectively. For the sinus dataset, we consider both the variable selection and feature learning settings. We conduct a total of $n_{\text{iter}} = 5$ iterations, and the grid used for cross-validation can be found in Appendix B.

To provide a comparison, we also include the performance of the state-of-the-art method **MAVE** [34], which is based on local averaging and does not use regularisation. In our implementation, we follow the recommended procedure for **MAVE**, which involves first training the Outer Product of Gradients (OPG) method to determine the effective dimensionality reduction (e.d.r) space. We use cross-validation to select the underlying dimension of the space and then retrain the model using **MARS** on the projected data. This allows us to compute the R^2 score. For the feature learning score, we compute it based on the learned

effective dimensionality reduction (e.d.r) space. Specifically, we choose $s = 2$ as the dimension of the subspace to compute the score, following the same approach as **RegFeaL**.

Additionally, we include the R^2 score for **Kernel Ridge**, which uses kernel ridge regression with the kernel $k(x, x') = \sum_{\alpha \in (\mathbb{N}^d)^*} c_{|\alpha|} H_\alpha(x) H_\alpha(x')$ and the hyperparameter λ, which we cross-validate over. To provide a comprehensive analysis, we also display the noise level, which represents the best achievable score considering the noise level σ. We repeat the entire experiment five times, each time with different data, and present the average results with error bars of $\pm \sigma_{\text{exp}}/\sqrt{5}$, where σ_{exp} is the standard deviation of the scores across the repetitions. The results of the experiment can be found in Figures 1, 2, and 3.



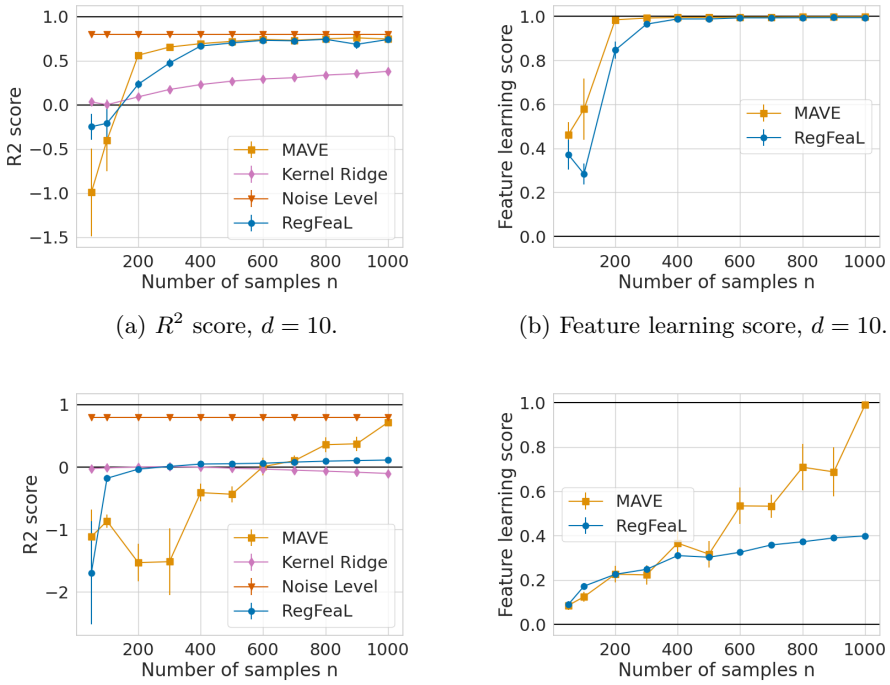
(a) R^2 score $d = 10$. (b) R^2 score, $d = 40$.

Fig 1: Performance dependency on d and n for the sinus dataset in the variable selection setting.

In all figures, we observe that the performance improves with a higher number of samples (n), which is expected, while it deteriorates with a larger dimension (d), which is typical behaviour.

In Figure 1, we focus on the R^2 score for the sinus dataset in the variable regression setting. We observe that **RegFeaL** performs well in both dimensions (10 and 40) without requiring a large number of samples. However, **Kernel Ridge** fails in dimension 40 as the kernel cannot effectively capture the dependency on only 2 variables. As for **MAVE**, it does not benefit from the knowledge that this is a variable selection problem, unlike **RegFeaL**, resulting in a higher sample requirement, particularly in dimension 40.

In Figure 2, we examine the R^2 score and the feature learning score for the sinus dataset in the feature learning setting. We observe that **MAVE** and **RegFeaL** exhibit similar behaviour in dimension 10, reaching the noise level for the R^2 score and achieving a perfect feature learning score with enough samples. However, in dimension 40, **MAVE** struggles significantly when the number of samples is low, while **RegFeaL** requires a substantially larger sample size to accurately learn the e.d.r. space. Our interpretation is that in this setting, where



(a) R^2 score, $d = 10$.

