Conditional regression for single-index models

ALESSANDRO LANTERI[1], MAURO MAGGIONI[2] and STEFANO VIGOGNA[3]

[1] *Department of Economics and Statistics, University of Torino and Collegio Carlo Alberto, Torino, Italy. E-mail:* alessandro.lanteri@unito.it

[2] *Department of Mathematics and Department of Applied Mathematics & Statistics, Johns Hopkins University, Baltimore, USA. E-mail:* mauromaggionijhu@icloud.com

[3] *Department of Computer Science, Bioengineering, Robotics and Systems Engineering, University of Genova, Genova, Italy. E-mail:* vigogna@dibris.unige.it

The single-index model is a statistical model for intrinsic regression where the responses are assumed to depend on a single yet unknown linear combination of the predictors, allowing to express the regression function as $\mathbb{E}[Y|X] = f(\langle v, X\rangle)$ for some unknown *index* vector v and *link* function f. Estimators converging at the 1-dimensional min-max rate exist, but their implementation has exponential cost in the ambient dimension. Recent attempts at mitigating the computational cost yield estimators that are computable in polynomial time, but do not achieve the optimal rate. Conditional methods estimate the index vector v by averaging moments of X conditioned on Y, but do not provide generalization bounds on f. In this paper we develop an extensive non-asymptotic analysis of several conditional methods, and propose a new one that combines some benefits of the existing approaches. In particular, we establish $\sqrt{n}$-consistency for all conditional methods considered. Moreover, we prove that polynomial partitioning estimates achieve the 1-dimensional min-max rate for regression of Hölder functions when combined to any $\sqrt{n}$-consistent index estimator. Overall this yields an estimator for dimension reduction and regression of single-index models that attains statistical and computational optimality, thereby closing the statistical-computational gap for this problem.

MSC 2010 subject classifications: Primary 62G05; secondary 62G08, 62H99..
Keywords: Single-index model, dimension reduction, nonparametric regression, finite sample bounds.

1. Introduction

Consider the standard regression problem of estimating a function $F : \mathbb{R}^d \to \mathbb{R}$ from n samples $\{(X_i, Y_i)\}_{i=1}^n$, where the X_i's are independent realizations of a predictor variable $X \in \mathbb{R}^d$,

$$Y_i = F(X_i) + \zeta_i\,, \qquad i = 1, \dots, n\,, \tag{1}$$

and the ζ_i's are realizations, independent among themselves and of the X_i's, of a random variable ζ modeling noise. Under rather general assumptions on ζ and the distribution ρ of X, if we only know that F is s-Hölder regular (and, say, compactly supported), it is well-known that the min-max nonparametric rate for estimating F in $L^2(\rho)$ is $n^{-s/(2s+d)}$ [22]. This is an instance of the *curse of dimensionality*: the rate slows down dramatically

as the dimension d increases. Many regression models have been introduced throughout the decades to circumvent this phenomenon; see, for example, the classical reference [45]. When the covariates are intrinsically low-dimensional, concentrating on an unknown low-dimensional set, several estimators have been proved to converge at rates that are optimal with respect to the intrinsic dimension [3, 32, 33, 39, 40]. In other models, the domain may be high-dimensional, but the function itself is assumed to depend only on a small number of features. A classical case is the so-called *single-index model*, where F has the structure

$$F(x) = f(\langle v, x\rangle) \tag{2}$$

for some *index vector* $v \in \mathbb{R}^d$ (that we may assume unitary without loss of generality) and *link function* $f : \mathbb{R} \to \mathbb{R}$. In this context one may consider different estimation problems, depending on whether f is known (e.g. in logistic regression) or both f and v are unknown. We are interested in the latter case. Clearly, if v was known we could learn f by solving a 1-dimensional regression problem, which may be done efficiently for large classes of functions f. So the question is: what is the price to pay for not knowing v?

It was conjectured in [45] that the min-max rate for regression of single-index models is $n^{-s/(2s+1)}$, that is, the min-max rate for univariate functions: no statistical cost would have to be paid. This rate was proved for pointwise convergence with kernel estimators in [25, Theorem 3.3] and [26, Section 2.5], where it was also observed that the index can be learned at the parametric rate $n^{-1/2}$. Based on these results or on similar heuristics, a wide part of literature focused on index estimation, setting aside the regression problem. From this perspective, the main point is that the estimation of the index v can be carried out at parametric rate in spite of the unknown nonparametric nonlinearity f. A proof of Stone's conjecture (for convergence in $L^2(\rho)$) can be found in [22, Corollary 22.1].

Granted that the estimation of the index does not entail additional *statistical* costs (in terms of regression rates), a different but no less important problem is determining the *computational* cost to implement a statistically optimal estimator for the single-index model. The rate in [22, Corollary 22.1] is obtained by a least squares joint minimization over v and f, but no executable algorithm is provided. [20] proposed an adaptive algorithm aggregating local polynomial estimators on a lattice of the unit sphere, yielding a universal min-max estimator, although at the expense of a possibly exponential number of operations $\Omega(n^{(d-1)/2})$. While a heuristic faster algorithm is therein also proposed, its statistical effectiveness is unknown.

Several other methods for the estimation of v or f were developed over the years. A first category includes semiparametric methods based on maximum likelihood estimation [28, 24, 14, 15, 16, 8, 9]. M-estimators produce $\sqrt{n}$-consistent index estimates under general assumptions, but their implementation is cumbersome and computationally demanding, in that it depends on sensitive bandwidth selections for kernel smoothing and relies on high-dimensional joint optimization. An attempt at avoiding the data sparsity problem was made by [13], which proposed a fixed-point iterative scheme only involving 1-dimensional nonparametric smoothers. Alternatively, methods such as the average derivative estimation (ADE [44, 25, 27]), the outer product of gradients estimation (OPG [50]) and the minimum average variance estimation (MAVE [50, 49]) directly estimate the

index vector exploiting its proportionality with the derivative of the regression function. Early versions of these methods suffer from the curse of dimensionality due to kernel estimation of the gradient, while later iterative modifications provided $\sqrt{n}$-consistency under mild assumptions, yet not eliminating the computational overhead. More recently, Isotron [30, 29] and SILO [21] achieved linear complexity, but the proven regression rate, even if independent of d, is not min-max (albeit SILO focuses on the $n \ll d$ regime, rather than the limit $n \to \infty$ as here and most past work). In a different yet related direction, the even more recent [1] showed that convex neural networks can adapt to a large variety of statistical models, including single-index; however, they do not match the optimal learning rate (even for the single-index case), and at the same time do not have associated fast algorithms. All in all, the literature on single-index models seems to express a trade-off between statistical optimality and computational efficiency.

A parallel line of research has been devoted to *sufficient dimension reduction* [35] in the so-called *multi-index model* (or a slight extension thereof), where F depends on multiple $k < d$ index directions spanning an unknown index subspace, thus generalizing the single-index case, and the aim is to estimate this k-dimensional subspace. Along this thread we can find sliced inverse regression (SIR [18, 38]), sliced average variance estimation (SAVE [10]), simple contour regression (SCR [37]) and its generalizations (e.g. GCR [37], DR [36]). We call such methods *conditional methods*, because the estimates they provide are derived from statistics obtained from the conditional distribution of the explanatory variable X given the response variable Y. Conditional methods are appealing for several reasons. Compared to semiparametric methods, their implementation is straightforward, consisting of noniterative computation of empirical moments and having only one "scale" parameter to tune. Moreover, they are computationally efficient and simple to analyze, enjoying $\sqrt{n}$-consistency and, in most cases, complexity linear in the sample size and quadratic in the ambient dimension. On the downside, this comes in general at the cost of stronger distributional assumptions, and with no known theoretically optimal choice of the scale parameter [12, p. 75]. While conditional methods offer a provable, efficient solution for sufficient dimension reduction, they do not address the problem of estimating the link function on the estimated index space. The very recent preprint [41] introduces a variation of GCR coupled with estimation of the link function; however the analysis in [41] appears fatally flawed in the key step of bounding the regression error conditioned on an estimated (multi-)index subspace, which is a regression problem with nonzero mean "noise" with dependencies on the samples. As we discuss momentarily, one of the important technical contributions of our work is to tackle this problem (via Wasserstein metrics), leading to a rigorous statistical analysis of the joint estimation of many single- or multi-index and link function estimators. We summarize the key properties for these and other aforementioned techniques in Table 1 below.

In this work we introduce a new estimator and a corresponding algorithm, called Smallest Vector Regression (SVR), that are optimal both in the statistical and in the computational sense. We also provide a unifying theoretical framework for single-index

Table 1. Proven rate (up to log factors) and computational cost of several methods for index estimation and/or regression in single-index models, together with salient assumptions on the model.

	Performance				Assumptions		
	Proven rate		Computational cost		X	f	ζ
	$\widehat{v}$	$\widehat{f}$	$\widehat{v}$	$\widehat{f}$			
SIR [38]	$n^{-1/2}$	–	$d^2 n \log n$	–	linear $\mathbb{E}[X\|v^T X]$	N/A	N/A
SAVE [10]	$n^{-1/2}$	–	$d^2 n \log n$	–	linear $\mathbb{E}[X\|v^T X]$, const $\mathrm{Cov}[X\|v^T X]$	N/A	N/A
SCR [37]	$n^{-1/2}$	–	$d^2 n^2 \log n$	–	linear $\mathbb{E}[X\|v^T X]$, const $\mathrm{Cov}[X\|v^T X]$	stochastically monotone	decreasing density of $\zeta - \widetilde{\zeta}$
DR [36]	$n^{-1/2}$	–	$d^2 n \log n$	–	linear $\mathbb{E}[X\|v^T X]$, const $\mathrm{Cov}[X\|v^T X]$	N/A	N/A
ADE [27]	$n^{-1/2}$	–	$d^2 n^2 \log n$	–	$\mathcal{C}^0$ positive density	$\mathcal{C}^2$	Gaussian
rMAVE [49]	$n^{-1/2}$	N/A	$d^2 n^2$ per iteration		$v^T X$ has $\mathcal{C}^3$ density, $\mathbb{E}\|X\|^6 < \infty$	$\mathcal{C}^3$	$\mathbb{E}\|Y\|^3 < \infty$
Aggregation [20]	–	$n^{-\frac{s}{2s+1}}$	$(n \log n)^d$		compact supported lower bounded density	$\mathcal{C}^s$	$\sigma(X)\mathcal{N}(0,1)$
SIIsotron [29]	N/A	$n^{-1/6}$	$(\frac{n}{d \log n})^{1/3} dn \log n$		bounded	monotone, Lipschitz	bounded
SILO [21]	$n^{-1/4}$	$n^{-1/8}$	dn	$n \log n$	Gaussian	monotone, Lipschitz	bounded
SVR	$n^{-1/2}$	$n^{-\frac{s}{2s+1}}$	$d^2 n \log n$	$n \log n$	linear $\mathbb{E}[X\|v^T X]$, $\mathrm{Var}[w^T X\|v^T X] \gtrsim 1$	coarsely monotone, $\mathcal{C}^s$	sub-Gaussian

models, from which it is easy to derive theoretical guarantees for methods (or slight modifications thereof) other than ours. Our dimension reduction technique falls in the category of conditional methods. Unlike existing studies for similar approaches, we are able to provide a characterization for the parameter selection, and bound both the index estimation and the regression errors. Since regression is performed using standard piecewise polynomial estimates on the projected samples after and independently of the index estimation step, our regression bounds hold conditioned to any index estimation method of sufficient accuracy. Our analysis yields that convergence by proving finite-sample bounds in high probability. The resulting statements are stronger compared to the ones in the available literature on conditional methods, where typically only asymptotic convergence, at most, is established. As a side note, SVR has been empirically tested with success also in the multi-index model, but our analysis, and therefore our exposition, will be restricted to the single-index case. In summary, the contributions of this work are:

1. We prove strong, finite-sample convergence bounds, both in probability and in expectation, of several conditional regression methods, existing and new.
2. We introduce a new conditional regression method that combines accuracy, robustness and low computational cost. This method is multiscale and sheds light on parameter choices that are important in theory and practice, and are mostly left unaddressed in other techniques.
3. We prove that polynomial partitioning estimates are Hölder continuous with high

probability with respect to the index estimation error. This allows to bridge the gap between a good estimator of the index subspace and the performance of regression on the estimated subspace.

4. We prove that all $\sqrt{n}$-consistent index estimation methods, and in particular all the conditional methods considered, lead to the min-max 1-dimensional rate of convergence when combined with polynomial partitioning estimates.
5. Using the above, we fill the gap between statistical and computational efficiency in single-index model regression, providing theoretical guarantees of optimal convergence in quasilinear time.

The paper is organized as follows. In Section 2 we review several conditional regression methods for single-index model regression, and introduce our new estimator; in Section 3 we analyze the converge of various methods, including ours; in Section 4 we conduct several numerical experiments, both validating the theory and exploring numerically aspects of various techniques that are not covered by theoretical results; in Section 5 and 6 we collect additional proofs of theorems and technical results.

Notation

symbol	definition	symbol	definition
C, c	positive absolute constants	$\|A\|$	spectral norm of matrix A
$a \lesssim b$	$a \leq Cb$ for some C	$\lambda_i(A)$	i-th largest eigenvalue of matrix A
$a \asymp b$	$a \lesssim b$ and $b \lesssim a$	$\|I\|$	Lebesgue measure of interval I
$\langle u, v\rangle$	inner product of vectors u and v	$\#S$	cardinality of set S
$\|u\|$	Euclidean norm of vector u	$\mathbb{1}\{E\}$	indicator function of event E
$B(x, r)$	Euclidean ball of center x and radius r	$X \mid Y$	r.v. X conditioned on r.v. Y

2. Conditional regression methods

We consider the regression problem as in (1), within the single-index model, with the definition and notation as in (2). When f is at least Lipschitz, (2) implies $\nabla F(x) \in \operatorname{span}\{v\}$ for a.e. x; this is the reason why we may refer to v as the gradient direction. Given n independent copies (X_i, Y_i), $i = 1, \dots, n$, of the random pair (X, Y), we will construct estimators $\widehat{v}$ and $\widehat{f}$, and derive separate and compound non-asymptotic error bounds in probability and expectation. Our method is conditional in two ways: 1) the estimator $\widehat{v}$ is a statistic of the conditional distributions of the X_i's given the Y_i's (restricted in suitable intervals); 2) the estimator $\widehat{f}$ is conditioned on the estimate $\widehat{v}$. Several conditional methods for step 1) have been previously introduced, see e.g. [38, 10, 37]. Our error bounds for step 2) are independent of the particular method used in 1), only requiring a minimal non-asymptotic convergence rate. For these reasons, we will introduce our own method for 1) along with other conditional methods, and establish for each one the convergence rate needed to pair it with 2).

The common idea of all conditional methods is to compute statistics of the predictor X conditioned on the response Y. Conditioning on Y, one forces the distribution of X

to reveal the index structure through its moments, be they means (SIR) or variances (SAVE, SCR).

Assumption. All the algorithms we consider include a preprocessing step where data are standardized to have 0 mean and isotropic covariance. Thus, when illustrating each method, we will assume such standardization.

2.1. Sliced Inverse Regression

Sliced Inverse Regression [38] (SIR) estimates the index vector by a principal component analysis of the inverse regression curve $\mathbb{E}[X|Y]$. Samples on this curve are obtained by slicing the range of the function and computing sample means of the corresponding approximate level sets. In the version of SIR we consider here, we take dyadic partitions $\{C_{l,h}\}_{h=1}^{2^l}$, $l \in \mathbb{Z}$, of the range of Y, where each $C_{l,h}$ is an interval of length $\asymp 2^{-l}$. After calculating the sample mean for each slice,

$$\widehat{\mu}_{l,h} = \frac{1}{\#C_{l,h}} \sum_i X_i \mathbb{1}\{Y_i \in C_{l,h}\}\,, \qquad h = 1, \dots, 2^l\,,$$

SIR outputs $\widehat{v}_l$ as the eigenvector of largest eigenvalue of the weighted covariance matrix

$$\widehat{M}_l = \sum_h \widehat{\mu}_{l,h}\widehat{\mu}_{l,h}^T \frac{\#C_{l,h}}{n}\,.$$

Note that the population limits of $\widehat{\mu}_{l,h}$ and $\widehat{M}_l$ are, respectively,

$$\mu_{l,h} = \mathbb{E}[X \mid Y \in C_{l,h}]\,, \qquad M_l = \sum_h \mu_{l,h}\mu_{l,h}^T \mathbb{P}\{Y \in C_{l,h}\}\,.$$

2.2. Sliced Average Variance Estimation

Sliced Average Variance Estimation [10] (SAVE) generalizes SIR to second order moments. After slicing the range of Y and computing the centers $\widehat{\mu}_{l,h}$'s, it goes further and construct the sample covariance on each slice:

$$\widehat{\Sigma}_{l,h} = \frac{1}{\#C_{l,h}} \sum_i (X_i - \widehat{\mu}_{l,h})(X_i - \widehat{\mu}_{l,h})^T \mathbb{1}\{Y_i \in C_{l,h}\}\,.$$

Then, it averages the $\widehat{\Sigma}_{l,h}$'s and defines $\widehat{v}_l$ as the eigenvector of largest eigenvalue of

$$\widehat{S}_l = \sum_h (I - \widehat{\Sigma}_{l,h})^2 \frac{\#C_{l,h}}{n}\,.$$

The matrices $\widehat{\Sigma}_{l,h}$ and $\widehat{S}_l$ are empirical estimates of

$$\Sigma_{l,h} = \mathrm{Cov}[X \mid Y \in C_{l,h}]\,, \qquad S_l = \sum_h (I - \Sigma_{l,h})^2 \mathbb{P}\{Y \in C_{l,h}\}\,.$$

2.3. Contour Regression

Simple Contour Regression [37] (SCR) seeks the directions of most functional variation estimating the smallest eigenvectors of

$$K_\delta = \mathbb{E}[(X - \widetilde{X})(X - \widetilde{X})^T \mid |Y - \widetilde{Y}| \le \delta]),$$

where $(\widetilde{X}, \widetilde{Y})$ is an independent copy of (X, Y). We shall use a dyadic scale $\delta = 2^{-l}$ and, with abuse of notation, write $K_l := K_\delta$ for such choice of δ. SCR uses the realizations $(X_i - X_{\tilde{\imath}})$ with $|Y_i - Y_{\tilde{\imath}}| \le \delta$ to generate approximations to K_δ:

$$\widehat{H}_\delta = \frac{\sum_{i<\tilde{\imath}} (X_i - X_{\tilde{\imath}})(X_i - X_{\tilde{\imath}})^T \mathbb{1}\{|Y_i - Y_{\tilde{\imath}}| \le \delta\}}{\sum_{i<\tilde{\imath}} \mathbb{1}\{|Y_i - Y_{\tilde{\imath}}| \le \delta\}}.$$

We let $\widehat{v}_l$ be the smallest eigenvector of $\widehat{H}_l$, where again $\widehat{H}_l := \widehat{H}_\delta$ with $\delta = 2^{-l}$.

