Nonlinear generalization of the single index model

Željko Kereta*, Timo Klock† and Valeriya Naumova‡

1 Simula Research Laboratory, Machine Intelligence Department, Oslo, Norway
2 SimulaMet, Machine Intelligence Department, Oslo, Norway

Abstract

Single index model is a powerful yet simple model, widely used in statistics, machine learning, and other scientific fields. It models the regression function as \( g(a, x) \), where \( a \) is an unknown index vector and \( x \) are the features. This paper deals with a nonlinear generalization of this framework to allow for a regressor that uses multiple index vectors, adapting to local changes in the responses. To do so we exploit the conditional distribution over function-driven partitions, and use linear regression to locally estimate index vectors. We then regress by applying a kNN type estimator that uses a localized proxy of the geodesic metric. We present theoretical guarantees for estimation of local index vectors and out-of-sample prediction, and demonstrate the performance of our method with experiments on synthetic and real-world data sets, comparing it with state-of-the-art methods.

Keywords: high-dimensional regression, dimension reduction, single index model, nonparametric regression, nonlinear methods

1 Introduction

Many problems in data analysis can be formulated as learning a function from a given data set in a high-dimensional space. Due to the curse of dimensionality, accurate regression on high-dimensional functions typically requires a number of samples that scales exponentially with the ambient dimension \[ \text{Sto82} \]. A common approach to mitigating these effects is to impose structural assumptions on the data. Indeed, a number of recent advances in data analysis and numerical simulation are based on the observation that high-dimensional, real-world data is inherently structured, and that the relationship between the features and the responses is of a lower dimensional nature \[ \text{AC09} \].

The most direct such model, which has become an important prior for many statistical and machine learning problems, assumes a 1-dimensional relationship, written as

\[
f(X) = g(\langle a, X \rangle),
\]

where the features \( X \in \mathbb{R}^D \) and responses \( Y \in \mathbb{R} \) are related through an unknown index vector \( a \in \mathbb{R}^D \), and an unknown monotonic function \( g \). The model (1), called the single index model (SIM), first appeared in economic and statistical communities in the early 90s \[ \text{Ich93, HH96} \]. Moreover, it provides a basis for more complex models such as multi-index models \[ \text{HTFF05, LŽC+05} \] and neural networks \[ \text{LBH15} \]. In particular, even though the monotonicity

*Email: zeljko@simula.no
†Email: timo@simula.no
‡Email: valeriya@simula.no
assumption on $g$ might seem restrictive, note that most neural networks use monotonic activation functions, such as ReLU or the sigmoid.

An assumption commonly shared by SIM and its generalizations is that there is a single lower dimensional linear space that accounts for the complexity in relating $X$ and $Y$. While simple, this assumption is only a first level approximation and is rarely observed in real-world regression problems. The goal of this paper is to relax the global linearity assumption present in model (1) in order to locally adapt to changes in the relationship between $X$ and $Y$. Specifically, we propose the nonlinear single index model (NSIM), defined by

$$f(X) = g(\pi_\gamma(X)),$$

where $\pi_\gamma$ represents the orthogonal projection a simple $C^3$-curve $\text{Im}(\gamma)$, defined by

$$\pi_\gamma(x) \in \arg\min_{z \in \text{Im}(\gamma)} \|x - z\|,$$

and $g$ is a bi-Lipschitz function. Since bi-Lipschitz functions are also monotonic, SIM is a special example of (2), where $\gamma(t) = at$.

Before formally describing our assumptions and detailing the approach, let us begin with a couple of comments. Recall that smooth curves can be locally approximated by affine approximations, i.e. $\pi_\gamma(x) \approx \langle a_j, x \rangle + c_j$, where $a_j$ is a local tangent vector of $\text{Im}(\gamma)$. Then (2) can be seen as a family of problems $f(x) \approx g_j(\langle a_j, x \rangle)$ where $j$ iterates over pieces of $\text{Im}(\gamma)$ that are approximately affine. Now notice that the model $f(x) = g(\pi_\gamma(x))$ and the monotonicity of $g$ imply an equivalence between $f(x) \approx f(x')$ and $\pi_\gamma(x) \approx \pi_\gamma(x')$. Therefore, instead of iterating over approximately affine pieces of $\text{Im}(\gamma)$, we can equivalently iterate over a partition of $\text{Im}(f)$ consisting of disjoint intervals $\mathcal{R}_j$, and assert that (2) can be split up into localized SIM problems

$$\mathbb{E}[Y|X, f(X) \in \mathcal{R}_j] \approx g_j(\langle a_j, X \rangle), \quad j = 1, \ldots, J,$$

where the tangent vectors $a_j$ now play the role of index vectors in (1).

In Figure 1 we study the effects of such an approach on several UCI data sets. Namely, for each data set we partition the data into $J$ sets, as detailed above, learn a SIM estimator

---

**Figure 1:** The generalization error for the SIM ($J = 1$) and NSIM models ($J > 1$) on 7 UCI data sets. We report the mean of 20 repetitions of the experiments for each $J$ and each data set. To create a partition into level sets we construct statistically equivalent blocks based on the ordered $Y$-sequence. The results for each data set are normalized so that the value at $J = 1$ is 1.0.
on each of the $J$ sets, and then plot the generalization error of the resulting estimator (i.e. the one described in this paper) as a function of the hyperparameter $J$. Given sufficient amount of data, we can observe that replacing (1) with its localized counterpart (4) often results in better estimation. For example, on the Yacht data set the generalization error improves by more than 30 percent for $J = 5$. Notice though that increasing the number of localized pieces does not always improve the performance. This can mostly be attributed to the fact that splitting the original data set into disjoint subgroups reduces the number of samples within each group, which has a detrimental effect on the variance of the estimator. In other words, we face a typical bias-variance trade-off, implying that the hyperparameter $J$ needs to be carefully selected.

We will now provide an overview of relevant literature and describe the details of our approach.

Related work. To the best of our knowledge, relaxations of SIM have not yet been considered in the given form. However, three research areas are highly relevant: linear single- and multi-index models; nonlinear sufficient dimension reduction; manifold regression. Below we provide a short summary and overview of the most significant and relevant achievements in each of these fields.

Single- and multi index models have been extensively researched, and we therefore, restrict ourselves to conceptually related work. Most studies focus only on the estimation of index vector(s). Along those lines, the most similar work is [HJS01] where the authors use iterative local linear regression to estimate the index vector $a$. The locality is enforced by kernel weights, which are initially set to be spherical around the estimation point, and then iteratively reshaped so that the isolines resemble level set boundaries of a strictly monotonous link function $g$. The approach has been extended to the multi-index case [DJS08].

Another relevant line of work are methods based on inverse regression that began with the introduction of sliced inverse regression (SIR) [Li91], and was followed up by [DC00] (SAVE), [Li02] (PHD), [XTL02] (MAVE), [LZC+05] (Contour regression), [LW07] (Directional regression), and others. These methods use estimates of inverse moments $E[X|Y]$, Cov $(X|Y)$ to estimate the index vector (or the sufficient dimension reduction space).

There are also several methods that simultaneously learn the link function and the index vector. We mention Isotron [KS09] and Slisotron [KKSK11], that iteratively update the link function using isotonic regression, and the vector using the function estimate; [GRB+17] that uses a similar approach but additionally assumes sparsity of the index vector; [KP16] that uses an iterative procedure and spline estimates; and [Rad15] that proposes higher dimensional splines.

The work and mathematical understanding of nonlinear sufficient dimension reduction is still in its early stages and there are many open questions. Most of the existing studies consider kernelized versions of linear estimators (such as SIR or SAVE) to globally linearize the problem in feature space, and then apply well-known methods regression methods, see [LAL+11, LLC+13, Wu08, YHL09].

Finally, model (2) can be considered from the viewpoint of manifold regression, where the goal is to estimate a function $f : M \rightarrow \mathbb{R}$ defined on the data. Manifold regression methods, such as [BL+07, Kpo11, MWZ+10], generally assume that the marginal distribution of $X$ is either supported on $M$ or in its close vicinity. As a consequence, Euclidean distances can be used to locally approximate the geodesic metric. This is a strong assumption which is implicitly or explicitly leveraged by all manifold regression techniques, and presents a breaking point for their effective use. In this work, we instead consider distributions that are spread in all directions of the ambient space around the curve $\gamma$. Therefore, geodesic proximity cannot be inferred from Euclidean distances and we need to use a localized proxy for the geodesic distance.

Main idea and estimation procedure for the NSIM model. The model (2) increases the flexibility of the ordinary SIM (1) by allowing for varying index vectors, corresponding to different regimes of the response $f(X)$. Consequently, a natural approach would be to partition the data into several groups, based on $Y$, and assume a SIM in each subgroup. Since SIM
estimators have been extensively studied we can rely on well-known methods to locally estimate the index vector. In particular, our approach follows three steps.

In the **first step** we partition the data set \( \{(X_1, Y_1), \ldots, (X_N, Y_N)\} \) into \( J \) sets \( \{X_j\}_{j=1}^J \) and \( \{Y_j\}_{j=1}^J \) through a disjoint union of the responses, \( \text{Range}(Y) = \bigcup_{j=1}^J \mathcal{R}_j \) for intervals \( \mathcal{R}_j \), and set

\[
\mathcal{Y}_j := Y \cap \mathcal{R}_j, \quad \mathcal{X}_j := \{X_i \in \mathcal{X} : Y_i \in \mathcal{Y}_j\}.
\]  

(5)

We refer to sets \( \mathcal{X}_j \) as **level sets**, since they can be defined as \( \mathcal{X}_j = \mathcal{X} \cap f^{-1}\left(\mathcal{R}_j\right) \) in the noise-free case. The optimal method for partitioning \( \text{Range}(Y) \) as \( \bigcup_{j=1}^J \mathcal{R}_j \) depends on the marginal distribution of \( Y \), and is best chosen after inspecting the empirical density. For example, we recommend using dyadic cells of \([\min Y; \max Y]\) if the density of \( Y \) is roughly uniform, and stochastically equivalent blocks\(^2\) if the probability mass is unevenly distributed. Also, it can be useful to first transform \( Y \)'s, e.g. using a log transform, to obtain more balanced level sets.

In the **second step** we compute estimates of local index vectors \( \{\hat{a}_j\}_{j=1}^J \) by using linear regression on \( \mathcal{X}_j \) and \( \mathcal{Y}_j \). Namely, let \( \hat{\Sigma}_j := \hat{E}_{\mathcal{X}_j \mathcal{Y}_j}(X - \hat{E}_{\mathcal{X}_j}X)(X - \hat{E}_{\mathcal{X}_j}X)^\top \) be the local empirical covariance matrix. Then, set \( \hat{a}_j = \hat{b}_j / \| \hat{b}_j \| \), where \( \hat{b}_j \) is defined by

\[
\hat{b}_j := \arg\min_{\{b_j \in \text{ker}(\hat{\Sigma}_j)\}_j} \left\| \hat{E}_{\mathcal{X}_j \mathcal{Y}_j} \left( Y - \hat{E}_{\mathcal{Y}_j}Y \right) - \left( \omega_j, X - \hat{E}_{\mathcal{X}_j}X \right) \right\|^2,
\]  

(6)

or equivalently,

\[
\hat{b}_j := \hat{\Sigma}_j \hat{E}_{\mathcal{X}_j \mathcal{Y}_j} \left( (Y - \hat{E}_{\mathcal{Y}_j}Y)(X - \hat{E}_{\mathcal{X}_j}X) \right).
\]  

(7)

We note that the slope vector of linear regression has been used in [HJS01] to estimate the index vector for SIM. Our analysis of \( \hat{a}_j \) is different since the nonlinearity of the geometry induces an additional approximation error. To reduce the approximation error the number of level sets \( J \) needs to grow with \( N \) and achieving the optimal error with \( \hat{a}_j \) requires a carefully balanced relationship \( J = J(N) \), see Section [3].

In the **final step** we use a kNN-type estimator to predict \( f(x^*) \) for out-of-sample \( x^* \). Since the make-or-break point of kNN-estimators regards how distances between \( x^* \) and training samples are measured, the critical point of this step is about the selection of an appropriate distance function. The issue is that the optimal choice (the geodesic metric on \( \text{Im}(\gamma) \)) is not available since \( \text{Im}(\gamma) \) is not known, and the naive choice (the Euclidean metric) generally leads to estimation rates that depend on the ambient dimension, and thus the curse of dimensionality.

---

**Algorithm 1** Summary of NSIM Estimator

**Learning local index vectors**

**Input:** \( \{(X_i, Y_i)\}_{i=1}^N, J \)

Split data into \( \{\mathcal{X}_j\}_{j=1}^J \) and \( \{\mathcal{Y}_j\}_{j=1}^J \) according to (5)

**for** \( j = 1, \ldots, J \) **do**

\( \hat{b}_j := \hat{\Sigma}_j \hat{E}_{\mathcal{X}_j \mathcal{Y}_j} \left( (Y - \hat{E}_{\mathcal{Y}_j}Y)(X - \hat{E}_{\mathcal{X}_j}X) \right) \)

\( \hat{a}_j := \hat{b}_j / \| \hat{b}_j \| \)

**end for**

**Output:** \( \hat{a}_j \) for \( j = 1, \ldots, J \)

**Out-of-sample prediction**

**Input:** sample \( x^* \), sets \( \{\mathcal{X}_j\}_{j=1}^J \), \( \{\mathcal{Y}_j\}_{j=1}^J \), index vectors \( \hat{a}_j \), parameters \( k, \eta \)

Compute \( \mathcal{N}_\Delta(x^*, k, \eta) \)

**Output:** \( \hat{f}_k(x^*) = \hat{E}_{\mathcal{N}_\Delta(x^*, k, \eta)}(Y) \)
To develop a proxy metric, consider now the ordinary SIM\(^1\). In this case the geodesic metric is equivalent to the Euclidean distance of projected samples, and training the kNN estimator on projected samples \(((\hat{a}, X_i), Y_i)\) achieves optimal univariate regression rates if the index vector \(\hat{a}\) approximates the true index vector with a sufficiently high rate. In other words, \(|(\hat{a}, x - x')|\) is a good proxy for the geodesic metric. The NSIM case is more challenging because we have \(J(N)\) different index vectors to choose from, and \(x^*\) cannot be \textit{a priori} assigned to any level set since we do not know \(f(x^*)\). Still, we can show that the quantity (referred to as proxy metric)

\[
\Delta(x^*, X_i) := \left| \langle \hat{a}_{j(X_i)}\rangle, x^* - X_i \right|, \quad \text{where } j(X_i) \text{ is such that } X_i \in \mathcal{X}_{j(X_i)},
\]

approximates the geodesic metric \(d_\gamma(\pi_\gamma(x^*), \pi_\gamma(X_i))\) reasonably well, provided that the piece of \(\text{Im}(\gamma)\) connecting \(\pi_\gamma(x^*)\) and \(\pi_\gamma(X_i)\) is linear enough, see \((19)\). This motivates the following estimator: let \(\mathcal{N}_\Delta(x^*, k, \eta)\) contain the \(k\) closest training samples of \(x^*\) with respect to \(\Delta(x^*, \cdot)\) within a Euclidean ball around \(x^*\) of radius \(\eta\). Then we set

\[
\hat{f}_k(x^*) := \frac{1}{k} \sum_{X_i \in \mathcal{N}_\Delta(x^*, k, \eta)} Y_i.
\]

As we will discuss in Section 3.4, the radius \(\eta\) plays a dual role. It needs to be large enough so that there are sufficiently many samples to choose neighbors from, but also small enough that so \((8)\) is still a good proxy for the geodesic metric. The entire estimation approach is summarized in Algorithm 1.

**Computational complexity.** The first two steps, \textit{i.e.} partitioning and computing tangents, are dominated by \(\mathcal{O}(\min\{JD^3, JND^2\} + ND^2)\), which is mostly due to forming covariance matrices and computing the generalized inverse. The out-of-sample prediction has a complexity of \(\mathcal{O}(N + JD)\) operations per evaluation.

**Contributions.** In this work, we provide a thorough introduction and theoretical analysis of NSIM model \((2)\) by considering the estimation problem from \(N\) data samples \(\{(X_i, Y_i)\}_{i=1}^N \sim_{\text{iid}} \rho^N\), where \(\rho\) is the unknown underlying distribution of \((X, Y)\). Our main result concerns the estimation of local index vectors, respectively the tangent field of \(\gamma\). Moreover, we provide a simple, and computationally efficient algorithm to estimate \(f\) that achieves nearly-optimal theoretical guarantees in some scenarios. Finally, we validate our theoretical results using synthetic data experiments, and we show that the additional flexibility of the NSIM allows superior performance over the SIM on some real data sets.

The theoretical results of this work can be summarized as follows:

- **Theorem 3** ensures accurate approximations of localized index vectors, namely we show \(\|\hat{a}_j - \gamma'(\mathbb{E}[\hat{a}_j Y \in \mathcal{R}_j])\| \lesssim |\mathcal{R}_j|^2 + |\mathcal{R}_j| |\mathcal{X}_j|^{-1/2}\) in the regime where \(\varepsilon\) is negligible compared to \(|\mathcal{R}_j|\), and the corresponding piece of \(\text{Im}(\gamma)\) can be well approximated by an affine space.

- **Corollary 10** ensures a consistent estimation of \(f\) with nearly-optimal univariate regression rate \(\mathcal{O}(\log^2(N)/N)\) for piecewise constant estimators in the noise-free setting \((Y = f(X))\).

- **Corollary 11** considers the case of bounded noise \(\|\varepsilon\| \leq \sigma_e\) and ensures \(|\hat{f}_k(x^*) - f(x^*)| \in \mathcal{O}(\sigma_e)\) in general. In case of an ordinary SIM however, we recover the optimal univariate regression rate \(\mathcal{O}(N^{-1/3})\) for piecewise constant estimators if the noise is sufficiently small.

\(^3t\) is the distribution induced by \(t = \gamma^{-1} \circ \pi_\gamma(X)\) for any arc-length parametrization \(\gamma\) of \(\text{Im}(\gamma)\).
Organization of the paper. We begin in Section 2 by introducing the model, the non-parametric assumptions on the distribution, data generating function, and function noise. In Section 3 we present the theoretical analysis of our estimators, starting with error rates for learning the geometry (i.e. the localized index vectors) and concluding with the guarantees for prediction accuracy. Most of the proofs needed for this section are postponed to the Appendix. In Section 4 we present numerical experiments on synthetic data (to validate the theoretical results) and real-world data sets (to compare with state-of-the-art estimators). Section 5 concludes with a discussion about our results and future directions.

General notation. \( \| \cdot \| \) denotes the standard Euclidean norm for vectors, and the spectral norm for matrices. \( d_\gamma \) denotes the geodesic metric on \( \text{Im}(\gamma) \). Provided that \( \gamma \) is an arc-length parametrisation, this means \( d_\gamma(\gamma(t_1),\gamma(t_2)) = |t_1 - t_2| \). For a discrete set of points \( A = \{x_1, \ldots, x_k\} \subset \mathbb{R}^D \) we use \( |A| \) to denote the number of elements: \( |A| = k \). On the other hand, if \( A \) is a connected subsegment of \( \text{Im}(\gamma) \), then \( |A| \) denotes the length of the segment \( A \); and if \( A \subset \mathcal{R} \) is an interval, \( |A| \) is the length of the interval. Here an interval \( A \subset \mathcal{R} \) always refers to a closed and connected subset of the real line. The Moore-Penrose inverse of a matrix \( M \) is denoted by \( M^\dagger \). \( B_m(x,R) \) denotes a ball of radius \( R \) around a point \( x \), with respect to a metric \( m \).