(b) Feature learning score, $d = 10$.

(c) R^2 score, $d = 40$.

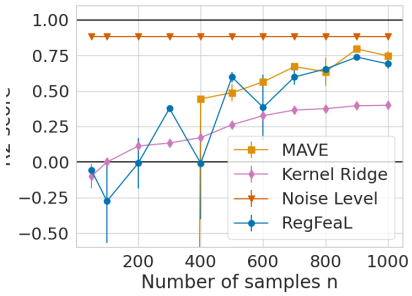
(d) Feature learning score, $d = 40$.

Fig 2: Performance dependency on d and n for the sinus dataset in the feature learning setting.

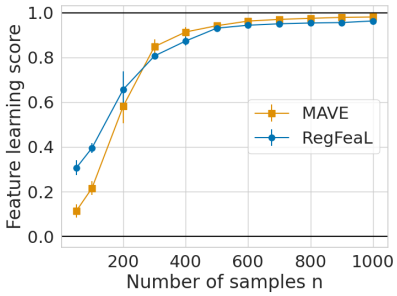
the true regression function uses a sinus, **RegFeaL** is hindered by its definition using a basis of polynomials.

In Figure 3, we investigate the R^2 score and the feature learning score for the polynomial dataset in the feature learning setting. The feature learning performance of **MAVE** and **RegFeaL** is similar in this scenario. Regarding the R^2 score, **Kernel Ridge** encounters difficulties in dimension 40 as it does not benefit from the underlying hidden structure. In dimension 10, **RegFeaL** performs similarly to **MAVE**, but in dimension 40, it outperforms **MAVE** as **MAVE** tends to be overly restrictive and consistently underestimates the number of linear features required to provide a good fit when the e.d.r. space is not perfectly learnt. In contrast, **RegFeaL** is less conservative, allowing us to leverage more features when the number of samples is too low to accurately estimate them.

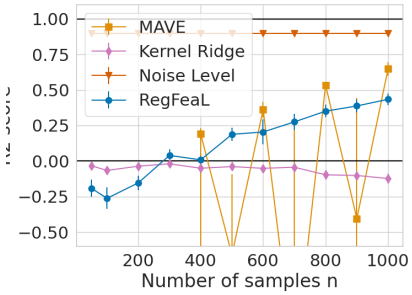
Experiment 2. In this experiment, we investigate the impact of the number of random features m (as discussed in Section 3.3) on the R^2 score and feature learning score for different values of n. The dimension d is fixed at 10, while the



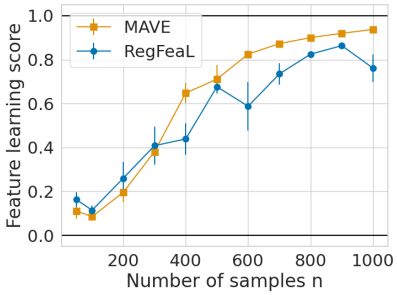
(a) R^2 score, $d = 20$.



(b) Feature learning score, $d = 40$.



(c) R^2 score, $d = 10$.



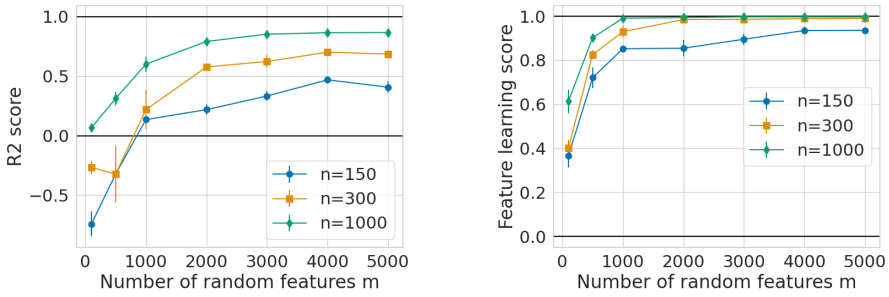
(d) Feature learning score, $d = 40$.

Fig 3: Performance dependency on d and n for the polynomial dataset in the feature learning setting.

true underlying dimension s is 2. We consider the noiseless setting $\sigma = 0$ and use the sinus dataset. The same methodology is applied for error bar computation as in Experiment 1. The results are presented in Figure 4.

We observe that both the R^2 score and feature learning score improve with an increase in the number of random features m. This observation aligns with the discussion in Section 3.3, where a larger value of m leads to a better approximation of the kernel k_Λ, and allows for a wider range of α and H_α, resulting in enhanced descriptive power and improved fit and prediction of the subspace. However, we note that beyond a certain value of m, the performance improvement levels off while computational costs continue to rise. This suggests that choosing excessively large values of m does not provide any significant benefit.