2.4. Smallest Vector Regression (SVR)

This is the new method we propose here. We perform a local principal component analysis on each approximate level set obtained by multiscale slices of Y: because of the special structure (2), each (approximate) level set should be narrow in the v-direction and spread out along the orthogonal directions, therefore the smallest principal component should approximate v. Once we have an estimate for v, we can project down the d-dimensional samples and perform nonparametric regression of the 1-dimensional function f. The method consists of the following steps:

1.a) Construct a multiscale family of dyadic partitions of $[\min_i Y_i, \max_i Y_i]$

$$\{C_{l,h}\}_{h=1}^{2^l}, \qquad l \in \mathbb{Z},$$

with $|C_{l,h}| = 2^{-l} |\max_i Y_i - \min_i Y_i|$.

1.b) Let $\mathcal{H}_l$ be the set of h's such that $\#C_{l,h} \ge 2^{-l} n$. For $h \in \mathcal{H}_l$, let $\widehat{v}_{l,h}$ be the eigenvector corresponding to the *smallest* eigenvalue of $\widehat{\Sigma}_{l,h}$.

1.c) Compute the eigenvector $\widehat{v}_l$ corresponding to the *largest* eigenvalue of

$$\widehat{V}_l = \frac{1}{\sum_{h \in \mathcal{H}_l} \#C_{l,h}} \sum_{h \in \mathcal{H}_l} \widehat{v}_{l,h} \widehat{v}_{l,h}^T \#C_{l,h}.$$

2) Regress f using a dyadic polynomial estimator $\widehat{f}_{|\widehat{v}_l}$, on the samples $(\langle \widehat{v}_l, X_i \rangle, Y_i)$, $i = 1, \dots, n$ (more details in Section 2.5). Return $\widehat{F}_{j|\widehat{v}_l}(x) := \widehat{f}_{|\widehat{v}_l}(\langle \widehat{v}_l, x \rangle)$.

While SVR shares step 1.a) with SIR, it differs from SIR in step 1.b), where it takes conditional (co)variance statistics in place of conditional means, and in step 1.c), where it averages smallest-variance directions rather than means. We may regard SAVE and SVR

as two different modifications of SIR to higher order statistics, which allows in general for better and more robust estimates (see Section 4.1). The fundamental difference between SVR and SAVE is that SVR computes local estimates of the index vector which then aggregates in a global estimate, while SAVE first aggregates local information and then computes a single global estimate. Similarly to SCR, SVR is a second order method and it searches for directions of minimal variance; however, SCR conditions on a level set sliding continuously on the range of the function, while SVR considers fixed conditional distributions on dyadic partitions, as SIR and SAVE. Lastly, the computational cost of SVR is quasilinear as for SIR and SAVE, compared to the quadratic time required by SCR.

2.5. Conditional partitioning estimators

In step 2) we use piecewise polynomial estimators in the spirit of [5, 4]: these techniques are based on partitioning the domain (here, in a multiscale fashion), and constructing a local polynomial on each element of the partition by solving a least squares fitting problem. A global estimator is then obtained by summing the local polynomials over a certain partition (possibly using a partition of unity to obtain smoothness across the boundaries of the partition elements). The degree of the local polynomials needed to obtain optimal rates depends on the regularity of the function, and may be chosen adaptively if such regularity is unknown. A proper partition (or scale) is then chosen to minimize the expected mean squared error (MSE), by classical bias-variance trade-off.

In detail, given an estimated direction $\widehat{v}$, our step 2) consists of:

2.a) construct a multiscale family of dyadic partitions of $[\min_i \langle \widehat{v}, X_i\rangle, \max_i \langle \widehat{v}, X_i\rangle]$: for each $j \in \mathbb{N}$, $\{I_{j,k|\widehat{v}}\}_{k\in\mathcal{K}_j}$ is a partition, with $|I_{j,k|\widehat{v}}| = 2^{-j}|\max_i \langle \widehat{v}, X_i\rangle - \min_i \langle \widehat{v}, X_i\rangle|$.

2.b) For each $I_{j,k|\widehat{v}}$, compute the best fitting polynomial

$$\widehat{f}_{j,k|\widehat{v}} = \operatorname*{arg\,min}_{\deg(p)\le m} \sum_i |Y_i - p(\langle \widehat{v}, X_i\rangle)|^2 \mathbb{1}\{\langle \widehat{v}, X_i\rangle \in I_{j,k|\widehat{v}}\}\,.$$

2.c) Sum all local estimates $\widehat{f}_{j,k|\widehat{v}}$ over the partition $\{I_{j,k|\widehat{v}}\}_{k\in\mathcal{K}_j}$:

$$\widehat{f}_{j|\widehat{v}}(t) = \sum_{k\in\mathcal{K}_j} \widehat{f}_{j,k|\widehat{v}}(t)\mathbb{1}\{t \in I_{j,k|\widehat{v}}\}\,.$$

The final estimator of F at scale j and conditioned on $\widehat{v}$ is given by

$$\widehat{F}_{j|\widehat{v}}(x) = \widehat{f}_{j|\widehat{v}}(\langle \widehat{v}, x\rangle)\,. \tag{3}$$

In SVR, step 2) is carried out on $\widehat{v} = \widehat{v}_l$, yielding for each l a multiscale family of partitions $\{I_{j,k|l}\}_{j,k}$, local polynomials $\{\widehat{f}_{j,k|l}|\}_{j,k}$ and global estimators $\{\widehat{f}_{j|l}\}_j$. However, we will prove results on the performance of 2) also when $\widehat{v}$ is the output of (our versions

of) SIR, SAVE and SCR (Corollary 1), and more in general by any estimator of v with $n^{-1/2}$ probabilistic convergence rate (Theorem 6).

Note that for SVR, but also for SIR, SAVE and SCR, the final estimator $\widehat{f}_{j|l}$ depends on two scale parameters: l controls the scale in the range, and j controls the scale of the 1-dimensional regression after projection onto $\widehat{v}_l$, and these two scales may be chosen independently. Our analysis yields optimal choices for these two scale parameters; the scale 2^{-l} at which the direction v is estimated will not be finer than the noise level, while a possibly finer partition with $j > l$ may be selected to improve the polynomial fit, allowing the estimator $\widehat{f}_{j|l}$ to de-noise its predictions, provided that enough training samples are available (see Figure 1).

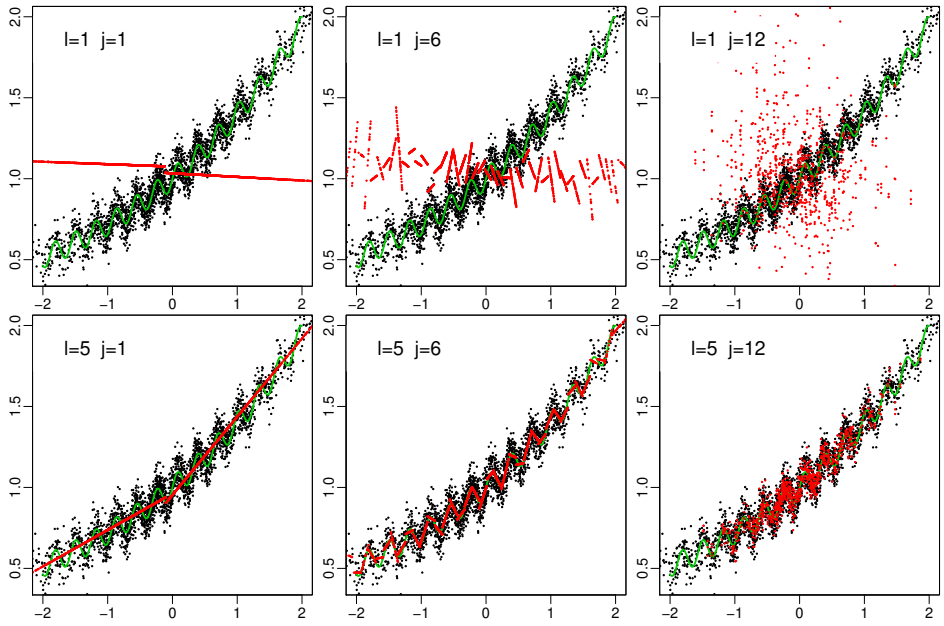

Figure 1: Local linear estimator (red) at different scales l, to regress the function f (green) from noisy samples (black). The horizontal axis is $\langle v, x\rangle$, while of course the estimator $\hat{f}_{j|l}$ is a function of $\langle \widehat{v}, x\rangle$ and may appear multi-valued in $\langle v, x\rangle$. For small l (top row) the error in the estimation of the index v is large, leading to poor regression estimates regardless of the regression scale j. For larger l (bottom row) a good accuracy for the index vector v is achieved, and the estimator is able to approximate the function even below the noise level and the non-monotonicity scale (e.g. for $j = 6$); overfitting occurs for j too large (e.g. $j = 12$ in this case).

We report below the complete sequence of steps run by SVR. The time complexity of the algorithm is shown in Table 2. Note that 2.c) has only an evaluation cost, i.e. $\widehat{f}_{j|\widehat{v}}$

does not need to be constructed, but only evaluated.

Algorithm: SVR

Input : samples $\{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$,
polynomial degree $m \in \mathbb{N}$.
Output: $\widehat{v}_l$ estimate of v, $\widehat{f}_{j|l}$ estimate of f.

standardize data to 0 mean and I_d covariance;

1.a) construct $\{C_{l,h}\}_{l,h}$, dyadic decomposition of $[\min_i Y_i, \max Y_i]$;

1.b) compute $\widehat{v}_{l,h}$, the eigenvector of $\frac{1}{\#C_{l,h}} \sum_i X_i X_i^T \mathbb{1}\{Y_i \in C_{l,h}\}$ corresponding to the smallest eigenvalue, for all $h \in \mathcal{H}_l = \{h : \#C_{l,h} \geq 2^{-l} n\}$;

1.c) compute $\widehat{v}_l$, the eigenvector of $\frac{1}{\sum_{h \in \mathcal{H}_l} \#C_{l,h}} \sum_{h \in \mathcal{H}_l} \widehat{v}_{l,h} \widehat{v}_{l,h}^T \#C_{l,h}$ corresponding to the largest eigenvalue;

2.a) construct $\{I_{j,k|l}\}_{j,k}$, dyadic decomposition of $[\min_i \langle \widehat{v}_l, X_i \rangle, \max_i \langle \widehat{v}_l, X_i \rangle]$;

2.b) compute $\widehat{f}_{j,k|l} = \arg\min_{\deg(p) \leq m} \sum_i |Y_i - p(\langle \widehat{v}_l, X_i \rangle)|^2 \mathbb{1}\{\langle \widehat{v}_l, X_i \rangle \in I_{j,k|l}\}$;

2.c) define $\widehat{f}_{j|l}(t) = \sum_k \widehat{f}_{j,k|l}(t) \mathbb{1}\{t \in I_{j,k|l}\}$.

Table 2. Computational cost breakdown for SVR.

	task	**computational cost**
	standardization	$O(d^2 n)$
1.a)	dyadic decomposition of the range	$O(n \log n)$
1.b)	PCA on level sets	$O(d^2 n \log n)$
1.c)	PCA of local directions	$O(d^2 n \log n)$
2.a)	dyadic decomposition of the domain	$O(n \log n)$
2.b)	m-order polynomial regression	$O(m^2 n \log n)$
	total	$O((d^2 + m^2) n \log n)$

3. Analysis of convergence

To carry out our analysis we shall make several assumptions on the distributions of X, Y and ζ:

(X) X has strictly sub-Gaussian distribution with $\mathrm{Cov}[X] = R^2 I_d$.

(Y) Y has strictly sub-Gaussian distribution with $\mathrm{Var}[Y] = R^2$.

(Z) ζ is strictly sub-Gaussian with $\mathrm{Var}[\zeta] \leq \sigma^2$.

(X), (Y) and (Z) are standard assumptions in regression analysis *tout court*. Note that we start from standardized data, that is, we will not be tracking the (negligible) error resulting from standardization based on data samples. The following is instead typical of single-index models:

(LCM) $\mathbb{E}[X \mid \langle v, X\rangle] = v\langle v, X\rangle$, or the stronger assumption
(LCM') X has symmetric distribution around v.

(LCM) is commonly referred to as the *linear conditional mean* assumption [38, Condition 3.1], because (for centralized X) it is equivalent to requiring $\mathbb{E}[X \mid \langle v, X\rangle]$ to be linear in $\langle v, X\rangle$ [35, Lemma 1.1]. Every spherical distribution, hence every elliptical distribution after standardization, satisfies (LCM) for every v [7, Corollary 5], and conversely [19]. While it does introduce some symmetry, it is less restrictive than it may seem. It has been shown to hold approximately in high dimension, where most low-dimensional projections are nearly normal [17, 23]. (LCM) is introduced to ensure that $\widehat{v}$ is an unbiased estimate of v [11, Theorem 1]. The restriction (LCM') is purely technical, and we impose it only in order to apply standard Bernstein inequalities for bounded variables [46, Lemma 2.2.9]. Since X and Y are in general unbounded, we will condition the statistics of interest in suitable balls of constant radius (see Section 3.1). Such conditioning would in general break (LCM), but not (LCM'). On the other hand, at the expense of a slightly more complicated analysis, one could directly use Bernstein inequalities for sub-exponential variables [46, Lemma 2.2.11], thus avoiding the conditioning and hence (LCM'). All boiling down to a technical distinction, we will not stress (LCM') versus (LCM) any further.

In addition to (LCM), second order methods usually require the so-called *constant conditional variance* assumption [10, p. 2117]:

(CCV) $\mathrm{Cov}[X \mid \langle v, X\rangle]$ is nonrandom.

Assuming (X) and (LCM), (CCV) is equivalent to $\mathrm{Cov}[X \mid \langle v, X\rangle] = R^2(I_d - vv^T)$ almost surely [35, Corollary 5.1]. (CCV) is true for the normal distribution [35, Proposition 5.1], and again approximately true in high dimension [17, 23]. Some care is required when assuming both (LCM) and (CCV): imposing (LCM) for every v is equivalent to assuming spherical symmetry [19], and the only spherical distribution satisfying (CCV) is the normal distribution [31, Theorem 7]. Two possible relaxations of (CCV) are the following upper and lower bounded conditional variance conditions:

(UCV) There is $\alpha \geq 1$ such that $\mathrm{Var}[\langle w, X\rangle \mid \langle v, X\rangle] \leq \alpha R^2$ almost surely for all $w \in \mathrm{span}\{v\}^\perp \cap \mathbb{S}^{d-1}$.

(LCV) There is $\alpha \geq 1$ such that $\mathrm{Var}[\langle w, X\rangle \mid \langle v, X\rangle] \geq R^2/\alpha$ almost surely for all $w \in \mathrm{span}\{v\}^\perp \cap \mathbb{S}^{d-1}$.

We present bounds separately on estimators for v in the next subsection, and on regression of f in subsection 3.2: this will be useful to understand properties of SIR, SAVE and SCR, besides SVR. Our main result, Theorem 6, will give near-optimal bounds

on the SVR estimator for F, in both probability and expectation, in the form

$$\mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - F(X)|^2] \le K' d^{2(s\vee 1)} \log d \log^{s\vee 1} n \left(\frac{\log n}{n}\right)^{\frac{2s}{2s+1}}$$

for $\widehat{F}_{j|\widehat{v}}$ as in (3), j selected according to Theorem 6, $f \in \mathcal{C}^s$, $s \in [\frac{1}{2}, \frac{3}{2}]$, K' a constant independent of n and d, and once assumptions (X), (Y), (Z), (Ω), (LCM'), (LCV) and (Θ) (the latter is discussed below) are satisfied.

3.1. Bounds on estimators of the index vector

For SIR, SAVE and SCR, all population statistics will be taken on the distribution of X conditioned on $X \in B(0, \sqrt{2d\log(4)}R)$. In view of assumption (X) and Lemma 3, neglecting events of probability lower than $2e^{-cn}$ we may assume $\|X_i\| \le \sqrt{2d\log(4)}R$ almost surely for at least $n/4$ many i's, thus losing only a constant fraction of the samples. Assumption (LCM') is not affected by such conditioning. Uniform boundedness of the samples is required for the application of the Bernstein inequality.

Theorem 1 (SIR). *Assume (X) and (LCM'), and that there are $l \ge 1$ and $\alpha \ge 1$ such that*

(I) $\sum_h |\langle v, \mathbb{E}[X \mid Y \in C_{l,h}]|^2 \mathbb{P}\{Y \in C_{l,h}\} \ge R^2/\alpha$.

Let $\widehat{v}_l$ be the direction estimated by SIR, as defined in Section 2.1. Then:

(a) $\|\widehat{v}_l - v\| \lesssim \alpha\sqrt{t + l + \log d}\; 2^{l/2} \frac{d}{\sqrt{n}}$ *with probability higher than* $1 - e^{-t}$;

(b) $\mathbb{E}[\|\widehat{v}_l - v\|^2] \lesssim \alpha^2 (l + \log d) 2^l \frac{d^2}{n}$.

Condition (I) says that the variance of the means of the level sets of F is comparable with the variance of X. Thus, it is satisfied whenever f is monotone, or at least "monotone at scale coarser than 2^{-l}" – this will reappear formally below, in particular in condition (Ω). Note that (I) is the quantized version of $\mathbb{E}[|\langle v, \mathbb{E}[X \mid Y]\rangle|^2] > 0$, which is equivalent to saying that $E[\langle v, X\rangle \mid Y]$ is nondegenerate, that is, non almost surely equal to a constant.

Theorem 2 (SAVE). *Assume (X), (LCM') and (UCV) with $R = \alpha = 1$, and that there are $l \ge 1$ and $\alpha \ge 1$ such that, for every $w \in \operatorname{span}\{v\}^\perp \cap \mathbb{S}^{d-1}$,*

(V) $\operatorname{Var}[\langle w, X\rangle \mid Y \in C_{l,h}] - \operatorname{Var}[\langle v, X\rangle \mid Y \in C_{l,h}] \ge R^2/\alpha, \qquad h = 1, \dots, 2^l$.

Let $\widehat{v}_l$ be the direction estimated by SAVE as described in Section 2.2. Then:

(a) $\|\widehat{v}_l - v\| \lesssim \alpha^2\sqrt{t + l + \log d}\; 2^{l/2} \frac{d^2}{\sqrt{n}}$ *with probability higher than* $1 - e^{-t}$;

(b) $\mathbb{E}[\|\widehat{v}_l - v\|^2] \lesssim \alpha^4 (l + \log d) 2^l \frac{d^4}{n}$.

Theorem 3 (SCR). *Assume (X) and (LCM'), and that there are $l \geq 1$ and $\alpha \geq 1$, such that, for every $w \in \mathrm{span}\{v\}^{\perp} \cap \mathbb{S}^{d-1}$,*

(C) $\mathrm{Var}[\langle w, X - \widetilde{X}\rangle \mid |Y - \widetilde{Y}| \leq 2^{-l}] - \mathrm{Var}[\langle v, X - \widetilde{X}\rangle \mid |Y - \widetilde{Y}| \leq 2^{-l}] \geq R^2/\alpha$.