We denote by \( \mathbb{E}X \) and \( \text{Cov}(X) \) the expectation and covariance matrix of a random variable \( X \), and by \( \bar{X} \) and \( \bar{\Sigma} \) the sample mean and sample covariance over all samples. \( \bar{X}_U \) denotes the sample mean over all samples that belong to a subset \( U \), that is, \( \bar{X}_U = \frac{1}{|U|} \sum_{X \in U} X \). We use \( \rho \) to denote the joint probability distribution of \((X,Y)\), and \( \rho_X \) to denote the marginal, i.e., the probability distribution of \( X \), and \( \text{supp}(\rho_X) \) is its support. The abbreviation \textit{a.s.} is used as a shorthand for \textit{almost sure} events (with respect to implicit random vectors).

2 Theoretical framework for the NSIM model

Whereas the description of the NSIM model in Section 1 may convey its basic motives, it is certainly not adequate for the purposes of a rigorous investigation. Due to NSIM’s scope, it is relatively easy to construct examples that fit the model but for which estimation from finite samples is not possible. Thus, our goal now is to define a framework that allows a theoretical analysis, yet is broad enough to encompass both the single index model and its nonlinear relaxation. In the following we describe the assumptions on the function class, the underlying nonlinearity, and on the distribution of the data set.

Regularity assumptions for \( f \) and \( \text{Im}(\gamma) \). We consider simple, simply-connected and \( C^3 \) smooth curves, \textit{i.e.} that admit an arc-length parametrisation \( \gamma : \mathcal{I} \to \text{Im}(\gamma) \) with \( \| \gamma'(t) \| = 1 \). We assume that its curvature and torsion are bounded and denote \( \kappa = \| \gamma'' \|_\infty \) and \( K = \max\{\| \gamma'' \|_\infty, \| \gamma''' \|_\infty \} < \infty \). We consider \( f \in C^2 \) with a bounded Hessian, \( \| \nabla^2 f \|_\infty \leq L_H \), that satisfy \( f(x) = g(\pi_\gamma(x)) \) for some \( L_f \)-bi-Lipschitz function \( g : \text{Im}(\gamma) \to \mathbb{R} \), that is

\[
L_f^{-1} d_\gamma(v,v') \leq |g(v) - g(v')| \leq L_fd_\gamma(v,v'), \text{ for all } v,v' \in \text{Im}(\gamma).
\] (10)

Through rescaling we can always assume \( \text{Im}(f) = [0,1] \). We can, without loss of generality, align \( \gamma \) with \( \nabla f \), \textit{i.e.} choose an orientation such that \( \langle \nabla f(v), \gamma'(v) \rangle > 0 \). Another important quantity is the reach \( \tau_\gamma \) of \( \text{Im}(\gamma) \) - the largest \( r > 0 \) such that any point at distance less than \( r \) from \( \text{Im}(\gamma) \) has a unique nearest point on \( \text{Im}(\gamma) \) \cite{Fed59}. This ensures that \( f(x) \) is well defined for all \( x \) within the reach, \textit{i.e.} such that \( \min_{z \in \text{Im}(\gamma)} \| x - z \| < \tau_\gamma \).

Distributional assumptions. The following assumptions \((A1) - (A5)\) characterize the family of distributions that are admissible to the analysis in the next section. They are, apart from
standard and mostly related to commonly used assumptions. At the end of the section we provide a family of distributions that satisfies all assumptions related to the marginal \( \rho_X \).

Our first assumption asserts that \( \text{Im}(\gamma) \) is centered in the middle of the distribution and reads

(A1) \( \mathbb{E}[X - \pi_\gamma(X)|f(X)] = 0, \ f(X) \ a.s. \)

By rewriting it as \( \mathbb{E}[X|\pi_\gamma(X)] = \pi_\gamma(X) \), we see that it is inspired by the linear conditional mean assumption\(^4\) from multi-index model literature \([\text{Li91, DC00}]\), which holds, for example, for any elliptical distribution. Moreover, it also implies the identifiability of \( \text{Im}(\gamma) \) by the distribution of \((X,Y)\) in the following sense.

**Lemma 1.** Let \( \mathcal{D}, \mathcal{D}' \subset \mathbb{R}^D \), with orthogonal projections \( \pi_\mathcal{D}, \pi_{\mathcal{D}'} \) defined according to \((3)\), and let \( X \) be a random vector such that \( \pi_\mathcal{D}(X) \) and \( \pi_{\mathcal{D}'}(X) \) are a.s. unique. Let \( g : \mathcal{D} \to \mathbb{R} \) and \( g' : \mathcal{D}' \to \mathbb{R} \) be measurable and injective. If \( f = g \circ \pi_\mathcal{D} = g' \circ \pi_{\mathcal{D}'} \), and \( \mathbb{E}[X - \pi_\mathcal{D}(X)|f(X)] = \mathbb{E}[X - \pi_{\mathcal{D}'}(X)|f(X)] \) a.s., then \( \pi_\mathcal{D}(X) = \pi_{\mathcal{D}'}(X) \) a.s.

**Proof.** The assumption implies that we have a.s.

\[
0 = \mathbb{E}[X - \pi_\mathcal{D}(X)|f(X)] - \mathbb{E}[X - \pi_{\mathcal{D}'}(X)|f(X)] = \mathbb{E}[\pi_{\mathcal{D}'}(X)|f(X)] - \mathbb{E}[\pi_\mathcal{D}(X)|f(X)].
\]

(11)

Since conditioning on an injective function of a random variable is equivalent with conditioning on the random variable itself\(^5\), we get

\[
\mathbb{E}[\pi_{\mathcal{D}'}(X)|f(X)] = \mathbb{E}[\pi_\mathcal{D}(X)|g(\pi_{\mathcal{D}'}(X))] = \mathbb{E}[\pi_\mathcal{D}'(X)|\pi_\mathcal{D}(X)] = \pi_{\mathcal{D}'}(X),
\]

and similarly \( \mathbb{E}[\pi_\mathcal{D}(X)|f(X)] = \pi_\mathcal{D}(X) \). Plugging into \((11)\) the claim follows. \(\square\)

Let us now split the random vector \( X \) into \( V := \pi_\gamma(X) \), its component on \( \text{Im}(\gamma) \), and \( W := X - \pi_\gamma(X) \), the component orthogonal to \( \text{Im}(\gamma) \), see Figure 2: Random vectors \( V \) and \( W \) are the content of next two assumptions.

(A2) \( \|W\| \leq B < \tau_\gamma \) holds almost surely.

(A3) There are \( 0 < c_V^- \leq c_V^+ \) such that \( c_V^- |\mathcal{S}| |\mathcal{I}|^{-1} < \mathbb{P}(V \in \mathcal{S}) < c_V^+ |\mathcal{S}| |\mathcal{I}|^{-1} \) holds for any \( \mathcal{S} \subset \text{Im}(\gamma) \). Denote \( c_V := c_V^-/c_V^+ \).

---

\(^4\)This asserts that for any \( b \in \mathbb{R}^D, \ E[b^\top X|AX] = cAX \) and some \( c \in \mathbb{R}^k \), where \( A \in \mathbb{R}^{D \times k} \) is the index space.

\(^5\)This can be proven by applying twice \( \sigma(h(U)) \subset \sigma(U) \) for \( h = g, U = \pi_{\mathcal{D}'}(X) \) and \( h = g^{-1}, U = g(\pi_{\mathcal{D}'}(X)) \), where \( \sigma(U) \) is the \( \sigma \)-algebra spanned by a random variable \( U \).
These assumptions are standard for estimation problems involving nonlinear geometries, such as manifold estimation or manifold regression. In particular, some version of \([A2]\) is required because the projection \(\pi_\gamma(x)\), and consequently the function \(f\), is in general not uniquely defined for points outside the reach of \(\text{Im}(\gamma)\).

For an interval \(\mathcal{R} \subset [0,1]\) we will in the rest of the paper denote conditional moments such as \(\text{Cov}(X|Y \in \mathcal{R})\) and \(\mathbb{E}[V|Y \in \mathcal{R}]\) by \(\text{Cov}(X|\mathcal{R})\) and \(\mathbb{E}[V|\mathcal{R}]\).

Our fourth assumption is the one that is most specific to the NSIM and reveals a limitation of our approach. Let \(t := \gamma^{-1}(\pi_\gamma(X))\) be the induced random variable, taking values in \(\mathcal{I}\), and for an interval \(\mathcal{R} \subset [0,1]\), define the mean \(\bar{t}_\mathcal{R} := \mathbb{E}[t|\mathcal{R}]\). The orthogonal projection onto the normal space at \(\gamma(t_\mathcal{R})\) is denoted as \(P_{\mathcal{R}_\perp} = \text{Id} - \gamma'(t_\mathcal{R})\gamma'(t_\mathcal{R})^\top\), see Figure 2. We assume there exist universal constants \(R_{\max}, C_L, C_W\) such that for all intervals \(\mathcal{R} \subset [0,1]\) with \(|\mathcal{R}| < R_{\max}\) we have

\[(A.4.1)\] For all \(v \in \text{Im}(\text{Cov}(X|\mathcal{R})) \cap \text{Im}(P_{\mathcal{R}_\perp}),\) with \(|v| = 1\), we have \(v^\top \text{Cov}(X|\mathcal{R}) v > C_\perp > 0,\)

\[(A.4.2)\] \(\mathbb{E}_V \left[ \|\text{Cov}(W|V) - \text{Cov}(W|\mathcal{R})\|^2 |\mathcal{R}\right] \leq C_W^2 \mathbb{E}[d_\gamma(V,V')^2 |\mathcal{R},Y,Y' \in \mathcal{R}]\).

Assumption \([A.4.1]\) defines a lower bound for non-zero eigenvalues of the normal component of the localized covariance matrix. This will be critical for ensuring that linear regression selects the local gradient (which is a tangent of \(\gamma\)) instead of the local curvature vector of \(\gamma\), if \(|\mathcal{R}|\) is small enough. Specifically, the requirement reads \(|\mathcal{R}| \lesssim \sqrt{C_\perp}/K\) (see Theorem 3), which implies that a larger \(C_\perp\) provides better differentiation between local gradients and local curvature vectors. We also note that for \(B\) fixed, \(C_\perp\) depends on the dimensionality because

\[B^2 \geq \mathbb{E}[\|W\|^2 |\mathcal{R}] = \text{Tr}(\text{Cov}(W|\mathcal{R})) \approx \text{Tr}(\text{Cov}(P_{\mathcal{R}_\perp}X|\mathcal{R})) > \text{rank}(\text{Cov}(P_{\mathcal{R}_\perp}X|\mathcal{R}))C_\perp.\]

For a full rank \(\text{Cov}(X|\mathcal{R})\) this implies \(C_\perp \approx B^2/(D-1)\).

On the other hand, assumption \([A.4.2]\) is a nonlinear relaxation of the commonly used \textit{constant conditional covariance assumption} for multi-index model estimation, cf. [DC00, LZC+05, LW07]. It is perhaps best understood as a Lipschitz-continuity condition of the sliced conditional covariance \(\text{Cov}(W|V)\), seen as a function of \(V\), with respect to \(\|\cdot\|\) and \(d_\gamma\). The parameter \(C_W\) acts as an amplifier of curvature effects, but does not play a crucial role in the analysis.

Our last assumption concerns the function noise \(\varepsilon\).

\[(A.5)\] Let \(Y = f(X) + \varepsilon\), where \(\varepsilon \perp (V,W)\), and \(|\varepsilon| \leq \sigma_\varepsilon\) almost surely. Furthermore, whenever \(|\mathcal{R}| > 2\sigma_\varepsilon\) we have \(\|\text{Cov}(V,\varepsilon|Y \in \mathcal{R})\| \leq C_L f |\mathcal{R}|^3\).

The joint independece implies \(\varepsilon \perp W|g(V)\), and with \([A1]\) we get \(\mathbb{E}[W|Y] = 0\). By the law of total covariance (see [42]) it follows that \(\text{Cov}(W,\varepsilon|\mathcal{R}) = 0\) for any interval \(\mathcal{R} \subset [0,1]\). The independence assumption also implies \(\text{Cov}(V,\varepsilon|f(V) \in \mathcal{R}) = 0\), but when conditioning on \(Y \in \mathcal{R}\) the random variables \(V\) and \(\varepsilon\) become increasingly dependent as \(|\mathcal{R}|\) tends to \(\sigma_\varepsilon\). Since \(\|\text{Cov}(V,\varepsilon|\mathcal{R})\| \leq L_f \sigma_\varepsilon |\mathcal{R}| + 2L_f \sigma_\varepsilon^2\) holds without any assumptions, \([A.5]\) is trivially satisfied when replacing \(|\mathcal{R}| > 2\sigma_\varepsilon\) by \(|\mathcal{R}|^2 > \sigma_\varepsilon\). However, this encompasses very general scenarios, e.g. neglecting any potential symmetries in the distribution of \(\varepsilon\), which is our main motivation to quantify the effect more precisely.

Finally, we provide a family of marginal distributions that satisfies \([A1]-[A4.2]\). The proof is in Appendix A.3

\[\text{Lemma 2.} \ Let \mathcal{F} : (\text{Im}(\gamma), d_\gamma) \to (\mathbb{R}^{D \times (D-1)}, \|\cdot\|) \text{ be a } L_F\text{-Lipschitz map such that the columns of } \mathcal{F}(v) \text{ form an orthonormal basis of the normal space of } \text{Im}(\gamma) \text{ at } v. \text{ Let } W = F(V)U, \text{ for some random vector } U \in \mathbb{R}^{D-1}, \text{ with } \mathbb{E}[U|V] = 0, \|U\| \leq B, \text{ and } \text{Cov}(U) \succ C_L^{-1}. \text{ Then } X = V + W \text{ satisfies } \[A1\]-\[A2\]-\[A4.1\]-\[A4.2\] \text{ with } \mathcal{R}_{\max} = \frac{C_L}{2L_F\|\text{Cov}(U)\|} - 2\sigma_\varepsilon, C_W = 2\|\text{Cov}(U)\| L_F, \text{ and } C_\perp = R_{\max} - |\mathcal{R}|.\]
3 Main results

In this section we conduct a theoretical analysis of the proposed estimator \( \hat{b}_j \). We begin in Section 3.2 by providing guarantees for the estimation of local index vectors in terms of the number of samples and \( J \), the number of level sets. In Section 3.3 we use these results to show that the estimator \( \hat{b}_j \) is accurate on almost linear pieces of the curve. To extend the estimator to the entire curve \( \text{Im}(\gamma) \) in Section 3.4 we establish results for the proxy metric \( \Delta(x^*, \cdot) \) in a restricted search space around an out-of-sample point \( x^* \).

3.1 Preliminary remarks

The main results are in Theorem 3 and in Corollaries 10 and 11. Generally speaking, our analysis naturally follows the steps that motivated the construction of the estimator:

**Step 1** Partition \( X \)‘s according to a dyadic partitioning of the range \( \text{Im}(f) = [0, 1], \) i.e.

let \( \mathcal{R}_j := \left[ \frac{j-1}{J}, \frac{j}{J} \right] \) and define \( Y_j := \mathcal{Y} \cap \mathcal{R}_j \), and \( X_j := \{ X_i \in \mathcal{X} : Y_i \in Y_j \} \).

**Step 2** Estimate local index vectors by using \( \hat{a}_j \) in each partition.

**Step 3** Regression on unseen data \( x^* \) using \( \hat{f}_k(x^*) = \hat{E}_{N\Delta(x^*, k, \eta)} Y_i \), where \( N\Delta(x^*, k, \eta) \) is a set denoting the \( k \)-nearest neighbors of \( x^* \) according to the proxy geodesic metric \( \overline{\eta}(x, \cdot) \).

Denote the tangent at \( \pi_j(X) \) by \( a(X) := \gamma^{-1} \circ \pi_j(X) \), and let \( a_j := a(\gamma(\overline{\eta}(x, \mathcal{R}_j))) \). Before providing details of the analysis we begin with a couple of comments regarding \( \hat{a}_j \):

- **First**, to balance bias and variance of the estimator the number of level sets in \( \mathcal{R}_j \) has to be sufficiently short. Recall that local index vector are defined as \( \hat{a}_j := \hat{b}_j/\|\hat{b}_j\| \), where \( \hat{b}_j \) is a solution of local linear regression \( \hat{E}_j \). The variance of the estimator is mainly due to the estimation error \( \|\hat{E}[\hat{a}_j] - \hat{a}_j\| \), and decreases as \( |\mathcal{R}_j| = J^{-1} \) and \( |\mathcal{X}_j|^{-1/2} \) decrease. The bias of the estimator is on one hand due to the approximation of \( \text{Im}(\gamma) \) by piecewise linear spaces, and on the other due to \( \hat{E}[\hat{a}_j] \neq a_j \) (first term in \( \gamma^2 \)). Both components decrease only with \( |\mathcal{R}_j| = J^{-1} \).

- **Second**, our analysis shows that \( J \approx \min\{N/\log^2(N), \sigma_\varepsilon^{-1}\} \) achieves the optimal balance, implying that there are two regimes. In the first regime, where \( J \ll \sigma_\varepsilon^{-1} \), increasing \( J \) will simultaneously reduce both the bias and the variance of the estimator. In the second regime the function noise precludes further decreasing \( |\mathcal{R}_j| \) (since we cannot decrease the length of the corresponding segment of \( \text{Im}(\gamma) \) that contains \( \pi_j(X_j) \)). In other words, the noise level \( \sigma_\varepsilon \) imposes a lower bound on \( |\mathcal{R}_j| \), and the bias does not completely vanish in the nonlinear case (cf. Lemma 6).

- **Step 2** is valid in a regime where \( |\mathcal{R}_j| \) satisfies certain demands:
  - **ensuring that the covariance in the direction of the tangent does not vanish**: we require \( a_j \in \text{Im} (\text{Cov}(X|Y \in \mathcal{R}_j)) \), which is guaranteed if \( 2\sigma_\varepsilon < |\mathcal{R}| < (2KL_f)^{-1} \) (Lemma 18). Otherwise, it would be impossible to approximate \( a_j \) using linear regression;
  - **solving nonlinear problems with linear methods**: we need to isolate sufficiently short portions of \( \text{Im}(\gamma) \) so that the corresponding local geometry can be well approximated by a linear space and that the influence of the Hessian \( \nabla^2 f \) normal to \( a_j \) is weak compared to the magnitude of the gradient \( \nabla f \) (which we implicitly use to estimate local index vectors).

