Journal of Machine Learning Research 27 (2026) 1-75 Submitted 11/24; Revised 8/26; Published 8/26

Conditional Regression for the Nonlinear Single-Variable Model

Yantao Wu YWU212@JHU.EDU
Department of Mathematics
Johns Hopkins University
Baltimore, MD 21218-2683, USA

Mauro Maggioni MAUROMAGGIONIJHU@ICLOUD.COM
Department of Mathematics and Department of Applied Mathematics & Statistics
Johns Hopkins University
Baltimore, MD 21218-2683, USA

Editor: Mladen Kolar

Abstract

Regressing a function F on $\mathbb{R}^d$ without incurring the statistical and computational curse of dimensionality requires exploitable structure. Compositional models $F = f \circ g$ in which g has a low-dimensional range include classical single- and multi-index models as well as certain neural networks; while the case of linear g is well understood, substantially less is known for nonlinear g. We study the model $F(X) = f(\Pi_\gamma X)$, where Π_γ is the closest-point coordinate associated with an unknown regular curve γ, and f is an unknown one-dimensional link function. The predictor X need not be intrinsically low-dimensional and may have full-dimensional variation throughout a tubular neighborhood of the curve. We construct a nonparametric estimator based on response slicing, local principal component analysis, data-adaptive slice assignment, and one-dimensional local polynomial regression. Under coarse monotonicity of f and sufficient variation normal to the curve relative to the observational noise and the coarse-monotonicity scale, the estimator attains, up to logarithmic factors, the minimax-optimal one-dimensional mean squared rate down to an explicit geometry- and noise-dependent saturation level. When the normal-variation condition is removed, we prove a complementary guarantee for the wide-slice regime. The estimator can be constructed in time $\mathcal{O}(d^2 n \log n)$, and the constants and sample-size thresholds in our bounds depend at most polynomially on the ambient dimension d.

Keywords: high-dimensional regression, nonparametric regression, compositional models, single-index model, inverse regression

1. Introduction

We consider the standard regression problem of estimating a function $F : \mathbb{R}^d \to \mathbb{R}$ from n samples $\{(X_i, Y_i)\}_{i=1}^n$, where $X_1, \ldots, X_n$ are i.i.d. realizations of a predictor variable $X \in \mathbb{R}^d$ with distribution ρ_X, and $\zeta_1, \ldots, \zeta_n$ are mutually independent realizations of a centered observational-noise variable ζ, independent of $X_1, \ldots, X_n$, with

$$Y_i = F(X_i) + \zeta_i \,.$$

In the general nonparametric, distribution-free setting, if we know only that $F \in \mathcal{C}^s(\mathbb{R}^d)$ is Hölder continuous with exponent $s > 0$ and, say, compactly supported, then the minimax

rate for estimating F in $L^2(\rho_X)$ is $n^{-\frac{s}{2s+d}}$; see (Binev et al., 2024; Györfi et al., 2002) and the references therein. Equivalently, the minimax squared $L^2(\rho_X)$ risk is of order $n^{-\frac{2s}{2s+d}}$. Kernel estimators attain this rate with a properly chosen bandwidth and kernel, as do a variety of other well-understood estimators based on Fourier or multiscale decompositions; see (Györfi et al., 2002; Binev et al., 2005, 2007) and the references therein. This rate deteriorates dramatically as the dimension d of the ambient space increases: this is an instance of the *curse of dimensionality*. Because no estimator can achieve a faster minimax rate over all such functions and d is very large in many applications, it is natural to consider structured regression problems in which the curse of dimensionality can be avoided.

In this work, we study the Nonlinear Single-Variable Model $F(x) = f(\Pi_\gamma x)$, where γ is an unknown curve in $\mathbb{R}^d$ and Π_γ returns the parameter of the closest point on γ. The predictor X may have full-dimensional variation in a neighborhood of the curve and is therefore not intrinsically low-dimensional. From response slices, our estimator learns local curve geometry by PCA, assigns each point to a nearby slice, and performs one-dimensional local polynomial regression along the resulting coordinate. It attains the minimax-optimal one-dimensional mean squared rate up to logarithmic factors, plus an explicit saturation term, and can be computed in $\mathcal{O}(d^2 n \log n)$ time.

Two key assumptions are made for our strongest convergence results: the link f is assumed to be coarsely monotone, so that response slices correspond to localized portions of the curve, and a lower conditional variance condition is imposed, requiring sufficient normal variation around the curve relative to the observational noise and the coarse-monotonicity scale of f. Under these assumptions, the estimator avoids both the statistical and the computational curse of dimensionality. We investigate the robustness of our estimator when these assumptions are relaxed. In particular, when the lower conditional variance condition is removed, we prove that the estimator achieves small error provided the observational noise is small and f is nearly monotone.

We first state the model, its main guarantees, and our contributions, and then place them in the context of related work. We view the present model as primarily of theoretical interest because its inner function is currently restricted to projection onto a one-dimensional manifold. Extending the construction to higher-dimensional manifolds is natural and may help explain why some functions can be learned efficiently even from data that appear far from any low-dimensional structure. We also give a stylized application to learning committor functions in high-dimensional stochastic systems, where a reaction-path coordinate can induce an approximate NSVM structure; see Section 4.3.

1.1 The Nonlinear Single-Variable Model and contributions

We call this the *Nonlinear Single-Variable Model* (NSVM). Our contributions are as follows:

(i) We study the *Nonlinear Single-Variable Model*, an intermediate model between the semiparametric single-index model and the nonparametric function-composition model: both the outer and inner functions are nonlinear, while the inner function is the closest-point coordinate associated with an unknown curve $\gamma \subseteq \mathbb{R}^d$, and the distribution of X has a corresponding tubular geometric structure;

(ii) For the coarse-monotone subclass of links, and under the lower conditional variance condition used in the main thin-slice theorem, we construct an estimator that achieves the one-dimensional minimax rate up to logarithmic factors and the explicit noise/curve-approximation term appearing in Theorem 3. When the lower variance condition is not satisfied, Theorem 6 shows that our estimator can also perform well, but with a saturation level that depends on the geometry of the response slices and the link function;

(iii) We give an efficient, near-linear-time algorithm that constructs the estimator from data. All constants in the learning bounds, sample-size thresholds, and computational costs scale as low-order polynomials in the ambient dimension d.