Experiment 3. In this experiment, we maintain the number of samples $n = 5000$, the number of random features $m = 2500$, the dimension $d = 10$, and the underlying dimension $s = 2$ fixed. We work with the noiseless sinus dataset, i.e., $\sigma = 0$, and examine the training behaviour of **RegFeaL** over the iterations. We train the model using cross-validation based on the R^2 score and set $n_{\text{iter}} = 10$.



(a) R^2 score (b) Feature learning score

Fig 4: Influence of the number of random features

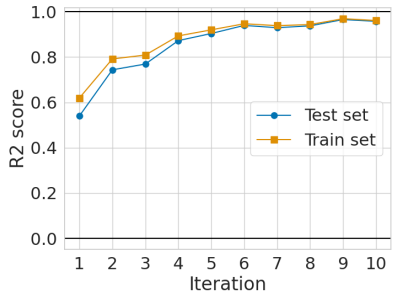
The results of this experiment are depicted in Figure 5.

In Figure 5a, we observe that the R^2 score improves across the iterations on both the test set and the training set. However, the behaviour is not strictly increasing on the training set. This can be attributed to the fact that the kernel approximation differs at each iteration, leading to variations in the fit.

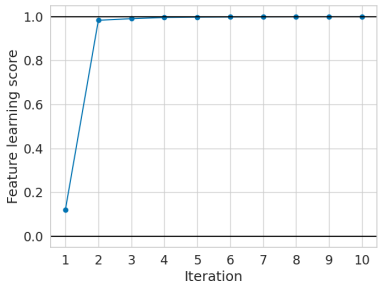
Figure 5b demonstrates that the features are learned more rapidly than the fit. It is important to note that the feature learning score assumes knowledge of the underlying dimension $s = 2$. Hence, an important question is whether the estimated value of s is accurate.

In Figure 5c, we observe the values of $\eta_a^{r/(2-r)}$ for all $a \in [d]$ across the iterations. Recall that $\sum_{a=1}^{d} \eta_a^{r/(2-r)} = 1$ and that $\eta_a^{r/(2-r)}$ represents the relative importance of feature $(R^\top x)_a$. Initially, all $\eta_a^{r/(2-r)}$ are equal to $1/d$. As the training progresses, most of the components of $\eta^{r/(2-r)}$ decrease, while two components increase, surpassing the threshold of $1/d$. These two components correspond to the relevant dimensions, indicating that the correct number of dimensions would be easily predicted. Additionally, we observe that these two components of η have relatively similar values, which aligns with the symmetry of the regression function in this example.

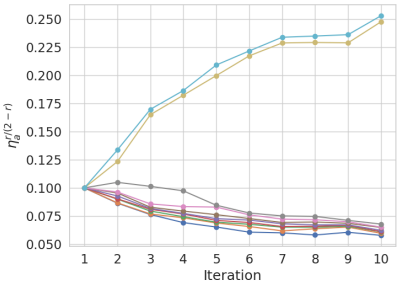
Figure 5d displays the empirical density (in log scale) of α_a for two different values of $a \in [d]$ (specifically, $a_{\text{small}} := \arg\min_{a\in[d]} \eta_a$ and $a_{\text{large}} := \arg\max_{a\in[d]} \eta_a$ for the final η) at two different iterations: the first and last iteration. During the first iteration, the distributions of α_a for a_{large} and a_{small} are equal, which aligns with the initialisation discussed in Section 3.3 (all components of η are equal). However, at the end of the optimisation, we observe that the distribution of $\alpha_{a_{\text{small}}}$, corresponding to a non-important linear feature, remains almost constant at 0. Conversely, the distribution of $\alpha_{a_{\text{large}}}$, representing an important linear feature, is more widely spread, which is beneficial to the fit.



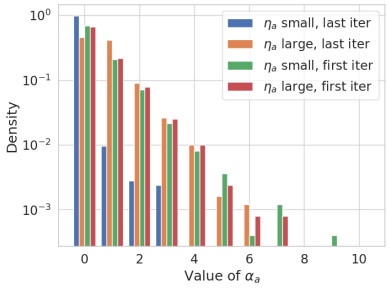
(a) R^2 score.



(b) Feature learning score.



(c) $\eta_a^{r/(2-r)}, \forall a \in [d]$.



(d) Empirical distribution of α.

Fig 5: Training behaviour.