Let $\widehat{v}_l$ be the direction estimated by SCR as described in Section 2.3. Then:

(a) $\|\widehat{v}_l - v\| \lesssim \frac{\alpha\sqrt{t+\log d}}{\mathbb{P}\{|\widetilde{Y}-Y|\leq 2^{-l}\}^{1/2}}\sqrt{\frac{d}{n}}$ *with probability higher than* $1 - e^{-t}$*;*

(b) $\mathbb{E}[\|\widehat{v}_l - v\|^2] \lesssim \frac{\alpha^2 \log d}{\mathbb{P}\{|\widetilde{Y}-Y|\leq 2^{-l}\}}\frac{d}{n}$.

The parameter l should be chosen sufficiently large to satisfy (C), but still small enough so that $\mathbb{P}\{|\widetilde{Y} - Y| \leq 2^{-l}\}$ is high. For conditions sufficient to imply (C), see [37, Theorem 2.3, Theorem 3.1]. The proofs of Theorems 1, 2 and 3 are postponed to Section 5.

We now turn to the analysis of SVR. It is possible to prove that SVR achieves

(a) $\|\widehat{v}_l - v\| \lesssim \alpha\sqrt{t + l + \log d}\; 2^{l/2}\sqrt{\frac{d}{n}}$ with probability higher than $1 - e^{-t}$,

(b) $\mathbb{E}[\|\widehat{v}_l - v\|^2] \lesssim \alpha^2(l + \log d)2^l \frac{d}{n}$,

under assumptions (X), (LCM') and (V). We prefer however to work with more interpretable conditions, that better decouple the geometry of the distribution of X and properties of the function f. For this purpose, we introduce:

(Ω) There are $\omega \geq 0$ and $\ell > 0$ such that, for every interval C with $|C| \geq \omega$, $f^{-1}(C)$ is an interval and $|f^{-1}(C)| \leq |C|/\ell$.

Assumption (Ω) may be regarded as a large scale sub-Lipschitz property for the set-valued function f^{-1}. Note that, if f is bi-Lipschitz, then (Ω) is satisfied with $\omega = 0$. However, (Ω) for $\omega > 0$ does not imply that f is monotone; it relaxes monotonicity to monotonicity "at scales larger than ω".

In the following, we will condition the statistics $\Sigma_{l,h}$ on $\|X\| \lesssim \sqrt{d}R$ and $|Y| \lesssim R$. In doing so, we only discard a constant fraction of X_i's and Y_i's with confidence $1 - 4e^{-cn}$, thanks to assumptions (X) and (Y) and Lemma 3, while not invalidating assumption (LCM').

We may now state the main result for the SVR estimator of v:

Theorem 4 (SVR). *Suppose (X), (Y), (Z), (Ω), (LCM') and (LCV) hold true. Let l be such that $|C_{l,h}| \gtrsim \max\{\sigma, \omega\}$, $h = 1, \ldots, 2^l$. Then, for n large enough so that $\frac{n}{\sqrt{\log n}} \gtrsim (t + l + \log d)2^{2l}$ we have*

(a) $\|\widehat{v}_l - v\| \lesssim \alpha\ell^{-1}\sqrt{t + l + \log d}\; 2^{-l/2}\sqrt{\frac{d}{n/\sqrt{\log n}}}$ *with probability higher than* $1 - e^{-t}$.

Moreover, if $\frac{n}{\log n\sqrt{\log n}} \gtrsim \alpha^2\ell^{-2}d2^l$, then

(b) $\mathbb{E}[\|\widehat{v}_l - v\|^2] \lesssim \alpha^2\ell^{-2}(l + \log d)2^{-l}\frac{d}{n/\sqrt{\log n}}$.

If, furthermore, $|\zeta| \le \sigma$ *a.s., then* (a) *and* (b) *hold with* $n/\sqrt{\log n}$ *replaced by* n.

Theorem 4 not only proves convergence for SVR, but also shows that finer scales give more accurate estimates, provided the number of local samples $\#C_{l,h}$ is not too small and we stay above the critical scales σ and ω, representing the noise and the non-monotonicity levels, respectively. Without assumption (LCM), both SIR and SVR provide biased estimates of the index vector; it is not known if such bias is removable. Nevertheless, Theorem 4 suggests that the estimation error of SVR could be driven to 0 by increasing l, only limited by the constraint of keeping the scale l larger than $\max\{\sigma, \omega\}$. On the other hand, for distributions not satisfying the assumptions above, the inverse regression curve can deviate considerably from the direction v, regardless of the size of the noise (see Figure 3). In SVR, assuming for a moment monotonicity ($\omega = 0$) and zero noise ($\sigma = 0$), choosing the scale parameter l according to the lower bound on n yields a $O(n^{-2})$ convergence rate for the MSE, disregarding log factors.

To prove Theorem 4, we first establish bounds on the local statistics involved in the computation of the estimator of v:

Proposition 1. *Suppose* (Z) *and* (Ω) *hold true. Let* C *be a bounded interval with* $|C| \ge \omega$. *Then:*

(a) *For every* X_i *such that* $Y_i \in C$, *and for every* $\tau \ge 1$,

$$\mathbb{P}\{|\langle v, X_i\rangle - \mathbb{E}[\langle v, X\rangle \mid Y \in C]| \gtrsim \ell^{-1}(|C| + \sqrt{\tau \log n}\ \sigma)\} \le 2n^{-\tau}\,.$$

If $|\zeta| \le \sigma$ *a.s., then*

$$\mathbb{P}\{|\langle v, X_i\rangle - \mathbb{E}[\langle v, X\rangle \mid Y \in C]| \lesssim \ell^{-1}(|C| + \sigma)\} = 1\,.$$

(b) $\mathrm{Var}[\langle v, X\rangle \mid Y \in C] \lesssim \ell^{-2}(|C|^2 + \sigma^2)$.

Proof. Let $Z_t = (-\sqrt{2(t+1)}\sigma, -\sqrt{2t}\sigma] \cup [\sqrt{2t}\sigma, \sqrt{2(t+1)}\sigma)$ for $t \in \mathbb{N}$. To prove (a) we first note that, thanks to (Z), we have $\zeta_i \in \bigcup_{t \le \tau \log n} Z_t$ for every i with probability higher than $1 - 2n^{-\tau}$. Conditioned on this event, $\langle v, X_i\rangle \in f^{-1}(C + \bigcup_{t \le \tau \log n} Z_t)$ if $Y_i \in C$. On the other hand, $\mathbb{E}[\langle v, X\rangle \mid Y \in C, \zeta \in Z_t] \in f^{-1}(C + Z_t)$. It follows from assumption (Ω) that

$$|\langle v, X_i\rangle - \mathbb{E}[\langle v, X\rangle \mid Y \in C, \zeta \in Z_t]| \lesssim \ell^{-1}(|C| + \sqrt{\max\{t, \tau \log n\}}\sigma).$$

Thus, by the law of total expectation,

$$\begin{aligned}
|\langle v, X_i\rangle - \mathbb{E}[\langle v, X\rangle \,|\, Y \in C]| &\le \sum_{t=0}^{\infty} |\langle v, X_i\rangle - \mathbb{E}[\langle v, X\rangle \,|\, Y \in C, \zeta \in Z_t]| \mathbb{P}\{\zeta \in Z_t\} \\
&\lesssim \ell^{-1}\left(|C| + \sqrt{\tau \log n}\sigma + \sigma \sum_{t > \tau} \sqrt{t} e^{-t}\right) \\
&\lesssim \ell^{-1}(|C| + \sqrt{\tau \log n}\ \sigma).
\end{aligned}$$

The case where $|\zeta| \le \sigma$ almost surely is similar and simpler. For (b), we write

$$\begin{aligned}\mathrm{Var}[\langle v, X\rangle \mid Y \in C] &= \mathbb{E}[(\langle v, X\rangle - \mathbb{E}[\langle v, X\rangle \,|\, Y \in C])^2 \,|\, Y \in C] \\ &= \sum_{t=0}^{\infty} \mathbb{E}[(\langle v, X\rangle - \mathbb{E}[\langle v, X\rangle \,|\, Y \in C])^2 \,|\, Y \in C, \zeta \in Z_t]\mathbb{P}\{\zeta \in Z_t\}.\end{aligned}$$

Conditioned on $\zeta \in Z_t$, assumption (Ω) gives

$$\begin{aligned}|\langle v, X\rangle - \mathbb{E}[\langle v, X\rangle \mid Y \in C]| &\le \sum_{s=0}^{\infty} |\langle v, X\rangle - \mathbb{E}[\langle v, X\rangle \mid Y \in C, \zeta \in Z_s]|\mathbb{P}\{\zeta \in Z_s\} \\ &\lesssim \ell^{-1}\left(|C| + \sqrt{t}\sigma + \sigma \sum_{s=0}^{\infty} \sqrt{s}e^{-s}\right) \\ &\lesssim \ell^{-1}(|C| + \sqrt{t}\sigma),\end{aligned}$$

whence

$$\mathrm{Var}[\langle v, X\rangle \mid Y \in C] \lesssim \ell^{-2}\left(|C|^2 + \sigma^2 \sum_{t=0}^{\infty} te^{-t}\right) \lesssim \ell^{-2}(|C|^2 + \sigma^2). \qquad \square$$

Proposition 2. *Suppose (X), (Y), (Z), (Ω), (LCM) and (LCV) hold true. Then, for every l such that $2^{-l} \lesssim \ell/\sqrt{\alpha}$ and $|C_{l,h}| \ge \max\{\sigma, \omega\}$, $h = 1, \dots, 2^l$, v is the eigenvector of smallest eigenvalue of $\Sigma_{l,h}$, and (V) holds true, that is,*

$$\lambda_{d-1}(\Sigma_{l,h}) - \lambda_d(\Sigma_{l,h}) \gtrsim R^2/\alpha\,,$$

with probability higher than $1 - 2e^{-cn}$.

Proof. First of all we condition on $C_{l,h}$, which is otherwise random since so are $\min_i Y_i$ and $\max_i Y_i$. We first lower bound $\lambda_{d-1}(\Sigma_{l,h})$. We have

$$\mathrm{Cov}[X \mid Y \in C_{l,h}] = \mathbb{E}[XX^T \mid Y \in C_{l,h}] - \mathbb{E}[X \mid Y \in C_{l,h}]\mathbb{E}[X^T \mid Y \in C_{l,h}].$$

Since X is independent of ζ, (2) implies that X is independent of Y given $\langle v, X\rangle$, hence

$$\begin{aligned}\mathbb{E}[XX^T \mid Y \in C_{l,h}] &= \mathbb{E}[\mathbb{E}[XX^T \mid \langle v, X\rangle, Y \in C_{l,h}] \mid Y \in C_{l,h}] \\ &= \mathbb{E}[\mathbb{E}[XX^T \mid \langle v, X\rangle] \mid Y \in C_{l,h}].\end{aligned}$$

For the same reason, and using assumption (LCM), we have

$$\begin{aligned}\mathbb{E}[X \mid Y \in C_{l,h}] &= \mathbb{E}[\mathbb{E}[X \mid \langle v, X\rangle, Y \in C_{l,h}] \mid Y \in C_{l,h}] \\ &= \mathbb{E}[\mathbb{E}[X \mid \langle v, X\rangle] \mid Y \in C_{l,h}] \\ &= \mathbb{E}[v\langle v, X\rangle \mid Y \in C_{l,h}].\end{aligned}$$

Now, let w be a unitary vector orthogonal to v. Then

$$w^T\mathbb{E}[X \mid Y \in C_{l,h}] = \mathbb{E}[\langle w, v\rangle\langle v, X\rangle \mid Y \in C_{l,h}] = 0,$$

while assumption (LCV) gives

$$w^T\mathbb{E}[XX^T \mid Y \in C_{l,h}]w = \mathbb{E}[\mathrm{Var}[\langle w, X\rangle \mid \langle v, X\rangle] \mid Y \in C_{l,h}] \geq R^2/\alpha.$$

Moreover, (LCM) implies by [11, Theorem 1.a] that v is an eigenvector of $\Sigma_{l,h}$. Therefore,

$$\lambda_{d-1}(\Sigma_{l,h}) = \min_{\substack{w\in\mathrm{span}\{v\}^\perp \\ \|w\|=1}} w^T\mathrm{Cov}[X \mid Y \in C_{l,h}]w \geq R^2/\alpha.$$

To upper bound $\lambda_d(\Sigma_{l,h})$ note that, conditioning on $|Y_i| \lesssim R$, we have $|C_{l,h}| \lesssim R2^{-l}$. Thus, assumption ($\Omega$) implies by Proposition 1(b) that

$$\lambda_d(\Sigma_{l,h}) \lesssim \ell^{-2}R^2 2^{-2l}.$$

We finally put together lower and upper bound. Taking $2^{-l} \lesssim \ell/\sqrt{\alpha}$ yields the desired inequality. □

We now establish convergence in probability for the local estimators $\widehat{v}_{l,h}$.

Proposition 3 (local SVR)**.** *Suppose (X), (Y), (Z), (Ω), (LCM') and (LCV) hold true. Then, conditioned on $\#C_{l,h}$, for every l such that $|C_{l,h}| \gtrsim \max\{\sigma, \omega\}$, $h = 1, \dots, 2^l$, for every $\varepsilon > 0$ and $\tau \geq 1$,*

$$\mathbb{P}\{\|\widehat{v}_{l,h} - v\| > \varepsilon\} \lesssim d\left[\exp\left(-\frac{c\#C_{l,h}\varepsilon^2}{\alpha^2\ell^{-2}d\sqrt{\tau\log n}(2^{-2l}+2^{-l}\varepsilon)}\right) + \exp\left(-\frac{c\#C_{l,h}}{\alpha^2 d}\right)\right] + n^{-\tau}.$$

If $|\zeta| \leq \sigma$ a.s., then

$$\mathbb{P}\{\|\widehat{v}_{l,h} - v\| > \varepsilon\} \lesssim d\left[\exp\left(-\frac{c\#C_{l,h}\varepsilon^2}{\alpha^2\ell^{-2}d(2^{-2l}+2^{-l}\varepsilon)}\right) + \exp\left(-\frac{c\#C_{l,h}}{\alpha^2 d}\right)\right].$$

Proof. Since $\|\widehat{v}_{l,h} - v\| \leq 1$, we can assume $\varepsilon^2 \leq \varepsilon \leq 1$ whenever needed. The Davis–Kahan Theorem [2, Theorem VII.3.1] together with Proposition 2 gives

$$\|\widehat{v}_{l,h} - v\| \leq \frac{\|v^T(\widehat{\Sigma} - \Sigma)\|}{|\lambda_{d-1}(\widehat{\Sigma}) - \lambda_d(\Sigma)|}$$

with $\Sigma = \Sigma_{l,h}$ and $\widehat{\Sigma} = \widehat{\Sigma}_{l,h}$. By Proposition 2 and the Weyl inequality we get

$$|\lambda_{d-1}(\widehat{\Sigma}) - \lambda_d(\Sigma)| \geq \lambda_d(\Sigma) - \lambda_{d-1}(\Sigma) - |\lambda_{d-1}(\widehat{\Sigma}) - \lambda_{d-1}(\Sigma)| \gtrsim R^2/\alpha - \|\widehat{\Sigma} - \Sigma\|.$$

We bound $\|\widehat{\Sigma} - \Sigma\|$ using the Bernstein inequality. First, we introduce the intermediate term

$$\widetilde{\Sigma} = \frac{1}{\#C}\sum_i (X_i - \mu)(X_i - \mu)^T \mathbb{1}\{Y_i \in C\},$$

and split $\widehat{\Sigma} - \Sigma$ into

$$\widehat{\Sigma} - \Sigma = \widetilde{\Sigma} - \Sigma - (\widehat{\mu} - \mu)(\widehat{\mu} - \mu)^T,$$

where $C = C_{l,h}$, $\mu = \mu_{l,h}$ and $\widehat{\mu} = \widehat{\mu}_{l,h}$. We have $\|X_i - \mu\|^2 \lesssim R^2 d$, hence

$$\mathbb{P}\{\|\widetilde{\Sigma} - \Sigma\| \gtrsim R^2/\alpha\} \lesssim d \exp\left(-c\frac{\#C}{\alpha^2 d}\right).$$

Moreover, $\|\widehat{\mu} - \mu\|^2 \lesssim \mathbb{R}^2/\alpha$ with same probability.