This model generalizes the single-index model by allowing g to be nonlinear (but with a particular geometric structure), while remaining amenable to estimation with strong statistical and computational guarantees that are not cursed by the ambient dimension and require no regularity assumptions that scale with the dimension.

Definition 1 (Nonlinear Single-Variable Model) *The regression function F admits the decomposition*

$$\mathbb{E}[Y|X] = F(X) = f(\Pi_\gamma X)$$

where the underlying curve $\gamma : [0, \mathrm{len}_\gamma] \to \mathbb{R}^d$ *is twice continuously differentiable, is parameterized by arclength without loss of generality, and has length* len_γ *, and the* link function $f : [0, \mathrm{len}_\gamma] \to \mathbb{R}^1$ *depends only on one variable. The random vector X is supported in some domain $\Omega_\gamma \subseteq \mathbb{R}^d$ containing the image of γ, such that the closest-point projection $\Pi_\gamma : \Omega_\gamma \to [0, \mathrm{len}_\gamma]$, with*

$$\Pi_\gamma(x) := \operatorname*{arg\,min}_{t \in [0, \mathrm{len}_\gamma]} \|x - \gamma(t)\| \ ,$$

is well defined on Ω_γ, i.e., the minimizer is unique.

We express the random vector X as its position along the curve and its displacement away from the curve:

$$X = \gamma(t) + M_{\gamma'(t)} \begin{pmatrix} Z_{d-1} \\ 0 \end{pmatrix} \tag{1}$$

with $t = \Pi_\gamma X$, t a random variable on $[0, \mathrm{len}_\gamma]$ and Z_{d-1} a centered random vector in $\mathbb{R}^{d-1}$. For each unit vector $v \in \mathbb{S}^{d-1}$, let $M_v \in \mathbb{O}(d) \subseteq \mathrm{GL}(d, \mathbb{R})$ be an orthogonal matrix that maps the dth canonical basis vector $\hat{e}_d$ to v. Therefore, conditioned on $t = t_0 \in [0, \mathrm{len}_\gamma]$, the random vector $M_{\gamma'(t_0)} \begin{pmatrix} Z_{d-1} \\ 0 \end{pmatrix} \in \mathrm{span}\{\gamma'(t_0)\}^\perp$ is the displacement of X away from γ.

The regression problem for the Nonlinear Single-Variable Model: Given pairs $\{(X_i, Y_i)\}_{i=1}^n$, where $X_1, \ldots, X_n$ are independent copies of X as above and $Y_i = F(X_i) + \zeta_i$, with F as in Definition 1 and observational noise ζ_i, the goal is to estimate the regression function F on the support of X. The variables $\zeta_1, \ldots, \zeta_n$ are assumed to be sub-Gaussian, mutually independent, and independent of $X_1, \ldots, X_n$.

The distribution of X, the link function f, and the underlying curve γ are all unknown. Figure 1 illustrates the model.

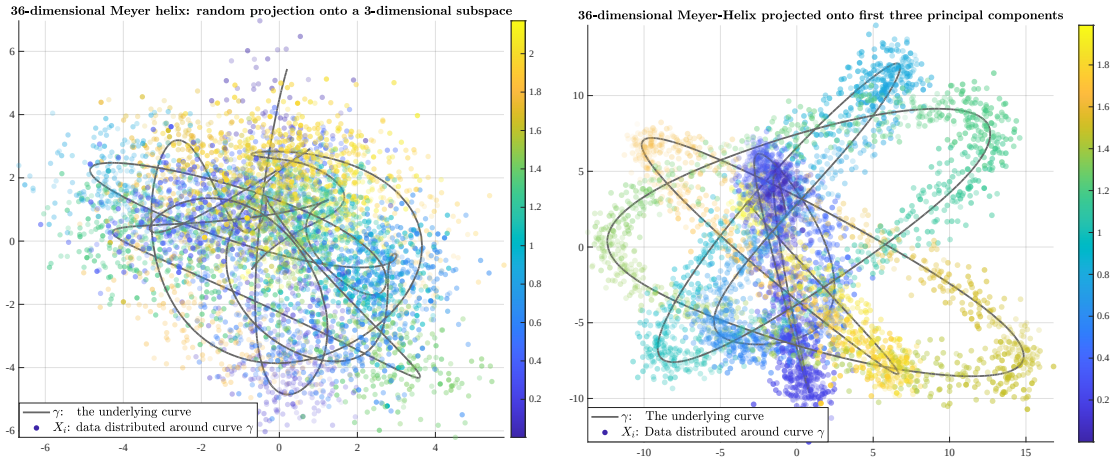

Figure 1: An example of the Nonlinear Single-Variable Model in Definition 1. The curve γ, shown in black, is a Meyer helix in $\mathbb{R}^{36}$ (see Appendix C) with $\sigma_\gamma = 0.5$; the link $f \in \mathcal{C}^{0.7}(\mathbb{R})$ is strictly monotone, and $\zeta \equiv 0$. We generate $n = 5000$ samples X_i in a tube of radius 6 around the curve and color them by $Y_i = F(X_i) = f(\Pi_\gamma X_i)$. Left: a random projection onto $\mathbb{R}^3$. Right: the orthogonal projection onto the first three principal components. The distribution does not appear to admit a low-dimensional linear embedding without increased geometric complexity; see Appendix C.