6. Conclusion

We addressed the challenge of prediction function estimation in multi-index models by proposing a novel approach **RegFeaL**. Our method combines empirical risk minimisation with derivative-based regularisation to simultaneously estimate the prediction function, the relevant linear transformation, and its dimension. By leveraging the orthogonality and rotation invariance properties of Hermite polynomials, **RegFeaL** captures the underlying structure of the data. Through alternative minimisation, we iteratively rotate the data to better align it with the leading dimensions.

Theoretical results support the statistical consistency of the expected risk of our estimator and provide explicit rates of convergence. We demonstrated the performance and effectiveness of our method through extensive empirical experiments on diverse datasets. One of the strengths of our approach is that it does not rely on strong assumptions about the distribution shape or prior knowledge of the subspace dimension.

However, we acknowledge that our method is still subject to the curse of dimensionality, as indicated by the theoretical results showing an exponential

dependence on the dimension of the covariates. Nonetheless, we believe that our findings will contribute to further developments in representation learning and high-dimensional data analysis. Regularisation is a versatile approach that can be applied to a wide range of problems where an empirical risk can be formulated, foregoing the limitations of some methods solely based on the square loss in supervised learning.

There are several interesting directions for future research. One possibility is exploring alternative bases other than Hermite polynomials. Additionally, investigating more efficient algorithms and strategies for handling high-dimensional data could be valuable. Furthermore, examining the applicability of our approach to various types of problems and datasets would also be worth pursuing.

Appendix A: Additional proofs and results

A.1. Proof of Lemma 2.2

Proof of Lemma 2.2. We denote by $\mathcal{N}(0, I_d)$ the d-dimensional normal distribution with mean $0 \in \mathbb{R}^d$ and covariance matrix I_d. For any $k \in \mathbb{N}$, $x, x' \in \mathbb{R}^d$, using $\forall z \in \mathbb{R}$, $h_k(z) = \frac{1}{\sqrt{k!}} \mathbb{E}_{Y \sim \mathcal{N}(0,1)} (z + iY)^k$ (which can be shown by recurrence), we have

$$\begin{aligned}
\sum_{|\alpha|=k} H_\alpha(x) H_\alpha(x') &= \sum_{|\alpha|=k} \prod_{a=1}^{d} h_{\alpha_a}(x_a) h_{\alpha_a}(x'_a) \\
&= \mathbb{E}_{Y,Y' \sim \mathcal{N}(0,I_d)} \left(\sum_{|\alpha|=k} \prod_{a=1}^{d} \frac{1}{\alpha_a!} (x_a + iY_a)^{\alpha_a} (x'_a + iY'_a)^{\alpha_a} \right) \\
&= \frac{1}{k!} \mathbb{E}_{Y,Y' \sim \mathcal{N}(0,I_d)} \left(\left(x^\top x' - Y^\top Y' + i(x^\top Y' + Y^\top x') \right)^k \right).
\end{aligned}$$

This shows rotational invariance, that is, for any orthogonal matrix $R \in O_d$,

$$\sum_{|\alpha|=k} H_\alpha(x) H_\alpha(x') = \sum_{|\alpha|=k} H_\alpha(Rx) H_\alpha(Rx').$$

□

A.2. Proof of Lemma 4.1

Proof of Lemma 4.1. Define $\mathcal{H} = \{h : (x, y) \in \mathcal{X} \times \mathcal{Y} \to \ell(y, f(x)), \text{ for } f \in \mathcal{G}\}$. We have that

$$\begin{aligned}
&\sup_{f \in \mathcal{G}} \left(\mathcal{R}(f) - \widehat{\mathcal{R}}(f) \right) + \sup_{f \in \mathcal{G}} \left(\widehat{\mathcal{R}}(f) - \mathcal{R}(f) \right) \\
&= \sup_{h \in \mathcal{H}} \left(\mathbb{E}(h(z)) - \frac{1}{n} \sum_{i=1}^{n} h(z^{(i)}) \right) + \sup_{h \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^{n} h(z^{(i)}) - \mathbb{E}(h(z)) \right).
\end{aligned}$$

We define the Rademacher complexity of the set $\mathcal{H}$ by

$$R_n(\mathcal{H}) = \mathbb{E}_{\mathcal{D},\varepsilon\sim\left(\mathcal{U}\{-1,1\}\right)^n}\left(\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i h(z^{(i)})\right),$$

where $\varepsilon \sim \left(\mathcal{U}\{-1,1\}\right)^n$ means that each component of ε is independent and follows the uniform distribution over the set $\{-1,1\}$.