We now apply the Bernstein inequality to concentrate $v^T(\widehat{\Sigma} - \Sigma)$. By Proposition 1(a) we have, with probability no lower than $1 - 2n^{-\tau}$,

$$|v^T(X_i - \mu)|\|X_i - \mu\| \lesssim \ell^{-1} R^2 \sqrt{d\tau \log n} 2^{-l},$$

or $|v^T(X_i - \mu)|\|X_i - \mu]\| \lesssim \ell^{-1} R^2 \sqrt{d} 2^{-l}$ when $|\zeta| \le \sigma$. Next, we estimate the variance. We have

$$\|v^T(\widetilde{\Sigma} - \Sigma)\|^2 = v^T(\widetilde{\Sigma} - \Sigma)^2 v = v^T\widetilde{\Sigma}^2 v - v^T\widetilde{\Sigma}\Sigma v - v^T\Sigma\widetilde{\Sigma} v + v^T\Sigma^2 v,$$

hence, taking the expectation (conditioned on C),

$$\mathbb{E}[\|v^T(\widetilde{\Sigma} - \Sigma)\|^2] = \mathbb{E}[v^T\widetilde{\Sigma}^2 v] - v^T\Sigma^2 v,$$

where

$$\begin{aligned}
\mathbb{E}[v^T\widetilde{\Sigma}^2 v] &= \frac{1}{(\#C)^2} v^T \mathbb{E}\left[\left(\sum_i (X_i - \mu)(X_i - \mu)^T\right)^2\right] v \\
&\le \frac{1}{\#C} v^T \mathbb{E}[(X - \mu)\|X - \mu\|^2 (X - \mu)^T] v + v^T\Sigma^2 v \\
&\le \frac{1}{\#C} dR^2 \mathbb{E}[(v^T(X - \mu))^2] + v^T\Sigma^2 v \\
&= \frac{1}{\#C} dR^2 \mathrm{Var}[v^T X] + v^T\Sigma^2 v.
\end{aligned}$$

Thus, Proposition 1(b) gives

$$\mathbb{E}[\|v^T(\widetilde{\Sigma} - \Sigma)\|^2] \le \frac{1}{\#C} \ell^{-2} dR^4 2^{-2l}.$$

We therefore obtain

$$\mathbb{P}\{\|v^T(\widehat{\Sigma} - \Sigma)\| > \alpha^{-1} R^2 \varepsilon\} \lesssim d \exp\left(-c\frac{\#C\varepsilon^2}{\alpha^2 \ell^{-2} d\sqrt{\tau \log n}\ (2^{-2l} + 2^{-l}\varepsilon)}\right),$$

without $\sqrt{\tau \log n}$ if $|\zeta| \le \sigma$. Same bounds hold for $v^T(\widehat{\mu} - \mu)(\widehat{\mu} - \mu)^T$, which completes the proof. □

Proof of Thm 4. The Davis–Kahan Theorem [43, Theorem 2] yields

$$\|\widehat{v}_l - v\| \lesssim \left\| \frac{1}{\sum_h \#C_{l,h}} \sum_h \widehat{v}_{l,h}\widehat{v}_{l,h}^T \#C_{l,h} - vv^T \right\| \lesssim \frac{1}{\sum_h \#C_{l,h}} \sum_h \|\widehat{v}_{l,h} - v\| \#C_{l,h}.$$

Applying Proposition 3 and taking the union bound over h gives now (a). For (b), we condition on $|\zeta_i| \le \sqrt{2\tau \log n}\sigma$ for all i's and calculate

$$\begin{aligned}
\mathbb{E}[\|\widehat{v}_l - v\|^2] - n^{-\tau} &\lesssim \int_0^1 \varepsilon\ \mathbb{P}\{\|\widehat{v}_l - v\| > \varepsilon\} d\varepsilon \\
&= \int_0^{2^{-l}} \varepsilon\ \mathbb{P}\{\|\widehat{v}_l - v\| > \varepsilon\} d\varepsilon + \int_{2^{-l}}^1 \varepsilon\ \mathbb{P}\{\|\widehat{v}_l - v\| > \varepsilon\} d\varepsilon \\
&\le \int_0^{2^{-l}} \min\left\{1, 2^l d \exp\left(-\tfrac{c(n/\sqrt{\log n})\varepsilon^2}{\alpha^2 \ell^{-2}\sqrt{\tau} d 2^{-l}}\right)\right\} \varepsilon d\varepsilon \\
&+ \int_{2^{-l}}^1 2^l d \exp\left(-\tfrac{c(n/\sqrt{\log n})}{\alpha^2 \ell^{-2}\sqrt{\tau} d 2^{l}}\right) \varepsilon d\varepsilon \\
&\lesssim \alpha^2 \ell^{-2}\sqrt{\tau} d \log(2^l d) \frac{2^{-l}}{n/\sqrt{\log n}} \\
&+ 2^l d \exp\left(-\tfrac{c(n/\sqrt{\log n}}{\alpha^2 \ell^{-2}\sqrt{\tau} d 2^{l}}\right),
\end{aligned}$$

where the last inequality follows from Lemma 4. For $\tau = 2$ and n large enough as in the first assumed lower bound, we obtain (b). Analogous computations for the case where $|\zeta| \le \sigma$ lead to the final claim. □

3.2. Conditional regression bounds

In this section we study how partitioning polynomial regression is affected by the projection onto an estimate $\widehat{v}$ of v. We view these as estimators conditioned on $\widehat{v}$; we first prove that, with high probability, conditional estimators as defined in Section 2.5 differ from an oracle estimator (possessing knowledge of v) by the angle between $\widehat{v}$ and v (Theorem 5). Then, we show that such estimators achieve the 1-dimensional min-max convergence rate (up to logarithmic factors) when conditioned on any $\sqrt{n}$-convergent estimate of v (Theorem 6), and thus, in particular, on the $\widehat{v}$ obtained with SIR, SAVE, SCR or SVR (Corollary 1).

To prove Theorem 5 and 6 we need to assume that the distribution ρ of X does not change too much when projected onto directions within a small angle. A version of this property may be formalized as follows:

(Θ) X has an upper bounded density ρ for which there are $\Theta \in (0, 2\pi]$ and $\delta > 0$ such that, for every $r > 0$, angle θ with $|\theta| \le \Theta$, interval I with $|I| \le \delta$, and $u \in \mathbb{S}^{d-1}$ with $\|u - v\| \le \theta$,

$$W_1(\rho(x \mid \|x\| \le r, \langle v, x\rangle \in I), \rho(x \mid \|x\| \le r, \langle u, x\rangle \in I)) \le Cr\theta\,,$$

where W_1 denotes the 1^{st} Wasserstein distance.

Intuitively, (Θ) says that the mass of ρ does not move too far when ρ is slightly rotated, and is a continuity property. Note that, if ρ is rotationally invariant, then (Θ) holds trivially with $\Theta = 2\pi$ and any $\delta > 0$.

We will also need to impose some regularity on the function f. We recall that a function $g : \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{C}^s$ Hölder continuous ($g \in \mathcal{C}^s$) if, for $s = k + \alpha$, $k \geq 0$ an integer and $\alpha \in (0, 1]$, g has continuous derivatives up to order k and

$$|g|_{\mathcal{C}^s} = \max_{|\lambda|=k} \sup_{x \neq z} \frac{\delta^\lambda g(x) - \delta^\lambda g(z)}{\|x - z\|^\alpha} < \infty.$$

Theorem 5. *Assume (X), (Y), (Z), (Θ) and $f \in \mathcal{C}^\alpha$ with $\alpha \in [\frac{1}{2}, 1]$. Let $\widehat{v}$ be an estimate of v. For $u \in \{v, \widehat{v}\}$, let $\widehat{F}_{j|u}$ be a piecewise constant ($\alpha < 1$) or linear ($\alpha = 1$) estimator of F at scale j conditioned on u as defined in Section 2.5. Then, for every $\varepsilon > 0$, $r \geq 1$ and j such that $2^{-j} \geq \|\widehat{v} - v\|/t$ for some $t \geq 1$, conditioned on $\|X_i\| \leq r$ for all i's, we have*

$$\begin{aligned}&\left(\mathbb{E}_X[|\widehat{F}_{j|\widehat{v}}(X) - \widehat{F}_{j|v}(X)|^2 \mid X \in B(0, r)]\right)^{\frac{1}{2}} \\ &\lesssim t|f|_{\mathcal{C}^\alpha}\left(r^{\frac{1}{2-\alpha}}\|\widehat{v} - v\|^{\frac{1}{2-\alpha}} + r^\alpha 2^{-j\alpha} 2^{\frac{j}{2}}\|\widehat{v} - v\|^{\frac{1}{2}}\right) + \varepsilon\end{aligned}$$

with probability higher than $1 - C\#\mathcal{K}_j \exp(-\frac{c\, n\varepsilon^2}{\#\mathcal{K}_j t^2 |f|^2_{\mathcal{C}^\alpha} r^{2\alpha}})$.

The proof of Theorem 5 is postponed to Section 5. The key tool to obtain the dependence on $\|\widehat{v} - v\|$ in the upper bound is the Wasserstein distance. It enables us to bound the difference between statistics computed on the conditional distribution given $\widehat{v}$ rather than v.

We can finally establish the intrinsic min-max convergence rate of a conditional partitioning polynomial estimator. We will focus on one standard class of priors for regression functions, namely the class $\mathcal{C}^s$ of Hölder continuous functions.

Theorem 6. *Assume (X), (Y), (Z), (Θ) and $f \in \mathcal{C}^s \cap \mathcal{C}^{s \wedge 1}$ with $s \in [\frac{1}{2}, \frac{3}{2}]$. Let $\widehat{v}$ be an estimator for v such that*

$(\widehat{\text{V}})$ $\mathbb{P}\{\|\widehat{v} - v\| > \varepsilon\} \leq A \exp(-n\varepsilon^2/B)$

for some $A, B \geq 1$ possibly dependent on d and specific parameters. Let $\widehat{F}_{j|\widehat{v}}$ be a piecewise constant ($s \leq 1$) or linear ($s > 1$) estimator of F at scale j conditioned on $\widehat{v}$, defined as in Section 2.5 on a ball of radius r, and 0 outside. Then, setting $2^{-j} \asymp \sqrt{B}(\log n/n)^{1/(2s+1)}$ and $r = \sqrt{2d \log n^{2s/(2s+1)}} R$ we have:

(a) *For every $\nu > 0$ there is $c_\nu(d, R, B, |f|_{\mathcal{C}^{s\wedge 1}}, s) \geq 1$ such that*
$$\mathbb{P}\left\{(\mathbb{E}_X[|\widehat{F}_{j|\widehat{v}}(X) - F(X)|^2])^{\frac{1}{2}} > (\kappa + c_\nu)\log^{\frac{s\vee 1}{2}} n \left(\frac{\log n}{n}\right)^{\frac{s}{2s+1}}\right\} \lesssim An^{-\nu}$$
for some $\kappa(d, R, B, F, |f|_{\mathcal{C}^{s\wedge 1}}, |f|_{\mathcal{C}^s}, s)$.

(b) $\mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - F(X)|^2] \le K \log^{s\vee 1} n \left(\frac{\log n}{n}\right)^{\frac{2s}{2s+1}}$
for some $K = K(d, R, A, B, F, |f|_{\mathcal{C}^{s\wedge 1}}, |f|_{\mathcal{C}^s}, s)$.

The dependence of all constants upon d, A *and* B *is polynomial.*

Proof. Let $\alpha = s \wedge 1$ so that $f \in \mathcal{C}^s \cap \mathcal{C}^\alpha$. We first prove (b). Let us start by isolating the error outside a ball $B(0, r)$:

$$\mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - F(X)|^2] \le \mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - F(X)|^2 \mid X \in B(0,r)] + \mathbb{E}[|F(X)|^2 \mathbb{1}\{X \notin B(0,r)\}],$$

where, in view of Lemma 5,

$$\mathbb{E}[|F(X)|^2 \mathbb{1}\{X \notin B(0,r)\}] \lesssim (|F(0)|^2 + d|f|^2_{\mathcal{C}^\alpha} R^2) \exp(-r^2/2dR^2). \tag{T}$$

Now we can focus on $\mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - F(X)|^2 \mid X \in B(0,r)]$. To lighten the notation, from this point we will spare writing the conditioning $X \in B(0,r)$. Let us split the mean squared error $\mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - F(X)|^2]$ into a bias and a variance term:

$$\mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - F(X)|^2] = \mathbb{E}[|F_{j|v}(X) - F(X)|^2] + \mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - F_{j|v}(X)|^2],$$

where $F_{j|v}$ is the population version of $\widehat{F}_{j|v}$. Since, by assumption, $f \in \mathcal{C}^s$, the bias term $\mathbb{E}[|F_{j|v}(X) - F(X)|^2]$ is bounded by

$$\mathbb{E}[|F_{j|v}(X) - F(X)|^2] \lesssim |f|^2_{\mathcal{C}^s} r^{2s} 2^{-2js} \tag{B}$$

(see [40, Section 3.2]). Now we need to bound the variance. We introduce the intermediate term $\widehat{F}_{j|v}$, the oracle estimator computed along the true direction v. Thus,

$$\mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - F_{j|v}(X)|^2] \lesssim \mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - \widehat{F}_{j|v}(X)|^2] + \mathbb{E}[|\widehat{F}_{j|v}(X) - F_{j|v}(X)|^2].$$

The second term can be bounded in expectation as in [40, Proposition 13]:

$$\mathbb{E}[|\widehat{F}_{j|v}(X) - F_{j|v}(X)|^2] \lesssim |f|^2_{\mathcal{C}^\alpha} \frac{j 2^j}{n}. \tag{V1}$$

To obtain a bound for the first term, we write

$$\begin{aligned}\mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - \widehat{F}_{j|v}(X)|^2] \le{}& \mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - \widehat{F}_{j|v}(X)|^2 \mid \|\widehat{v} - v\| \le 2^{-j}] \\ &+ \mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - \widehat{F}_{j|v}(X)|^2 \mid r\|\widehat{v} - v\| > 2^{-j}] \, \mathbb{P}\{\|\widehat{v} - v\| > 2^{-j}\}.\end{aligned}$$

By assumption $(\widehat{\mathrm{V}})$ we get

$$\mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - \widehat{F}_{j|v}(X)|^2 \mid \|\widehat{v} - v\| > 2^{-j}] \, \mathbb{P}\{\|\widehat{v} - v\| > 2^{-j}\} \lesssim A|f|^2_{\mathcal{C}^\alpha} r^2 \exp(-n2^{-2j}/B). \tag{V2}$$

On the other hand, Theorem 5 along with assumption ($\widehat{\mathrm{V}}$) gives

$$\begin{aligned}&\mathbb{P}\{\mathbb{E}_X[|\widehat{F}_{j|\widehat{v}}(X)-\widehat{F}_{j|v}(X)|^2]>\varepsilon^2 \mid \|\widehat{v}-v\|\le 2^{-j}\}\\ &\lesssim \#\mathcal{K}_j \exp\Big(-c\tfrac{n\varepsilon^2}{\#\mathcal{K}_j|f|^2_{\mathcal{C}^\alpha} r^{2\alpha}}\Big)+A\exp\Big(-\tfrac{n\varepsilon^{2(2-\alpha)}}{B|f|^{2(2-\alpha)}_{\mathcal{C}^\alpha} r^2}\Big)+A\exp\Big(-\tfrac{n\varepsilon^4}{B|f|^4_{\mathcal{C}^\alpha} r^{4\alpha}2^{-4j\alpha}2^{2j}}\Big),\end{aligned}$$

hence, using Lemma 4,

$$\begin{aligned}&\mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X)-\widehat{F}_{j|v}(X)|^2 \mid \|\widehat{v}-v\|\le 2^{-j}] \qquad \text{(V3)}\\ &\lesssim |f|^2_{\mathcal{C}^\alpha} r^{2\alpha}\frac{j2^j}{n}+(\log(A)B)^{\frac{1}{2-\alpha}}|f|^2_{\mathcal{C}^\alpha} r^{\frac{2}{2-\alpha}} n^{-\frac{1}{2-\alpha}}+(\log(A)B)^{\frac12}|f|^2_{\mathcal{C}^\alpha} r^{2\alpha}2^{-2j\alpha}2^j n^{-\frac12}.\end{aligned}$$

In order to balance the tail (T), the bias (B) and variance terms (V1), (V2) and (V3), we choose

$$r=\sqrt{2d\log n^{2s/(2s+1)}}R\,,\qquad 2^{-j}\asymp \sqrt{B}(\log n/n)^{1/(2s+1)},$$

which leads to

$$\begin{aligned}\mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X)-F(X)|^2] &\lesssim (|F(0)|^2+|f|^2_{\mathcal{C}^\alpha}R^2 d)\left(\frac1n\right)^{\frac{2s}{2s+1}}\\ &\quad+|f|^2_{\mathcal{C}^s}R^{2s}B^s d^s\log^s n\left(\frac{\log n}{n}\right)^{\frac{2s}{2s+1}}\\ &\quad+|f|^2_{\mathcal{C}^\alpha}\left(\frac{\log n}{n}\right)^{\frac{2s}{2s+1}}\\ &\quad+|f|^2_{\mathcal{C}^\alpha}R^2 Ad\left(\frac{\log n}{n}\right)\\ &\quad+|f|^2_{\mathcal{C}^\alpha}R^{\frac{2}{2-\alpha}}(\log(A)B)^{\frac{1}{2-\alpha}}d^{\frac{1}{2-\alpha}}\log^{\frac{1}{2-\alpha}}n\left(\frac{\log n}{n}\right)^{\frac{2s}{2s+1}}.\end{aligned}$$

(For (V2), $\exp(-n2^{-2j}/B)=\exp(-n(\log n/n)^{\frac{2}{2s+1}})$, bounded by n^{-1} for $s\ge\frac12$. For (V3), $n^{-\frac{1}{2-\alpha}}\le n^{-\frac{2s}{2s+1}}$ for all $\alpha=s\in(0,1]$ and for $\alpha=1$ and all $s>0$; moreover, $2^{-j(2\alpha)}2^j n^{-\frac12}\asymp B^{\alpha-\frac12}(\log n/n)^{\frac{2\alpha-1}{2s+1}}n^{-\frac12}$ where $n^{-\frac{2\alpha-1}{2s+1}}n^{-\frac12}=n^{-\frac{4\alpha+2s-1}{2(2s+1)}}\le n^{-\frac{2s}{2s+1}}$ for all $\alpha=s\ge\frac12$ and for $\alpha=1$ and all $s\le\frac32$.) Collecting the constants we obtain (b).

We now turn to (a). Outside $B(0,r)$, (T) reads

$$(\mathbb{E}_X[|\widehat{F}_{j|\widehat{v}}(X)-F(X)|^2\mathbb{1}\{X\notin B(0,r)\}])^{\frac12}\lesssim(|F(0)|+\sqrt{d}|f|_{\mathcal{C}^\alpha}R)n^{-\frac{s}{2s+1}}.$$

On $X\in B(0,r)$, again skipping the conditioning, we have

$$\begin{aligned}(\mathbb{E}_X[|\widehat{F}_{j|\widehat{v}}(X)-F(X)|^2])^{\frac12}&\le(\mathbb{E}_X[|\widehat{F}_{j|\widehat{v}}(X)-\widehat{F}_{j|v}(X)|^2])^{\frac12}\\ &\quad+(\mathbb{E}_X[|\widehat{F}_{j|v}(X)-F_{j|v}(X)|^2])^{\frac12}\\ &\quad+(\mathbb{E}[|F_{j|v}(X)-F(X)|^2])^{\frac12},\end{aligned}$$

where the last term is bounded by $|f|_{\mathcal{C}^s} R^s B^{\frac{s}{2}} d^{\frac{s}{2}} \log^{\frac{s}{2}} n (\log n/n)^{s/(2s+1)}$ by (B). The middle term can be concentrated with standard calculations (see e.g. [40]), to obtain

$$\mathbb{P}\{(\mathbb{E}_X[|\widehat{F}_{j|v}(X) - F_{j|v}(X)|^2])^{\frac{1}{2}} > \varepsilon\} \lesssim \#\mathcal{K}_j \exp\left(-c\tfrac{n\varepsilon^2}{\#\mathcal{K}_j |f|^2_{\mathcal{C}^\alpha}}\right).$$

Setting $\varepsilon = c_\nu \log^{\frac{s}{2}} n \, (\log n/n)^{s/(2s+1)}$, the right hand side becomes

$$B^{-\frac{1}{2}} \left(\frac{\log n}{n}\right)^{-\frac{1}{2s+1}} \exp\left(-c\frac{n c_\nu^2 \log^s n \left(\frac{\log n}{n}\right)^{\frac{2s}{2s+1}}}{B^{-\frac{1}{2}}\left(\frac{\log n}{n}\right)^{-\frac{1}{2s+1}} |f|^2_{\mathcal{C}^\alpha}}\right) \leq n^{-\left(c\frac{c_\nu^2\sqrt{B}}{|f|^2_{\mathcal{C}^\alpha}} - \frac{1}{2s+1}\right)}.$$

Finally, in view of Theorem 5 and assumption ($\widehat{\mathrm{V}}$), for $Z = (\mathbb{E}_X[|\widehat{F}_{j|\widehat{v}}(X) - \widehat{F}_{j|v}(X)|^2])^{\frac{1}{2}}$ we have

$$\begin{aligned}\mathbb{P}\{Z > \varepsilon\} &\leq \mathbb{P}\{Z > \varepsilon \mid \|\widehat{v} - v\| \leq \sqrt{c_\nu} 2^{-j}\} + \mathbb{P}\{\|\widehat{v} - v\| > \sqrt{c_\nu} 2^{-j}\} \\ &\lesssim n^{-\left(c\frac{c_\nu\sqrt{B}}{|f|^2_{\mathcal{C}^\alpha} R^{2\alpha} d^\alpha} - \frac{1}{2s+1}\right)} + An^{-\frac{c_\nu^2}{|f|^2_{\mathcal{C}^\alpha} R^2 B d}} + An^{-\frac{c_\nu^4}{|f|^4_{\mathcal{C}^\alpha} R^4 B^2 d^2}} + An^{-\frac{c_\nu}{B}}.\end{aligned}$$

The bound (a) follows by taking c_ν large enough. □

The additional logarithmic factors in (a) and (b) are exclusively due to the unboundedness of the distribution and can be avoided in the bounded case.