To quantify the first requirement we introduce a constant \( C_T \) which is, up to absolute constants (see Lemma 28 for full details), defined by

\[
C_T \approx \frac{1}{(1 - 2KL_f |\mathcal{R}|)^2} \left( \frac{|\mathcal{R}|}{|\mathcal{R}| - 2\sigma_\varepsilon} \right)^2.
\]

\(^6\) Technically, we ought to use \( \mathcal{R}_1 = [-\sigma_\varepsilon, J^{-1}] \), and \( \mathcal{R}_j = [\sigma_\varepsilon, J, 1 + \sigma_\varepsilon] \) to account for noise at the boundaries, but for the sake of simplicity we assume \( Y \) is thresholded to \( [0, 1] \), such that \( |\mathcal{R}_j| = J^{-1} \) for all \( j \).
The first factor is a technical requirement that arises from the nonlinearity of the problem. The second factor on the other hand reflects the fact that if $|\mathcal{R}| \leq 2\sigma_\varepsilon$ the conditional marginal measure (induced by $Y \in \mathcal{R}$) could be supported on a set $\mathcal{S} \subset \text{Im}(\gamma)$ of Lebesgue measure zero.

In what follows we focus our attention on how our estimates depend on fundamental parameters in the model: the sample size $N$, length of $\mathcal{R}$ (resp. number of level sets $J^{-1} = |\mathcal{R}|$), curvature and torsion, $\mathcal{C}_\lambda$ and $\mathcal{C}_\perp$. We use $\lesssim$ and $\gtrsim$ to absorb positive multiplicative factors for terms that decay with the total sample size $N$ or with the local sample size $|\mathcal{X}_j|$, and will mostly omit other constants such as the Lipschitz constant $L_f$ or $c_V$. Details regarding other constants that affect the model, and higher order terms in $|\mathcal{R}|$ or in $N$, are deferred to the Appendix.

For $\mathcal{S} \subset \text{Im}(\gamma)$ we set $\kappa(\mathcal{S}) := \max_{t \in I_{\gamma^{-1}(\mathcal{S})}} \|\gamma''(t)\|$, $K(\mathcal{S}) := \max\{\kappa(\mathcal{S}), \max_{t \in I_{\gamma^{-1}(\mathcal{S})}} \|\gamma''''(t)\|\}$. For an interval $\mathcal{R} \subset [0,1]$ we define $K_\mathcal{R} = K(\mathcal{S}(\mathcal{R}))$, and $K_\mathcal{R} = K(\mathcal{S}(\mathcal{R}))$, where $\mathcal{S}(\mathcal{R}) \subset \text{Im}(\gamma)$ is the infimum of all connected pieces of $\text{Im}(\gamma)$ such that $\mathbb{P}(V \in \mathcal{S}(\mathcal{R})|\mathcal{R}) = 1$.

### 3.2 Local direction approximation via linear regression

Define $\hat{a}(X_i) := \hat{a}_{j(X_i)}$ where $j(X_i) \in \{1, \ldots, J\}$ is the a.s. unique level set with $X_i \in \mathcal{X}_{j(X_i)}$ and $\hat{a}_{j(X_i)}$ is given by (7). The error in estimating the local index vectors can be decomposed as

$$
\|\hat{a}(X_i) - a(X_i)\| \leq \|\hat{a}_{j(X_i)} - a_{j(X_i)}\| + \|a_{j(X_i)} - a(X_i)\| \leq \|\hat{a}_{j(X_i)} - a_{j(X_i)}\| + K_R \|S(\mathcal{R}_j)\|. 
$$

(13)

Note that assuming the distribution of $\pi_\gamma(X)$ is equivalent to a uniform distribution (i.e. [A3] holds), the second term, which we refer to as the quantization error, cannot be improved. We will now bound the first term, which improves as $|\mathcal{R}|$ decreases and as the number of corresponding data samples increases.

**Theorem 3.** Let $u > 1$, and $\mathcal{R} \subset [0,1]$ be a closed interval with

$$
2\sigma_\varepsilon < |\mathcal{R}| \lesssim \max \left\{ \frac{1}{2K_\mathcal{R} L_f}, \frac{1}{C_T^2(L_H + C_\varepsilon)}, \frac{1}{K_\mathcal{R}} \sqrt{\frac{C_\perp}{C_T}}, R_{\max} \right\}
$$

(14)

Assume we have $\sqrt{N_\mathcal{R}} \gtrsim \max\{C_T^2, C_\perp^{-1}\} \log(D) u^2$ i.i.d. copies of $(X,Y)$, where $Y \in \mathcal{R}$. Denote $\hat{\Sigma}_\mathcal{R} = \hat{E}(X - \hat{E}X)(X - \hat{E}X)^\top$, $\hat{\gamma}_\mathcal{R} = \hat{E}(Y - \hat{E}Y)(X - \hat{E}X)$ and $\tilde{\gamma}_\mathcal{R} = \gamma(\mathbb{E}[t|\mathcal{R}])$. For $\hat{a} = \hat{b}/\|\hat{b}\|$, with $\hat{b} = \hat{\Sigma}_\mathcal{R}^{-1} \hat{\gamma}_\mathcal{R}$, we have with probability of at least $1 - \exp(-u^2)$

$$
\|\hat{a} - a(\tilde{\gamma}_\mathcal{R})\| \lesssim \frac{C_T^2}{C_\perp} \left( K_\mathcal{R} \|\mathcal{R}\|^2 + u^2 \frac{|\mathcal{R}|}{\sqrt{N_\mathcal{R}}} \right).
$$

(15)

**Remark 4** (Special cases of Theorem 3).

- $K_\mathcal{R} = 0$: If $\mathcal{R}$ corresponds to a flat piece of the curve the first term in (15) vanishes. Therefore, provided that $\sigma_\varepsilon$ is small, $\hat{a}$ is an unbiased estimator of $a(\tilde{\gamma}_\mathcal{R})$, with convergence rate $N_\mathcal{R}^{-1/2}$. This result covers the usual single index model, and our estimation rate matches that of known estimators (e.g. [HJS01]).

- $\sigma_\varepsilon = 0$: In the noise-free case we can remove the lower bound restriction on $|\mathcal{R}|$ in (14). Theorem 3 then implies $\|\hat{a} - a(\tilde{\gamma}_\mathcal{R})\| = O(N_\mathcal{R}^{-1/4})$, provided $|\mathcal{R}| \propto N_\mathcal{R}^{-1/2}$. Note that a similar rate is achieved in the noisy case, in the regime $|\mathcal{R}| \propto N_\mathcal{R}^{-1/2} \gg \sigma_\varepsilon$.

We now use a union bound argument and Theorem 3 to bound $\|\hat{a}(X_i) - a(X_i)\|$. The first step is to ensure that there are sufficiently many samples (simultaneously) in all sets $\mathcal{X}_j$.

---

7The precise requirement on $\mathcal{R}$ is restated in Theorem 29 in the Appendix, and $C_T$ is defined in (72).
Lemma 5. Assume we have $N$ i.i.d. copies of $(X, Y)$. Let $\{X_j\}_{j=1}^J$ and $\{Y_j\}_{j=1}^J$ be a partitioning according to (12). Then for any $0 < u^2 < N$ we have with probability at least $1 - \exp(-u^2)$

$$\min_{j=1, \ldots, J} |X_j| \gtrsim \frac{(J-1 - 2\sigma_e) N}{|I| u^2}. \quad (16)$$

Proof. We first show that for each set $X_j$ there exists a segment $|S_j| \subset \text{Im}(\gamma)$ of a minimal length such that $V \in S_j$ implies $X \in X_j$ almost surely. Then we use Lemma 15 to conclude the result. So let $j \in \{1, \ldots, J\}$ arbitrary and denote $R_j = [a_j, b_j]$, $R_j^- = [a_j + \sigma_e, b_j - \sigma_e]$. Then $\mathbb{P}(Y \in R_j f(X) \in R_j^-) = 1$, and thus there exists a segment $S_j \subset \text{Im}(\gamma)$ with $|S_j| \geq L_j^{-1}(|R_j| - 2\sigma_e)$ such that $\mathbb{P}(Y \in R|V \in S) = 1$. Now we use Lemma 15 to get

$$\mathbb{P}\left(\min_{j \in J}|X_j| \geq \frac{J-1 - 2\sigma_e}{L_j} \frac{c_v^{-1} N}{|I| u^2} - 2 \right) \geq \mathbb{P}\left(\{V_i\}_{i=1}^N \text{ is a } |I| u^2 c_v^{-1} N \text{-net} \right) \geq 1 - \exp(-u^2).$$

We will now combine Lemma 5 and Theorem 3 to bound the estimation error for local index vectors in all level sets.

Lemma 6. Assume we have $N$ i.i.d. copies of $(X, Y)$. Let $\{X_j\}_{j=1}^J$ and $\{Y_j\}_{j=1}^J$ be a partitioning according to (12) such that $|R_j| = J^{-1}$ satisfies (14). Moreover, assume $N$ satisfies

$$\sqrt{N_{LB}} := \sqrt{(J-1 - 2\sigma_e) N/|I| \log^2(J)} \gtrsim \max\left\{C_T^2, \frac{1}{C_{\perp}}\right\} \log(D) u^3. \quad (17)$$

Then for $1 < u^2 < N$ we have with probability at least $1 - \exp(-u^2)$

$$\max_{i=1, \ldots, N} \|\hat{a}(X_i) - a(X_i)\| \lesssim \frac{C_T^2}{C_{\perp}} \frac{u^3}{\sqrt{N_{LB} J}} + \frac{K}{J}. \quad (18)$$

Let us make a couple of comments about the implications of this lemma before writing down its proof. First, the terms in the bound on the right hand side of (18) can also be written in a local form, i.e., global curvature and torsion bounds can be replaced with curvature and torsion bounds on segments of the curve that correspond to a specific sample $X_i$. In other words, the learning of index vectors is consistent on locally linear pieces.

Second, (17) and (15) suggest that the choice of $J$ that achieves the optimal error bound is $J = C \min\{N/L \log^2(N), N^{\frac{1}{2}}\}$, where $C > 0$ is large enough so that (17) is satisfied. Looking at (18), this implies that in order to decrease the error we ought to increase $J$ as long as $J \ll \sigma_e^{-1}$ (i.e., subdivide the data set into a larger number of subsets), while keeping the number of samples within each subset constant (up to log-factors and lower bounded by $N_{LB}$). The rationale behind this is that further subdividing the data set not only reduces the approximation error (which is caused by the curvature), but it also reduces the variance in the linear regression part of the problem, i.e., when estimating $a_j(X_i)$ by $\hat{a}_j(X_i)$. Note also that in (18), compared to (15), we lose an order in $J^{-1}$, which is due to the last term in (13), i.e., using quantization to approximate the entire tangent field $\gamma'(t)$ over the respective level set.

Proof of Lemma 5. Lemma 5 and (17) imply that with probability at least $1 - \exp(-u^2)$ we have

$$\min_{j=1, \ldots, J} \sqrt{|X_j|} \gtrsim \frac{\sqrt{N_{LB}} \log(J)}{u} \gtrsim \max\left\{C_T^2, \frac{1}{C_{\perp}}\right\} \log(D) \log(J) u^2.$$

Therefore, $\sqrt{|X_j|}$ satisfies the conditions of Theorem 3 (replacing $u^2$ by $\log(J) u^2$) for all $j = 1, \ldots, J$ with probability $1 - \exp(-u^2)$. We now have

$$\mathbb{P}\left(\forall i: \|\hat{a}(X_i) - a(X_i)\| \lesssim \frac{C_T^2}{C_{\perp}} \frac{u^3}{\sqrt{N_{LB} J}} + \frac{K}{J}\right) \geq \mathbb{P}\left(\forall j: \|\hat{a}_j - a_j\| \lesssim \frac{C_T^2}{C_{\perp}} \frac{u^3}{\sqrt{N_{LB} J}}\right),$$

11
where we used $K \geq \kappa$, and \[^{13}\] Note also that we absorbed the $K/J^2$ term from \[^{15}\] into $K/J$. Conditioning on $\min_{j} \sqrt{X_j} \geq \sqrt{\frac{N_LB}{J}} u^{-1}$, and using Theorem \[^{3}\] (noting $|\mathcal{R}_j| = J^{-1}$) with the union bound we have

$$P \left( \forall j : \|\hat{a}_j - a_j\| \leq \frac{C_T^2}{C_7} \frac{u^3}{\sqrt{\frac{N_LB}{J}}} \right) \geq P \left( \forall j : \|\hat{a}_j - a_j\| \leq \frac{C_T^2}{C_7} \frac{u^3 \log(J)}{\sqrt{X_j} / J} \min_j \sqrt{X_j} \geq \frac{\sqrt{\frac{N_LB \log(J)}{u}}}{1 - \exp(-u^2)} \right) \geq (1 - J \exp(-\log(J) u^2)) \geq 1 - 2 \exp(-u^2).$$

\[\square\]

### 3.3 Function estimation on almost linear segments

The goal is now to use the results on direction estimation to analyze the regression accuracy of the proposed estimator. Firstly, as mentioned earlier, linear approximations are only sensible on patches of $\text{Im}(\gamma)$ that are short and linear-enough in a certain sense. Indeed, Theorem \[^{3}\] and Lemma \[^{10}\] make this relationship explicit. We will now define the notion of almost linearity that we shall adhere to from now on and which attempts to quantify the amount of nonlinearity in a given piece of $\text{Im}(\gamma)$.

**Definition 7.** Let $\mathcal{I}$ be an interval and $\gamma : \mathcal{I} \rightarrow \mathbb{R}^D$ a $C^1(\mathcal{I})$ curve that is parametrized by arc-length. Let $0 < \theta \leq 1$. We say $\gamma$ is $\theta$-almost linear if $(\gamma'(t), \gamma'(s)) > \theta$ for all $t, s \in \mathcal{I}$.

Definition \[^{7}\] implies that if $\theta$ is close to 1 then the Euclidean distance approximates the geodesic distance well, i.e. $\|v - v'\| \approx d_\gamma(v, v')$ for any $v, v' \in \text{Im}(\gamma)$. This allows us to establish how well the proxy metric $\Delta(x^*, \cdot)$ approximates $d_\gamma(x^*, \cdot)$. Namely, if $\theta > \kappa B$, then

$$\frac{\Delta(x^*, X_i) - (|I| + 2B) \|\hat{a}_i(X_i) - a(X_i)\|}{\theta - \kappa B} \leq d_\gamma(v^*, V_i) \leq \frac{\Delta(x^*, X_i) + (|I| + 2B) \|\hat{a}_i(X_i) - a(X_i)\|}{\theta - \kappa B},$$

for $v^* = \pi_\gamma(x^*)$, and $X_i \in X$. The proof is in Section \[^{A.4}\] in the Appendix.

Let now $\mathcal{N}_\Delta(x^*, k) = \mathcal{N}_\Delta(x^*, k, \infty)$ be the set of $k$-closest neighbors with respect to $\Delta(x^*, \cdot)$ without restriction on the search space, and let $\mathcal{N}_{d_\gamma}(x^*, k)$ be the set of $k$-closest neighbors with respect to $d_\gamma(\pi_\gamma(x^*), \pi_\gamma(\cdot))$. To provide guarantees for the estimator we need bounds on $d_\gamma(\pi_\gamma(x^*), V_i)$ for $X_i \in \mathcal{N}_\Delta(x^*, k)$.

**Lemma 8.** Assume $\gamma$ is $\theta$-almost linear for some $\theta > \kappa B$, and the training set $\{V_i\}_{i=1}^N$ forms a $\delta$-net on $\text{Im}(\gamma)$. Let $x^* \in \text{supp}(\rho_X)$ arbitrary. Then

$$\max_{X_i \in \mathcal{N}_\Delta(x^*, k)} d_\gamma(\pi_\gamma(x^*), \pi_\gamma(X_i)) \leq \frac{2}{\theta - \kappa B} \left( \delta k + (|I| + 2B) \max_{i=1,\ldots,N} \|\hat{a}_i(X_i) - a(X_i)\| \right). \quad (20)$$

**Proof.** Let $\{\tilde{X}_1, \ldots, \tilde{X}_k\} = \mathcal{N}_{d_\gamma}(x^*, k)$ and $\{\tilde{X}_1, \ldots, \tilde{X}_k\} = \mathcal{N}_\Delta(x^*, k)$. Since $\{V_i\}_{i=1}^N$ is by assumption a $\delta$-net it follows that the set $\pi_\gamma(\mathcal{N}_{d_\gamma}(x^*, k))$ is contained within a connected subset of $\text{Im}(\gamma)$ of length at most $k\delta$. Using the first inequality in \[^{19}\], and $\kappa B < 1$ (which follows from $\kappa^{-1} \geq \gamma_\tau$), we get

$$\Delta(x^*, \tilde{X}_j) \leq \max_{i=1,\ldots,k} \Delta(x^*, \tilde{X}_i) \leq 2\delta k + (|I| + 2B) \max_{X_i \in \mathcal{N}_{d_\gamma}(x^*, k)} \|\hat{a}_i(X_i) - a(X_i)\|.$$

The second inequality in \[^{19}\] now gives a bound on $d_\gamma(\pi_\gamma(x^*), \pi_\gamma(\tilde{X}_j))$ by

$$d_\gamma(\pi_\gamma(x^*), \pi_\gamma(\tilde{X}_j)) \leq \frac{\Delta(x^*, \tilde{X}_j) + (|I| + 2B) \max_{\tilde{X}_i \in \mathcal{N}_\Delta(x^*, k)} \|\hat{a}_i(\tilde{X}_i) - a(\tilde{X}_i)\|}{\theta - \kappa B} \leq \frac{2}{\theta - \kappa B} \left( \delta k + (|I| + 2B) \max_{X_j \in \mathcal{N}_\Delta(x^*, k) \cup \mathcal{N}_{d_\gamma}(x^*, k)} \|\hat{a}_i(X_i) - a(X_i)\| \right).$$
A guarantee on the performance of our estimator $\hat{f}_k$ now follows immediately.

**Lemma 9.** Assume the conditions of Lemma 6 are satisfied and that $\gamma$ is $\theta$-almost linear for some $\theta > \kappa B$. Then for any $x^* \in \text{supp}(\rho_X)$, we have with probability at least $1 - \exp(-u^2)$

$$\left| \hat{f}_k(x^*) - f(x^*) \right| \lesssim \left( \mathbb{E}_{\Delta(\gamma)}^N \right) + \frac{u^2}{\theta - \kappa B} N + \frac{1}{\theta - \kappa B} \left( \frac{C_T}{C_\perp} \frac{u^3}{\sqrt{N}} \right),$$

(21)

where $N_{LB}$ is defined in [17].

**Proof.** We first decompose the left-hand side of (22) as

$$\left| \hat{f}_k(x^*) - f(x^*) \right| = \left| \mathbb{E}_{\Delta(\gamma)}^N Y - f(x^*) \right| \leq \mathbb{E}_{\Delta(\gamma)}^N f(X) - f(x^*) \right|.$$

To bound the second term we use the Lipschitz property and then apply Lemma 8. Then, since $\{V_i\}_{i=1}^N$ is a $|I| u^2 c_{\perp}^2 N^{-1}$ net with probability at least $1 - \exp(-u^2)$ (due to Lemma 15), we have

$$\left| \mathbb{E}_{\Delta(\gamma)}^N f(X) - f(x^*) \right| \leq 2L f \mathbb{E}_{\Delta(\gamma)}^N \max_{i_1, \ldots, i_N} \| \hat{a}(X_i) - a(X_i) \|. $$

The statement now follows by applying Lemma 6 and $\delta = |I| u^2 c_{\perp}^2 N^{-1}$. $\square$

Note that Lemma 9 actually obeys a local, instead of a global, error bound. That is, considering Lemma 8 we can replace the global curvature parameter $\kappa$ with $\kappa' = \kappa(S(x^*, k))$ where $S(x^*, k) \subset \text{Im}(\gamma)$ is the shortest sub-segment of $\text{Im}(\gamma)$ containing all $X_i \in N_{\Delta(\gamma)}(x^*, k)$, i.e.

$$S(x^*, k) = B_{d_\gamma}(x^*, \delta k) \cup \bigcup_{X_i \in N_{\Delta(\gamma)}(x^*, k)} B_{d_\gamma}(X_i, |S(R_j)|).$$

(22)

Thus, the estimator is consistent if $K' = 0$, i.e., if $x^*$ belongs to a locally linear piece of a $\theta$-almost linear curve (under an additional technical assumption on $\varepsilon$, see Corollary 11).