Using Proposition 4.2 from [4], we obtain

$$\mathbb{E}_{\mathcal{D}}\left(\sup_{h\in\mathcal{H}}\mathbb{E}(h(z)) - \frac{1}{n}\sum_{i=1}^{n}h(z^{(i)})\right) \leq 2R_n(\mathcal{H})$$
$$\mathbb{E}_{\mathcal{D}}\left(\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}h(z^{(i)}) - \mathbb{E}(h(z))\right) \leq 2R_n(\mathcal{H}).$$

Now from Assumption 4.2.2 and using Proposition 4.3 from [4]

$$R_n(\mathcal{H}) \leq G\cdot R_n(\mathcal{G}),$$

with

$$R_n(\mathcal{G}) = \mathbb{E}_{\mathcal{D},\varepsilon\sim\left(\mathcal{U}\{-1,1\}\right)^n}\left(\sup_{f\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x^{(i)})\right).$$

We have from Exercise 4.9 from [4] that $R_n(\mathcal{G}) \leq \sqrt{\frac{\pi}{2}}G_n(\mathcal{G})$. Combining all inequalities yields the desired result. □

A.3. Lemma A.1 and its proof

Lemma A.1. *Under Assumption 4.1, Assumptions 4.2.1, 4.2.2, 4.2.3, with* $D \geq \Omega(f^*)$, *and* $f^D := \arg\min_{f\in\mathcal{F},\ \Omega(f)\leq D}\widehat{\mathcal{R}}(f)$, *for any* $\delta\in(0,1)$, *with probability larger than* $1-\delta$

$$\mathcal{R}(f^D) \leq \mathcal{R}(f^*) + \frac{4GD}{\sqrt{n}}\sqrt{\frac{\pi}{2}}\sqrt{1+\sum_{\alpha\in(\mathbb{N}^d)^*}\frac{c_{|\alpha|}}{|\alpha|}\mathbb{E}_X\left(H_\alpha(X)^2\right)} + \frac{\ell_\infty 2\sqrt{2}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}.$$

Proof of Lemma A.1. Define $\mathcal{G} := \{f\in\mathcal{F},\ \Omega(f)\leq D\}$. We apply McDiarmid's inequality [7] to $\sup_{f\in\mathcal{G}}\mathcal{R}(f)-\widehat{\mathcal{R}}(f)+\sup_{f\in\mathcal{G}}\widehat{\mathcal{R}}(f)-\mathcal{R}(f)$, which has bounded variation with constant $4\ell_\infty/n$, yielding that for all $\delta\in(0,1)$

$$\mathbb{P}_{\mathcal{D}}\Bigg(\sup_{f\in\mathcal{G}}\mathcal{R}(f)-\widehat{\mathcal{R}}(f)+\sup_{f\in\mathcal{G}}\widehat{\mathcal{R}}(f)-\mathcal{R}(f) \leq$$
$$\mathbb{E}\big(\sup_{f\in\mathcal{G}}\mathcal{R}(f)-\widehat{\mathcal{R}}(f)+\sup_{f\in\mathcal{F}}\widehat{\mathcal{R}}(f)-\mathcal{R}(f)\big) + \frac{\ell_\infty 2\sqrt{2}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}\Bigg) \geq 1-\delta.$$

We recall that

$$\mathcal{R}(f^D)-\mathcal{R}(f^*) \leq \sup_{f\in\mathcal{G}}\mathcal{R}(f)-\widehat{\mathcal{R}}(f)+\sup_{f\in\mathcal{G}}\widehat{\mathcal{R}}(f)-\mathcal{R}(f)$$

and from the proof of Lemma 4.3

$$\mathbb{E}\left(\sup_{f\in\mathcal{G}}\left(\mathcal{R}(f)-\widehat{\mathcal{R}}(f)\right)+\sup_{f\in\mathcal{G}}\left(\widehat{\mathcal{R}}(f)-\mathcal{R}(f)\right)\right)$$
$$\leq \frac{4GD}{\sqrt{n}}\sqrt{\frac{\pi}{2}}\sqrt{1+\sum_{\alpha\in(\mathbb{N}^d)^*}\frac{c_{|\alpha|}}{|\alpha|}\mathbb{E}_X\left(H_\alpha(X)^2\right)},$$

yielding the final result. □

A.4. Proof of Lemma 4.4

Proof of Lemma 4.4. The result for centred normal data with identity covariance matrix is by the construction of the Hermite polynomials [15].