The domain $\Omega_\gamma \subseteq \mathbb{R}^d$ must be chosen so that the closest-point projection Π_γ is well defined; this condition is related to the reach of γ (Federer, 1959). Define the distance function $\mathrm{dist}_\gamma(x) : \mathbb{R}^d \to [0,\infty)$, $\mathrm{dist}_\gamma(x) = \inf\{\|x - \gamma(t_0)\| : \ t_0 \in [0, \mathrm{len}_\gamma]\}$, the domain Unp_γ is defined as the set of all points $x \in \mathbb{R}^d$ for which there is a unique point $\gamma(t_0)$ closest to x. The map $\Pi_\gamma : Unp_\gamma \to [0, \mathrm{len}_\gamma]$ maps $x \in Unp_\gamma$ to the unique $t \in [0, \mathrm{len}_\gamma]$ such that $\mathrm{dist}_\gamma(x) = \|x - \gamma(t)\|$. For $t_0 \in [0, \mathrm{len}_\gamma]$, the local reach is $\mathrm{reach}_\gamma(t_0) := \sup\{r : B(\gamma(t_0), r) \subseteq Unp_\gamma\}$, and the global reach is $\mathrm{reach}_\gamma := \inf\{\mathrm{reach}_\gamma(t_0) : \ t_0 \in [0, \mathrm{len}_\gamma]\}$. Both $\mathrm{reach}_\gamma(t_0)$ and reach_γ take values in $[0,\infty]$. With these definitions, Ω_γ could be as large as $\bigcup_{t_0 \in [0,\mathrm{len}_\gamma]} B(\gamma(t_0), \mathrm{reach}_\gamma(t_0))$. Thus, unlike the manifold-regression models discussed above, ρ_X need not be supported on, or highly concentrated near, γ. As in single-index models, the data may have substantial variance in directions normal to the index; here we allow this variance to be as large as possible while keeping Π_γ, the nonlinear analogue of orthogonal projection onto the index line, well defined.

For the regularity range used in the formal guarantees, let $0 < s \leqslant 1$. A bounded function $f : \mathbb{R}^m \to \mathbb{R}$ belongs to $\mathcal{C}^s(\mathbb{R}^m)$ if

$$[f]_{\mathcal{C}^s} := \sup_{x \neq y} \frac{|f(x) - f(y)|}{\|x - y\|^s} < \infty.$$

The quantity $[f]_{\mathcal{C}^s}$ is the s-Hölder seminorm of f. Throughout the formal guarantees below, $s \in [\frac{1}{2}, 1]$.

We are now ready to state an informal version of the main theorem:

Theorem 2 (Informal) *Suppose that $f \in \mathcal{C}^s(\mathbb{R}^1)$ for some $s \in [\frac{1}{2}, 1]$, that f is coarsely monotone, and that the lower conditional variance condition $\sigma_\gamma \geqslant C_{\mathrm{LCV}} C_f \max(\sigma_\zeta, \omega_f)$*

holds for a sufficiently large universal constant $C_{\mathrm{LCV}} > 0$. With some additional assumptions on the underlying curve γ, the distribution ρ_X of the random variable X, and the variance σ_ζ^2 of the noise ζ, if the sample size satisfies $n \geqslant \mathrm{poly}(d, \mathrm{len}_\gamma)$, then the estimator $\widehat{F}$ constructed by Algorithm 1 satisfies

$$\mathbb{E}\left[\left|\widehat{F}(X) - F(X)\right|^2\right] \lesssim C_1(f, \gamma, \rho_X, \sigma_\zeta, d) n^{-\frac{2s}{2s+1}} \log n + C_2(f, \gamma, \rho_X, \sigma_\zeta, d)\,.$$

The dependence of the constants C_1, C_2 on d is a low-order polynomial. The estimator $\widehat{F}$ can be constructed by an algorithm that runs in time $\mathcal{O}(d^2 n \log n)$.

The bound on the expected $L^2(\rho_X)$ error contains two terms. The first gives a near-optimal learning rate, up to logarithmic factors, matching the minimax rate for one-dimensional nonparametric regression as if γ were known. The second is a saturation term caused by approximating γ piecewise linearly only at scales above σ_ζ. The estimator in fact produces an estimate of the *underlying curve* γ, of the (nonlinear) closest-point projection Π_γ, and of the *link* function f, thereby providing an interpretable estimate of each component in the compositional structure of F. Further remarks may be found after the formal statements of the main results in Section 2.

Table 1 collects the notation specific to the NSVM and its estimator.

symbol	definition	symbol	definition
Π_γ	closest-point coordinate on γ	$\mathrm{len}_\gamma, \mathrm{reach}_\gamma$	length and reach of γ
$F = f \circ \Pi_\gamma$	regression function	f	one-dimensional link function
$R_{l,h}$	hth response interval	$S_{l,h}, \widehat{S}_{l,h}$	population and empirical slices
$\mu_{l,h}, \widehat{\mu}_{l,h}$	population and empirical slice centers	$\Sigma_{l,h}, \widehat{\Sigma}_{l,h}$	population and empirical slice covariances
$G_{l,h}, \widehat{G}_{l,h}$	population and empirical slice-shape scores	$v_{l,h}, \widehat{v}_{l,h}$	population and empirical significant vectors
$\mathcal{H}_l$	retained heavy slice indices	n_{loc}	minimum count for a retained slice
$\mathrm{dist}(x,h), \widehat{\mathrm{dist}}(x,h)$	population and empirical slice distances	$h_x, \widehat{h}_x$	population and empirical assigned indices
l, j	slice and local-regression resolutions	$I^{(l,h)}, A_{l,h}$	projected interval and accepted region

Notation

1.2 Related work

Intrinsically low-dimensional models. One approach is to impose geometric assumptions on the distribution ρ_X of the predictor X, for example, that ρ_X is supported on a low-dimensional manifold $\mathcal{M}$ of dimension $r \ll d$, while F is a generic s-Hölder function on $\mathcal{M}$. In these settings, some estimators attain the optimal rate with respect to the intrinsic dimension r and admit fast algorithms; see (Bickel and Li, 2007; Kpotufe, 2011; Kpotufe and Garg, 2013; Liao et al., 2016, 2022; Liu et al., 2025a). For example, the estimators in (Liao et al., 2016, 2022) can be constructed in time $\mathcal{O}(C^r d^s n \log n)$ and adaptively achieve a minimax-optimal rate, up to logarithmic factors, over a large family of function spaces with unknown regularity that may vary across locations and scales. The model considered here extends some of the works above by allowing the data not to lie exactly on a low-dimensional manifold, for example because of factors irrelevant to predicting the values of F or because of noise.