We will now consider two special cases: the noise-free case, and the asymptotic regime $N \to \infty$ of the noisy case. If $\sigma_\varepsilon = 0$, Lemma 8 implies $|\hat{f}_k(x^*) - f(x^*)| \asymp \log^2(N)/N$, which is optimal (up to log factors) when using a piecewise constant estimator, such as kNN, see [Koh14].

**Corollary 10** (Noise free case). Let $1 < u^2 < N$. Assume the setting of Lemma 6 and let $\sigma_\varepsilon = 0$. Choose $k = 1$, $J = CN/(\log^2(N))$, where $C$ and $N$ are such that the conditions of Lemma 6 are satisfied. We then have with probability $1 - \exp(-u^2)$

$$\left| \hat{f}_k(x^*) - f(x^*) \right| \leq \frac{u^2}{\theta - \kappa B} N + \frac{1}{\theta - \kappa B} \left( \frac{C_T}{C_\perp} u^3 \right) \frac{\log^2(N)}{N}.$$  

(23)

Thus, the estimator is consistent in the noise free case. On the other hand, in the noisy case we have, regardless of $k$, a biased estimator, i.e. $|\hat{f}_k(x^*) - f(x^*)| \in O(\sigma_\varepsilon)$.

**Corollary 11** (Noisy case). Let $1 < u^2 < N$. Assume the setting of Lemma 6 with $N \gg \sigma_\varepsilon^{-1}$, and let $J = C\sigma_\varepsilon^{-1}$ be fixed for some $C \in O(1)$. Assume $\sigma_\varepsilon$ is small so that $|R| = J^{-1}$ satisfies [14]. Then with $k = N^2$, for any $\omega < 1$, and with probability at least $1 - \exp(-u^2)$, we have

$$\left| \hat{f}_k(x^*) - f(x^*) \right| \lesssim \left( 1 + \frac{K}{\theta - \kappa B} \right) \sigma_\varepsilon.$$  

(24)

If additionally $k = 0$ and if $\{\varepsilon_i : X_i \in N_{\Delta(\gamma)}(x^*, k)\}$ is a collection of i.i.d. centered random variables, then with $k \asymp N^{2/3}$, and probability at least $1 - \exp(-u^2)$, we have

$$\left| \hat{f}_k(x^*) - f(x^*) \right| \lesssim \sigma_\varepsilon u + \frac{u^2}{\theta - \kappa B} N^{-1/3}.$$  

(25)
Proof. Let us first show \(24\). Since \(J = \mathcal{O}(\sigma^{-1})\) is fixed, we have \(\sqrt{N\log} \in \mathcal{O}(\sqrt{N})\). Since \(kN^{-1} \in \mathcal{O}(N^{-1+\omega})\), and \(|\varepsilon| \leq \sigma\) holds almost surely, the result follows from \(21\) by treating \(\mathcal{O}(N^{-1+\omega} + N^{-1/2} + \sigma^2)\) as higher order terms.

For \(25\) we only need to show that \(\hat{\mathbb{E}}_{N\Delta}(x^*, k)\varepsilon| \approx N^{-1/3}\) holds with high probability, since the rest follows from \(21\) and \(\sqrt{N\log} \in \mathcal{O}(\sqrt{N})\). Using the assumption on \(\varepsilon_i\) for \(X_i \in N\Delta(x^*, k)\) and Hoeffding’s inequality we have that

\[
\hat{\mathbb{E}}_{N\Delta}(x^*, k)\varepsilon = \frac{1}{k} \sum_{X_i \in N\Delta(x^*, k)} \varepsilon_i \leq \frac{1}{k} \sum_{X_i \in N\Delta(x^*, k)} (\varepsilon_i - \mathbb{E}[\varepsilon_i | X_i \in N\Delta(x^*, k)]) \lesssim \sigma \varepsilon u \frac{N}{k} \approx \sigma \varepsilon u N^{-1/3},
\]

holds with probability \(1 - \exp(-u^2)\), as desired.

\(\square\)

Remark 12 (\(\varepsilon\)-assumption in Corollary \(11\)). The dependency of random variables \(\{\varepsilon_i\}_{i \in J(x^*)}\) for \(J(x^*) = \{i: X_i \in N\Delta(x^*, k)\}\), is an inconvenience arising from our algorithmic approach. It can be avoided by randomly partitioning \(X\) in two equisized sets \(X_1\) and \(X_2\), creating the corresponding \(Y_1\) and \(Y_2\), and then using \((X_1, Y_1)\) to learn local index vectors, and \((X_2, Y_2)\) for function regression. More precisely, after learning index vectors \(\hat{a}_j(X)\) for \(X \in X_1\), for \(X' \in X_2\) we assign index vectors according to \(j(X') = j(\arg\min_{X_j \in X_1} \Delta(X', X_j))\). Then, we use only \(X_2\) for the selection of nearest neighbors and regression which renders the corresponding \(\varepsilon_i\) independent as they do not influence the selection of nearest neighbors.

Note that \(25\) implies that our estimator achieves optimal estimation rates in case of the ordinary SIM, \(i.e.,\) when \(\gamma(t) = at\). Moreover, since Theorem \(8\) is fundamentally a result about local approximations of index vectors, we could track local versions of \(\kappa\) and \(K\) throughout the analysis and replace \(K\) by \(K^*\) in Corollary \(11\). Since in the asymptotic regime \(N \to \infty\) and \(J \approx \sigma^{-1}\), we have by Lemma \(8\)

\[
\max_{\theta} \frac{d_\theta(v^*, V_i)}{\theta - \kappa B} = \frac{K}{\theta - \kappa B} \mathcal{O}(\sigma),
\]

and \(|S(x^*, k)| = \mathcal{O}(K\sigma)\). This implies that we achieve consistent estimation with an \(N^{-1/3}\)-rate for \(x^* \in \text{supp}(\rho_X)\) provided curvature vanishes in a \(\mathcal{O}(K\sigma)\) neighborhood around \(\pi_\gamma(x^*)\).

3.4 Extension to general curves

The results of the previous section are valid only for \(\theta\)-almost linear curves. The issue in extending these results to general curves is that even though subdividing the data set into a larger number of subsets results (for the most part) in pieces of \(\text{Im}(\gamma)\) that are increasingly more linear, the proxy metric \(\Delta(x^*, \cdot)\) will not be a reliable estimator of nearest neighbors in the geodesic metric. In other words, the fact that \(\Delta(x^*, X_i)\) is small does not necessarily imply that \(d_\theta(\pi_\gamma(x^*), \pi_\gamma(X_i))\) is also small. To see this point more vividly, let \(\gamma\) be a segment of the unit circle that contains two antipodal points \(\pi_\gamma(x)\) and \(\pi_\gamma(x')\). In this case \(\Delta(x, x') = 0\) but \(d_\theta(\pi_\gamma(x), \pi_\gamma(x')) = \pi\) and the computation of nearest neighbors for an out-of-sample point \(x^*\) is not reliable.

In order to handle general curves we will thus restrict the space in which we search for the nearest neighbors to \(N\Delta(x^*, k, \eta)\) for a suitably chosen \(\eta\), where we recall that \(N\Delta(x^*, k, \eta)\) denotes the set of \(k\) nearest neighbors of \(x^*\) with respect to \(\Delta(x^*, \cdot)\) in the set \(\mathcal{X} \cap B_{\|\|}(x^*, \eta)\). To conduct the analysis, we require a stronger version of assumption \(25\) namely,

(A2S) Assume \(|W| \leq B := q \pi_\gamma\) holds almost surely for some \(q < 1/2\).

Under \(A2S\) if we set \(\eta\) to be slightly larger than \(2B\) we can isolate points whose projections correspond to the same almost linear piece of \(\text{Im}(\gamma)\) and at the same time guarantee \(\pi_\gamma^{-1}(x^*) \cap \text{supp}(\rho_X) \subset B_{\|\|}(x^*, \eta)\). Namely, this implies that there are sufficiently many \(X_i\) in \(B_{\|\|}(x^*, \eta)\) needed to select good-enough neighbors for kNN.

14
Then we have (see Section 4) using \( \eta \) which implies the choice \( \eta \). Therefore, we have provided that \( \sigma \) and if \( \omega < N \) the asymptotic regime where \( \sigma \), i.e. \( \omega \), grows in \( 1/q \) according to Lemma 13. Moreover, usually \( k \in \mathcal{O}(N^{2/3}) \) and \( \delta \in \mathcal{O}(N^{-1}) \), which implies the choice \( \eta = 2B + \mathcal{O}(N^{-1/3}) \). We add that in most reasonable testing scenarios (see Section 4) using \( \eta = 2B \) works just as well (provided \( B \) is given or reliably estimated).

4 Numerical Experiments

In this section, we present experimental results of the proposed estimator in two settings. First, we conduct synthetic experiments to validate theoretical results of Section 3. Second,
we benchmark the proposed estimator against other commonly used methods on a selection of real-world data sets. The source code for Algorithm 1 and synthetic experiments is available at https://github.com/soply/nsim. Moreover, real-world data sets, code for their pre-processing, and implementations of competing estimators (or references, if publicly available source code is used) are readily available at https://github.com/soply/local_sim_experiments.

### 4.1 Experiments with synthetic data

**General setup.** We consider the Helix curve $\gamma_{\text{Helix}}(\sqrt{2}t) = (\cos(t), \sin(t), t)$, for $t \in [0, \pi]$, and embed it into $\mathbb{R}^D$ for varying $D$. We set $X = V + F(V)U$, where $V$ is sampled uniformly on $\text{Im}(\gamma)$, $U$ is sampled uniformly on $\{ U \in \mathbb{R}^{D-1} : \|U\| \leq 0.25 \}$, and $F(V) \in \mathbb{R}^{D \times (D-1)}$ contains an orthonormal basis of the normal space of $\text{Im}(\gamma)$ at $V$. Figure 4a gives an example of a random sample from such a distribution. The target function $g \circ \gamma^{-1}$ is a strictly monotonic, piecewise quadratic polynomial, as in Figure 4b. Note that as soon as $t \geq \pi \sqrt{2}/2$ there is no $\theta > 0$ such that $\gamma_{\text{Helix}}(t)$ is globally almost-linear (Definition 7). Hence, the search for nearest neighbors of an out-of-sample point $x^*$ is restricted to samples within a Euclidean ball $B + \eta$, as discussed in Section 3.4. Parameters $\eta = 0$ and $B = 0.5$ determine the radius of the intersecting ball. In all experiments we report the root mean squared error (RMSE) averaged over all $J$ level sets for the tangent error, respectively over 1000 test samples for the function error. The results are averaged over 20 repetitions of the same experiment. The standard deviation is indicated by vertical bars.

**Experiment 1:** $\varepsilon = 0$ (noise-free case). For $k = 1$ and $J \asymp N$ we expect an $O(N^{-1})$ error rate for both the tangent estimation error and the function estimation error (Theorem 3 and Corollary 10). In the experiments we use $J = (15D)^{-1}N$, thus ensuring that each level set has $O(D)$ samples for any $N$. We also use $D \in \{6, 12, 24\}$ to study the influence of the ambient dimension.

The results are presented in Figures 5a and 5b. In both cases the experimental error rates match the asserted $N^{-1}$ rate. Moreover, ambient dimension affects the error only through a multiplicative constant. Thus, the estimator does not suffer from the curse of dimensionality. For the sake of comparison, we include the standard kNN estimator in Figure 5b (dashed lines),
in which case, as expected, the error decay deteriorates as $D$ increases.

**Experiment 2:** $\varepsilon \sim \text{Uni}[−\sigma_ε, \sigma_ε]$ (noisy case). The noisy case is used to validate the results of Corollary 1. Since the optimal choice of $J$ is less obvious if $Y$ is affected by noise, we choose it by cross-validation on 5% of the data. $k$ is chosen as $k = 0.5N^{2/3}$. We analyze the case $D = 12$, and $\sigma_ε = c\Delta g$ where $\Delta g := (\max g − \min g)/\pi$ is the average slope, and $c \in \{0.125, 0.25, 0.5, 1.0\}$.

In addition to the experiments for $\gamma_{\text{Helix}}$, we also illustrate results when $\gamma_{\text{Helix}}$ is replaced by $\gamma_{\text{SIM}}(t) := (1, \ldots, 1)^\top t$, which corresponds to an ordinary SIM with $K = 0$. Thus, using the NSIM estimator we expect consistent estimation.

The results for the function error are in Figure 5c. For $\gamma_{\text{SIM}}$, we achieve consistent estimation at the asserted optimal rate $N^{-1/3}$. For $\gamma_{\text{Helix}}$ however, we see a $N^{-1/3}$ decay in the beginning which continues until the noise starts to dominate, after which point the error flattens out. The size of the bias increases linearly in $\sigma_ε$.

### 4.2 Real data

We will now test the NSIM algorithm and compare it to other commonly used algorithms on a variety of real-worlds data sets. We report the mean RMSE and its standard deviation over 30 repetitions of each experiment. In each run, we use 15% of the data as the test set, and we tune hyper-parameters for each estimator using 5-fold cross-validation on exhaustive parameter grids.

**Data sets.** We use 6 UCI data sets (Air Quality, Boston Housing, Concrete Compressive Strength, Istanbul Stock Exchange, Skillcraft1, Yacht) and the Ames Housing data set in our study. For each data set, the components of $X$ are standardized and we exclude clearly irrelevant features. Moreover, if the marginal of $\hat{Y} = \log(Y)$ resembles the uniform distribution better (compared to $Y$), we use $\hat{Y}$ instead of $Y$. The preprocessed the data sets are readily available at [https://github.com/soply/db_hand](https://github.com/soply/db_hand).

**Estimators.**

- **NSIM-dyad**, respectively, **NSIM-stat** refer to Algorithm 1 using a dyadic partition, respectively statistically equivalent blocks. $k$ and $J$ are chosen via cross-validation. The radius is the intersecting Euclidean ball is determined by $\eta = \infty$.
- **Lin-Reg** and **kNN** are standard linear regression and kNN-regression.
Table 1: RMSE, standard deviation, and cross-validated hyper-parameters, over 30 repetitions for several estimators and real-world data sets. Values for $k$, $J$, and for numbers of iterations and nodes, are averages over different runs of each experiment. First 5 rows describe the data sets and their characteristics, and the remaining rows contain the results. For a simplified presentation, we divide the mean and STD of RMSE, and the mean and STD of the data (5th row) by the value in row Factor.

| Characteristics | Yacht | Istanbul | Ames | Concrete | Air Quality | Boston | Skillcraft |
|-----------------|-------|----------|------|----------|-------------|--------|------------|
| log-TF          | Yes   | No       | Yes  | No       | Yes         | Yes    | Yes        |
| $D, N$          | 6, 307| 7, 536   | 7, 1197| 8, 1030 | 11, 7393    | 12, 506 | 16, 3338   |
| Factor          | $10^1$| $10^{-2}$| $10^5$| $10^1$   | $10^{-1}$   | $10^1$  | $10^2$     |
| $\bar{Y}$ ± STD(Y) | 1.05 ± 0.51 | 0.16 ± 2.11 | 1.74 ± 0.67 | 3.58 ± 1.67 | 9.95 ± 4.03 | 1.27 ± 0.71 | 1.15 ± 0.48 |

Method

| Method          | NSIM-dyad $k$ | SNN-Sig $k$ | Isotron $k$ | SIR-kNN $k$ | kNN $k$ | SIR-Sig Nodes | ELM-Sig Nodes | SNN-Tan Nodes | SNN-Sig Nodes |
|-----------------|---------------|-------------|-------------|-------------|--------|---------------|---------------|---------------|---------------|
| $\bar{Y}$ ± STD(Y) | 1.05 ± 0.04 | 1.52 ± 0.14 | 0.23 ± 0.04 | 0.9 ± 0.06  | 0.82 ± 0.04 | 0.42 ± 0.06 | 0.8 ± 0.01   | 0.14 ± 0.03   | 0.14 ± 0.01  |
| $J$             | 2.4 | 1.1 | 2.4 | 3.9 | 5.6 | 1.0 | 4.1 | 5.2 | 9.8 |
| $\bar{Y}$ ± STD(Y) | 1.02 ± 0.03 | 1.39 ± 0.18 | 0.23 ± 0.03 | 0.97 ± 0.06 | 0.80 ± 0.02 | 0.42 ± 0.04 | 0.08 ± 0.01 | 0.14 ± 0.03 | 0.17 ± 0.01 |
| $J$             | 8.6 | 19.3 | 18.2 | 41.6 | 69.3 | 43.0 | 17.7 | 5.2 | 9.8 |

Discussion. The results are presented in Table 1. It is helpful to divide these estimators into two groups. The first group consists of simple estimators (kNN and linear regression) and of estimators that use a reduced (1D) representation of the data (NSIM, SIR and Isotron). The second group are shallow neural networks which search for an estimator in a considerably richer
class of functions. Among the first group, NSIM variants achieve very convincing results as they always belong to the best performing group of estimators. Moreover, experiments suggest that our approach adapts well to the complexity of a given data set. For example, on a data set where linear regression performs best (Istanbul), NSIM achieves roughly the same performance, and automatically chooses (most of the time) \( J = 1 \). On the other hand, for the Concrete data set, where all models that use a single index vector perform rather poorly, the added model flexibility of the NSIM approach proves beneficial, and we achieve the same performance as kNN, despite reducing the dimensionality. This is not the case for SIR-kNN and Isotron, both of which use a linear 1D projection. Finally, on Air Quality and Yacht, NSIM-stat achieves superior performance while leveraging the enhanced model flexibility with \( J \approx 5 \) level sets.

Estimators in the second group enjoy a greater model flexibility, but are at the same time more prone to overfitting. For data sets with a lot of samples (Air Quality, Concrete, and Skillcraft), these methods are better than the estimators in the first group. On the other hand for data sets with smaller sample sizes (Istanbul and Yacht), the model can not be fitted easily, and we observe exactly the opposite effect. Considering the results for the Ames data set, all estimators perform roughly the same.

5 Conclusions

In this paper we propose a nonlinear relaxation of the single index model for data sets with inherent monotonicity between features and outputs. We propose to estimate the model by combining localization through level set partitioning, local linear regression and a kNN type regressor for out-of-sample prediction. Our theoretical results provide guarantees on the error of the quantization of the tangent field of \( \text{Im}(\gamma) \), and yield guarantees for out-of-sample prediction. In the noise free case we provide optimal learning rates, while in the noisy case we generally have a biased estimator. If the NSIM reduces to a SIM, i.e. if \( \text{Im}(\gamma) \) is a straight line, we recover the optimal learning rates for estimating the SIM also in the noisy case.