If $\|X\|_2$ is bounded by R, using the bound from Equation (2.3), we get that

$$\mathbb{E}_X(H_\alpha(X)^2)\leq\mathbb{E}(e^{\|X\|^2/2})\leq\mathbb{E}_X\left(e^{R^2/2}\right)\leq e^{\frac{R^2}{2}}.$$

If X is such that $\|X\|$ is subgaussian with variance proxy σ^2, we know that $\forall\lambda\leq 1/(6\sqrt{2e}\sigma)$, then $\mathbb{E}_X(e^{\|X\|^2\lambda^2})\leq e^{72e\lambda^2\sigma^2}$ [30, Proposition 2.5.2]. Therefore, using the bound from Equation (2.3), we have

$$\mathbb{E}_X(H_\alpha(X)^2)\leq\mathbb{E}(e^{\|X\|^2/2})\leq e^{36e\sigma^2}\leq e$$

This concludes the study of $\mathbb{E}_X\left(H_\alpha(X)^2\right)$. □

A.5. Proof of Lemma 4.5

Proof of Lemma 4.5. Using d-dimensional geometric random variables, we know that

$$\sum_{\alpha\in\mathbb{N}^d}(1-\rho)^d\rho^{|\alpha|}=1,\text{ and therefore }\sum_{\alpha\in(\mathbb{N}^d)^*}\frac{\rho^{|\alpha|}}{|\alpha|}\leq\frac{1}{(1-\rho)^d}.$$

For the other setting,

$$\sum_{\alpha\in(\mathbb{N}^d)^*,|\alpha|\leq M}\frac{1}{|\alpha|}=\sum_{k=1}^{M}\frac{1}{k}\binom{d-1+k}{d-1}\leq\frac{M+1}{d}\binom{M+d}{M+1},$$

which concludes the proof. □

A.6. Proof of Corollary 4.1

Proof of Corollary 4.1. First, we note from Lemma 4.4 that for any $\alpha\in\mathbb{N}^d$, we have $\mathbb{E}_X\left(H_\alpha(X)^2\right)\leq e^{R^2/2}$. Additionally, from Lemma 4.5, we know that $\sum_{\alpha\in(\mathbb{N}^d)^*}\frac{c_{|\alpha|}}{|\alpha|}\leq\frac{1}{(1-\rho)^d}$.

Next, we aim to improve the use of McDiarmid's inequality by bounding the deviation of $\sup_{f\in\mathcal{F},\Omega(f)\leq D}\mathcal{R}(f)-\widehat{\mathcal{R}}(f)+\sup_{f\in\mathcal{F},\Omega(f)\leq D}\widehat{\mathcal{R}}(f)-\mathcal{R}(f)$ when a single data point $(x^{(i)},y^{(i)})$ is changed to $(\tilde{x}^{(i)},\tilde{y}^{(i)})$ changing the dataset from $\mathcal{D}$ to $\tilde{\mathcal{D}}$. In the original proof of Theorem 4.1, we used $4l_\infty/n$ as our bound, but we can provide a tighter bound. We write $\widehat{\mathcal{R}}_{\mathcal{D}}(f)$ to specify the dependency on the dataset. We also write $\mathcal{G}:=\{f\in\mathcal{F},\ \Omega(f)\leq D\}$. Specifically, we have

$$\begin{aligned}
&\sup_{f\in\mathcal{G}}\mathcal{R}(f)-\widehat{\mathcal{R}}_{\mathcal{D}}(f)-\sup_{f\in\mathcal{G}}\mathcal{R}(f)-\widehat{\mathcal{R}}_{\tilde{\mathcal{D}}}(f)\\
&=\sup_{f\in\mathcal{G}}\mathcal{R}(f)-\widehat{\mathcal{R}}_{\tilde{\mathcal{D}}}(f)+\frac{1}{n}\ell(\tilde{y}^{(i)},f(\tilde{x}^{(i)}))-\frac{1}{n}\ell(y^{(i)},f(x^{(i)}))-\sup_{f\in\mathcal{G}}\mathcal{R}(f)-\widehat{\mathcal{R}}_{\tilde{\mathcal{D}}}(f)\\
&\leq\sup_{f\in\mathcal{G}}\frac{1}{n}\ell(\tilde{y}^{(i)},f(\tilde{x}^{(i)}))-\frac{1}{n}\ell(y^{(i)},f(x^{(i)})),
\end{aligned}$$

and similarly

$$\sup_{f\in\mathcal{G}}\widehat{\mathcal{R}}_{\mathcal{D}}(f)-\mathcal{R}(f)-\sup_{f\in\mathcal{G}}\widehat{\mathcal{R}}_{\tilde{\mathcal{D}}}(f)-\mathcal{R}(f)\leq\sup_{f\in\mathcal{G}}\frac{1}{n}\ell(y^{(i)},f(x^{(i)}))-\frac{1}{n}\ell(\tilde{y}^{(i)},f(\tilde{x}^{(i)}).$$