Highly regular function classes. In a different direction, one may consider function classes with smaller complexity than s-Hölder functions on $\mathbb{R}^d$. One straightforward but highly restrictive assumption is that the Hölder exponent s scales proportionally with d, making the minimax rate above independent of d. Larger function classes can be obtained by imposing either strong mixed smoothness (Strömberg, 1998), such as membership in highly anisotropic Besov spaces (with $\mathcal{O}(1)$ axis-aligned directions of $\mathcal{O}(1)$ smoothness and $\mathcal{O}(d)$ smoothness in all remaining directions), or integrability of the Fourier transform (Barron, 1993). In the context of deep neural networks, Suzuki and Nitanda (2021) demonstrate that functions in this class can be approximated well, with suitable architectures, in a way that avoids the curse of dimensionality if the degree of anisotropic smoothness is sufficiently large ("$\mathcal{O}(d)$"), while leaving open the aspects of learning and optimization. The condition of belonging to a Barron's space requires sufficiently fast decay of the Fourier transform of ∇F, which is a condition that gets stronger with the dimension, except for very special cases, among which, notably the single- and multi-index models that we discuss momentarily, where the Fourier transform of ∇F (and F itself) is a singular measure. This requirement excludes intrinsically low-dimensional nonlinear models from the Barron's space since the Fourier transform of a measure supported on (or near) a general nonlinear curve has slow decay (Arkhipov et al., 2004; Brandolini et al., 2007). Our main model here is therefore typically not included in the Barron's space, though both the Barron space and our main model include the single-index model as a special case.

Compositional models. Other function classes arise from structural assumptions, often expressed through compositional models. For example, suppose that

$$F = f \circ g \quad , \text{ with } g : \mathbb{R}^d \to \mathbb{R}^r \text{ and } f : \mathbb{R}^r \to \mathbb{R}^1$$

where f and g have suitable Hölder regularity and $r \leqslant d$. For general compositions, Juditsky et al. (2009) prove that sufficiently smooth inner functions can improve the minimax nonparametric rate for estimating $F = f \circ g$, although the resulting rate may still suffer from the curse of dimensionality. Recent works combine anisotropic smoothness with composability, especially in the context of approximators or estimators constructed with deep neural networks. In this direction, (Schmidt-Hieber, 2020; Shen et al., 2021) consider spaces of functions that are compositions of low-dimensional functions with anisotropic smoothness conditions, studying the dependence of approximability and learning risk on the dimensionality of the spaces involved and the smoothness of the corresponding functions. Unfortunately the innermost such function is defined on the original high-dimensional space $\mathbb{R}^d$, so in general these results do not circumvent the curse of dimensionality, unless again the (anisotropic, in general) regularity of that function scales with d. Another direction is to design different types of neural networks that exploit nonlinear compositions, see (Lai and Shen, 2021; Liu et al., 2025b). Most of these results address only whether a function can be approximated well by a neural network; they neither analyze the learning problem nor address computational questions, such as whether an optimization algorithm can efficiently find the parameters of the desired estimator. Another aspect that is often left unaddressed is the dependence of constants on the ambient dimension, which is unfortunately often exponential: see Shamir (2020) for a discussion of some of these aspects, and Shen et al. (2021) for bounds in particular cases. In the parametric setting, models in which g is a polynomial and f is a function in a known finite-dimensional space have been consid-

ered; for example Wang et al. (2024); Lee et al. (2024) (and related work referenced therein) consider special classes of polynomials g for which a customized layer-by-layer training of a neural network leads to estimators of F that are provably not cursed by the dimensionality.

Single- and multi-index models. Classical examples of structural assumptions based on function composition for which the curse of dimensionality can provably be avoided include single- and multi-index models, as well as generalized linear models (Stone, 1982; Hastie and Tibshirani, 1990; Breiman and Friedman, 1985; Horowitz and Mammen, 2007). When $g = G : \mathbb{R}^d \to \mathbb{R}^r$ is linear, the resulting model is called the multi-index model: $F \in \mathcal{C}^s(\mathbb{R}^d)$ depends on only r linear features, $F(x) = f(Gx)$ for some *link* function $f \in \mathcal{C}^s(\mathbb{R}^r)$, matrix $G : \mathbb{R}^d \to \mathbb{R}^r$, and the projection of X onto the range of G is sufficient for regression. The special case $r = 1$ is the single-index model, in which $F(x) = f(\langle v, x\rangle)$ for an unknown index vector $v \in \mathbb{S}^{d-1}$ and an unknown link function $f \in \mathcal{C}^s(\mathbb{R})$. These models have been intensively studied: Stone (1982) conjectured that the minimax regression rate for the single-index model is $n^{-\frac{s}{2s+1}}$ (and $n^{-\frac{s}{2s+r}}$ for multi-index models), coinciding with the one-dimensional minimax nonparametric rate, thereby escaping the curse of dimensionality. This rate is achieved with kernel estimators in (Härdle and Stoker, 1989, Theorem 3.3) and (Horowitz, 1998, Section 2.5), and the index v can be estimated at the parametric rate $\mathcal{O}(n^{-\frac{1}{2}})$ under suitable assumptions. The existence of an estimator converging at rate $n^{-\frac{s}{2s+1}}$ is shown in (Györfi et al., 2002, Corollary 22.1), and (Gaïffas and Lecué, 2007, Theorem 2) demonstrates that $n^{-\frac{s}{2s+1}}$ is indeed the minimax rate. Unfortunately, these estimators often lack implementations that provably avoid the curse of dimensionality; for example, their constants may be exponential in the dimension. Liu and Liao (2024) is a recent approach, based on contour regression analysis, for multi-index models; however, the dependence of its constants on d is not made explicit, and the available analysis does not establish polynomial dependence on d. These models have also attracted recent attention through their connection to "feature learning" in neural networks. See, for example, Lee et al. (2024); Alberti et al. (2021) and the references therein on learning single-index models with suitably trained neural networks or gradient descent, respectively. Related work suggests that deep neural networks exploit low-dimensional linear subspaces for classification and prediction (Radhakrishnan et al., 2024).