As a direct consequence of our findings on index estimation and conditional regression, we obtain:

Corollary 1. *Let $\widehat{v}$ be estimated by SIR, SAVE, SCR or SVR under the assumptions of Theorems 1, 2, 3 or 4, respectively. Then $\widehat{v}$ satisfies assumption ($\widehat{\mathrm{V}}$), and therefore the assertions (a) and (b) of Theorem 6 hold true for each of the four methods, with constants depending polynomially on d.*

4. Numerical experiments

In this section we conduct numerical experiments to demonstrate that the theoretical results above have practical relevance and to investigate how relaxations of the assumptions affect the estimators. In order to highlight specific aspects of different algorithms we use three different functions to conduct our experiments. The first two are

$$F_1(x) = \exp(\langle v, x\rangle/3)), \qquad F_2(x) = F_1(x) + \sin(20\langle v, x\rangle)/15.$$

Both functions are smooth. F_1 is monotone and thus we may choose $\omega = 0$, while F_2 is non-monotone, thus condition (Ω) is satisfied only for $\omega > 0$. This allows us to explore the behavior of $\widehat{v}$ under monotonicity or lack thereof, and how the estimators are effected by the choice of the scales l and j. To investigate the convergence rate of the regression estimator $\widehat{F}$, we use a monotone function F_3 which is piecewise quadratic on a random partition and continuous. The domain of x, and its dimension d, will be specified in each experiment. F_1, F_2, F_3 are shown in Figure 2.

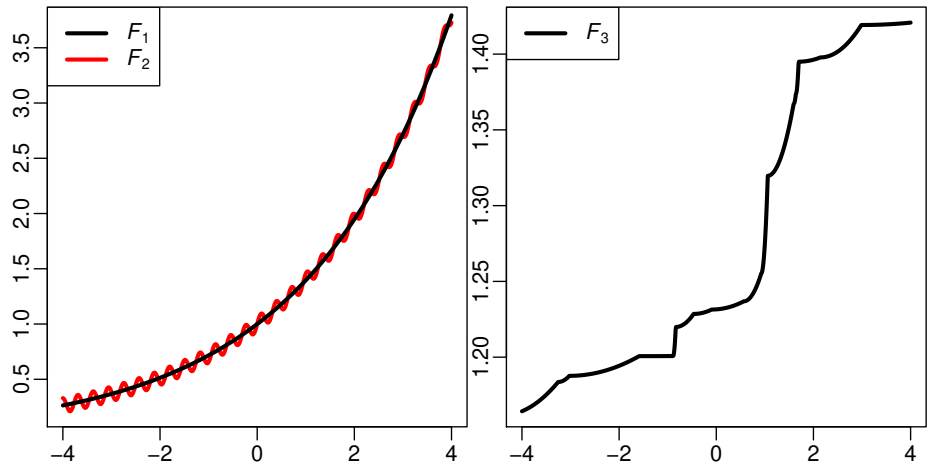

Figure 2: Different functions used in the experiments, with horizontal axis representing $\langle v, x\rangle$.

4.1. Estimating the index vector v

Here we compare the performances of SIR, SAVE, SCR and SVR in estimating the index vector v. We consider two settings S1, S2, corresponding to two different non-elliptical distributions for X: $\rho_{X,1}, \rho_{X,2}$; $\rho_{X,1}$ is a standard normal $\mathcal{N}(0,1)$ in one coordinate, and a skewed normal with shape parameter $\alpha = 5$ in the other coordinate; $\rho_{X,2}$ is uniform on the triangle with vertices (0,0),(1,1),(0,1); all distributions are normalized to have zero mean and standard deviation equal to one. Note that in both settings condition (LCM) is not satisfied. For each setting we draw $n = 1000$ i.i.d. samples and generate the response variable $Y_i = F(X_i) + \zeta_i$ using functions F_1 and F_2, where $\zeta_i \sim \mathcal{N}(0, \sigma^2)$. We use different levels of noise setting σ equal to the 0%, 1% and 2% of $|f(-4) - f(4)|$. We chose $v = (1/\sqrt{5}, 2/\sqrt{5})$ for setting S1, and $v = (1, 0)$ for setting S2. The results in Table 3 show the detailed performance of SIR, SAVE, SCR and SVR for all settings, functions, and noise levels. We notice that SCR has overall good performances, especially in setting S1, but its quadratic computational cost makes the average computational time 2 to 3 orders of magnitude larger than the one of the other methods. SVR and SAVE have similar performance, although SVR produces most of the times slightly better estimates, especially in S2 where it also outperforms SCR. Note that the cases of F_1 with 1% noise and F_2 with zero noise produce similar results. This is consistent with the intuition that noise and non-monotonicity levels play a similar role in the accuracy of the estimators.

In Figure 3 we show graphically how the empirical inverse regression curve may drift away from v, resulting in a poor SIR estimate. On the other hand, the local gradients

Table 3. Performance of the different algorithms in different settings, with err $= \log_{10}(\|\widehat{v} - v\|^2)$, corresponding standard error, and average computational time in seconds/100.

			S1			S2		
	σ		err	se	time	err	se	time
F_1	0%	**SIR**	-3.04	-2.83	0.60	-0.68	-1.83	0.60
		SAVE	-5.97	-1.06	1.30	-7.41	-7.14	1.40
		SCR	-8.42	-8.16	143.80	-8.50	-8.30	118.20
		SVR	-6.42	-6.06	1.00	-7.60	-6.65	0.70
F_2		**SIR**	-3.11	-2.99	0.50	-0.67	-1.83	0.50
		SAVE	-4.39	-4.10	1.30	-4.13	-4.01	1.30
		SCR	-4.80	-4.40	150.70	-4.00	-3.77	115.40
		SVR	-4.41	-4.08	0.90	-4.16	-3.92	0.70
F_1	1%	**SIR**	-2.97	-2.69	0.50	-0.68	-1.81	0.50
		SAVE	-4.57	-4.24	1.40	-4.16	-4.03	1.30
		SCR	-4.85	-4.55	152.70	-4.20	-4.05	117.20
		SVR	-4.58	-4.28	1.00	-4.12	-3.75	0.80
F_2		**SIR**	-2.94	-2.72	0.50	-0.68	-1.77	0.50
		SAVE	-4.06	-3.57	1.40	-3.42	-3.45	1.40
		SCR	-4.41	-4.07	150.80	-3.43	-3.44	117.30
		SVR	-4.08	-3.87	0.90	-3.45	-3.46	0.80
F_1	2%	**SIR**	-2.92	-2.67	0.50	-0.68	-1.38	0.60
		SAVE	-3.92	-3.58	1.30	-3.08	-3.01	1.40
		SCR	-4.20	-3.87	149.90	-3.09	-3.24	119.10
		SVR	-3.91	-3.61	0.90	-3.21	-3.06	0.80
F_2		**SIR**	-2.87	-2.64	0.60	-0.68	-1.55	0.60
		SAVE	-3.64	-1.98	1.40	-2.90	-2.85	1.40
		SCR	-4.00	-3.68	150.40	-2.91	-3.10	119.00
		SVR	-3.66	-3.45	0.90	-3.02	-2.80	0.80

used by SVR provide good local estimates.

To investigate more extensively the performance of SVR in estimating v, we perform another experiment: we draw X from a 10-dimensional standard normal distribution, and to generate the response variable we use function F_2 plus an additive Gaussian noise with standard deviation $\sigma = 0.01|f_2(-4) - f_2(4)|$. We repeat the experiment for different values of the sample size n. Results are shown in Figure 4. The left inset shows that the error in $\widehat{v}$ stabilizes at scales comparable to the noise level σ, which suggests that the assumption $|C_{l,h}| \gtrsim \sigma$ is needed. The right plot shows that the rate of the error of $\widehat{v}$, for scales l coarser than the noise level, is approximately $-\frac{1}{2}$, which is again consistent with Theorem 4.

4.2. Estimating the regression function F

In this section we perform some experiments to support our theoretical results regarding the regression estimator obtained with SVR. The first experiment we perform consists on

drawing X_i, $i = 1, ..., n$, from a d-dimensional standard Normal distribution and obtain $Y_i = F_3(X_i) + \zeta_i$ where $\zeta_i \sim \mathcal{N}(0, \sigma^2)$. Here we use function F_3 because we want to limit the function smoothness in order to obtain concentration rates comparable with the min-max rate with $s = 1$. We vary the dimension $d = 5, 10, 50, 100$, the size of the noise σ, equal to the 5% and 10% of $|f_3(-4) - f_3(4)|$. To investigate the convergence rates of the estimator we repeat each experiment for different sample sizes n. In Figure 5 we show the empirical MSE, averaged over 10 repetitions, as a function of the sample size, in logarithmic scale, for both our estimator and the k-Nearest-Neighbor (kNN) regression. We see that the MSE of the SVR estimator decays with a rate slightly better than the optimal value $-2/3$, independently from the dimension d and the noise level σ: this is all

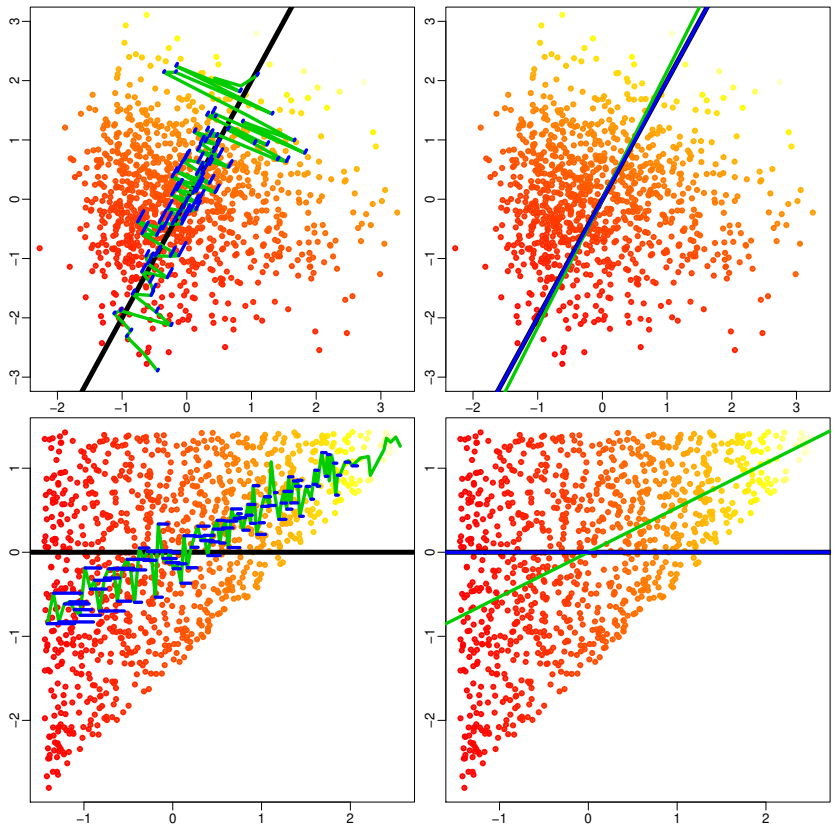

Figure 3: Left column displays the ingredients for the estimates: the empirical inverse regression curve (green) used by SIR and the local gradients (blue), with length proportional to the number of samples in the corresponding level set, used by SVR. Estimates of v using SVR (blue) and SIR (green) are displayed on the right column. The methods are applied on setting S1 (top row) and S2 (bottom row). The black line indicates v, while data points are colored according to the value of the corresponding response variable, generated with F_2 and $\sigma = 0$, using a red-to-yellow color scale.

consistent with Theorem 6. As expected, kNN-regression has a convergence rate which severely deteriorates with the dimension (curse of dimensionality). We can also notice that the MSE drops far below the noise level, which confirms the de-noising feature of the SVR estimator.

To explore the behavior of the empirical MSE as a function of the scales l and j we conduct another experiment: we draw X from a 10-dimensional standard normal distribution, and obtain the response variable $Y = F_2(X) + \zeta$, with ζ Gaussian noise with standard deviation $\sigma = 0.01|f_2(-4) - f_2(4)|$. Figure 6 shows the behavior of the $\log_{10}$(MSE), obtained with SVR, for different values of l, j and n. To obtain robust estimates in regions with high Monte Carlo variability, in regimes where our results do not hold, the errors are averaged over 50 repetition of each setting with a 10% trimming. By observing each row, we notice that the MSE reaches its minimum for low values of l and stays constant for larger l. By looking at the plot column-wise, we observe the bias variance trade-off, with coarse scales giving rough estimates, and fine scales resulting in overfitting. As expected, as the sample size grows, the optimal scale j increases.

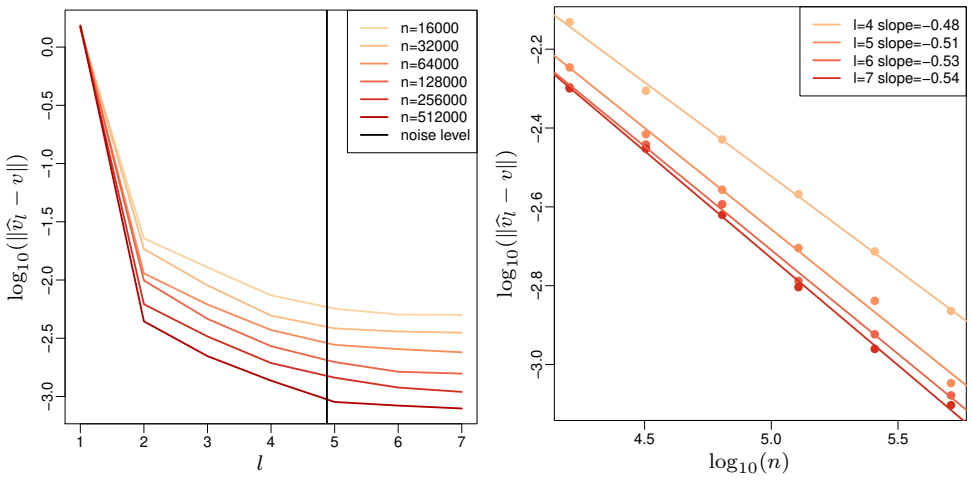

Figure 4: Behavior of the SVR estimate $\widehat{v}_l$ with respect to scale and sample size, for regression of F_2 (see text). Left: error versus scale l. Right: error versus sample size n.

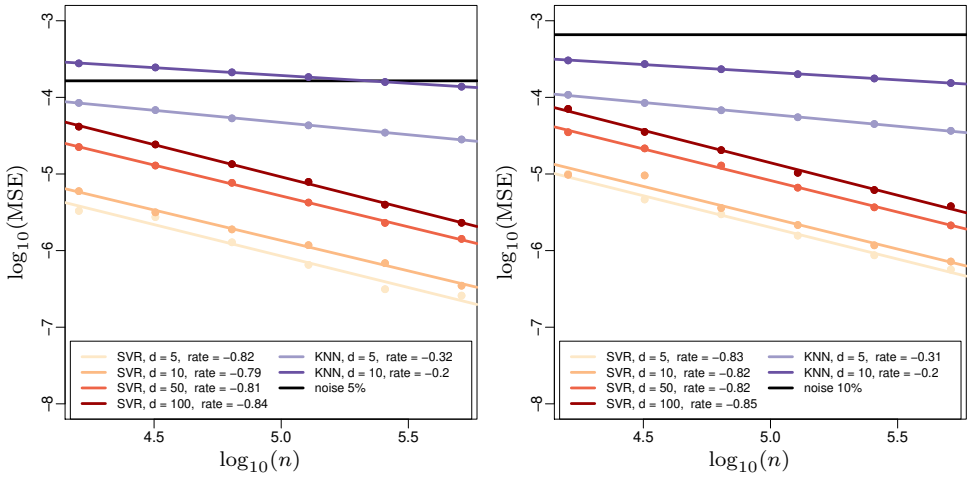

Figure 5: Comparison of convergence rates for the regression estimator with SVR and KNN-regression in different settings.