Combining both and taking the argmax functions f_1 and f_2, we obtain

$$\begin{aligned}
&\sup_{f\in\mathcal{G}}\mathcal{R}(f)-\widehat{\mathcal{R}}_{\mathcal{D}}(f)-\sup_{f\in\mathcal{G}}\mathcal{R}(f)-\widehat{\mathcal{R}}_{\tilde{\mathcal{D}}}(f)+\sup_{f\in\mathcal{G}}\widehat{\mathcal{R}}_{\mathcal{D}}(f)-\mathcal{R}(f)-\sup_{f\in\mathcal{G}}\widehat{\mathcal{R}}_{\tilde{\mathcal{D}}}(f)-\mathcal{R}(f)\\
&\quad\leq\frac{1}{n}\ell(\tilde{y}^{(i)},f_1(\tilde{x}^{(i)})-\frac{1}{n}\ell(y^{(i)},f_1(x^{(i)}))+\frac{1}{n}\ell(y^{(i)},f_2(x^{(i)}))-\frac{1}{n}\ell(\tilde{y}^{(i)},f_2(\tilde{x}^{(i)})\\
&\quad\leq\frac{G}{n}(|(f_1-f_2)(x^{(i)}|+|(f_1-f_2)(\tilde{x}^{(i)})|)\\
&\quad\leq\frac{4}{n}G\sup_{f\in\mathcal{F},\Omega(f)\leq D,x\in\mathbb{R}^d,\|x\|_2\leq R}|f(x)|\\
&\quad\leq\frac{4}{n}GD\sup_{x\in\mathbb{R}^d,\|x\|_2\leq R}\Omega^*((H_\alpha(x))_\alpha)\\
&\quad\leq\frac{4}{n}GD\sup_{x\in\mathbb{R}^d,\|x\|_2\leq R}\sqrt{1+\sum_{\alpha\in(\mathbb{N}^d)^*}\frac{c_{|\alpha|}}{|\alpha|}H_\alpha(x)^2}\\
&\quad\leq\frac{4}{n}GD\sqrt{1+\frac{e^{R^2/2}}{(1-\rho)^d}}.
\end{aligned}$$

We can obtain the same exact bound for the opposite quantity of

$$\sup_{f\in\mathcal{G}}\mathcal{R}(f)-\widehat{\mathcal{R}}_{\mathcal{D}}(f)-\sup_{f\in\mathcal{G}}\mathcal{R}(f)-\widehat{\mathcal{R}}_{\tilde{\mathcal{D}}}(f)+\sup_{f\in\mathcal{G}}\widehat{\mathcal{R}}_{\mathcal{D}}(f)-\mathcal{R}(f)-\sup_{f\in\mathcal{G}}\widehat{\mathcal{R}}_{\tilde{\mathcal{D}}}(f)-\mathcal{R}(f)$$

by using the same arguments. We use this bound for $D=2\Omega(f^*)$. The result follows by employing the proof of Theorem 4.1. □

Appendix B: Technical details of the numerical experiments

Experiment 1. For **MAVE** and **RegFeaL**, the **MARS** final training used the default parameters provided by the py-earth python package (https://contrib.scikit-learn.org/py-earth/), except for the maximum degree, which was taken as the estimated dimension for both methods. **MAVE** was run using the provided CRAN package in R (https://cran.r-project.org/web/packages/MAVE/index.html) and the default parameters.

The number of iterations n_{iter} was set to 5. For **RegFeaL**, the cross-validation for $\rho \times \mu$ was done over the grid defined by $(0.01, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8)$ for ρ and $(1000, 100, 10, 1, 0.1, 0.01, 0.001)/d^{((2-r)/r)}$ for μ.

The cross-validation for **Kernel Ridge** was done on parameter λ, with the set of values $(1000, 100, 10, 1, 0.1, 0.01, 0.001)/d^{((2-r)/r)}$. The score of the noise level was estimated by $1 - \frac{n\sigma^2}{\sum_{i=1}^{n}(y_{\text{test}}^{(i)} - \bar{y}_{\text{test}})^2}$.

Experiment 2. For each value of n, we used cross-validation for the largest value of m and then used the selected ρ and λ for all other values of m. The cross-validation was done over the grid defined by $(0.2, 0.4, 0.6, 0.8, 1.0)$ for ρ and $(100, 1, 0.1, 0.01, 0.001)/d^{((2-r)/r)}$ for μ. The number of iterations n_{iter} was 3.

Experiment 3. The cross-validation for $\rho \times \mu$ was done over the grid defined by $(0.2, 0.4, 0.6, 0.8, 1.0)$ for ρ and $(100, 1, 0.1, 0.01, 0.001)/d^{((2-r)/r)}$ for μ.