Functions in our nonlinear single-variable model can be approximated efficiently by wavelets, local Fourier expansions, or neural networks adapted to the geometry of γ, for example through local tensor approximations whose first component is tangent to the curve. What remains unclear is how to estimate and compute such representations from data without incurring the curse of dimensionality. Related work includes kernel methods composed with a linear projection (Huang et al., 2025), which provide some algorithmic guarantees, and neural-network approximations (Shen et al., 2025), for which no corresponding algorithmic guarantee is given.

Many methods have been developed for jointly estimating the index vector v (or matrix G) and the link f, with different tradeoffs between statistical assumptions and computational cost; see Lanteri et al. (2022) for an extended discussion. One category includes semiparametric methods based on maximum likelihood estimation (Ichimura, 1993; Härdle et al., 1993; Delecroix et al., 2003; Delecroix and Hristache, 1999; Delecroix et al., 2006; Carroll et al., 1997) and M-estimators that produce $\sqrt{n}$-consistent index estimates under

general assumptions, but with computationally demanding implementations, relying on sensitive bandwidth selections for kernel smoothing and on high-dimensional joint optimization of f and v. Another category includes direct methods. For example, Average Derivative Estimation (Stoker, 1986; Härdle and Stoker, 1989) estimates the index vector v by exploiting its proportionality to ∇F. Early implementations of this idea suffer from the curse of dimensionality due to the use of kernel estimation for the gradient. Xia (2006) uses local kernel estimators for estimating the direction of the gradient, and the idea could in principle be applied to our model. However, our theoretical analysis would not apply, and it is unclear whether the resulting method would avoid the curse of dimensionality. The method of Hristache et al. (2001) uses an iterative scheme to adapt an elongated neighborhood window gradually, but it requires a sufficiently accurate initialization. Gradient-based methods (stochastic and non-stochastic) have been studied in the context of neural networks (Lee et al., 2024; Bietti et al., 2025, 2022; Damian et al., 2024).

A category of techniques particularly relevant to the present work includes *conditional (or inverse-regression) methods*, which derive their estimators from statistics of the conditional distribution of the explanatory variable X given observations of the (noisy) response variable Y. Prominent examples include sliced inverse regression (Duan and Li, 1991; Li, 1991), sliced average variance estimation (Cook, 2000), simple contour regression (Li et al., 2005; Coudret et al., 2014), and its multi-index analysis in Li and Wang (2007), which yields a near-optimal estimator for the multi-index model. Conditional methods partition the range of the response variable Y into small intervals and consider their preimages, which, when f is monotone, are slices whose thinnest direction is the index direction v. They are often straightforward to implement, requiring the computation of conditional empirical moments or other statistics related to the level sets of f, and may have only one tuning parameter: the slice width. Unfortunately, several of the works above, including (Hristache et al., 2001; Li and Wang, 2007), while obtaining asymptotic learning rates that avoid the curse of dimensionality and are in some cases minimax optimal or near-optimal, require a minimum sample size that is exponential in the ambient dimension d (either explicitly, or through constants in the bounds).

Beyond the statistical fact that single- and multi-index models need not incur a cost cursed by the ambient dimension, another critical problem is to design algorithms with reasonable computational cost for implementing statistically optimal estimators. For example, even in the case of single-index models, M-estimators typically require high-dimensional nonconvex optimization in v and f; methods based on optimizing over v, even when combined with random sampling, may scale exponentially in d because of the need to obtain points inside a narrow cone around the unknown v. Conditional methods can often be implemented efficiently, with running time $\mathcal{O}(C_d n)$ up to logarithmic factors (except for contour-regression methods, which scale quadratically in n), where C_d is a low-order polynomial in d. Lanteri et al. (2022) discuss these techniques in detail for the single-index model and introduce Smallest Vector Regression, which attains optimal statistical guarantees up to logarithmic factors, provides a theoretically optimal rule for selecting the slice width (a crucial parameter whose choice is often not discussed; see Coudret et al., 2014, p. 75), and admits an implementation with cost $\mathcal{O}(d^2 n \log n)$. Neither the rate exponents nor the constants in the statistical, sample-size, and computational bounds incur the curse of dimensionality.

At the level of statistical models, the manuscript of Lanteri et al. (2018) introduced a version of the nonlinear single-index model and attempted a first study of its properties using the ideas in Lanteri et al. (2022), which had put the focus on the development of inverse regression techniques under assumptions on the geometry of sufficient statistics for regression, and on using low-order moments to estimate the geometry of response slices. The contribution here lies in the slice-geometry estimator, its geometric adaptivity, and the accompanying finite-sample and computational guarantees rather than in the curve-projection model alone. The ideas in Lanteri et al. (2018) were further developed, by a subset of the authors, in Kereta et al. (2021) (a work of which we became aware when preparing the final revision of this paper), to estimate local index vectors by linear regression within response slices and predict with a k-nearest-neighbor estimator based on a localized proxy for the geodesic metric on the curve. The main results in Kereta et al. (2021) for bi-Lipschitz link functions, under the stated assumptions, are not proven therein, due to gaps in the proof[1]. The proofs as given only work by restricting the results to the case of monotone $\mathcal{C}^2$ functions with sufficiently small Lipschitz constant and Hessian, as well as small observation noise. Under these stronger assumptions the results do not yield optimal learning rates. On the computational side, the construction therein proposed is straightforward to implement when the curve is nearly linear, but in the general case, it requires prior knowledge of a critical scale parameter lying in a narrow interval determined by the reach, which is itself difficult to estimate. The numerical experiments reported in Kereta et al. (2021) do not establish that all assumptions required by the theory hold in those settings.

Our approach differs in geometry, adaptivity, and theory. Response-slice PCA identifies significant vectors, with the selected principal component adapting to the local geometry of the response slices along and normal to the curve. Points are assigned to slices by a data-adaptive local distance, reducing reliance on unknown parameters such as the observational-noise level and the reach of the curve. Prediction is then performed by one-dimensional local polynomial regression at the appropriate scale on the estimated tangent spaces. We cover the Hölder–Lipschitz range $\mathcal{C}^s$, $s \in [1/2, 1]$, and prove a finite-sample MSE bound in Theorem 3, with the one-dimensional rate up to logarithmic factors, an explicit saturation term, and polynomial dependence of all constants on the ambient dimension.