5. Proofs of Theorems 1, 2, 3 and 5

Proof of Thm 1. By the Davis–Kahan Theorem [43, Theorem 2] we have, up to possibly a choice of sign for $\widehat{v}_l$, which will subsume here and in all that follows,

$$\|\widehat{v}_l - v\| \lesssim \frac{\|\widehat{M}_l - M_l\|}{\lambda_1(M_l) - \lambda_2(M_l)} \leq \frac{\alpha}{R^2}\|\widehat{M}_l - M_l\|\,,$$

since assumption (LCM) implies $\lambda_2(M_l) = 0$, while $\lambda_1(M_l) \geq R^2/\alpha$ by assumption (I). Moreover,

$$\begin{aligned}\|\widehat{M}_l - M_l\| &\leq \sum_h \|\widehat{\mu}_{l,h}\|^2 |\#C_{l,h}/n - \mathbb{P}\{Y \in C_{l,h}\}| \\ &\quad + \sum_h (\|\widehat{\mu}_{l,h}\| + \|\mu_{l,h}\|)\|\widehat{\mu}_{l,h} - \mu_{l,h}\|\mathbb{P}\{Y \in C_{l,h}\} \\ &\lesssim dR^2 \sum_h |\#C_{l,h}/n - \mathbb{P}\{Y \in C_{l,h}\}| \\ &\quad + \sqrt{d}R \sum_h \|\widehat{\mu}_{l,h} - \mu_{l,h}\|\mathbb{P}\{Y \in C_{l,h}\}.\end{aligned}$$

Thus,

$$\|\widehat{v}_l - v\| \lesssim \alpha d \sum_h |\#C_{l,h}/n - \mathbb{P}\{Y \in C_{l,h}\}|$$

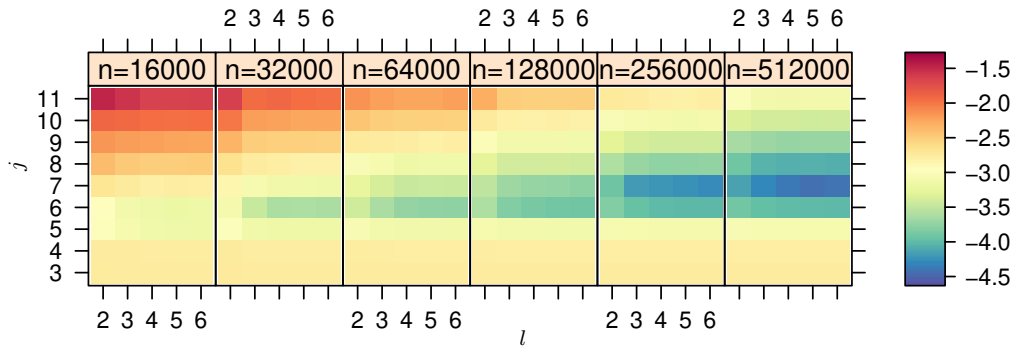

Figure 6: Empirical MSE versus sample size n and scales l and j.

$$+ \alpha\sqrt{d}R^{-1}\sum_h \|\widehat{\mu}_{l,h} - \mu_{l,h}\|\mathbb{P}\{Y \in C_{l,h}\}.$$

Notice that, since $\|\widehat{v}_l - v\| \leq 1$, we may assume $\varepsilon^2 \leq \varepsilon \leq 1$, and we will make these substitutions in the probability exponential bounds wherever is convenient. For the first term, we use Lemma 1 with $t = \varepsilon/2^l\alpha d$ for the h's such that $\mathbb{P}\{Y \in C_{l,h}\} \leq 2^{-l}$, and with $t = \sqrt{\mathbb{P}\{Y \in C_{l,h}\}}\varepsilon/2^{l/2}\alpha d$ for the h's such that $\mathbb{P}\{Y \in C_{l,h}\} > 2^{-l}$. This gives that the first term is bounded by ε with probability $1 - C2^l\exp(-cn\varepsilon^2/\alpha^2 2^l d^2)$. We now deal with the second term. For all h's s.t. $\mathbb{P}\{Y \in C_{l,h}\} \leq \varepsilon/\alpha 2^l d$ we have directly

$$\alpha R^{-1}\sqrt{d}\sum_h \|\widehat{\mu}_{l,h} - \mu_{l,h}\|\mathbb{P}\{Y \in C_{l,h}\} \leq \varepsilon.$$

For all other h's, we condition on $\#C_{l,h} \geq \frac{1}{2}\mathbb{E}\#C_{l,h} = \frac{1}{2}n\mathbb{P}\{Y \in C_{l,h}\}$ up to events of probability lower than

$$C\exp(-cn\mathbb{P}\{Y \in C_{l,h}\}) \leq C\exp(-cn\varepsilon/\alpha 2^l d),$$

thanks to Lemma 1. We finally apply the Bernstein inequality: for $\mathbb{P}\{Y \in C_{l,h}\} \in (\varepsilon/\alpha 2^l d, 1/2^l]$, we have

$$\mathbb{P}\left\{\|\widehat{\mu}_{l,h} - \mu_{l,h}\| > \frac{R\varepsilon}{\sqrt{2^l\mathbb{P}\{Y \in C_{l,h}\}}\alpha\sqrt{d}}\right\} \lesssim d\exp\left(-c\frac{n\varepsilon^2}{\alpha^2 2^l d^2}\right);$$

and for $\mathbb{P}\{Y \in C_{l,h}\} > 1/2^l$

$$P\left\{\|\widehat{\mu}_{l,h} - \mu_{l,h}\| > \frac{R\varepsilon}{\alpha\sqrt{d}}\right\} \lesssim d\exp\left(-c\frac{n\varepsilon^2}{\alpha^2 2^l d^2}\right).$$

The union bound over h gives (a), while (b) follows from (a) and Lemma 4. □

Proof of Thm 2. Assumption (LCM) implies by [11, Theorem 1.a] that v is an eigenvector of S_l. Moreover,

$$v^T S_l v = \sum_h \|(I-\Sigma_{l,h})v\|^2 \mathbb{P}\{Y\in C_{l,h}\} = \sum_h (1-\lambda_d(\Sigma_{l,h}))^2 \mathbb{P}\{Y\in C_{l,h}\},$$

while

$$\begin{aligned}\max_{\substack{w\in\mathrm{span}\{v\}^\perp\\ \|w\|=1}} w^T S_l w &\le \sum_h \max_{\substack{w_h\in\mathrm{span}\{v\}^\perp\\ \|w_h\|=1}} \|(I-\Sigma_{l,h})w_h\|^2 \mathbb{P}\{Y\in C_{l,h}\}\\ &= \sum_h (1-\lambda_{d-1}(\Sigma_{l,h}))^2 \mathbb{P}\{Y\in C_{l,h}\},\end{aligned}$$

hence, thanks to (UCV) and (V), v is the eigenvector of largest eigenvalue of S_l. Now, the Davis–Kahan theorem [43, Theorem 2] gives

$$\|\widehat{v}_l - v\| \lesssim \frac{\|\widehat{S}_l - S_l\|}{\lambda_1(S_l)-\lambda_2(S_l)},$$

where, using (V) and (UCV),

$$\lambda_1(S_l)-\lambda_2(S_l) \ge \sum_h [(1-\lambda_d(\Sigma_{l,h}))^2 - (1-\lambda_{d-1}(\Sigma_{l,h}))^2]\mathbb{P}\{Y\in C_{l,h}\} \ge \alpha^{-2}.$$

Therefore, $\|\widehat{v}_l - v\| \lesssim \alpha^2 \|\widehat{S}_l - S_l\|$, where

$$\|\widehat{S}_l - S_l\| \lesssim d^2 \sum_h |\#C_{l,h}/n - \mathbb{P}\{Y\in C_{l,h}\}| + \sum_h \|(I-\widehat{\Sigma}_{l,h})^2 - (I-\Sigma_{l,h})^2\| \mathbb{P}\{Y\in C_{l,h}\}$$

with

$$\begin{aligned}\|(I-\widehat{\Sigma}_{l,h})^2 - (I-\Sigma_{l,h})^2\| &= \|\widehat{\Sigma}_{l,h}^2 - \Sigma_{l,h}^2 - 2(\widehat{\Sigma}_{l,h}-\Sigma_{l,h})\|\\ &= \|\widehat{\Sigma}_{l,h}(\widehat{\Sigma}_{l,h}-\Sigma_{l,h}) + (\widehat{\Sigma}_{l,h}-\Sigma_{l,h})\Sigma_{l,h} - 2(\widehat{\Sigma}_{l,h}-\Sigma_{l,h})\|\\ &\lesssim d\|\widehat{\Sigma}_{l,h}-\Sigma_{l,h}\|.\end{aligned}$$

Recall that the maximum error of $\widehat{v}_l$ is 1, hence we can always take $\varepsilon^2 \le \varepsilon \le 1$. Applying Lemma 1 with $t = \varepsilon/2^l\alpha^2 d^2$ when $\mathbb{P}\{Y\in C_{l,h}\} \le 2^{-l}$ and with $t = \sqrt{\mathbb{P}\{Y\in C_{l,h}\}}\varepsilon/2^{l/2}\alpha^2 d^2$ when $\mathbb{P}\{Y\in C_{l,h}\} > 2^{-l}$, we obtain

$$\mathbb{P}\{d^2 \sum_h |\#C_{l,h}/n - \mathbb{P}\{Y\in C_{l,h}\}| > \varepsilon\} \lesssim 2^l \exp(-cn\varepsilon^2/\alpha^4 d^4 2^l).$$

For the h's with $\mathbb{P}\{Y\in C_{l,h}\} \le \varepsilon/\alpha^2 d^2 2^l$ we already have

$$\sum_h \|\widehat{\Sigma}_{l,h} - \Sigma_{l,h}\| \mathbb{P}\{Y\in C_{l,h}\} \le \varepsilon.$$

Thus we can assume $\mathbb{P}\{Y \in C_{l,h}\} > \varepsilon/\alpha^2 d^2 2^l$, and condition on $\#C_{l,h} > n\mathbb{P}\{Y \in C_{l,h}\}$ with confidence $1 - C\exp(-cn\varepsilon/\alpha^2 d^2 2^l)$, thanks to Lemma 1. We split $\widehat{\Sigma}_{l,h} - \Sigma_{l,h} = \widetilde{\Sigma}_{l,h} - \Sigma_{l,h} - (\widehat{\mu}_{l,h} - \mu_{l,h})(\widehat{\mu}_{l,h} - \mu_{l,h})^T$ with

$$\widetilde{\Sigma}_{l,h} = \frac{1}{\#C_{l,h}} \sum_h (X_i = \mu_{l,h})(X_i - \mu_{l,h})^T \mathbb{1}\{Y_i \in C_{l,h}\},$$

and use the Bernstein inequality to concentrate $\widetilde{\Sigma}_{lh} - \Sigma_{l,h}$ and $\widehat{\mu}_{l,h} - \mu_{l,h}$. For $\mathbb{P}\{Y \in C_{l,h}\} \in (\varepsilon/\alpha^2 d^2 2^l, 1/2^l]$ we get

$$\mathbb{P}\left\{\|\widetilde{\Sigma}_{lh} - \Sigma_{l,h}\| > \frac{\varepsilon}{\sqrt{2^l \mathbb{P}\{Y \in C_{l,h}\}}\alpha^2 d}\right\} \lesssim d\exp\left(-c\frac{n\varepsilon^2}{\alpha^4 2^l d^4}\right);$$

and for $\mathbb{P}\{Y \in C_{l,h}\} > 1/2^l$

$$P\left\{\|\widetilde{\Sigma}_{lh} - \Sigma_{l,h}\| > \frac{\varepsilon}{\alpha^2 d}\right\} \lesssim d\exp\left(-c\frac{n\varepsilon^2}{\alpha^4 2^l d^4}\right).$$

Similarly for $\widehat{\mu}_{l,h} - \mu_{l,h}$. Summing these bounds over h yields (a), and (b) follows by Lemma 4. □

Proof of Thm 3. The calculations in [35, Theorem 6.1] show that under (LCM) (and even without (CCV)) one has

$$K_l = PK_lP + 2P^\perp \mathbb{E}[\mathrm{Cov}[X \mid Y] \mid |Y - \widetilde{Y}| < 2^{-l}]P^\perp,$$

where P is the orthogonal projection onto $\mathrm{span}\{v\}$. Hence, v is an eigenvector of K_l, and, by assumption (C), it is the eigenvector of smallest eigenvalue. Moreover, the matrix

$$H_l = \mathbb{E}[(X - \widetilde{X})(X - \widetilde{X})^T \mathbb{1}\{|Y - \widetilde{Y}| \le 2^{-l}\}]$$

differs from K_l only by the factor $\rho_l = \mathbb{P}\{|Y - \widetilde{Y}| \le 2^{-l}\}$. Using the Davis–Kahan theorem [43, Theorem 2] and assumption (C) we obtain

$$\|\widehat{v}_l - v\| \lesssim \frac{\|\widehat{H}_l - H_l\|}{\lambda_{d-1}(H_l) - \lambda_d(H_l)} \le \frac{\alpha}{R^2 \rho_l}\|\widehat{H}_l - H_l\|.$$

Concentration inequalities for U-statistics (see [42]) now imply

$$\mathbb{P}\left\{\|\widehat{H}_l - H_l\| > \alpha^{-1} R^2 \rho_l \varepsilon\right\} \lesssim d\exp\left(-c\frac{n\rho_l \varepsilon^2}{\alpha^2 d}\right).$$

This gives (a), which in turn implies (b) by Lemma 4. □

Proof of Thm 5. First, we set out some notation and exclude some low-probability events. In all expressions $\mathbb{E}_X[\cdot]$ we will drop the random variable X and simply write $\mathbb{E}[\cdot]$. We define ρ to be the distribution of X, and $\rho(\cdot \mid E)$ the conditional distribution of X given $X \in E$. All integrations in $d\rho(x)$ are implicitly taken on $x \in B(0,r)$. Let $u \in \{v, \widehat{v}\}$; when a property is stated for u, it is meant to hold for both v and $\widehat{v}$. Abusing the notation, we write $I_{j,k|u}$ for $\{x \in \mathbb{R}^d : \langle u, x\rangle \in I_{j,k|u}\}$.

Thanks to the Hölder continuity of F, we can restrict to the sets $I_{j,k|u}$ with

$$\rho(I_{j,k|u}) \gtrsim \varepsilon^2/\#\mathcal{K}_j |f|^2_{\mathcal{C}^\alpha} r^{2\alpha}. \tag{E1}$$

We further condition on the event

$$\#I_{j,k|u} \gtrsim n\rho(I_{j,k|u}) \text{ for all } k\text{'s}, \tag{E2}$$

which has probability at least $1 - C\#\mathcal{K}_j \exp(-c\ n\varepsilon^2/\#\mathcal{K}_j |f|^2_{\mathcal{C}^\alpha} r^{2\alpha})$, thanks to Lemma 1. For two probability measures μ and ν, we define the Kantorovich distance

$$K_\alpha(\mu,\nu) = \sup_{g \in \mathcal{C}^\alpha, |g|_{\mathcal{C}^\alpha} \le 1} \int g(x) d(\mu - \nu)(x).$$

We can now work to establish the main bound.

We start with decomposing $\mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - \widehat{F}_{j|v}(X)|^2] \le A + B$ with

$$A = \sum_{k \in \mathcal{K}_j} \mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - \widehat{F}_{j|v}(X)|^2 \mid X \in I_{j,k|v} \cap I_{j,k|\widehat{v}}]$$
$$B = \sum_{k \in \mathcal{K}_j} \mathbb{E}[|\widehat{F}_{j|\widehat{v}}(X) - \widehat{F}_{j|v}(X)|^2 \mid X \in I_{j,k|v} \setminus I_{j,k|\widehat{v}}]\rho(I_{j,k|v} \setminus I_{j,k|\widehat{v}}).$$

Let us first focus on B. Observe that, for all k' with $\rho(I_{j,k|v} \cap I_{j,k'|\widehat{v}}) > 0$, $|v^T(I_{j,k|v} \cup I_{j,k'|\widehat{v}})| \lesssim r(2^{-j} + \|\widehat{v} - v\|) \lesssim tr2^{-j}$, hence, using the Hölder continuity of F and the sub-Gaussian tail inequality [47, Proposition 5.10] (recall (E2)), we have

$$|\widehat{F}_{j|\widehat{v}}(X) - \widehat{F}_{j|v}(X)|^2 \lesssim t^2 |f|^2_{\mathcal{C}^\alpha} r^{2\alpha} 2^{-2j\alpha} + \frac{\varepsilon^2}{\#\mathcal{K}_j \rho(I_{j,k|v} \setminus I_{j,k|\widehat{v}})}$$

with probability higher than $1 - C\exp(-c\ n\varepsilon^2/\#\mathcal{K}_j\sigma^2)$. Thus, since $\rho(I_{j,k|v} \setminus I_{j,k|\widehat{v}}) \lesssim \|\widehat{v} - v\|$ and $\#\mathcal{K}_j = 2^j$, we obtain

$$B \lesssim \sum_{k \in \mathcal{K}_j} t^2 |f|^2_{\mathcal{C}^\alpha} r^{2\alpha} 2^{-2j\alpha} \rho(I_{j,k|v} \setminus I_{j,k|\widehat{v}}) + \varepsilon^2 \lesssim t^2 |f|^2_{\mathcal{C}^\alpha} r^{2\alpha} 2^{-2j\alpha} 2^j \|\widehat{v} - v\| + \varepsilon^2.$$

We now pass to A. For $x \in I_{j,k|v} \cap I_{j,k|\widehat{v}}$ we can write

$$\widehat{F}_{j|\widehat{v}}(x) - \widehat{F}_{j|v}(x) = \widehat{f}_{j,k|\widehat{v}}(\langle \widehat{v}, x\rangle) - \widehat{f}_{j,k|v}(v^T x),$$

where

$$\widehat{f}_{j,k|\widehat{v}}(t) = \widehat{b}_{j,k|\widehat{v}} + \widehat{m}_{j,k|\widehat{v}}t, \qquad \widehat{f}_{j,k|v}(t) = \widehat{b}_{j,k|v} + \widehat{m}_{j,k|v}t$$

are our local empirical estimators with respect to the estimated direction $\widehat{v}$ and the oracle direction v, respectively. Let us separate the constant and the linear components. Defining

$$\widehat{x}_{j,k|u} = \frac{1}{\#I_{j,k|u}} \sum_i X_i \mathbb{1}\{X_i \in I_{j,k|u}\}$$

we have

$$\begin{aligned}|\widehat{f}_{j,k|\widehat{v}}(\langle \widehat{v}, x\rangle) - \widehat{f}_{j,k|v}(v^T x)| \le |\widehat{b}_{j,k|\widehat{v}} - \widehat{b}_{j,k|v}| \\ + |\widehat{m}_{j,k|\widehat{v}}\widehat{v}^T(x - \widehat{x}_{j,k|\widehat{v}}) - \widehat{m}_{j,k|v}v^T(x - \widehat{x}_{j,k|v})|.\end{aligned}$$

We first approach the constant part: $|\widehat{b}_{j,k|\widehat{v}} - \widehat{b}_{j,k|v}| \le C_1 + C_2 + C_3$ with

$$\begin{aligned}C_1 &= \left|\frac{1}{\#I_{j,k|\widehat{v}}} \sum_i Y_i \mathbb{1}\{X_i \in I_{j,k|\widehat{v}}\} - \mathbb{E}[F(X) \mid X \in I_{j,k|\widehat{v}}]\right| \\ C_2 &= |\mathbb{E}[F(X) \mid X \in I_{j,k|\widehat{v}}] - \mathbb{E}[F(X) \mid X \in I_{j,k|v}]| \\ C_3 &= \left|\mathbb{E}[F(X) \mid X \in I_{j,k|v}] - \frac{1}{\#I_{j,k|v}} \sum_i Y_i \mathbb{1}\{X_i \in I_{j,k|v}\}\right|.\end{aligned}$$

C_1 can be further split by $C_1 \le C_{11} + C_{12}$ where

$$\begin{aligned}C_{11} &= \left|\frac{1}{\#I_{j,k|\widehat{v}}} \sum_i F(X_i) \mathbb{1}\{X_i \in I_{j,k|\widehat{v}}\} - \mathbb{E}[F(X) \mid X \in I_{j,k|\widehat{v}}]\right| \\ C_{12} &= \left|\frac{1}{\#I_{j,k|\widehat{v}}} \sum_i \zeta_i \mathbb{1}\{X_i \in I_{j,k|\widehat{v}}\}\right|.\end{aligned}$$

By the Bernstein inequality (exploiting (E1) and (E2)), we have

$$\begin{aligned}\mathbb{P}\{C_{11} > \varepsilon/\sqrt{\#\mathcal{K}_j \rho(I_{j,k|\widehat{v}})}\} &\le C \exp\left(-c \frac{n\rho(I_{j,k|\widehat{v}}) \frac{\varepsilon^2}{\#\mathcal{K}_j \rho(I_{j,k|\widehat{v}})}}{|f|^2_{\mathcal{C}^\alpha} r^{2\alpha} + |f|_{\mathcal{C}^\alpha} r^\alpha \frac{\varepsilon}{\sqrt{\#\mathcal{K}_j \rho(I_{j,k|\widehat{v}})}}}\right) \\ &= C \exp\left(-c \frac{n\varepsilon^2/\#\mathcal{K}_j}{|f|^2_{\mathcal{C}^\alpha} r^{2\alpha} + |f|_{\mathcal{C}^\alpha} r^\alpha \frac{\varepsilon}{\sqrt{\#\mathcal{K}_j \rho(I_{j,k|\widehat{v}})}}}\right) \\ &\le C \exp(-c\, n\varepsilon^2/\#\mathcal{K}_j |f|^2_{\mathcal{C}^\alpha} r^{2\alpha})\,.\end{aligned}$$

Also, $\mathbb{P}\{C_{12} > \varepsilon/\sqrt{\#\mathcal{K}_j \rho(I_{j,k|\widehat{v}})}\} \le C \exp(-c\, n\varepsilon^2/\#\mathcal{K}_j \sigma^2)$ by [47, Proposition 5.10]. C_3 can be concentrated in the same way as C_1. For C_2, in view of [34, Proposition 1.2.6] (bounding K_α in terms of W_1) and Kantorovich-Rubinstein duality for W_1, we have

$$C_2 = \left|\frac{1}{\rho(I_{j,k|\widehat{v}})} \int_{I_{j,k|\widehat{v}}} F(x) d\rho(x) - \frac{1}{\rho(I_{j,k|v})} \int_{I_{j,k|v}} F(x) d\rho(x)\right|$$

$$\begin{aligned}
&\leq |f|_{\mathcal{C}^\alpha} K_\alpha(\rho(\cdot \mid I_{j,k|\widehat{v}}), \rho(\cdot \mid I_{j,k|v})) \\
&\leq |f|_{\mathcal{C}^\alpha} \left(W_1(\rho(\cdot \mid I_{j,k|\widehat{v}}), \rho(\cdot \mid I_{j,k|v}))\right)^{\frac{1}{2-\alpha}} \\
&\lesssim |f|_{\mathcal{C}^\alpha} r^{\frac{1}{2-\alpha}} \|\widehat{v} - v\|^{\frac{1}{2-\alpha}},
\end{aligned}$$

where in the last inequality we have used assumption (Θ).