Acknowledgements

The author thanks Lawrence Stewart, Antonin Brossollet and Oumayma Bounou for fruitful discussions related to this work. The authors are grateful to the CLEPS infrastructure from the Inria of Paris for providing resources and support, particularly Simon Legrand (https://paris-cluster-2019.gitlabpages.inria.fr/cleps/cleps-userguide/index.html). This work is funded in part by the French government under the management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support from the European Research Council (grants SEQUOIA 724063 and DYNASTY 101039676).

References

[1] Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics* **19** 563-570.

[2] Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society* **68** 337-404.

[3] Babichev, D. and Bach, F. (2018). Slice inverse regression with score functions. *Electronic Journal of Statistics* **12** 1507-1543.

[4] Bach, F. (2024). *Learning Theory from First Principles*. MIT Press to appear.
[5] Bach, F., Jenatton, R., Mairal, J. and Obozinski, G. (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning* **4** 1-106.
[6] Bartlett, P. and Mendelson, S. (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research* **3** 463-482.
[7] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.
[8] Brillinger, D. (2012). A Generalized Linear Model With "Gaussian" Regressor Variables. In *Selected Works of David Brillinger* 589-606. Springer.
[9] Cabannes, V., Pillaud-Vivien, L., Bach, F. and Rudi, A. (2021). Overcoming the curse of dimensionality with Laplacian regularization in semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)* **34** 30439-30451.
[10] Cook, R. D. and Weisberg, S. (1991). Sliced Inverse Regression for Dimension Reduction: Comment. *Journal of the American Statistical Association* **86** 328–332.
[11] Dalalyan, A., Juditsky, A. and Spokoiny, V. (2008). A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research* **9** 1647-1678.
[12] Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics* **19** 1-67.
[13] Fukumizu, K., Bach, F. and Jordan, M. (2009). Kernel dimension reduction in regression. *The Annals of Statistics* **37** 1871-1905.
[14] Györfi, L., Kohler, M., Krzyzak, A. and Walk, H. (2002). *A Distribution-free Theory of Nonparametric Regression. Springer Series in Statistics*. Springer.
[15] Hermite, C. (2009). *Sur un nouveau développement en série des fonctions*. In *Œuvres de Charles Hermite. Cambridge Library Collection - Mathematics* **2** 293–308. Cambridge University Press.
[16] Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *The Annals of Statistics* **29** 593-623.
[17] Jenatton, R., Audibert, J.-Y. and Bach, F. (2011). Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research* **12** 2777-2824.
[18] Jenatton, R., Obozinski, G. and Bach, F. (2010). Structured sparse principal component analysis. In *International Conference on Artificial Intelligence and Statistics (ICML)* 366-373.
[19] Jing Zeng, Q. M. and Zhang, X. (2024). Subspace Estimation with Automatic Dimension and Variable Selection in Sufficient Dimension Reduction. *Journal of the American Statistical Association* **119** 343–355.
[20] Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction. *Jour-*

nal of the American Statistical Association **86** 316-327.

[21] Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association* **87** 1025-1039.

[22] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Édouard Duchesnay (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12** 2825-2830.

[23] Recht, B., Fazel, M. and Parrilo, P. A. (2010). Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review* **52** 471-501.

[24] Rosasco, L., Villa, S., Mosci, S., Santoro, M. and Verri, A. (2013). Nonparametric sparsity and regularization. *Journal of Machine Learning Research* **14** 1665-1714.

[25] Rudi, A., Camoriano, R. and Rosasco, L. (2015). Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems (NeurIPS)* **28**.

[26] Stenseth, N. C., Tao, Y., Zhang, C., Bramanti, B., Büntgen, U., Cong, X., Cui, Y., Zhou, H., Dawson, L. A., Mooney, S. J., Li, D., Fell, H. G., Cohn, S., Sebbane, F., Slavin, P., Liang, W., Tong, H., Yang, R. and Xu, L. (2022). No evidence for persistent natural plague reservoirs in historical and modern Europe. *Proceedings of the National Academy of Sciences of the United States of America* **119**.

[27] Stoker, T. M. (1986). Consistent estimation of scaled coefficient. *Econometrica* **54** 1461-1481.

[28] Szegő, G. (1939). *Orthogonal Polynomials. American Mathematical Society Colloquium Publications*. American Mathematical Society.

[29] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **58** 267-288.

[30] Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

[31] Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics.

[32] Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20** 557-585.

[33] Xia, Y. (2008). A multiple-index model and dimension reduction. *Journal of the American Statistical Association* **103** 1631-1640.

[34] Xia, Y., Tong, H., Li, W. K. and Zhu, L.-X. (2002). An Adaptive Estimation of Dimension Reduction Space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64** 363-410.

[35] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B*

(Statistical Methodology) **68** 49-67.

[36] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894-942.