Our numerical experiments show that the NSIM estimator yields superior results when compared to estimators of similar model complexity. Moreover, the estimator outperforms shallow neural network models on data sets with rather few samples. On the other hand, if the data sets are sufficiently rich to properly fit shallow networks models, their additional flexibility pays off and NSIM does not achieve similar predictive accuracy. Consequently, our future research direction aims at further enhancing the model space of our estimator, by replacing kNN with more sophisticated regressors and learning multiple index vectors, i.e. multi index models, in each level set.

Acknowledgements

TK thanks Prof. Mauro Maggioni, Stefano Vigogna and Alessandro Lanteri for helpful discussions about the project. The authors acknowledge the support of project No 251149/O70 ‘Function-driven Data Learning in High Dimensions’ (FunDaHD) funded by the Research Council of Norway.

A Appendix

A.1 Auxiliary results

Lemma 14. Take an interval \( \mathcal{R} \subset \text{Im}(f) \) and let \( \mathcal{S} \subset \text{Im}(\gamma) \) be the shortest segment such that \( \mathbb{P}(V \in \mathcal{S}|Y \in \mathcal{R}) = 1 \). Then \( L_f^{-1}(|\mathcal{R}| - 2\sigma_\varepsilon) \leq |\mathcal{S}| \leq L_f(|\mathcal{R}| + 2\sigma_\varepsilon). \)
Proof. For any \(V, V' \in \mathcal{S}\) we have \(|Y - Y' - 2\sigma| \leq |f(V) - f(V')| \leq |Y - Y'| + 2\sigma\), almost surely, where \(Y, Y'\) are such that \(Y = f(V) + \epsilon, Y' = f(V') + \epsilon'\). Using (10) we have
\[
d_{\gamma}(V, V') \leq L_f \left| f(V) - f(V') \right| \leq L_f \left( |Y - Y'| + 2\sigma \right) \leq L_f (|\gamma| + 2\sigma) ,
\]
and the upper bound follows after taking the supremum over \(V, V\). For the converse, taking \((X, Y), (X', Y')\) be such that \(|Y - Y'| = |\gamma|\), we have
\[
d_{\gamma}(V, V') \geq L_f^{-1} \left| f(V) - f(V') \right| \geq L_f^{-1} (|\gamma| - 2\sigma).
\]
\[\square\]

Lemma 15. Let \(X_1, \ldots, X_N\), with \(N > 1\), be i.i.d. copies of \(X\). For \(u < N\) we then have
\[
\mathbb{P}\left( \left\{ V_i \right\}_{i=1}^N \text{ is a } \frac{|I| u}{c_V N}\text{-net wrt. } d_{\gamma}(\cdot, \cdot) \right) \geq 1 - \exp(-u) .
\] (28)

Proof. Let \(\epsilon = \frac{|I| u}{c_V N}\), and \(V \in \text{Im}(\gamma)\). Since (A3) implies \(\mathbb{P}(V' \in B_{d_{\gamma}}(V, \epsilon)) > c_{\gamma, \epsilon} |I|^{-1}\), where \(V'\) is an independent copy of \(V\), we have
\[
\mathbb{P}\left( \left\{ V_i \right\}_{i=1}^N \text{ is a } \frac{|I| u}{c_V N}\text{-net w.r.t. } d_{\gamma} \right) = 1 - \mathbb{P}(\exists V : (\forall i) V \notin B_{d_{\gamma}}(V_i, \epsilon))
= 1 - \prod_{i=1}^N \left( 1 - \mathbb{P}(V \in B_{d_{\gamma}}(V_i, \epsilon)) \right) \geq 1 - \exp(-u).
\]
\[\square\]

Lemma 16. Let \(\gamma : \mathbb{R}^D \rightarrow \mathbb{R}^D\) be a \(\theta\)-almost linear curve. Then for \(t' \geq t\) and \(\hat{t}\) arbitrary
\[
\theta d_{\gamma}(\gamma(t), \gamma(t')) \leq \langle \gamma'(\hat{t}), \gamma(t') - \gamma(t) \rangle \leq d_{\gamma}(\gamma(t), \gamma(t')).
\]

Proof. The upper bound follows by Cauchy-Schwartz, \(|\gamma(t) - \gamma(t')| \leq d_{\gamma}(\gamma(t), \gamma(t'))\) and \(\|\gamma'(\hat{t})\| = 1\). For the lower bound the fundamental theorem of calculus gives
\[
\langle \gamma'(\hat{t}), \gamma(t') - \gamma(t) \rangle = \left\langle \gamma'(\hat{t}), \int_t^{t'} \gamma'(s) ds \right\rangle = \int_t^{t'} \langle \gamma'(\hat{t}), \gamma'(s) \rangle ds \geq \theta (t' - t) = \theta d_{\gamma}(\gamma(t), \gamma(t')).
\]
\[\square\]

Lemma 17. Let \(A, B, P_1, P_2 \in \mathbb{R}^{D \times D}\) where \(A\) and \(B\) are Hermitian with \(\text{Im}(A) = \text{Im}(B)\), and \(P_1\) and \(P_2\) are idempotent with \(P_1 + P_2 = \text{Id}\). Then,
\[
P_1 B^\dagger P_1 = \left( P_1 A^\dagger P_1 - P_1 B^\dagger P_2 (B - A) A^\dagger P_1 \right) \left( \text{Id} + P_1 (B - A) A^\dagger P_1 \right)^{-1} ,
\] (29)
\[
P_1 B^\dagger P_2 = \left( P_1 A^\dagger P_2 - P_1 B^\dagger P_1 (B - A) A^\dagger P_2 \right) \left( \text{Id} + P_2 (B - A) A^\dagger P_2 \right)^{-1} ,
\] (30)
provided that the inverses exist.

Proof. Since \(\text{Im}(A) = \text{Im}(B)\), and \(A, B\) are Hermitian, we can use the identity
\[
B^\dagger - A^\dagger = -B^\dagger (B - A) A^\dagger ,
\] (31)
according to [Wed73]. Then (29) follows from
\[
P_1 \left( B^\dagger - A^\dagger \right) P_1 = -P_1 B^\dagger (B - A) A^\dagger P_1 = -P_1 B^\dagger P_1 (B - A) A^\dagger P_1 - P_1 B^\dagger P_2 (B - A) A^\dagger P_1
\Rightarrow P_1 B^\dagger P_1 \left( \text{Id} + P_1 (B - A) A^\dagger P_1 \right) = P_1 A^\dagger P_1 - P_1 B^\dagger P_2 (B - A) A^\dagger P_1.
\]
Expression (30) follows by the same argument. \[\square\]
Covariance bounds. In this part we collect basic bounds on the spectral norm of covariance matrices of $V = \pi_\gamma(X)$ and $W = X - \pi_\gamma(X)$, given that $Y \in \mathcal{R}$ for some closed interval $\mathcal{R} \subset [0, 1]$, such that $|\mathcal{R}| < R_{\text{max}}$. These will be used throughout. We define $\beta = E[t|\mathcal{R}]$, $a = \gamma'(\beta)$, and let $P_\beta = aa^\top$, and $P_\perp = \text{Id} - P_\beta$ be orthogonal projections onto local tangent and normal spaces.

Lemma 18. Assume $2\sigma_x < |\mathcal{R}| < (2KL_f)^{-1}$, and denote by $\mathcal{S} \subset \text{Im}(\gamma)$ the shortest segment such that $P(V \in \mathcal{S}|Y \in \mathcal{R}) = 1$. Then the following hold

\[
\frac{(1-2KL_f)^2 |\mathcal{R}|^2}{27L_f^2} (|\mathcal{R}| - 2\sigma_x)^2 \leq \text{Cov}(\langle a, V \rangle |\mathcal{R}) \leq 1/2 |\mathcal{S}|^2, \tag{32}
\]

\[
\|\text{Cov}(P_\beta W |\mathcal{R})\| \leq |\mathcal{S}|^2, \tag{33}
\]

\[
\|\text{Cov}(P_\perp V |\mathcal{R})\| \leq E[\|P_\perp (V - E[V|\mathcal{R}])\|^2 |\mathcal{R}] \leq 1/2K^2 |\mathcal{S}|^4, \tag{34}
\]

\[
\|\text{Cov}(V, P_\perp V |\mathcal{R})\| \leq 1/2K |\mathcal{S}|^3, \tag{35}
\]

\[
\|\text{Cov}(P_\beta W, P_\perp W |\mathcal{R})\| \leq (C_W + B^2/2K + B^2K^2) |\mathcal{S}|^2, \tag{36}
\]

\[
\|E[a(V) - a|\mathcal{R}|| \leq 1/4K |\mathcal{S}|^2, \tag{37}
\]

\[
\|X - \mu_X\| \leq B + |\mathcal{S}|, \text{ and } |\langle a, X - \mu_X\rangle| \leq 2 |\mathcal{S}|, \ Y \in \mathcal{R}. \tag{38}
\]

Proof. We begin with the lower bound in (32). Using the mean value theorem and the identity $E[|Z - E[Z]|^2] = 1/2E[|Z - Z'|^2]$ we have

\[
\text{Cov}(\langle a, V \rangle |\mathcal{R}) = E[\langle a, V - E[V|\mathcal{R}] \rangle^2 |\mathcal{R}] = \frac{1}{2}E[\langle a, V - V' \rangle^2 |\mathcal{R}] = \frac{1}{2}E[(t - t')^2 \langle a, \gamma'(t) \rangle^2 |\mathcal{R}]
\]

\[
\geq \frac{1}{2} \min_{s: \gamma(s) \in \mathcal{S}} \langle a, \gamma'(s) \rangle^2 E[(t - t')^2 |\mathcal{R}] = \min_{s: \gamma(s) \in \mathcal{S}} \langle a, \gamma'(s) \rangle^2 E[(t - t')^2 |\mathcal{R}]. \tag{39}
\]

For the first term note that $\langle a, \gamma'(s) \rangle \geq 1 - K |\mathcal{S}|$, and thus $\langle a, \gamma'(s) \rangle \geq 1 - 2KL_f |\mathcal{R}|$ because of Lemma 14 and $|\mathcal{R}| > 2\sigma_x$. For the second term the law of total expectation, for any $c > 0$, gives

\[
E[(t - t')^2 |\mathcal{R}] \geq E[(t - \beta)^2 |t - \beta] > c |\mathcal{R}| P(t - \beta > c |\mathcal{R}) \geq c^2 P(t - \beta > c |\mathcal{R}).
\]

Let $\mathcal{I} := \gamma^{-1}(f^{-1}([\text{inf}\mathcal{R} + \sigma_x, \text{sup}\mathcal{R} - \sigma_x]))$, so that $|\mathcal{I}| > L_f^{-1} |\mathcal{R} - 2\sigma_x|$. Thus

\[
P(|t - \beta] > c |\mathcal{R}) = 1 - \frac{P(t - \beta < c)}{P(Y \in \mathcal{R})} \geq 1 - \frac{2cc^2\beta}{c^2 |\mathcal{I}|} \geq 1 - \frac{2cL_f}{cV(|\mathcal{R} - 2\sigma_x|)}.
\]

Setting $c = 1/3cv L_f^{-1}(|\mathcal{R} - 2\sigma_x)$ yields the desired bound. The upper bound, on the other hand, follows from

\[
\text{Cov}(\langle a, V \rangle |\mathcal{R}) \leq E[|V - E[V|\mathcal{R}]|^2 |\mathcal{R}] \leq \frac{1}{2}E[|V - V'|^2 |\mathcal{R}].
\]

For (33) we have by the law of total variance, $\kappa B < 1$, and (A1)

\[
|\text{Cov}(\langle a, W \rangle |\mathcal{R})| = |E_V[\text{Cov}(\langle a, W \rangle |\mathcal{R})]| \leq E_V \left[ \langle \text{Cov}(W|V), (a - a(V))(a - a(V))^\top \rangle |\mathcal{R} \right] \leq E_V \left[ \|\text{Cov}(W|V)\| |a - a(V)||^2 |\mathcal{R} \right] \leq B^2 \kappa^2 |\mathcal{S}|^2 \leq |\mathcal{S}|^2.
\]

For (34), we have by the fundamental theorem of calculus

\[
\|\text{Cov}(P_\perp V |\mathcal{R})\| \leq \frac{1}{2}E[\|P_\perp (V - V')\|^2 |\mathcal{R}] \leq \frac{1}{2}E \left[ \int_\beta^\beta \left( \|P_\perp (\gamma'(s) - \gamma'(t))\| ds \right)^2 |\mathcal{R} \right] \leq \frac{1}{2}K^2 |\mathcal{S}|^4,
\]

\footnote{This is short hand for $E[\gamma^{-1}(\gamma(x_\gamma))]<Y \in \mathcal{R}]$}
where we used $P_\perp \perp \gamma'(\bar{t})$. Moreover, \([35]\) follows by Cauchy-Schwartz, \([32]\) and \([34]\), whereas \([37]\) follows from Taylor’s theorem and the fact that $\bar{t} = E[\bar{t}|\mathcal{R}]$.

For \([36]\), note that $\|P\| W \| \leq |S|$ would yield only a linear order in $|S|$. Instead, we will use \([A4.2]\) to achieve the quadratic order. We begin by using the law of total covariance and $E[W|V] = 0$. For $P_{\perp}\perp := a(V)a(V)^T$ this yields
\[
\text{Cov} \left( P_{\perp} W, P_{\perp} W | \mathcal{R} \right) = E_V [\text{Cov} \left( P_{\perp} W, P_{\perp} W | V \right) | \mathcal{R}] = E_V [(P_{\perp} - P_{\perp}\perp) \text{Cov} \left( W | V \right) P_{\perp} | \mathcal{R}],
\]
where we used $a(V) \perp W$ in the last expression. Adding and subtracting $\text{Cov} \left( W | \mathcal{R} \right)$ now yields
\[
E_V [(P_{\perp} - P_{\perp}\perp) \text{Cov} \left( W | V \right) P_{\perp} | \mathcal{R}] = E_V [(P_{\perp} - P_{\perp}\perp) \left( \text{Cov} \left( W | V \right) - \text{Cov} \left( W | \mathcal{R} \right) \right) P_{\perp} | \mathcal{R}] + E_V [(P_{\perp} - P_{\perp}\perp) | \mathcal{R} \text{Cov} \left( W | \mathcal{R} \right) P_{\perp}].
\]

The Lipschitz assumption on the sliced covariance \([A4.2]\) can be used for the first term
\[
E_V \left[ \| (P_{\perp} - P_{\perp}\perp) \left( \text{Cov} \left( W | V \right) - \text{Cov} \left( W | \mathcal{R} \right) \right) P_{\perp} | \mathcal{R} \right] \leq \sqrt{E_V \left[ \| P_{\perp} - P_{\perp}\perp \|^2 | \mathcal{R} \right]} \sqrt{E_V \left[ \| \text{Cov} \left( W | V \right) - \text{Cov} \left( W | \mathcal{R} \right) \|^2 | \mathcal{R} \right]} \leq K C_W |S|^2.
\]

For the second term in \([40]\), factoring the difference of outer products gives
\[
E_V \left[ \| a a^T - a(V)a(V)^T \| \right] \leq E[\| a - a(V) \|^2 | \mathcal{R}] + 2 \| E[a - a(V)|\mathcal{R}] \|= \leq B^2 |S|^2.
\]

Concentration results. Standard concentration results used in the proof of Theorem 3

**Lemma 19.** Let $A \in \mathbb{R}^{d_A \times D}$ and $B \in \mathbb{R}^{D \times d_B}$, and assume $\|A(X - E X)\| \leq C_A, \|B^T(X - E X)\| \leq C_B$ almost surely. Let $\hat{\Sigma}$ be the sample mean, and $\Sigma$ the sample covariance from $N$ i.i.d. copies of $X$. For any $u > 1$, we have with probability at least $1 - \exp(-u^2)$
\[
\| A (E X - \hat{E}X) \| \leq C_A u / \sqrt{N}, \quad \text{and} \quad \| A (\Sigma - \hat{\Sigma}) B \| \leq \frac{\log(D) C_A C_B u^2}{\sqrt{N}} + O(C_A C_B N^{-1}).
\]

**Proof.** The first bound is a standard result that follows from the bounded differences inequality \([33]\). For \([41]\) denote $\Sigma = \E (X - EX)(X - EX)^T$. Then, \([32]\) and Remark \([33]\) give
\[
\| A (\Sigma - \hat{\Sigma}) B \| \leq \| A (\Sigma - \hat{\Sigma}) B \| + \| A (\hat{E}X - EX) \| \| (\hat{E}X - EX)^T B \|.
\]

By the first result in \([41]\), the second term is of order $O(C_A C_B N^{-1})$ with probability $1 - 2 \exp(-u^2)$. Denote $S_k := \frac{1}{N} A \hat{X}_k \hat{X}_k^T B - \frac{1}{N} A \Sigma B$ and $S := \sum_{k=1}^N S_k$, where $\hat{X}_k = X_k - EX$. Since $\E [\hat{X}_k \hat{X}_k^T] = \Sigma$ we have $\E[S] = 0$, and since $\hat{X}_k$ and $\hat{X}_j$ are independent for $k \neq j$ we get $\E[S_k S_j^T] = \E[S_k] \E[S_j^T] = 0$. Thus,
\[
\E[S S^T] = \sum_{k=1}^N \E[S_k S_k^T] + \sum_{k \neq j} \E[S_k S_j^T] = \sum_{k=1}^N \E[S_k S_k^T] + \Omega(N^{-1/2} C_A C_B).
\]

Since $\|S_k\| \leq 2N^{-1}C_A C_B$ holds almost surely we have $\|E S S^T\| \leq 4N^{-1}C_A^2 C_B^2$ and by an analogous argument we have the same bound for $\|E S^T S\|$. Thus, the variance statistic (cf. Remark \([33]\) satisfies $\sqrt{m(S)} \leq 2N^{-1/2} C_A C_B$ and Theorem \([32]\) yields the desired result. \(\square\)
Cross-covariance of $X$ and $Y$. We adopt the notation of the paragraph Covariance bounds, and for a sample $(X_1, Y_1), \ldots, (X_N, Y_N)$, such that $Y_i \in \mathcal{R}$ a.s., we define the sample cross-covariance $\hat{r} := \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$, and the population version $r := \text{Cov}(X, Y|\mathcal{R})$. In the next two Lemmas, we decompose $r$ into gradient, curvature, and noise contributions, and bound the estimation error $\|r - \hat{r}\|$. The results will be used in the proof of Theorem 3.