We now take care of the linear part, for which we can assume $\alpha = 1$. We have

$$\begin{aligned}
&|\widehat{m}_{j,k|\widehat{v}} \widehat{v}^T(x - \widehat{x}_{j,k|\widehat{v}}) - \widehat{m}_{j,k|v} v^T(x - \widehat{x}_{j,k|v})| \\
&\leq |\widehat{m}_{j,k|\widehat{v}}||\widehat{v}^T(x - \widehat{x}_{j,k|\widehat{v}}) - v^T(x - \widehat{x}_{j,k|v})| + |\widehat{m}_{j,k|\widehat{v}} - \widehat{m}_{j,k|v}||v^T(x - \widehat{x}_{j,k|v})| \\
&\lesssim |\widehat{m}_{j,k|\widehat{v}}| \min\{\|\widehat{x}_{j,k|\widehat{v}} - \widehat{x}_{j,k|v}\| + r\|\widehat{v} - v\|), r2^{-j}\} + |\widehat{m}_{j,k|\widehat{v}} - \widehat{m}_{j,k|v}| r 2^{-j}.
\end{aligned}$$

Defining

$$\begin{aligned}
\widehat{y}_{j,k|u} &= \frac{1}{\#I_{j,k|u}} \sum_i F(X_i) \mathbb{1}\{X_i \in I_{j,k|u}\} \\
\widehat{\mathrm{Cov}}_{j,k|u} &= \frac{1}{\#I_{j,k|u}} \sum_i u^T(X_i - \widehat{x}_{j,k|u}) \left(F(X_i) - \widehat{y}_{j,k|u}\right) \mathbb{1}\{X_i \in I_{j,k|u}\} \\
\widehat{\mathrm{Var}}_{j,k|u} &= \frac{1}{\#I_{j,k|u}} \sum_i |u^T(X_i - \widehat{x}_{j,k|u})|^2 \mathbb{1}\{X_i \in I_{j,k|u}\},
\end{aligned}$$

we have

$$|\widehat{m}_{j,k|\widehat{v}}| \leq \widehat{\mathrm{Var}}^{-1}_{j,k|\widehat{v}} \left(|\widehat{\mathrm{Cov}}_{j,k|\widehat{v}}| + \left| \frac{1}{\#I_{j,k|\widehat{v}}} \sum_i \widehat{v}^T(X_i - \widehat{x}_{j,k|\widehat{v}}) \zeta_i \mathbb{1}\{X_i \in I_{j,k|\widehat{v}}\} \right| \right),$$

$$\begin{aligned}
|\widehat{m}_{j,k|\widehat{v}} - \widehat{m}_{j,k|v}| \leq\ & \widehat{\mathrm{Var}}^{-1}_{j,k|\widehat{v}} |\widehat{\mathrm{Cov}}_{j,k|\widehat{v}} - \widehat{\mathrm{Cov}}_{j,k|v}| \\
&+ \widehat{\mathrm{Var}}^{-1}_{j,k|\widehat{v}} \widehat{\mathrm{Var}}^{-1}_{j,k|v} |\widehat{\mathrm{Var}}_{j,k|\widehat{v}} - \widehat{\mathrm{Var}}_{j,k|v}||\widehat{\mathrm{Cov}}_{j,k|v}| \\
&+ \widehat{\mathrm{Var}}^{-1}_{j,k|\widehat{v}} \left| \frac{1}{\#I_{j,k|\widehat{v}}} \sum_i \widehat{v}^T(X_i - \widehat{x}_{j,k|\widehat{v}}) \zeta_i \mathbb{1}\{X_i \in I_{j,k|\widehat{v}}\} \right| \\
&+ \widehat{\mathrm{Var}}^{-1}_{j,k|v} \left| \frac{1}{\#I_{j,k|v}} \sum_i v^T(X_i - \widehat{x}_{j,k|v}) \zeta_i \mathbb{1}\{X_i \in I_{j,k|v}\} \right|.
\end{aligned}$$

Note that $|\widehat{\mathrm{Var}}_{j,k|u}| \leq r^2 2^{-2j}$, $|\widehat{\mathrm{Cov}}_{j,k|v}| \leq |f|_{\mathcal{C}^1} r^2 2^{-2j}$ and $|\widehat{\mathrm{Cov}}_{j,k|\widehat{v}}| \leq t|f|_{\mathcal{C}^1} r^2 2^{-2j}$. Introducing

$$\begin{aligned}
\overline{x}_{j,k|u} &= \mathbb{E}[X \mid X \in I_{j,k|u}] \\
\mathrm{Var}_{j,k|u} &= \mathrm{Var}[\langle v, X\rangle \mid X \in I_{j,k|u}] \\
\widetilde{\mathrm{Var}}_{j,k|u} &= \frac{1}{\#I_{j,k|u}} \sum_i |v^T(X_i - \overline{x}_{j,k|u})|^2 \mathbb{1}\{X_i \in I_{j,k|u}\},
\end{aligned}$$

we get

$$\begin{aligned}|\widehat{\mathrm{Var}}_{j,k|u} - \mathrm{Var}_{j,k|u}| &= |\widetilde{\mathrm{Var}}_{j,k|u} - \mathrm{Var}_{j,k|u} - |u^T(\widehat{x}_{j,k|u} - \overline{x}_{j,k|u})|^2| \\ &\le |\widetilde{\mathrm{Var}}_{j,k|u} - \mathrm{Var}_{j,k|u}| + r^2 2^{-2j}.\end{aligned}$$

The Bernstein inequality (together with (E1) and (E2)) yields

$$\mathbb{P}\{|\widetilde{\mathrm{Var}}_{j,k|u} - \mathrm{Var}_{j,k|u}| \gtrsim r^2 2^{-2j}\} \le C\exp(-c\, n\rho(I_{j,k|v})) \le C\exp(-c\, n\varepsilon^2/\#\mathcal{K}_j|f|_{\mathcal{C}^1}^2 r^2).$$

Thus $|\widehat{\mathrm{Var}}_{j,k|u} - \mathrm{Var}_{j,k|u}| \lesssim r^2 2^{-2j}$, and hence $\widehat{\mathrm{Var}}_{j,k|u} \gtrsim r^2 2^{-2j}$, with probability higher than $1 - C\exp(-c\, n\varepsilon^2/\#\mathcal{K}_j|f|_{\mathcal{C}^1}^2 r^2)$. Moreover, thanks to [6, Theorem 3.1],

$$\left|\frac{1}{\#I_{j,k|u}}\sum_i u^T(X_i - \widehat{x}_{j,k|u})\zeta_i \mathbb{1}\{X_i \in I_{j,k|u}\}\right| \lesssim r2^{-j}\frac{\varepsilon}{\sqrt{\#\mathcal{K}_j\rho(I_{j,k|u})}}$$

with probability higher than $1 - C\exp(-c\, n\varepsilon^2/\#\mathcal{K}_j\sigma^2)$. Therefore,

$$\begin{aligned}&|\widehat{m}_{j,k|\widehat{v}}\widehat{v}^T(x - \widehat{x}_{j,k|\widehat{v}}) - \widehat{m}_{j,k|v}v^T(x - \widehat{x}_{j,k|v})| \\ &\lesssim t|f|_{\mathcal{C}^1}(\|\widehat{x}_{j,k|\widehat{v}} - \widehat{x}_{j,k|v}\| + r\|\widehat{v} - v\|) \\ &+ r^{-1}2^j\left(|\widehat{\mathrm{Cov}}_{j,k|\widehat{v}} - \widehat{\mathrm{Cov}}_{j,k|v}| + |f|_{\mathcal{C}^1}|\widehat{\mathrm{Var}}_{j,k|\widehat{v}} - \widehat{\mathrm{Var}}_{j,k|v}|\right) \\ &+ \frac{\varepsilon}{\sqrt{\#\mathcal{K}_j\rho(I_{j,k|v}) \wedge \rho(I_{j,k|\widehat{v}})}}\end{aligned}$$

with probability higher than $1 - C\exp(-c\, n\varepsilon^2/\#\mathcal{K}_j|f|_{\mathcal{C}^1}^2 r^2\})$. Now,

$$|\widehat{x}_{j,k|\widehat{v}} - \widehat{x}_{j,k|v}| \le \|\widehat{x}_{j,k|\widehat{v}} - \overline{x}_{j,k|\widehat{v}}\| + \|\overline{x}_{j,k|\widehat{v}} - \overline{x}_{j,k|v}\| + \|\overline{x}_{j,k|v} - \widehat{x}_{j,k|v}\|.$$

The middle term is bounded by

$$\|\overline{x}_{j,k|\widehat{v}} - \overline{x}_{j,k|v}\| \le W_1(\rho(\cdot \mid I_{j,k|\widehat{v}}), \rho(\cdot \mid I_{j,k|v})) \lesssim r\|\widehat{v} - v\|,$$

thanks to assumption (Θ). For the first and third terms, applying the Bernstein inequality (and (E1),(E2)) we get

$$\|\widehat{x}_{j,k|u} - \overline{x}_{j,k|u}\| \le t^{-1}|f|_{\mathcal{C}^1}^{-1}\frac{\varepsilon}{\sqrt{\#\mathcal{K}_j\rho(I_{j,k|u})}}$$

with probability higher than $1 - C\exp\left(-c\, n\varepsilon^2/\#\mathcal{K}_j t^2|f|_{\mathcal{C}^1}^2 r^2\right)$.

We are now left to estimate $|\widehat{\mathrm{Cov}}_{j,k|\widehat{v}} - \widehat{\mathrm{Cov}}_{j,k|v}|$ and $|\widehat{\mathrm{Var}}_{j,k|\widehat{v}} - \widehat{\mathrm{Var}}_{j,k|v}|$. First, we break down $|\widehat{\mathrm{Cov}}_{j,k|\widehat{v}} - \widehat{\mathrm{Cov}}_{j,k|v}| \le \sum_{a=1}^8 T_a$ where, defining

$$\overline{y}_{j,k|u} = \mathbb{E}[F(X) \mid X \in I_{j,k|u}],$$

$$T_1 = \left|\frac{1}{\#I_{j,k|\widehat{v}}}\sum_i \langle \widehat{v}, \overline{x}_{j,k|v} - \widehat{x}_{j,k|\widehat{v}}\rangle(F(X_i) - \widehat{y}_{j,k|\widehat{v}})\mathbb{1}\{X_i \in I_{j,k|\widehat{v}}\}\right|$$

$$
\begin{aligned}
T_2 =& \left| \frac{1}{\#I_{j,k|\widehat{v}}} \sum_i \langle \widehat{v} - v, X_i - \overline{x}_{j,k|v} \rangle (F(X_i) - \widehat{y}_{j,k|\widehat{v}}) \mathbb{1}\{X_i \in I_{j,k|\widehat{v}}\} \right| \\
T_3 =& \left| \frac{1}{\#I_{j,k|\widehat{v}}} \sum_i \langle v, X_i - \overline{x}_{j,k|v} \rangle (\overline{y}_{j,k|v} - \widehat{y}_{j,k|\widehat{v}}) \mathbb{1}\{X_i \in I_{j,k|\widehat{v}}\} \right| \\
T_4 =& \left| \frac{1}{\#I_{j,k|\widehat{v}}} \sum_i \langle v, X_i - \overline{x}_{j,k|v} \rangle (F(X_i) - \overline{y}_{j,k|v}) \mathbb{1}\{X_i \in I_{j,k|\widehat{v}}\} \right. \\
& \left. - \mathbb{E}[\langle v, X - \overline{x}_{j,k|v} \rangle (F(X) - \overline{y}_{j,k|v}) \mid X \in I_{j,k|\widehat{v}}] \right| \\
T_5 =& |\mathbb{E}[\langle v, X - \overline{x}_{j,k|v} \rangle (F(X) - \overline{y}_{j,k|v}) \mid X \in I_{j,k|\widehat{v}}] \\
& - \mathbb{E}[\langle v, X - \overline{x}_{j,k|v} \rangle (F(X) - \overline{y}_{j,k|v}) \mid X \in I_{j,k|v}]| \\
T_6 =& \left| \mathbb{E}[\langle v, X - \overline{x}_{j,k|v} \rangle (F(X) - \overline{y}_{j,k|v}) \mid X \in I_{j,k|v}] \right. \\
& \left. - \frac{1}{\#I_{j,k|v}} \sum_i \langle v, X_i - \overline{x}_{j,k|v} \rangle (F(X_i) - \overline{y}_{j,k|v}) \mathbb{1}\{X_i \in I_{j,k|v}\} \right| \\
T_7 =& \left| \frac{1}{\#I_{j,k|v}} \sum_i \langle v, \widehat{x}_{j,k|v} - \overline{x}_{j,k|v} \rangle (F(X_i) - \overline{y}_{j,k|v}) \mathbb{1}\{X_i \in I_{j,k|v}\} \right| \\
T_8 =& \left| \frac{1}{\#I_{j,k|v}} \sum_i \langle v, X_i - \widehat{x}_{j,k|v} \rangle (\widehat{y}_{j,k|v} - \overline{y}_{j,k|v}) \mathbb{1}\{X_i \in I_{j,k|v}\} \right|.
\end{aligned}
$$

We bound the terms T_i's as follows.

$\boldsymbol{T}_1$.

$$
T_1 \le \|\overline{x}_{j,k|v} - \widehat{x}_{j,k|\widehat{v}}\| \frac{1}{\#I_{j,k|\widehat{v}}} \sum_i |F(X_i) - \widehat{y}_{j,k|\widehat{v}}| \mathbb{1}\{X_i \in I_{j,k|\widehat{v}}\}
$$

with

$$
|F(X_i) - \widehat{y}_{j,k|\widehat{v}}| \lesssim |f|_{\mathcal{C}^1} r (2^{-j} + \|\widehat{v} - v\|) \lesssim t |f|_{\mathcal{C}^1} r 2^{-j}.
$$

Hence

$$
r^{-1} 2^j T_1 \le t |f|_{\mathcal{C}^1} \|\widehat{x}_{j,k|\widehat{v}} - \overline{x}_{j,k|v}\| \le t |f|_{\mathcal{C}^1} (\|\widehat{x}_{j,k|\widehat{v}} - \overline{x}_{j,k|\widehat{v}}\| + \|\overline{x}_{j,k|\widehat{v}} - \overline{x}_{j,k|v}\|),
$$

where

$$
\mathbb{P}\{\|\widehat{x}_{j,k|\widehat{v}} - \overline{x}_{j,k|\widehat{v}}\| > t^{-1} |f|_{\mathcal{C}^1}^{-1} \tfrac{\varepsilon}{\sqrt{\#\mathcal{K}_j \rho(I_{j,k|\widehat{v}})}}\} \le C \exp\left(-c\, n\varepsilon^2 / \#\mathcal{K}_j t^2 |f|_{\mathcal{C}^1}^2 r^2\right)
$$

by the Bernstein inequality, and

$$
\|\overline{x}_{j,k|\widehat{v}} - \overline{x}_{j,k|v}\| \le W_1(\rho(\cdot \mid I_{j,k|\widehat{v}}), \rho(\cdot \mid I_{j,k|v})) \lesssim r \|\widehat{v} - v\|
$$

by assumption (Θ).

$\boldsymbol{T}_2$.

$$
T_2 \lesssim \|\widehat{v} - v\| r |f|_{\mathcal{C}^1} r (2^{-j} + \|\widehat{v} - v\|) \lesssim 2^{-j} t |f|_{\mathcal{C}^1} r^2 \|\widehat{v} - v\|,
$$

hence

$$r^{-1}2^j T_2 \le t|f|_{\mathcal{C}^1} r\|\widehat{v} - v\|.$$

$\boldsymbol{T}_3$.

$$T_3 \le \frac{1}{\#I_{j,k|\widehat{v}}} \sum_i |\langle v, X_i - \overline{x}_{j,k|v}\rangle||\overline{y}_{j,k|v} - \widehat{y}_{j,k|\widehat{v}}|\mathbb{1}\{X_i \in I_{j,k|\widehat{v}}\}$$

with $|\langle v, X_i - \overline{x}_{j,k|v}\rangle| \lesssim r(2^{-j} + \|\widehat{v} - v\|) \lesssim tr2^{-j}$. Hence

$$r^{-1}2^j T_3 \le t|\widehat{y}_{j,k|\widehat{v}} - \overline{y}_{j,k|v}| \le t(|\widehat{y}_{j,k|\widehat{v}} - \overline{y}_{j,k|\widehat{v}}| + |\overline{y}_{j,k|\widehat{v}} - \overline{y}_{j,k|v}|),$$

where

$$\mathbb{P}\{|\widehat{y}_{j,k|\widehat{v}} - \overline{y}_{j,k|\widehat{v}}| > t^{-1}\tfrac{\varepsilon}{\sqrt{\#\mathcal{K}_j \rho(I_{j,k|\widehat{v}})}}\} \le C\exp\left(-c\, n\varepsilon^2/\#\mathcal{K}_j t^2 |f|^2_{\mathcal{C}^1} r^2\right)$$

by the Bernstein inequality, and

$$|\overline{y}_{j,k|\widehat{v}} - \overline{y}_{j,k|v}| \le |f|_{\mathcal{C}^1} W_1(\rho(\cdot \mid I_{j,k|\widehat{v}}), \rho(\cdot \mid I_{j,k|v})) \lesssim |f|_{\mathcal{C}^1} r\|\widehat{v} - v\|$$

by assumption (Θ).