In conclusion, while the situation is rather well-understood for compositional models $F = f \circ g$ with g linear, much remains open when g is nonlinear: one needs to identify which geometric and statistical assumptions allow the compositional structure to circumvent the curse of dimensionality, beyond settings that require very high regularity of g, regularity increasing with d, or membership in special polynomial families. In the curve-projection setting treated here, the remaining challenge is not only to specify a nonlinear index, but to learn from data the relevant local geometry, the nonlinear coordinate, and the one-dimensional link with explicit risk and computational guarantees.

1. The positivity of $\sigma_{j,Y}$, used pervasively in all the main results (Corollary 3.1; the main Theorem 4.3, Remark 4.1 (the headline rates (4.4) and the noise-free (4.5)), and Proposition 4.5), requires an upper bound on the Hessian, and therefore, in particular, $f \in \mathcal{C}^2$; once that assumption is added, one further needs to assume a sufficiently small upper bound on σ_ϵ, $||Hf||_\infty \, ||f||_{\mathrm{Lip}}$ (in the notation of the present work) to guarantee that a suitable scale for which $\sigma_{j,Y}$ is positive in fact exists. Other errors in the proofs appear either repairable or of lesser consequence.

2. An estimator for the Nonlinear Single-Variable Model

We propose an inverse-regression estimator of F for the Nonlinear Single-Variable Model, together with an efficient algorithm that also produces a sketch of the curve γ and an estimate of the closest-point projection Π_γ.

Step 1: extract geometric features of the underlying curve γ. Given data $\{(X_i, Y_i)\}_{i=1}^n$, randomly split and relabel the sample into two halves, and split the first half again into two quarters. Using only $i \leqslant n/4$, let R be an interval containing the corresponding Y_i's, and let $\{R_{l,h}\}_{h=1}^l$ be the uniform partition of R into l intervals. We form the empirical slice pairs $\{(\widehat{S}_{l,h}, \widehat{R}_{l,h})\}_{h=1}^l$ from the first quarter, where $\widehat{R}_{l,h} := \{Y_i : i \leqslant n/4,\ Y_i \in R_{l,h}\}$ and $\widehat{S}_{l,h} := \{X_i : i \leqslant n/4,\ Y_i \in R_{l,h}\}$. Each empirical slice $\widehat{S}_{l,h}$ is the empirical pre-image of the interval $R_{l,h}$ in the output variable Y. For each l, h, let $S_{l,h}$ denote the conditional distribution $X \mid Y \in R_{l,h}$. Then $\widehat{S}_{l,h}$ is the empirical version of $S_{l,h}$, and its moments approximate those of $S_{l,h}$ when the slice contains sufficiently many observations. We perform principal component analysis (PCA) on each $\widehat{S}_{l,h}$ to obtain its mean $\widehat{\mu}_{l,h}$ and its "significant vector" $\widehat{v}_{l,h}$: $\widehat{\mu}_{l,h}$ is expected to lie near γ, and $\widehat{v}_{l,h}$ is expected to be tangent to γ near $\widehat{\mu}_{l,h}$, yielding a local first-order approximation to the curve.

Step 2: design a distance function $\mathrm{dist}(x, h)$ (and its empirical version $\widehat{\mathrm{dist}}(x, h)$), based on the geometric shape of $S_{l,h}$ (and $\widehat{S}_{l,h}$ respectively), that measures the distance from x to the slice $S_{l,h}$ ($\widehat{S}_{l,h}$, respectively). By assigning each $x \in \Omega_\gamma$ to the "nearest" slice according to this distance function, we partition the domain Ω_γ into several local neighborhoods. The same rule assigns each input in the second half to a local neighborhood using its X-value only. Because the significant vector $\widehat{v}_{l,h}$ is approximately tangential to the curve γ, the local linear projection approximates the nonlinear projection Π_γ.

Step 3: perform one-dimensional piecewise-polynomial regression using the observations in the second half assigned to each local neighborhood. This, together with the "nearest" slice assignment in the second step, gives a global estimator of F on Ω_γ.

2.1 Extracting the geometric features of the underlying curve γ

For each $t \in [0, \mathrm{len}_\gamma]$, let $\hat{n}(t) := \frac{\gamma''(t)}{\|\gamma''(t)\|}$ be the unit normal vector to γ, pointing toward the center of the osculating circle. For $x \in \Omega_\gamma$, define the signed projected distance from x to γ by $\tilde{d}(x, \gamma) := \langle x - \gamma(\Pi_\gamma x), \hat{n}(\Pi_\gamma x)\rangle_{\mathbb{R}^d}$. It is zero when x lies on γ, positive when x lies inside the circle of curvature, and negative otherwise. A direct calculation shows that, when γ is parameterized by arclength, the compositional structure $F = f \circ \Pi_\gamma$ in Definition 1 implies, for $x \in \Omega_\gamma$,

$$\nabla F(x) = \frac{f'(\Pi_\gamma x)}{1 - \|\gamma''(\Pi_\gamma x)\|\, \tilde{d}(x, \gamma)} \gamma'(\Pi_\gamma x)\,. \tag{2}$$

Therefore, the gradients at points in each level set $\Pi_\gamma^{-1}(t_0) = \{x : \Pi_\gamma x = t_0\}$ are all parallel to the unit vector $\gamma'(t_0)$, although their magnitudes depend on the relative position of x and γ. Consequently, each $\Pi_\gamma^{-1}(t_0)$ is contained in a hyperplane. If we could perform a singular value decomposition of points on each level set $\Pi_\gamma^{-1}(t_0)$, the singular vector corresponding to the 0 singular value would be parallel to $\gamma'(t_0)$.