**Lemma 20.** Let $\gamma := \gamma(\bar{t})$ and assume $|\mathcal{R}| > 2\sigma_z$. We can write $r = T_1 + T_2 + T_3$ with

$$T_1 = \|\nabla f(\gamma)\| \text{Cov}(V|\mathcal{R}) \ a, T_2 = \frac{1}{2} \text{Cov} \left( V, (V - \gamma)^\top \nabla^2 f(\eta)(V - \gamma)|\mathcal{R} \right), T_3 = \text{Cov}(V, \varepsilon|\mathcal{R}),$$

where $\eta \in [V, \gamma]$. Moreover, the following bounds hold

$$\|T_1\| \leq \frac{1}{2}L_f |\mathcal{S}|^2, \quad \|T_2\| \leq \frac{1}{\sqrt{2}} L_H |\mathcal{S}|^3, \quad \|T_3\| \leq C_2 L_f |\mathcal{R}|^3, \quad \|P_{\perp}r\| \leq 2KL_f^2 |\mathcal{R}|^3.$$

**Proof.** Since $\mathbb{E}[W|Y] = 0$ holds a.s. by assumptions [A1] and [A5] we have, by the law of total covariance,

$$\text{Cov}(W, Y|\mathcal{R}) = \mathbb{E}_Y[\text{Cov}(W, Y|Y)|\mathcal{R}] + \text{Cov}_Y(\mathbb{E}[W|Y], Y|\mathcal{R}) = 0. \quad (42)$$

Therefore, we can write $r = \text{Cov}(X, Y|\mathcal{R}) = \text{Cov}(V, Y|\mathcal{R}) = \text{Cov}(V, f(V)|\mathcal{R}) + \text{Cov}(V, \varepsilon|\mathcal{R})$, where the last term is $T_3$. Next, we note that the model $f(x) = f(\pi_s(x))$ implies for $n \in \text{Im}(P_{\perp})$

$$\langle \nabla f(\gamma), n \rangle = \lim_{h \to 0} \frac{f(\gamma + hn) - f(\gamma)}{h} = \lim_{h \to 0} \frac{f(\pi_s(\gamma + hn)) - f(\gamma)}{h} = \lim_{h \to 0} \frac{f(\gamma) - f(\gamma)}{h} = 0,$$

and therefore $\nabla f(\gamma) \in \text{Im}(P_{\perp})$, respectively. $\nabla f(\gamma) = \|\nabla f(\gamma)\| a$. Using this, $\text{Cov}(V, f(V)|\mathcal{R}) = \text{Cov}(V, f(V) - f(\gamma)|\mathcal{R})$ and Taylor's Theorem we get

$$\text{Cov}(V, f(V) - f(\gamma)|\mathcal{R}) = \text{Cov} \left( V, (V - \gamma)^\top \nabla f(\gamma)|\mathcal{R} \right) + \frac{1}{2} \text{Cov} \left( V, (V - \gamma)^\top \nabla^2 f(\eta)(V - \gamma)|\mathcal{R} \right)$$

$$= \|\nabla f(\gamma)\| \text{Cov}(V|\mathcal{R}) a + \frac{1}{2} \text{Cov} \left( V, (V - \gamma)^\top \nabla^2 f(\eta)(V - \gamma)|\mathcal{R} \right) = T_1 + T_2,$$

where $\eta = \eta(V, \gamma)$. The bound on $T_1$ is implied by $\|\nabla f(\gamma)\| \leq L_f$, and $\|\text{Cov}(V|\mathcal{R})\| \leq 1/2 |\mathcal{S}|^2$, which has been shown in [32]. For $T_2$ on the other hand, we write out the covariance and use Cauchy-Schwartz together with $\|\nabla^2 f(x)\| \leq L_H$ to get

$$\|T_2\| \leq \frac{1}{2}L_H |\mathcal{S}|^2 \sqrt{\mathbb{E}[\|V - \mathbb{E}[V]|\mathcal{R}\|] |\mathcal{S}|^2} = \frac{1}{2}L_H |\mathcal{S}|^2 \sqrt{1/2 \mathbb{E}[\|V - \mathbb{E}[V]|\mathcal{R}\|] |\mathcal{S}|^2} \leq \frac{1}{\sqrt{8}} L_H |\mathcal{S}|^3.$$

The bound on $T_3$ follows from assumption [A5]. Finally, for $P_{\perp}r$, we get by Lemma 14 and [34]

$$\|P_{\perp}r\| \leq \|\text{Cov}(P_{\perp}V, Y|\mathcal{R})\| \leq \sqrt{\mathbb{E}[\|P_{\perp}(V - \mathbb{E}[V]|\mathcal{R})|\mathcal{R}\|]} \sqrt{\mathbb{E}[\|Y - \mathbb{E}[Y]|\mathcal{R}\|]} \leq 2KL_f^2 |\mathcal{R}|^3.$$

**Lemma 21.** Let $Y \in \mathcal{R}$ and $\|A(X - \mathbb{E}X)\| \leq C_A$ for some $A \in \mathbb{R}^{d \times d}$ almost surely. For any $u > 1$ we then have $\|A(r - \hat{r})\| \leq \frac{uC_A |\mathcal{R}|}{N^{1/2}} + \mathcal{O}(C_A |\mathcal{R}| N^{-1})$, with probability at least $1 - 2 \exp(-u^2)$.

**Proof.** With $Z_i := (X_i - \mathbb{E}X)(Y_i - \mathbb{E}Y) - \text{Cov}(X, Y)$, we can rewrite

$$A(r - \hat{r}) = \mathbb{E}AZ_i + (\mathbb{E}AX_i - \mathbb{E}AX)(\mathbb{E}Y - \mathbb{E}Y_i).$$

We can neglect the second term because it can be bounded with concentration inequalities in Lemma 19 by $\mathcal{O}(C_A |\mathcal{R}| N^{-1})$. For the first term, we notice that $\mathbb{E}Z_i = 0$, and that

$$\|AZ_i\| \leq \|A(X_i - \mathbb{E}X)\| \|Y_i - \mathbb{E}Y\| + \sqrt{\mathbb{E}[\|Y - \mathbb{E}Y\|^2]} \sqrt{\mathbb{E}[\|A(X - \mathbb{E}X)\|^2]} \leq 2C_A |\mathcal{R}|.$$

The results follows from [11] in Lemma 19. 

23
A.2 Proof of Theorem 3

Throughout this proof, \( \mathcal{R} \subset [0,1] \) is fixed, with \( |\mathcal{R}| < R_{\text{max}} \) and we denote \( \bar{t} := \mathbb{E}[t|\mathcal{R}], \ a = \gamma'(\bar{t}), \ P_0 = aa^T, \ P_1 = \mathrm{Id} - P_0, \ S \subset \text{Im}(\gamma) \) is the infimum of all connected pieces of \( \text{Im}(\gamma) \) such that \( \mathbb{P}(V \in S|\mathcal{R}) = 1 \). Moreover we drop the subscript \( \mathcal{R} \) from all quantities (e.g. from \( N, \Sigma, r, \hat{r}, \kappa \)) to ease the notation, and we use \( \lesssim \) and \( \gtrsim \) to absorb universal factors for terms that decay with the sample size \( N \).

As a reminder, in Theorem 3 we want a bound on index vector estimation, i.e. \( \|\hat{a} - a\| \), where \( a = \gamma'(\bar{t}) \) and \( \hat{a} = \frac{b}{\|b\|} \), for \( b = \Sigma^\dagger \hat{r} \). A direct calculation (77) yields
\[
\|\hat{a} - a\| \leq \sqrt{2\|P_\parallel \hat{b}\| - 1\|P_\perp \hat{b}\|}.
\]

Thus, it suffices to bound the tangential component \( \|P_\parallel \hat{b}\| \) from below and bound the normal component \( \|P_\perp \hat{b}\| \) from above. A meticulous analysis of these bounds is a technically tedious task. Specifically, we have to decompose these two factors into separate terms that reflect the spectral behaviour of \( \Sigma \) on \( \text{Im}(P_0) \) and \( \text{Im}(P_\perp) \), which provide different dependencies on \( |\mathcal{R}| \).

To analyze \( \hat{\Sigma} \) we will first analyze its population version \( \Sigma \). We can then transfer the bounds to the finite sample case with the help of concentration inequalities, and Lemma 17. To ease the notation, we use \( \lambda := \text{Cov} (\langle a, X \rangle |\mathcal{R}) \). Moreover, \( C_\Delta := 2(C_W + B^2 + |S|)k + 2B^2k^2 \) will be used to characterize the cumulative effect of curvature on our results, and \( B_\perp := B + |S| \) is the maximum distance of \( X \) from \( \mu_X \).

Write now the cross-components of \( \Sigma \) as
\[
\Sigma - \Sigma_P = \text{Cov} (P_\parallel X, P_\perp X|\mathcal{R}) + \text{Cov} (P_\perp X, P_\parallel X|\mathcal{R}) ,
\]
where
\[
\Sigma_P := \text{Cov} (P_\parallel X|\mathcal{R}) + \text{Cov} (P_\perp X|\mathcal{R}) .
\]

Lemma 22. We have \( \|\Sigma - \Sigma_P\| \leq C_\Delta |S|^2 \).

Proof. \([A1]\) implies \( \text{Cov} (V, W) = 0 \) by the law of total covariance. Therefore,
\[
\text{Cov} (P_\parallel X, P_\perp X|\mathcal{R}) = \text{Cov} (P_\parallel W, P_\perp W|\mathcal{R}) + \text{Cov} (P_\parallel V, P_\perp V|\mathcal{R}) .
\]

Using equations (35) and (36) we thus have
\[
\|\Sigma - \Sigma_P\| \leq 2 \left( \left( C_W + \frac{B^2}{2} \right) k + B^2k^2 \right) |S|^2 + k |S|^3 \leq C_\Delta |S|^2 .
\]

\( \square \)

Lemma 23. If \( 2\sigma_x < |\mathcal{R}| < (2kL_f)^{-1}, \) then \( \text{Im}(\Sigma) = \text{Im}(\Sigma_P) \).

Proof. First note that \( \text{Im}(\Sigma_P) = \text{Im}(P_\parallel \Sigma P_\parallel) \oplus \text{Im}(P_\perp \Sigma P_\perp) \subset \text{Im}(P_\parallel \Sigma) \oplus \text{Im}(P_\perp \Sigma) = \text{Im}(\Sigma) \). Therefore, it suffices to show \( \text{rank}(\Sigma_P) = \text{rank}(\Sigma) \). Since \( \text{Cov} (\langle a, V \rangle |\mathcal{R}) > 0 \) by (32), we have \( \text{rank}(\Sigma_P) = \text{rank}(P_\parallel \Sigma_P P_\parallel) + \text{rank}(P_\perp \Sigma_P P_\perp) = 1 + \text{rank}(P_\perp \Sigma_P P_\perp) \). To find a lower bound for \( \text{rank}(P_\perp \Sigma_P P_\perp) \), we note that, by (A4.1), any unit norm \( v \in \text{Im}(\Sigma) \cap \text{Im}(P_\perp) \) obeys
\[
v^T P_\perp \Sigma_P P_\perp v = v^T \Sigma_P v = v^T \Sigma v > C_\perp .
\]

Therefore, \( \text{rank}(P_\perp \Sigma_P P_\perp) \geq \dim(\text{Im}(\Sigma) \cap \text{Im}(P_\perp)) \). The result now follows due to \( \dim(\text{Im}(\Sigma) \cap \text{Im}(P_\perp)) = \text{rank}(\Sigma) - \dim(\text{Im}(\Sigma) \cap \text{Im}(P_\parallel)) \geq \text{rank}(\Sigma) - 1 \).

\( \square \)

Lemma 24. If \( 2\sigma_x < |\mathcal{R}| < (2kL_f)^{-1}, \) we have \( P_\parallel \Sigma_P P_\parallel = 0, \) and
\[
\left\| P_\parallel \Sigma_P P_\parallel \right\| = \frac{1}{\lambda}, \quad \left\| P_\perp \Sigma_P P_\perp \right\| \leq \frac{1}{C_\perp} .
\]

(46)
Thus, $A$ with where $\{u_2, \ldots, u_d\}$ is the eigen-system for $P_\perp \Sigma P_\perp$. As $\Sigma^\perp_p$ admits the same eigen-decomposition as $\Sigma_p$, with the eigenvalues inverted, we have $P_\parallel \Sigma^\perp_p P_\perp = 0$. Furthermore, $\|P_\parallel \Sigma^\perp_p P\| = 1/\lambda$ and $\|P_\perp \Sigma^\perp_p P_\perp\| \leq 1/C_\perp$ by (A4.1).

**Lemma 25.** If $2\sigma_\varepsilon < |R| < (2KL_f)^{-1}$, and $|S|^4 < C_\perp \lambda/C_\Delta$, we have

$$
\begin{align*}
\|P_\perp \Sigma^\dagger P\| &= \frac{1}{\lambda} + \mathcal{O}(\kappa), \\
\|P_\perp \Sigma^\dagger P_\perp\| &\leq \frac{1}{C_\perp} + \mathcal{O}(\kappa |S|^2), \quad \text{and} \quad \|P_\parallel \Sigma^\dagger P\| \leq \frac{C_\Delta |S|^2}{\lambda} + \mathcal{O}(\kappa |S|^2).
\end{align*}
$$

**Proof.** First note that by (44) we have $P_\perp (\Sigma - \Sigma P) P_\parallel = 0$ and $P_\perp (\Sigma - \Sigma P) P_\parallel = 0$. Moreover, $P_\parallel \Sigma^\dagger_p P_\perp = 0$ holds by Lemma 24. Since $\text{Im}(\Sigma P) = \text{Im}(\Sigma)$ by Lemma 23, we can use Lemma 17 with $A = \Sigma_p$ and $B = \Sigma$, yielding

$$
\begin{align*}
P_\perp \Sigma^\dagger P_\parallel &= P_\perp \Sigma^\dagger P_\perp - P_\parallel \Sigma^\dagger P_\parallel (\Sigma - \Sigma P) \Sigma^\dagger P_\perp, \\
P_\parallel \Sigma^\dagger P_\parallel &= -P_\parallel \Sigma^\dagger P_\parallel (\Sigma - \Sigma P) \Sigma^\dagger P_\parallel, \\
P_\parallel \Sigma^\dagger P_\parallel &= P_\parallel \Sigma^\dagger P_\parallel (\Sigma - \Sigma P) \Sigma^\dagger P_\parallel.
\end{align*}
$$

Consider first (48). By plugging (51) into (52) and rearranging the terms, we get

$$
P_\parallel \Sigma^\dagger P_\parallel \left( \text{Id} + (\Sigma - \Sigma P) P_\perp \Sigma^\dagger P_\parallel (\Sigma - \Sigma P) P_\parallel \Sigma^\dagger P_\parallel \right) = P_\parallel \Sigma^\dagger P_\parallel.
$$

Moreover, using Lemma 22 and 24 it follows

$$
\|H\| = \| (\Sigma - \Sigma P) P_\perp \Sigma^\dagger P_\parallel (\Sigma - \Sigma P) P_\parallel \Sigma^\dagger P_\parallel \| \leq \| \Sigma - \Sigma P \|^2 \| P_\perp \Sigma^\dagger P_\perp \| \| P_\parallel \Sigma^\dagger P_\parallel \| \leq \frac{C_\Delta^2 |S|^4}{C_\perp \lambda}
$$

Thus, $\text{Id} + H$ is invertible due to the assumption, and (48) follows after inverting, using (46) and $\lambda \in \mathcal{O}(|S|^2)$ according to Lemma 14 and 18. For the remaining bounds, we first use (51) with Lemma 22 and 24 to get

$$
\|P_\parallel \Sigma^\dagger P_\parallel\| \leq \|P_\parallel \Sigma^\dagger P_\parallel\| \| \Sigma - \Sigma P \| \| \Sigma^\dagger P_\parallel \| \leq \frac{C_\Delta |S|^2}{\lambda} + \mathcal{O}(\kappa |S|^2),
$$

and then (50) with Lemma 22 and 24 and $\lambda \in \mathcal{O}(|S|^2)$ imply

$$
\|P_\perp \Sigma^\dagger P_\perp\| \leq \|P_\perp \Sigma^\dagger P_\perp\| + \|P_\parallel \Sigma^\dagger P_\parallel\| \| \Sigma - \Sigma P \| \| \Sigma^\dagger P_\parallel \| \leq \frac{1}{C_\perp} + \mathcal{O}(\kappa |S|^2).
$$

We shall now transfer these bounds to the sample version $\Sigma^\dagger$ by following a similar argumentation. A minor difficulty is that the estimation error $\hat{\Sigma} - \Sigma$ has different spectral bounds, with respect to $|S|$ and $N$, in tangential and in normal directions. Balancing these contributions is critical for Theorem 3.
Lemma 26. Let $\text{Im}(\hat{\Sigma}) = \text{Im}(\Sigma)$, and $u > 1$. If $2\sigma_k < |R| < (2KLf)^{-1}$, $|S|^4 < C_\perp \lambda / C_\Delta$ and

$$N^{1/2} \gtrsim \max \left\{ \frac{|S|^2}{\lambda}, \frac{B_+^2 (1 + C_\Delta |S|^2 \lambda^{-1})}{C_\perp} \right\} \log(D) u^2,$$

we have, with probability at least $1 - \exp(-u^2)$

$$\left\| P_\perp \hat{\Sigma}^\dagger P_\perp \right\| - \lambda^{-1} \lesssim \frac{\log(D) u^2 |S|^2}{N^{1/2}} + O(K + K |S|^{-1} N^{-1/2}),$$

(53)

$$\left\| P_\perp \hat{\Sigma}^\dagger P_\perp \right\| \leq \frac{1}{C_\perp} + O(K |S|^2 + N^{-1/2}),$$

(54)

$$\left\| P_\parallel \hat{\Sigma}^\dagger P_\parallel \right\| \lesssim \frac{C_\Delta |S|^2}{\lambda} + \frac{\log(D) B_+ u^2 \left(1 + C_\Delta |S|^2 \lambda^{-1}\right) |S|}{C_\perp \lambda N^{1/2}} + O(K |S| + KN^{-1/2}).$$

(55)

Proof. As before, we apply Lemma 17, this time with $A = \Sigma$ and $B = \hat{\Sigma}$, yielding

$$P_\parallel \hat{\Sigma}^\dagger P_\parallel \left( \text{Id} + P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel \right) = P_\parallel \Sigma^\dagger P_\parallel + P_\parallel \hat{\Sigma}^\dagger P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel,$$

(56)

$$P_\parallel \hat{\Sigma}^\dagger P_\parallel \left( \text{Id} + P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel \right) = P_\parallel \Sigma^\dagger P_\parallel + P_\parallel \hat{\Sigma}^\dagger P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel,$$

(57)

$$P_\parallel \hat{\Sigma}^\dagger P_\parallel \left( \text{Id} + P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel \right) = P_\parallel \Sigma^\dagger P_\parallel + P_\parallel \hat{\Sigma}^\dagger P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel.$$

(58)

Before bounding intermediate terms, we collect a list of intermediate bounds that will be needed. Using Lemma 19 and Lemma 25, we have with probability at least $1 - \exp(-u^2)$

$$\left\| \Sigma^\dagger P_\perp \right\| \leq \frac{1 + C_\Delta |S|^2 \lambda^{-1}}{C_\perp} + O(K |S|^2),$$

(59)

$$\left\| P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel \right\| \lesssim \frac{\log(D) u^2 |S|^2}{N^{1/2}} + O(N^{-1}),$$

(60)

$$\left\| \hat{\Sigma} - \Sigma \right\| \lesssim \frac{\log(D) B_+ u^2 |S|^2}{N^{1/2}} + O(N^{-1}),$$

(61)

We begin with (53). The idea is to substitute $P_\parallel \hat{\Sigma}^\dagger P_\parallel$ in (56) with the left-hand side term in (57) and thus find $P_\parallel \hat{\Sigma}^\dagger P_\parallel$. First, note that (59) and (61) imply $\|P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel \leq O(N^{-1/2})$. Thus, we can invert $\text{Id} + P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel$ (since $N$ is large enough). Therefore, (57) yields

$$P_\parallel \hat{\Sigma}^\dagger P_\parallel = \left( P_\parallel \Sigma^\dagger P_\parallel + P_\parallel \hat{\Sigma}^\dagger P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel \right)^{-1},$$

(62)

with $\| (\text{Id} + P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel)^{-1} \| = 1 + O(N^{-1/2})$, by a Neumann series argument. Plugging this into (56), and rearranging the terms we get

$$P_\parallel \hat{\Sigma}^\dagger P_\parallel (\text{Id} + H) = P_\parallel \Sigma^\dagger P_\parallel + P_\parallel \hat{\Sigma}^\dagger P_\parallel \left( \text{Id} + P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel \right)^{-1} (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel,$$

(63)

with $H := P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel + P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel \left( \text{Id} + P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel \right)^{-1} (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel$.