$\boldsymbol{T}_4$. We apply the Bernstein inequality. Since

$$|v^T(X_i - \overline{x}_{j,k|v})(F(X_i) - \overline{y}_{j,k|v})| \lesssim |f|_{\mathcal{C}^1} r^2(2^{-j} + \|\widehat{v} - v\|) \lesssim t|f|_{\mathcal{C}^1} r^2 2^{-j},$$

we obtain

$$\mathbb{P}\{r^{-1}2^j T_4 > \tfrac{\varepsilon}{\sqrt{\#\mathcal{K}_j \rho(I_{j,k|\widehat{v}})}}\} \le C\exp\left(-c\, n\varepsilon^2/\#\mathcal{K}_j t^2 |f|^2_{\mathcal{C}^1} r^2\right).$$

$\boldsymbol{T}_5$.

$$T_5 = \left| \int G(x)(d\rho(x \mid I_{j,k|\widehat{v}}) - d\rho(x \mid I_{j,k|v})) \right|$$

where $G(x) = v^T(x - \overline{x}_{j,k|v})(F(x) - \overline{y}_{j,k|v})$, $x \in I_{j,k|v} \cup I_{j,k|\widehat{v}}$, is Lipschitz of constant $\lesssim t|f|_{\mathcal{C}^1} 2^{-j} r$:

$$\begin{aligned} |G(x) - G(z)| &\le |v^T(x - z)||F(x) - \overline{y}_{j,k|v}| + |v^T(z - \overline{x}_{j,k|v})||F(x) - F(z)| \\ &\lesssim |f|_{\mathcal{C}^1} r(2^{-j} + \|\widehat{v} - v\|)\|x - z\| \\ &\lesssim t|f|_{\mathcal{C}^1} r2^{-j}\|x - z\|. \end{aligned}$$

Thus, by assumption (Θ),

$$r^{-1}2^j T_5 \lesssim t|f|_{\mathcal{C}^1} W_1(\rho(\cdot \mid I_{j,k|\widehat{v}}) - \rho(\cdot \mid I_{j,k|v})) \lesssim t|f|_{\mathcal{C}^1} r\|\widehat{v} - v\|.$$

$\boldsymbol{T}_6$. As for T_4.

$\boldsymbol{T}_7$.

$$T_7 \le \|\widehat{x}_{j,k|v} - \overline{x}_{j,k|v}\||f|_{\mathcal{C}^1} r2^{-j},$$

where, by the Bernstein inequality,

$$\mathbb{P}\{\|\widehat{x}_{j,k|v} - \overline{x}_{j,k|v}\| > |f|_{\mathcal{C}^1}^{-1} \tfrac{\varepsilon}{\sqrt{\#\mathcal{K}_j \rho(I_{j,k|v})}}\} \leq C \exp\left(-c\, n\varepsilon^2/\#\mathcal{K}_j |f|_{\mathcal{C}^1}^2 r^2\right).$$

$\boldsymbol{T}_8$.

$$T_8 \leq r 2^{-j} |\widehat{y}_{j,k|v} - \overline{y}_{j,k|v}|,$$

where, by the Bernstein inequality,

$$\mathbb{P}\{|\widehat{y}_{j,k|v} - \overline{y}_{j,k|v}| > \tfrac{\varepsilon}{\sqrt{\#\mathcal{K}_j \rho(I_{j,k|v})}}\} \leq C \exp\left(-c\, n\varepsilon^2/\#\mathcal{K}_j |f|_{\mathcal{C}^1}^2 r^2\right).$$

The quantity $|\widehat{\mathrm{Var}}_{j,k|\widehat{v}} - \widehat{\mathrm{Var}}_{j,k|v}|$ can be estimated through an analogous decomposition. We can finally put all the terms together, and taking the union bound over the $\mathcal{K}_j$'s completes the proof. □

6. Proofs of technical results

In our proofs, we make use of the following Lemma to ensure that we have enough local samples, or to concentrate the empirical measure on the underlying distribution.

Lemma 1. *Let X be a random variable, and let $X_1, \ldots, X_n$ be independent copies of X. Given a measurable set E, define $\rho(E) = \mathbb{P}\{X \in E\}$ and $\widehat{\rho}(E) = n^{-1} \sum_i \mathbb{1}\{X_i \in E\}$. Then*

$$\mathbb{P}\{|\widehat{\rho}(E) - \rho(E)| > t\} \leq 2 \exp\left(-\frac{nt^2/2}{\rho(E) + t/3}\right).$$

In particular, for $t = \rho(E)/2$ we have

$$\mathbb{P}\left\{\widehat{\rho}(E) \notin \left[\frac{1}{2}\rho(E), \frac{3}{2}\rho(E)\right]\right\} \leq \mathbb{P}\left\{|\widehat{\rho}(E) - \rho(E)| > \frac{1}{2}\rho(E)\right\} \leq 2 \exp\left(-\tfrac{3}{28} n\rho(E)\right).$$

Proof. The bound follows by a direct application of the Bernstein inequality to the random variables $\mathbb{1}\{X_i \in E\}$. □

When working with possibly unbounded distributions, we need some control on their tails. A common choice is to assume sub-Gaussian decay. We recall that a random variable X is sub-Gaussian of variance proxy R^2 if

$$\mathbb{P}\{|X| > t\} \leq 2 \exp\left(-\frac{t^2}{2R^2}\right).$$

A random vector $X \in \mathbb{R}^d$ is sub-Gaussian if $\langle u, X \rangle$ is sub-Gaussian for every $u \in \mathbb{S}^{d-1}$. In particular, bounded and normal distributions are sub-Gaussian.

Lemma 2. *Let $X \in \mathbb{R}^d$ be a sub-Gaussian vector with variance proxy R^2. Then*

$$\mathbb{P}\{\|X\| > t\} \le 2\exp\left(-\frac{t^2}{2dR^2}\right).$$

Proof. Let X_k be the k-th coordinate of X. Then

$$\mathbb{E}\left[\exp\left(\frac{\|X\|^2}{2dR^2}\right)\right] = \mathbb{E}\left[\prod_{k=1}^{d}\exp\left(\frac{|X_k|^2}{2dR^2}\right)\right] \le \left(\prod_{k=1}^{d}\mathbb{E}\left[\exp\left(\frac{|X_k|^2}{2R^2}\right)\right]\right)^{1/d} \le 2\,.$$

The result follows from [48, Proposition 2.5.2]. □

The lemma below shows that most samples from a d-dimensional sub-Gaussian distribution of variance proxy R^2 fall into a ball of radius $\sqrt{d}R$.

Lemma 3. *Let $X_1, \dots, X_n$ be independent copies of a sub-Gaussian vector $X \in \mathbb{R}^d$ with variance proxy R^2. Then, for every $\alpha \ge 2$ and $\beta \in (0,1)$,*

$$\mathbb{P}\left\{\#B(0, \sqrt{2d\log(2\alpha)}R)) < \left(1-\tfrac{1}{\alpha}\right)\beta n\right\} \le 2e^{-\left(1-\frac{1}{\alpha}\right)\frac{(1-\beta)^2/2}{1+(1-\beta)/3}n}.$$

Proof. Let $B = B(0, \sqrt{2d\log(2\alpha)}R))$ and $\rho(B) = \mathbb{P}\{X \in B\}$. Lemma 2 gives

$$\rho(B) \ge 1 - 2\exp(-\log(2\alpha)) = \left(1-\tfrac{1}{\alpha}\right);$$

an application of Lemma 1 with $t = (1-\beta)\rho(B)$ yields

$$\begin{aligned}&\mathbb{P}\left\{\#B < \left(1-\tfrac{1}{\alpha}\right)\beta n\right\} \le \mathbb{P}\left\{\#B < \beta\rho(B)n\right\}\\ &\le 2\exp\left(-\tfrac{(1-\beta)^2/2}{1+(1-\beta)/3}\rho(B)n\right) \le 2\exp\left(-\left(1-\tfrac{1}{\alpha}\right)\tfrac{(1-\beta)^2/2}{1+(1-\beta)/3}n\right).\end{aligned}$$

□

We often carry out the following integration to obtain expectation bounds from bounds in probability.

Lemma 4. *Let X be a random variable. Suppose there are $p \in [1,2]$, $a \ge e$ and $b > 0$ such that $\mathbb{P}\{|X| > \varepsilon\} \le ae^{-b\varepsilon^{2p}}$ for every $\varepsilon > 0$. Then $\mathbb{E}|X|^2 \le \left(\frac{\log a}{b}\right)^{1/p}$.*

Proof. Integrating over $\varepsilon > 0$ we get $\mathbb{E}|X|^2 \le \int_0^{\varepsilon_0}\varepsilon\, d\varepsilon + \int_{\varepsilon_0}^{\infty} ae^{-b\varepsilon^{2p}}\varepsilon d\varepsilon$ with $\varepsilon_0 = (\log a/b)^{1/2p}$. The first integral is equal to $\frac{1}{2}(\log a/b)^{1/p}$, while the substitution $b\varepsilon^{2p} \to \varepsilon$ in the second integral gives

$$\frac{a}{2p}\int_{\log a}^{\infty}\varepsilon^{1/p-1}e^{-\varepsilon}d\varepsilon\left(\frac{1}{b}\right)^{1/p} \le \frac{a}{2}\int_{\log a}^{\infty}e^{-\varepsilon}d\varepsilon\left(\frac{1}{b}\right)^{1/p} = \frac{1}{2}\left(\frac{1}{b}\right)^{1/p}.$$

□

The following Lemma describes the expected decay of a Hölder function of a sub-Gaussian vector outside a large ball. This result is useful to bound the mean squared error on the tails of a sub-Gaussian distribution, where the strong decay of the measure can offset poor pointwise predictions.

Lemma 5. *Let X be a sub-Gaussian vector in $\mathbb{R}^d$ with variance proxy R^2, and let $F : \mathbb{R}^d \to \mathbb{R}$ be a $\mathcal{C}^\alpha$ Hölder continuous function with $\alpha \in (0, 1]$. Then, for every $r \geq 1$,*

$$\mathbb{E}[|F(X)|^2] \lesssim \mathbb{E}[|F(X)|^2 \mid X \in B(0, r)] + (|F(0)|^2 + d|f|^2_{\mathcal{C}^\alpha} R^2) \exp(-r^2/2dR^2).$$

Proof. We split the left hand side into

$$\mathbb{E}[|F(X)|^2 \mid X \in B(0, r)] + \mathbb{E}[|F(X)|^2 \mathbb{1}\{X \notin B(0, r)\}]$$

and bound the second term. The Hölder continuity of F entails

$$\mathbb{E}[|F(X)|^2 \mathbb{1}\{X \notin B(0, r)\}] \lesssim |F(0)|^2 \mathbb{P}\{\|X\| > r\} + |f|^2_{\mathcal{C}^\alpha} \mathbb{E}[\|X\|^2 \mathbb{1}\{X \notin B(0, r)\}].$$

Using Lemma 2 we get $\mathbb{P}\{\|X\| > r\} \leq 2 \exp(-r^2/2dR^2)$ and

$$\begin{aligned}
\mathbb{E}[\|X\|^2 \mathbb{1}\{X \notin B(0, r)\}] &= \int_0^\infty \mathbb{P}\{\|X\|^2 \mathbb{1}\{X \notin B(0, r)\} > t\} dt \\
&= \int_{r^2}^\infty \mathbb{P}\{\|X\|^2 \mathbb{1}\{X \notin B(0, r)\} > t\} dt \leq \int_{r^2}^\infty \mathbb{P}\{\|X\|^2 > t\} dt \\
&= 2 \int_r^\infty \mathbb{P}\{\|X\| > t\} t dt \leq 4 \int_r^\infty \exp(-t^2/2dR^2) t dt \\
&= 4dR^2 \exp(-r^2/2dR^2).
\end{aligned}$$

□

Acknowledgements. This research was partially supported by AFOSR FA9550-17-1-0280, NSF-DMS-1821211, NSF-ATD-1737984. SV thanks Timo Klock for the discussion and the useful exchange of views about this and related problems.

References

[1] F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journ. of Mach. Learn. Res.*, 19(18):1–53, 2017.

[2] R. Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, 1997.

[3] P. J. Bickel and B. Li. Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series*, 54:177–186, 2007.

[4] P. Binev, A. Cohen, W. Dahmen, and R. A. DeVore. Universal Algorithms for Learning Theory. Part II: Piecewise Polynomial Functions. *Constructive Approximation*, 26(2):127–152, 2007.

[5] P. Binev, A. Cohen, W. Dahmen, R. A. DeVore, and V. N. Temlyakov. Universal Algorithms for Learning Theory Part I: Piecewise Constant Functions. *Journal of Machine Learning Research*, 6(1):1297–1321, 2005.

[6] V. Buldygin and E. Pechuk. Inequalities for the distributions of functionals of sub-gaussian vectors. *Theory of Probability and Mathematical Statistics*, 80:25–36, 2010.

[7] S. Cambanis, S. Huang, and G. Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385, 1981.

[8] R. J. Carroll, J. Fan, I. Gijbels, and M. P. Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997.

[9] R. J. Carroll, D. Ruppert, and A. H. Welsh. Local estimating equations. *Journal of the American Statistical Association*, 93(441):214–227, 1998.

[10] R. D. Cook. Save: a method for dimension reduction and graphics in regression. *Communications in Statistics - Theory and Methods*, 29(9-10):2109–2121, 2000.

[11] R. D. Cook and H. Lee. Dimension reduction in binary response regression. *Journal of the American Statistical Association*, 94(448):1187–1200, 1999.

[12] R. Coudret, B. Liquet, and J. Saracco. Comparison of sliced inverse regression approaches for underdetermined cases. *J. SFdS*, 155(2):72–96, 2014.

[13] X. Cui, W. K. Härdle, and L. Zhu. The EFM approach for single-index models. *Ann. Statist.*, 39(3):1658–1688, 06 2011.

[14] M. Delecroix, W. Härdle, and M. Hristache. Efficient estimation in single-index regression. Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes. SFB 373 Discussion Paper 37, Humboldt University of Berlin, 1997.

[15] M. Delecroix and M. Hristache. M-estimateurs semi-paramtriques dans les modles direction rvlatrice unique. *Bull. Belg. Math. Soc. Simon Stevin*, 6(2):161–185, 1999.

[16] M. Delecroix, M. Hristache, and V. Patilea. On semiparametric M-estimation in single-index regression. *Journal of Statistical Planning and Inference*, 136(3):730–769, 2006.

[17] P. Diaconis and D. Freedman. Asymptotics of graphical projection pursuit. *Ann. Statist.*, 12(3):793–815, 09 1984.

[18] N. Duan and K.-C. Li. Slicing regression: a link-free regression method. *Ann. Statist.*, 19(2):505–530, 1991.

[19] M. L. Eaton. A characterization of spherical distributions. *Journal of Multivariate Analysis*, 20(2):272–276, 1986.

[20] S. Gaïffas and G. Lecué. Optimal rates and adaptation in the single-index model using aggregation. *Electron. J. Statist.*, 1:538–573, 2007.

[21] R. Ganti, N. Rao, R. M. Willett, and R. Nowak. Learning single index models in high dimensions. arXiv:1506.08910, 2015.

[22] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer, 2002.

[23] P. Hall and K.-C. Li. On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.*, 21(2):867–889, 06 1993.

[24] W. Härdle, P. Hall, and H. Ichimura. Optimal smoothing in single-index models. *Ann. Statist.*, 21(1):157–178, 03 1993.

[25] W. Härdle and T. M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408):986–995, 1989.

[26] J. L. Horowitz. *Semiparametric Methods in Econometrics*. Lecture Notes in Statistics. Springer New York, 1998.

[27] M. Hristache, A. Juditsky, and V. Spokoiny. Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, 29(3):593–623, 06 2001.

[28] H. Ichimura. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58(1):71–120, 1993.

[29] S. M. Kakade, V. Kanade, O. Shamir, and A. T. Kalai. Efficient learning of generalized linear and single index models with isotonic regression. *Advances in Neural Information Processing Systems 24*, pages 927–935, 2011.

[30] A. T. Kalai and R. Sastry. The isotron algorithm: High-dimensional isotonic regression. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

[31] D. Kelker. Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 32(4):419–430, 1970.

[32] S. Kpotufe. k-nn regression adapts to local intrinsic dimension. *Advances in Neural Information Processing Systems 24*, pages 729–737, 2011.

[33] S. Kpotufe and V. Garg. Adaptivity to local smoothness and dimension in kernel regression. *Advances in Neural Information Processing Systems 26*, pages 3075–3083, 2013.

[34] S. Kuksin and A. Shirikyan. *Mathematics of Two-Dimensional Turbulence*. Cambridge Tracts in Mathematics. Cambridge University Press, 2012.

[35] B. Li. *Sufficient Dimension Reduction: Methods and Applications with R*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press, 2018.

[36] B. Li and S. Wang. On Directional Regression for Dimension Reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007.

[37] B. Li, H. Zha, and F. Chiaromonte. Contour regression: A general approach to dimension reduction. *The Annals of Statistics*, 33(4):1580–1616, 2005.

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

[39] W. Liao, M. Maggioni, and S. Vigogna. Learning adaptive multiscale approximations to data and functions near low-dimensional sets. In *2016 IEEE Information Theory Workshop (ITW)*, pages 226–230. IEEE, 2016.

[40] W. Liao, M. Maggioni, and S. Vigogna. Multiscale regression on unknown manifolds. *ArXiv e-prints*, 2020.

[41] H. Liu and W. Liao. Learning functions varying along an active subspace. arXiv:2001.07883, 2020.

[42] Y. Pitcan. A Note on Concentration Inequalities for U-Statistics. arXiv:1712.06160, 2017.

[43] R. J. Samworth, T. Wang, and Y. Yu. A useful variant of the DavisKahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2014.

[44] T. M. Stoker. Consistent estimation of scaled coefficient. *Econometrica*, pages 1461–1481, 1986.

[45] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 12 1982.

[46] A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996.

[47] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing*, pages 210–268. Cambridge University Press, 2012.

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

[49] Y. Xia. Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, 22(6):11121137, 2006.

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