The geometry of the level sets plays a central role in our estimator, as it does in earlier work on the single-index model, including (Stoker, 1986; Härdle and Stoker, 1989), and in recent refinements such as the Smallest Vector Regression estimator of Lanteri et al. (2022). In the single-index model, where the *underlying curve* γ is a line segment with direction v, any level set $\{x : F(x) = c\}$ is a hyperplane perpendicular to v. As in (Lanteri et al., 2022), using only the first quarter of the sample, we take a uniform partition of the empirical range $R := [\min_{i \leqslant n/4} Y_i, \max_{i \leqslant n/4} Y_i]$ into intervals $\{R_{l,h}\}_{h=1}^{l}$, indexed by $h = 1, \ldots, l$; for each $R_{l,h}$, we consider the empirical slice

$$\widehat{S}_{l,h} := \{X_i : i \leqslant n/4,\ Y_i \in R_{l,h}\}\,,$$

which is a sample from the conditional distribution $X \mid Y \in R_{l,h}$. Each empirical slice $\widehat{S}_{l,h}$ is utilized in Lanteri et al. (2022) as an approximation to a level set $\Pi_\gamma^{-1}(s_0)$, and $\widehat{S}_{l,h}$ should be "thin" along the direction of v (and therefore ∇F) and "wide" on directions orthogonal to v, at least under suitable assumptions on F and the noise level σ_ζ. Lanteri et al. (2022) then perform local PCA on each $\widehat{S}_{l,h}$ and use the smallest principal component to approximate the direction of v: since γ is a line with direction v, these smallest principal components should all be independent estimators of the index vector v, and these per-slice estimates can be suitably averaged across all slices to obtain an estimator for v. Once v is estimated, the input data is projected onto this line, and one-dimensional nonparametric regression yields an estimator for f.

In the NSVM, we still use empirical slices, but here we cannot perform an aggregation of estimated vectors because the tangent vectors to γ are not constant due to the curvature of γ. Instead, we can only rely on local information to estimate the closest-point projection onto γ, which is now a nonlinear function. Under Assumption **(LCV)** below, each $\widehat{S}_{l,h}$ still approximates a level set $\Pi_\gamma^{-1}(t_0)$: roughly speaking, when $R_{l,h}$ has sufficiently small diameter, $\widehat{S}_{l,h}$ is "thin" along the tangential direction $\gamma'(t_0)$ and "wide" along directions orthogonal to $\gamma'(t_0)$. Consequently, local PCA on each empirical slice $\widehat{S}_{l,h}$ yields:

(i) the empirical mean $\widehat{\mu}_{l,h}$, which approximates the conditional expectation $\mu_{l,h} := \mathbb{E}[X \mid Y \in R_{l,h}]$ and is expected to lie near γ;

(ii) the empirical covariance matrix $\widehat{\Sigma}_{l,h}$, which approximates the conditional covariance matrix $\Sigma_{l,h} := \mathbb{E}[(X - \mu_{l,h})(X - \mu_{l,h})^\intercal \mid Y \in R_{l,h}]$;

(iii) the smallest principal component $\widehat{v}_{l,h}$ of $\widehat{\Sigma}_{l,h}$, which approximates the smallest principal component $v_{l,h}$ of $\Sigma_{l,h}$ and is expected to be tangent to the curve.

Together, these quantities yield a local first-order approximation of γ near $\widehat{S}_{l,h}$. Figure 2 depicts an example where the sample mean is approximately on the curve, and the smallest principal component vector is approximately tangential to the curve.

This argument relies on Assumption **(LCV)**, which constrains the geometry of γ, the radius of Ω_γ around the curve, the regression function F, and the noise level σ_ζ. Without such an assumption, it might not be the case that slices are "thin" in the direction tangent to the curve for various reasons: (i) with an insufficient number of samples, choosing very fine partitions of the range R is not optimal: smaller intervals will contain fewer samples,

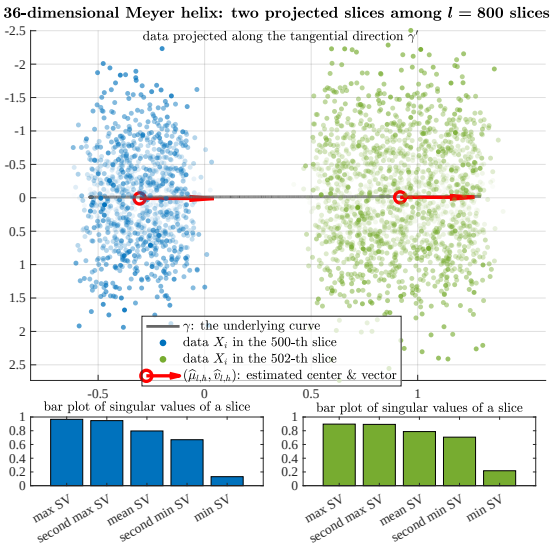

Figure 2: In the same setup as in Figure 1, we partition the range uniformly into $l = 800$ intervals, and consider two slices. Top: a visualization of the two empirical slices, where we show only 2000 samples per slice (in green and blue), with γ in black. The red circles and vectors are the sample means and smallest principal components of the two empirical slices. Bottom: bar plots with the largest, second largest, average, second smallest, and smallest singular value of the empirical slices. The smallest singular value is *significantly* smaller than the others.

leading to high variance in estimating the slice means and principal components; (ii) the observational noise ζ in the outputs forces a lower bound on the diameter of the partitioning intervals $R_{l,h}$ of the range R: subdividing the range into intervals with size smaller than the noise level σ_ζ does not improve estimation, nor does it make the slices thinner; (iii) the reach reach_γ may be small because of the complexity of the underlying curve γ; (iv) the distribution of X might strongly concentrate around γ, i.e., σ_γ, the standard deviation of Z_{d-1}, is small. In these situations, we might encounter slices that are "wide", i.e., much wider in the curve's local tangent direction than in all other directions—essentially the opposite of being "thin". In this case, the largest principal component can be significantly larger than the remaining components and is the one that aligns with the tangent direction of γ. Figure 3 illustrates how, in these situations, the sample mean and the *largest* principal component of a slice may provide an approximation of the underlying curve γ. In this setting, the largest singular value is significantly larger than the others.