We now repeat the same trick, i.e. show that $\text{Id} + H$ is invertible. Using (48), (49), (59), (61), we have $\|P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel || = O(|S|^{-1} N^{-1/2})$ and

$$\left\| (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel \right\| \leq \left\| (\hat{\Sigma} - \Sigma) \Sigma^\dagger \right\| + \left\| (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel \right\| = O(|S|^{-1} N^{-1/2}) + O(N^{-1/2}),$$

which implies $H = P_\parallel (\hat{\Sigma} - \Sigma) \Sigma^\dagger P_\parallel + O(N^{-1})$. Therefore, (60) implies the eigenvalue concentration

$$\|\text{Id} + H\| - 1 \leq \frac{\log(D) u^2 |S|^2}{N^{1/2}} + O(N^{-1}).$$
Under the assumption on $N$ in the claim, we can invert $\text{Id} + H$ in equation (63) to get

$$
\left\|P_\parallel \hat{\Sigma} P_\parallel \right\| = \left\|P_\parallel \Sigma P_\parallel (\text{Id} + H)^{-1} \right\| + \left\|P_\parallel \Sigma P_\parallel \left( (\hat{\Sigma} - \Sigma) \Sigma P_\parallel \right) \right\| = O(1 + N^{-1/2}).
$$

(64)

For the second term in (64), we can use (48), (49), (61), and $P_\perp + P_\parallel = \text{Id}$ to find

$$
\left\|P_\parallel \Sigma P_\parallel \left( (\hat{\Sigma} - \Sigma) \Sigma P_\parallel \right) \right\| = O(k) \left( O(|S| N^{-1/2}) O(|S|^{-1}) + O(N^{-1/2}) O(k) \right),
$$

(65)

Plugging (48) and (65) into (64) the desired bound follows.

For the bound (65), we plug (49), (53), (63), (61), and $\|\text{Id} + P_\parallel (\hat{\Sigma} - \Sigma) \Sigma P_\parallel^{-1}\| = 1 + O(N^{-1/2})$ into Equation (62) to get

$$
\left\|P_\parallel \hat{\Sigma} P_\parallel \right\| \leq \left( \left\|P_\parallel \Sigma P_\parallel \right\| + \left\|P_\parallel \hat{\Sigma} P_\parallel \right\| \left\|P_\parallel (\hat{\Sigma} - \Sigma) \Sigma P_\parallel \right\| \right) \left( 1 + O(N^{-1/2}) \right)
$$

$$
\leq C_\Delta |S|^2 \frac{2 \log(D) B \Delta u^2}{C_\perp N^{1/2} \Delta} + O(k |S| + KN^{-1/2}).
$$

Consider now (54). Since $\|\text{Id} + P_\perp (\hat{\Sigma} - \Sigma) \Sigma P_\perp^{-1}\| \leq 1 + O(N^{-1/2})$, equation (58) implies

$$
\left\|P_\perp \hat{\Sigma} P_\perp \right\| \leq \left( \left\|P_\perp \Sigma P_\perp \right\| + \left\|P_\perp \hat{\Sigma} P_\perp \right\| \left\|P_\perp (\hat{\Sigma} - \Sigma) \Sigma P_\perp \right\| \right) \left( 1 + O(N^{-1/2}) \right),
$$

(66)

Using (55), (59), and (61) for the last term, and (49) for the first term, we have

$$
\left\|P_\perp \hat{\Sigma} P_\perp \right\| \leq \left( \frac{1}{C_\perp} + O(k |S|^2 + K |S| N^{-2}) \right) \left( 1 + O(N^{-1/2}) \right) \leq \frac{1}{C_\perp} + O(k |S|^2 + N^{-1/2}).
$$

With Lemma 20, 21, 26 we have all tools to analyze the decay of $\|P_\perp \hat{u}\|$ with respect to $|R|$ and $N$, and show that $a \cdot b$ is bounded from below in the regime considered in the Theorem. To apply Lemma 26, we first need to ensure that $\Sigma$ and $\hat{\Sigma}$ have similar ranges.

**Lemma 27.** Let $u > 1$. Assume $2 \sigma_\varepsilon < |R| < (2KL_f)^{-1}$, and

$$
\sqrt{N} \geq \max \left\{ \frac{B^2}{C_\perp}, \left( \frac{27L_f^2}{(1 - 2KL_f |R|)^2 \varepsilon^2} \right)^2 \frac{|R|}{|R| - 2 \sigma_\varepsilon} \right\} \frac{\log(D) u^2}{2}.
$$

(67)

Then we have $\text{Im}(\Sigma) = \text{Im}(\hat{\Sigma})$ with probability at least $1 - \exp(-u^2)$.

**Proof.** Since $\text{Im}(\Sigma)$ is the minimal dimensional vector space in which $X - \mathbb{E}[X | R]$ is contained almost surely, it follows that $\text{Im}(\hat{\Sigma}) \subset \text{Im}(\Sigma)$. For the other direction, let $U_1 = \text{Im}(P_\perp) \cap \text{Im}(\Sigma)$ and $U_2 = \text{Im}(P_\parallel) \cap \text{Im}(\Sigma)$. Assuming $U_1 \cap \text{Ker}(\hat{\Sigma}) = \{0\}$ we would have

$$
\dim(U_i) = \dim \left( \left( U_i \cap \text{Ker}(\hat{\Sigma}) \right) \oplus \left( U_i \cap \text{Ker}(\hat{\Sigma})^\perp \right) \right) = \dim(U_i \cap \text{Ker}(\hat{\Sigma})^\perp).
$$

Therefore, $U_i \subset \text{Ker}(\hat{\Sigma})^\perp$ giving $\text{Im}(\Sigma) = U_1 \oplus U_2 \subset \text{Ker}(\hat{\Sigma})^\perp = \text{Im}(\hat{\Sigma})$. Take now an arbitrary unit norm $v \in U_1$. We can use (A4.1) and Lemma 19 to get

$$
v^\top \Sigma v = v^\top \Sigma v - v^\top (\Sigma - \hat{\Sigma}) v \geq v^\top \Sigma v - \frac{\log(D) B^2_\perp u^2}{\sqrt{N}} + O(N^{-1})
$$

$$
= v^\top \text{Cov}(P_\perp X | R) v - \frac{\log(D) B^2_\perp u^2}{\sqrt{N}} + O(N^{-1}) \geq C_\perp - \frac{\log(D) B^2_\perp u^2}{\sqrt{N}} + O(N^{-1}) > 0.
$$
Thus $v \not\in \ker(\hat{\Sigma})$. Take now unit norm $w \in U_2$. Equations (41) and (38) applied to $\langle w, X \rangle$ imply

$$w^\top \hat{\Sigma} w = w^\top \Sigma w - w^\top (\Sigma - \hat{\Sigma}) w \geq \text{Cov} (\langle w, X \rangle | R) - \frac{\log(D) L_f^2 |R|^2 u^2}{\sqrt{N}} + \mathcal{O}(|R|^2 N^{-1}).$$

Moreover, since $w^\top \Sigma w > \text{Cov} (w^\top V | R)$, a lower bound for $w^\top \Sigma w$ is provided in Equation (32). Together with (67), this ensures

$$w^\top \Sigma w > \left( \frac{1 - 2KL_f |R|^2 c_f^2}{27 L_f^2} \left( \frac{|R| - 2\sigma_c}{|R|} \right)^2 - \frac{\log(D) L_f^2 u^2}{\sqrt{N}} \right) |R|^2 + \mathcal{O}(|R|^2 N^{-1}) > 0. \quad (68)$$

Therefore, $w \not\in \ker(\Sigma)$ and the claim follows. $\square$

Now we can prove the key Lemma from which Theorem 8 will follow easily. Combining both bounds in the result, it states that the local linear regression vector $\hat{b}$ does not vanish and is increasingly aligned with $a$ as $|R|$ decreases and $N$ increases.

**Lemma 28.** Let $R \subset [0, 1]$ be an interval with $2\sigma_c < |R| < (2KL_f)^{-1}$, $|S|^t < C_{\perp} \lambda / C_{\Delta}^2$ and $u \geq 1$ be a fixed confidence parameter. Assume that

$$N^{1/2} \geq \max \left\{ L_2^2 C_T, \frac{B_2^2 (1 + L_3^2 C_T C_{\Delta})}{C_{\perp}} \right\} \log(D) u^2. \quad (69)$$

Then, with probability at least $1 - \exp(-u^2)$, we have

$$\langle a, \hat{b} \rangle \geq C_1 - C_2 |R| - C_3 u^2 / N^{1/2} + \mathcal{O}(\kappa |R|^2 + \kappa |R| N^{-1/2} + N^{-1}), \quad (70)$$

$$\| P_\| \hat{b} \| \leq C_4 |R|^2 + C_5 u^2 / N^{1/2} + \mathcal{O}(\kappa |R|^3 + \kappa |R|^2 N^{-1/2} + |R| N^{-1}). \quad (71)$$

where the constants are given by

$$C_1 := \frac{1}{L_f + 4L_f^2 C_T}, \quad C_2 := C_T L_f \left( \sqrt{8L_f L_f^2} + C_2 \right), \quad C_3 := L_f C_T \log(D), \quad C_4 := \frac{16L_f^2 C_T C_{\Delta}}{C_{\perp}},$$

$$C_5 := \frac{B_2}{C_{\perp}} \left( 1 + C_T \log(D) L_f^2 (1 + L_3^2 C_T C_{\Delta}) \right), \quad C_T := \frac{27L_f^2}{(1 - 2KL_f |R|^2 c_f^2)^2 \left( \frac{|R|}{|R| - 2\sigma_c} \right)^2}. \quad (72)$$

**Proof.** We first note $\text{Im}(\Sigma) = \text{Im}(\hat{\Sigma})$ with probability at least $1 - \exp(-u^2)$ according to Lemma 27 so we can use Lemma 26 by conditioning on that event. Now let us first consider (70). Using Lemma 20 to write $r = T_1 + T_2 + T_3$, and writing $P_\| = aa^\top$, we can decompose

$$\langle a, \hat{b} \rangle = a^\top \hat{\Sigma}^\dagger \hat{r} \geq a^\top \hat{\Sigma}^\dagger r - \| P_\| \hat{\Sigma}^\dagger \| T_1 - \| P_\| \hat{\Sigma}^\dagger P_\| \| T_2 \| + \| T_3 \|)$$

$$- \| P_\| \hat{\Sigma}^\dagger P_\| \| T_1 \|, \quad (73)$$

By the law of total covariance and [32], [33], we furthermore have

$$\frac{\text{Cov} (\langle a, V \rangle | R)}{\lambda} = \frac{\text{Cov} (\langle a, V \rangle | R) + \text{Cov} (\langle a, W \rangle | R)}{\lambda} \geq \frac{C_T^{-1} |R|^2 (\text{Cov} (\langle a, V \rangle | R) + \text{Cov} (\langle a, W \rangle | R)}{C_T^{-1} |R|^2 + 4L_f |R|^2} \geq \frac{1}{1 + 4L_f^2 C_T}. \quad (74)$$

$$28$$
Therefore, using (73), \(\|\nabla f(\hat{\gamma})\| \geq L_f^{-1}\) and \(\lambda \geq C_T^{-1}|R|^2\), we obtain

\[
\text{Cov}((a, V) | R) \|\nabla f(\hat{\gamma})\| \bigg\| P_\perp \hat{\Sigma}^\perp P_\perp \bigg\| \geq \frac{1}{L_f} \left( \frac{1}{1 + 4L_f^2 C_T} - \log(D)u^2 \frac{|S|^2}{\lambda} \right) \geq C_1 - \frac{L_f C_T \log(D)u^2}{N^{1/2}},
\]

where we neglect terms of order \(O(\kappa |R|^2 + \kappa |R| N^{-1/2} + N^{-1})\) from \(\|P_\perp \hat{\Sigma}^\perp P_\perp\|\). The remaining terms in (73) can be bounded using Lemmas 14, 20, 21 and 26 and \(\lambda \geq C_T^{-1}|R|^2\) as

\[
\bigg\| P_\perp \hat{\Sigma}^\perp P_\perp \bigg\| \leq C_T \log(D) \left( \sqrt{8} L_H L_f^2 + C_\varepsilon \right) |R| + O(\kappa |R|^2 + |R| N^{-1/2}),
\]

\[
\bigg\| P_\perp \hat{\Sigma}^\perp P_\perp \bigg\| \leq (\lambda^{-1} + O(|R|^{-2} N^{-1/2} + \kappa)) \bigg\| P_\perp (\hat{r} - \bar{r}) \bigg\| \leq \frac{u C_T L_f^2}{N^{1/2}} + O(N^{-1})
\]

\[
\bigg\| P_\perp \hat{\Sigma}^\perp P_\perp \bigg\| \leq (\kappa |R|^2 + \kappa |R| N^{-1/2}),
\]

\[
\bigg\| P_\perp \hat{\Sigma}^\perp P_\perp \bigg\| \leq O(\kappa |R| N^{-1/2} + N^{-1}),
\]

For (71), we write

\[
\bigg\| P_\perp \hat{b} \bigg\| = \bigg\| P_\perp \hat{\Sigma}^\perp \bar{r} \bigg\| \leq \bigg\| P_\perp \hat{\Sigma}^\perp \bar{r} \bigg\| + \bigg\| P_\perp \hat{\Sigma}^\perp (\hat{r} - \bar{r}) \bigg\|
\]

\[
\leq \bigg\| P_\perp \hat{\Sigma}^\perp P_\perp \bigg\| \| T_1 \| + \bigg\| P_\perp \hat{\Sigma}^\perp P_\perp \bigg\| \| T_2 \| + \| T_3 \| + \bigg\| P_\perp \hat{\Sigma}^\perp P_\perp \bigg\| \| P_\perp \bar{r} \|
\]

\[
+ \bigg\| P_\perp \hat{\Sigma}^\perp P_\perp \bigg\| \| P_\perp (\hat{r} - \bar{r}) \| + \bigg\| P_\perp \hat{\Sigma}^\perp P_\perp \bigg\| \| \hat{r} - \bar{r} \|
\]

Again, applying Lemmas 14, 20, 21 and 26 we get the bounds

\[
\bigg\| P_\perp \hat{\Sigma}^\perp P_\perp \bigg\| \| T_1 \| \leq \frac{L_f C_\Delta |S|^4 C_\perp}{\lambda} + \frac{\log(D) L_f B_+ u^2 (1 + C_\Delta |S|^2 \lambda^{-1}) |S|^3}{C_\perp N^{1/2}} \leq \frac{16 L_f^5 C_T C_\Delta}{C_\perp} \| R \|^2 + \frac{C_T \log(D) L_f^2 B_+ u^2 (1 + L_f^2 C_T C_\Delta)}{C_\perp} \| R \| \frac{N^{1/2}}{N^{1/2}}
\]

\[
\bigg\| P_\perp \hat{\Sigma}^\perp P_\perp \bigg\| \| T_2 \| + \| T_3 \| = O(\kappa |R|^2 + \kappa |R|^2 N^{-1/2}),
\]

\[
\bigg\| P_\perp \hat{\Sigma}^\perp P_\perp \bigg\| \| P_\perp \bar{r} \| = O(\kappa |R|^3),
\]

\[
\bigg\| P_\perp \hat{\Sigma}^\perp P_\perp \bigg\| \| P_\perp (\hat{r} - \bar{r}) \| = O((\kappa + |R|^{-1} N^{-1/2}) \log(|R|^2 N^{-1/2} + |R| N^{-1})),
\]

\[
\bigg\| P_\perp \hat{\Sigma}^\perp P_\perp \bigg\| \| \hat{r} - \bar{r} \| \leq \frac{B_+ u |R|}{C_\perp N^{1/2}} + O(|R|^{-1} + \kappa |R|^2 N^{-1/2}).
\]

Plugging all of these into (74) concludes the proof.

\[\square\]

**Theorem 29 (Theorem 3 - restated).** Let \(u > 1\), and \(R \subset [0, 1]\) be a closed interval with

\[
2 \sigma_\varepsilon < |R| < \max\left\{ \frac{1}{2K_R L_f}, \frac{L_f^2 C_T (1 + 4L_f^2 C_T) (\sqrt{8} L_H L_f^2 + C_\varepsilon)}{\sqrt{32} C_\Delta \sqrt{C_T L_f}}, R_{\text{max}} \right\}
\]

and \(C_T\) as in (72). Assume we have \(N_R\) copies of \((X, Y)\), where \(Y \in R\), and \(N_R\) satisfies

\[
N_R^{1/2} \geq \max\left\{ \frac{L_f^2 C_T^2}{C_\perp}, \frac{B_+^2 (1 + L_f^2 C_T C_\Delta)}{C_\perp} \right\} \log(D)u^2.
\]

Denote \(\hat{\Sigma}_R = \hat{E}(X - \hat{E}X)(X - \hat{E}X)^\top, \hat{r}_R = \hat{E}(Y - \hat{E}Y)(X - \hat{E}X)\). For \(\hat{a} = \hat{b} / \|\hat{b}\|\), with \(\hat{b} = \hat{\Sigma}_R^{\perp} \hat{r}_R\) we have with probability of at least \(1 - \exp(-u^2)\)

\[
\|\hat{a} - a(\text{E}[|R|])\| \leq \frac{C_T^2}{C_\perp} \left( \kappa_R |R|^2 + u^2 \frac{|R|}{\sqrt{N_R}} \right).
\]

(76)
Proof of Theorem 3. First note that $|\mathcal{R}| < \sqrt{C_{\Delta}}/\sqrt{2C_{\Delta}}T L_f^2$ implies $|S|^4 < C_\perp \lambda/C_{\Delta}$ by taking the square, using $\lambda \geq C T^{-1} |\mathcal{R}|^2$, and using Lemma 14. Moreover $C T \geq 1$ (since $L_f \geq 1$, $c \leq 1$), such that (69) is satisfied. Therefore, we can apply Lemma 28 which ensures that $\langle \hat{a}, \hat{b} \rangle > 0$. Using this, we can rewrite $a = \| P_1 \hat{b} \|^{-1} P_1 \hat{b} a$, and furthermore

$$\|P_\perp \hat{a} - a\| = \left\| \frac{P_\perp \hat{b}}{\| \hat{b} \|} - \frac{P_\perp \hat{b}}{\| P_1 \hat{b} \|} \right\| = \left\| \frac{P_\perp \hat{b}}{\| \hat{b} \|} \right\| \frac{1}{\| \hat{b} \|} - \left\| \frac{P_\perp \hat{b}}{\| P_1 \hat{b} \|} \right\| \leq \left\| \frac{P_\perp \hat{b}}{\| \hat{b} \|} \right\|,$$

we obtain the bound

$$\| \hat{a} - a \| = \sqrt{\|P_\perp \hat{a} - a\|^2 + \|P_\perp \hat{a}\|^2} \leq \sqrt{2\|\hat{b}\|^2 - \|P_\perp \hat{b}\|^2} \leq \sqrt{2\|\hat{b}\|^2} \leq \sqrt{2\|P_1 \hat{b}\|} \quad (77)$$

Furthermore, in the notation of Lemma 28 we have

$$\| P_\perp \hat{b} \| \geq C_1 - C_2 |\mathcal{R}| - C_3 u^2 N^{-1/2} = C_1 (1 - C_2 C_1^{-1} |\mathcal{R}|) - C_3 u^2 N^{-1/2} = C_1 + O(|\mathcal{R}| + N^{-1/2})$$

by using the restriction on $|\mathcal{R}|$ (second term in the max of (75)), and $N_{\mathcal{R}} \geq L_f^2 C_2^2 \log(D) u^2$. Therefore, the desired bound follows directly from Lemma 28 applied to $\| P_\perp \hat{b} \|$ and $C_1^{-1} \geq C_T$. □

A.3 Proof of Lemma 2

Proof. (A1) follows from $E[W|V] = E[F(V)U|V] = F(V)E[U|V] = 0$, whereas (A2) is a direct consequence of $\|U\| \leq B$. For (A4.2) we use the law of total covariance and $E[W|V] = 0$ to get

$$E_{\nu} \left[ ||\text{Cov} (W|V) - \text{Cov} (W|\mathcal{R})||^2 \right] = E_{\nu} \left[ \left\| \text{E}_{\nu} \left[ \text{Cov} (W|V) - \text{Cov} (W|\mathcal{R}) \right] \right\|^2 \right] \leq 4 \text{Cov} (U)^2 L_f^2 E \left[ d_{\nu} (V', V'')^2 |\mathcal{R}| \right].$$

Let $z := \gamma (I_{\mathcal{R}})$ and consider $v \in \text{Cov} (X|\mathcal{R}) \cap \text{Im} (P_{\mathcal{R}, \perp})$, with $P_{\mathcal{R}, \perp}$ as in (A4.1) Then there exists $u \in \mathbb{R}^{D-1}$, with $v = F(z) u$, and we can use $F(V)^T F(V) = \text{Id}_{D-1}$ and Lemma 14 to get

$$\|v^T \text{Cov} (W|\mathcal{R}) v - u^T \text{Cov} (U) u\| = \|u^T \text{Cov} \left( F(z)^T F(V) U |\mathcal{R} \right) u - u^T \text{Cov} (U) u\| \leq 2 \text{Cov} (U) \|E_{\nu} \| F(z) - F(V) \| |\mathcal{R}| \leq 2 L_f \|\text{Cov} (U)\| L_F (|\mathcal{R}| + 2 \sigma_z).$$

Assumption (A4.1) then follows by the law of total covariance according to

$$v^T \text{Cov} (X|\mathcal{R}) v = v^T \left( \text{Cov} (V|\mathcal{R}) + \text{Cov} (W|\mathcal{R}) \right) v \geq C_1^* - 2 L_f \|\text{Cov} (U)\| L_F (|\mathcal{R}| + 2 \sigma_z).$$

□

A.4 Proof of Equation (19)

Proof. We begin with an intermediate result. Let $v = x + w$, $x' = v' + w' \in \text{esssup} (\rho_X)$, where $v = \pi_v (x)$ and $v' = \pi_v (x')$, and let $S(v, v')$ be the curve segment between $v$ and $v'$. Assume $\gamma |S(v, v')|$ is $\theta$-almost linear for $\theta > \kappa (S(v, v')) B$. We will show that for any $p \in \mathbb{R}^D$ we have

$$\left| \langle p, x - x' \rangle - \| x - x' \| \| p - a(v') \| \right| \leq d_{\gamma} (v, v') \leq \left| \langle p, x - x' \rangle + \| x - x' \| \| p - a(v') \| \right| \leq \frac{\kappa (S(v, v')) B d_{\gamma} (v, v')}{\theta - \kappa (S(v, v'))} \quad (78)$$

For the first inequality we have $|\langle p, x - x' \rangle| \leq \| x - x' \| \| p - a(v') \| + |\langle a(v'), x - x' \rangle|$, by Cauchy-Schwarz. The fundamental theorem of calculus and $a(v) \perp w$, $a(v') \perp w'$ then yield

$$|\langle a(v'), x - x' \rangle| \leq |\langle a(v'), v - v' \rangle| + |\langle a(v') - a(v), w \rangle| \leq d_{\gamma} (v, v') + \kappa (S(v, v')) B d_{\gamma} (v, v').$$
where we used Lemma 16 in the last step. Thus, the bound follows after dividing by \( \kappa(S(v, v')) B \).

For the second inequality in (78) using Lemma 16 and again the fact that \( w' \perp a(v') \), to get

\[
\theta d_\gamma(v, v') < |\langle v - v', a(v') \rangle| \leq |\langle x - x', a(v') \rangle| + |\langle w, a(v') \rangle| \\
\leq |\langle x - x', p \rangle| + \|x - x'\| \|p - a(v')\| + \kappa(S(v, v')) B d_\gamma(v, v')
\]

Collecting terms with \( d_\gamma(v, v') \) and dividing through by \( \theta - \kappa(S(v, v')) B \) yields the desired bound. With (78) proven, (19) follows, since

\[
\|x - X_i\| \leq \|v - V_i\| + \|w - W_i\| \leq |I| + 2B, \text{ and } 1 + B\kappa(S(v, v')) \leq 2.
\]

\( \Box \)

A.5 Proof of Lemma 13

We will need the following lemma in the proof.

**Lemma 30.** Assume \( x \in \mathbb{R}^D \) has a unique projection \( v := \pi_\gamma(x) \), satisfying \( \|x - v\| \leq B \). Furthermore, let \( v' \neq v \in \text{Im}(\gamma) \) with \( \langle a(v'), x - v' \rangle = 0 \). Then \( \|x - v'\| \geq 2\tau_\gamma - B \).

**Proof.** First note that by the properties of \( \tau_\gamma \) we know that for all \( z \in \mathbb{R}^D \), such that \( \text{dist}(\text{Im}(\gamma); z) < \tau_\gamma \), there is only one \( v_z \in \text{Im}(\gamma) \) such that \( \langle a(v_z), z - v_z \rangle = 0 \) and \( \|z - v_z\| < \tau_\gamma \). Thus, \( \|x - v'\| \geq \tau_\gamma \). Moreover, for the line \( W(t) = v + ts \), where \( s = (x - v')/\|X - v'\| \), we have \( \text{dist}(\text{Im}(\gamma); W(t)) = \|W(t) - v'\| = t \), for all \( t \in (0, \tau_\gamma) \) and \( \text{dist}(\text{Im}(\gamma); W(t)) = \tau_\gamma \) holds for at least one \( t^* \in [\tau_\gamma, \|x - v'\|] \).

We now want to show that \( \|W(t^*) - x\| \geq \tau_\gamma - B \). Assume the contrary. Then

\[
\|W(t^*) - v\| \leq \|W(t^*) - x\| + \|x - v\| < \tau_\gamma,
\]

which is a contradiction with \( \text{dist}(\text{Im}(\gamma); W(t)) = \tau_\gamma \). Since \( W(t^*) \) lies on a line between \( v' \) and \( x \) we have

\[
\|x - v'\| = \|x - W(t^*)\| + \|W(t^*) - v'\| \geq 2\tau_\gamma - B.
\]

\( \Box \)

**Proof of Lemma 13.** We will first find an upper bound for \( \Delta(x^*, X_i) \) for \( X_i \in \mathcal{D}(x^*, k) \cup \mathcal{N}_\Delta(x^*, k, \delta k) \) and then argue (by contradiction) that \( d_\gamma(\pi_\gamma(x^*), \pi_\gamma(X_i)) \) satisfy the desired bound. Note first that since \( \{\pi_\gamma(X_i) : i = 1, \ldots, N\} \) form a \( \delta \)-net on \( \text{Im}(\gamma) \) there exist samples \( \tilde{X}_1, \ldots, \tilde{X}_k \) such that \( d_\gamma(v^*, \pi_\gamma(\tilde{X}_i)) \leq \delta k \). Moreover, \( \|\tilde{X}_i - x^*\| \leq 2B + \delta k \) by triangle inequality. Using (78) we then have

\[
\Delta(x^*, \tilde{X}_i) \leq 2d_\gamma(v^*, \pi_\gamma(\tilde{X}_i)) + (\|I\| + 2B)\varepsilon_a, \quad \forall i = 1, \ldots, k,
\]

and thus

\[
\max_{X_i \in \mathcal{D}(x^*, k) \cup \mathcal{N}_\Delta(x^*, k, \delta k)} \Delta(x^*, X_i) \leq 2\delta k + (\|I\| + 2B)\varepsilon_a.
\]

Assume that the contrary holds, i.e., there exists an \( X_i \) such that \( d_\gamma(V_i, v^*) > 4\mathfrak{A} \) where \( \mathfrak{A} = \nu + 4B\varepsilon_a \). Let \( \tilde{x} := V_i + (1d - a(V_i)a(V_i)^T)(x^* - V_i) \), so that \( \langle a(V_i), x - \tilde{x} \rangle = 0 \) and

\[
\|x^* - \tilde{x}\| = |\langle a(V_i), x^* - X_i \rangle| + |\langle a(V_i) - a(V_i), x^* - X_i \rangle| \leq \mathfrak{A}
\]

By Lemma 31 we have \( \|v^* - \tilde{v}\| \leq \frac{1}{1-q}\mathfrak{A} \), where \( \tilde{v} := \pi_\gamma(\tilde{x}) \), and thus \( \|\tilde{x} - \tilde{v}\| \leq \frac{2-q}{1-q}\mathfrak{A} + B \). The goal now is to use Lemma 30 to bound \( \|\tilde{x} - V_i\| \). To do that we need to show \( V_i \neq \tilde{v} \) and
\[ \|\tilde{x} - \tilde{v}\| \leq \tau_\gamma \] (which implies that \(\pi_\gamma(\tilde{x})\) is uniquely defined). For the second claim, using \(q < \frac{1}{2}\), we have

\[ \|\tilde{x} - \tilde{v}\| \leq \frac{2 - q}{1 - q} A + R \leq \frac{2 - q}{1 - q} \nu + \tau_\gamma \left(4q - \frac{2 - q}{1 - q} \epsilon_a + q\right) \leq 3 \nu + \tau_\gamma \left(\frac{1}{2} + 6 \epsilon_a\right). \]

Thus, \(\|\tilde{x} - \tilde{v}\| < \tau_\gamma\) by (27). For the first claim use \[\text{NSW08 Proposition 6.3}\], which gives

\[ d_\gamma(v^*, \tilde{v}) \leq \tau_\gamma \left(1 - \sqrt{1 - \frac{1}{1 - q} \frac{2A}{\tau_\gamma}}\right) \leq \frac{1}{1 - q} \left(1 + \frac{2}{\tau_\gamma} A\right). \]

Therefore,

\[ d_\gamma(v^*, V_i) \geq d_\gamma(V_i, v^*) - d_\gamma(v^*, \tilde{v}) \geq 6A \left(1 - \frac{\Delta}{1 - q} A\right) \geq \Delta \left(4 - \frac{2A}{\tau_\gamma}\right). \]

Since \(\Delta < \nu + \frac{3}{2} \epsilon_a\) we have \(\Delta < \frac{3}{2}\) by (27). Thus, \(d_\gamma(v^*, V_i) > 0\), and consequently \(v^* \neq V_i\), as desired. We can now apply Lemma [30] with \(x = \tilde{x}, v = \tilde{v}\) and \(v' = V_i\), yielding \(\|\tilde{x} - V_i\| \geq 2 \tau_\gamma - \|\tilde{x} - \tilde{v}\|\).

Since \(\|X_i - x^*\| \leq 2B + \delta k\), and \(\|X_i - x^*\| \geq \|\tilde{x} - V_i\| - \|V_i - X_i\| - \|\tilde{x} - x^*\|\), we have

\[ 2B + \delta k \geq 2 \tau_\gamma - \Delta \left(\frac{2 - q}{1 - q} + 1\right) - 2B. \]

We now use the fact that \(\frac{3 - 2q}{1 - q} \leq \frac{7}{2} + q\) holds for \(q < 1/2\) and \(B \leq q \tau_\gamma\) to solve the above inequality for \(q\). This gives

\[ q \geq \frac{\tau_\gamma (2 - 24 \epsilon_a) - 12 \nu}{\tau_\gamma (4 + 2 \epsilon_a) + \nu}, \]

which is a contradiction with (27).

\[
A.6 \quad \text{Referenced results}
\]

**Lemma 31** (Theorem 4.8 (8) in [Fed59]). If \(x, x'\) satisfy \(\max\{\|x - \pi_\gamma(x)\|, \|x' - \pi_\gamma(x')\|\} \leq B < \tau_\gamma\), then

\[ \|\pi_\gamma(x) - \pi_\gamma(x')\| \leq \frac{\tau_\gamma}{\tau_\gamma - B} \|x - x'\|. \]

**Theorem 32** (Matrix Bernstein, 6.1.1. in [TRO13]). Consider a finite sequence \(S_k\) of independent, random matrices, with common dimension \(d_1 \times d_2\) and assume that \(E[S_k] = 0\), and \(\|S_k\| \leq L, \forall k\). Define the random matrix \(S = \sum_{k=1}^N S_k\), and the matrix variance statistic

\[ m(S) = \max \left(\|E[SS^\top]\|, \|E[S^\top S]\|\right). \]

Then for all \(\epsilon \geq 0\) we have the tail bound

\[ \mathbb{P}(\|S\| \geq \epsilon) \leq (d_1 + d_2) \exp \left(-\frac{\epsilon^2}{2(m(S) + L \epsilon/3)}\right). \]

**Remark 33.** Let us make a short comment regarding Theorem 32. Jensen’s inequality gives

\[ m(S) \leq \mathbb{E} \max \|SS^\top\|, \|S^\top S\| = \mathbb{E} \|S\|^2. \]

Hence, it is sufficient to bound \(\mathbb{E} \|S\|^2\). Moreover, (80) holds if we replace \(m(S)\) with its upper bound \(\mu \geq m(S)\). Rewriting now the right hand side of (80) as

\[ \exp \left(\log(d_1 + d_2) - \frac{\epsilon^2}{2(\nu + L \epsilon/3)}\right) =: \exp(-u), \]

32
for $u > 0$, leads to a quadratic equation for $\epsilon$, the solution of which is given as

$$
\epsilon = \frac{1}{3} \left( \sqrt{L^2 (u + \log(d_1 + d_2))^2 + 18\nu(u + \log(d_1 + d_2)) + L(u + \log(d_1 + d_2))} \right). 
$$

This can be bounded as follows

$$
\epsilon \leq \frac{1}{3} \max(L, \sqrt{\nu}) \left( \sqrt{(u + \log(d_1 + d_2))(u + \log(d_1 + d_2) + 18) + (u + \log(d_1 + d_2))} \right)
\leq \frac{2}{3} \max(L, \sqrt{\nu}) \left( u + \log(d_1 + d_2) + \frac{9}{2} \right),
$$

where we use the inequality of arithmetic and geometric means in the second line. Finally, monotonicity of probability gives $P(\|S\| \geq \epsilon) \geq P(\|S\| \geq \epsilon')$ for $\epsilon \leq \epsilon'$. Thus, for every $u > 0$

$$
P(\|S\| \geq C) \leq \exp(-u), \text{ where } C = \frac{2}{3} \max(L, \sqrt{\nu}) \left( u + \log(d_1 + d_2) + \frac{9}{2} \right).
$$

References

[AC09] Kofi P. Adragni and R. Dennis Cook. Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society*, 367:4385–4405, 2009.

[BL+07] Peter J Bickel, Bo Li, et al. Local polynomial regression on unknown manifolds. In *Complex Datasets and Inverse Problems*, pages 177–186. Institute of Mathematical Statistics, 2007.

[DC00] R Dennis Cook. Save: a method for dimension reduction and graphics in regression. *Communications in statistics-Theory and methods*, 29(9-10):2109–2121, 2000.

[DJS08] Arnak S Dalalyan, Anatoly Juditsky, and Vladimir Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research*, 9(Aug):1647–1678, 2008.

[Fed59] Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

[GRB+17] Ravi Ganti, Nikhil S Rao, Laura Balzano, Rebecca Willett, and Robert D Nowak. On learning high dimensional structured single index models. In *AAAI*, pages 1898–1904, 2017.

[HH96] J. L. Horowitz and W. Härdle. Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91(436):1632–1640, 1996.

[HJS01] Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001.

[HTFF05] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[HZS06] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[Ich93] H. Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120, 1993.
[KKS11] Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pages 927–935, 2011.

[Koh14] Michael Kohler. Optimal global rates of convergence for noiseless regression estimation problems with adaptively chosen design. *Journal of Multivariate Analysis*, 132:197–208, 2014.

[KP16] Arun Kumar Kuchibhotla and Rohit Kumar Patra. Efficient estimation in single index models through smoothing splines. *arXiv preprint arXiv:1612.00068*, 2016.

[Kpo11] Samory Kpotufe. k-nn regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems*, pages 729–737, 2011.

[KS09] Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*. Citeseer, 2009.

[LAL+11] Bing Li, Andreas Artemiou, Lexin Li, et al. Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, 39(6):3182–3210, 2011.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[Li91] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[Li92] Ker-Chau Li. On principal hessian directions for data visualization and dimension reduction: Another application of stein’s lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992.

[LLC+13] Kuang-Yao Lee, Bing Li, Francesca Chiaromonte, et al. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics*, 41(1):221–249, 2013.

[LW07] Bing Li and Shaoli Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007.

[LZC+05] Bing Li, Hongyuan Zha, Francesca Chiaromonte, et al. Contour regression: a general approach to dimension reduction. *The Annals of Statistics*, 33(4):1580–1616, 2005.

[McD89] Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[MWZ+10] Sayan Mukherjee, Qiang Wu, Ding-Xuan Zhou, et al. Learning gradients on manifolds. *Bernoulli*, 16(1):181–207, 2010.

[NSW08] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008.

[Rad15] Peter Radchenko. High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282, 2015.

[RWD88] Tim Robertson, F. T. Wright, and R. L. Dykstra. *Order Restricted Statistical Inference*. John Wiley and Sons, 1988.
[Sto82] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

[Tro15] J. Tropp. An introduction to matrix concentration inequalities. *arXiv:1501.01571*, 2015.

[Wed73] Per-Åke Wedin. Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13(2):217–232, 1973.

[Wu08] Han-Ming Wu. Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 17(3):590–610, 2008.

[XTLZ02] Yingcun Xia, Howell Tong, WK Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002.

[YHL09] Yi-Ren Yeh, Su-Yun Huang, and Yuh-Jye Lee. Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Transactions on Knowledge and Data Engineering*, 21(11):1590–1603, 2009.