We design the algorithm to adapt to both the thin- and wide-slice regimes. Consider a slice $S_{l,h}$ with conditional mean $\mu_{l,h}$ and conditional covariance matrix $\Sigma_{l,h}$, together with their empirical counterparts $\widehat{S}_{l,h}$, $\widehat{\mu}_{l,h}$, and $\widehat{\Sigma}_{l,h}$. We define $G_{l,h}$ and $\widehat{G}_{l,h}$ as

$$G_{l,h} := \log\left(\frac{\lambda_{\mathrm{mid}}(\Sigma_{l,h})^2}{\lambda_1(\Sigma_{l,h})\lambda_d(\Sigma_{l,h})}\right), \quad \widehat{G}_{l,h} := \log\left(\frac{\lambda_{\mathrm{mid}}(\widehat{\Sigma}_{l,h})^2}{\lambda_1(\widehat{\Sigma}_{l,h})\lambda_d(\widehat{\Sigma}_{l,h})}\right),$$

where for any $d \times d$ square matrix M, $\lambda_1(M) \geqslant \lambda_2(M) \geqslant \cdots \geqslant \lambda_{d-1}(M) \geqslant \lambda_d(M)$ are the eigenvalues of M in descending order, and $\lambda_{\mathrm{mid}}(M) = (\lambda_2(M) \times \cdots \times \lambda_{d-1}(M))^{1/(d-2)}$ is the geometric mean of the eigenvalues excluding the largest and the smallest ones. A slice $S_{l,h}$ is "thin" when $\lambda_d(\Sigma_{l,h}) \ll \lambda_1(\Sigma_{l,h}) \approx \lambda_2(\Sigma_{l,h}) \approx \cdots \approx \lambda_{d-1}(\Sigma_{l,h})$, and hence $G_{l,h} > 0$; the larger $G_{l,h}$ is, the "thinner" $S_{l,h}$ is. A slice $S_{l,h}$ is "wide" when $\lambda_1(\Sigma_{l,h}) \gg \lambda_2(\Sigma_{l,h}) \approx \cdots \approx \lambda_d(\Sigma_{l,h})$, and hence $G_{l,h} < 0$; the more negative $G_{l,h}$ is, the "wider" $S_{l,h}$ is. If $G_{l,h}$ is close to zero, then the geometric shape of the slice $S_{l,h}$ is undetermined: it may be roughly isotropic or may have both very large $\lambda_1(\Sigma_{l,h})$ and very small $\lambda_d(\Sigma_{l,h})$. We define the *significant*

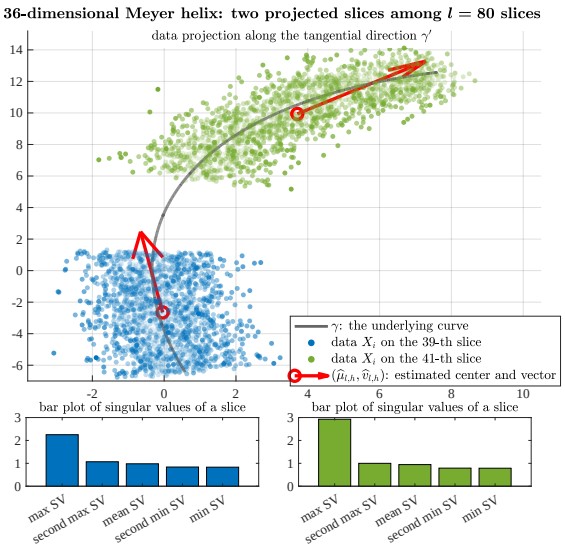

Figure 3: In the same setup and visual conventions of Figure 1, but with the range R split uniformly into 80 intervals, and a different pair of slices. Top: the slices are now elongated along the curve, rather than perpendicularly to it. Bottom: the *largest* singular value is now significantly larger than the remaining ones.

vector $v_{l,h}$ and the *empirical significant vector* $\widehat{v}_{l,h}$ as

$$v_{l,h} := \begin{cases} v_d(\Sigma_{l,h}) & \text{if } G_{l,h} > 0 \\ v_1(\Sigma_{l,h}) & \text{if } G_{l,h} < 0 \end{cases}, \quad \widehat{v}_{l,h} := \begin{cases} v_d(\widehat{\Sigma}_{l,h}) & \text{if } \widehat{G}_{l,h} > 0 \\ v_1(\widehat{\Sigma}_{l,h}) & \text{if } \widehat{G}_{l,h} < 0 \end{cases}.$$

The significant vector $v_{l,h}$ is used to estimate the tangent vector γ' in both the "thin" and "wide" slice scenarios. When $G_{l,h} \approx 0$, we only expect to use the sample mean $\mu_{l,h}$ as a local 0-th order approximation to γ, as the slice has no preferred direction. Crucially, while in the "thin" scenario we expect $\widehat{v}_{l,h}$ to be a good approximation of the tangential vector, in the "wide" scenario the curvature of γ can have a significant effect in our estimation of the local direction of the curve.

2.2 Estimating the nonlinear projection Π_γ by assigning points to the "nearest" slice

Before regressing f, we construct an estimator for $\Pi_\gamma(x)$, which is the value of the parameter in $[0, \mathrm{len}_\gamma]$ of the curve such that $\gamma(\Pi_\gamma(x))$ is the closest point on the curve to a point $x \in \Omega_\gamma$. We design a distance function $\mathrm{dist}(x, h)$ between x and slice $S_{l,h}$ and assign x to the "nearest" slice under this distance function. This assignment maps Ω_γ onto $\{1, \ldots, l\}$ and thus can be interpreted as a zeroth-order approximation of Π_γ. After this assignment, we use $\mu_{l,h}$ and $v_{l,h}$ to obtain a first-order approximation of $\Pi_\gamma(\cdot)$ locally on each slice $S_{l,h}$.

Our choice of distance function $\mathrm{dist}(x, h)$ is dictated by two purposes. First, it should be "local", i.e., the distance between x and the center of a slice should play a role. Second, it should be anisotropic: on any level set $\{x : \Pi_\gamma x = t_i\}$ we have $\langle x - \gamma(t_i), \gamma'(t_i)\rangle = 0$, so the distance to the hyperplane normal to $\gamma'(t_i)$ should play a prominent role, but cannot be too dominant, as there may be multiple slices, even far away from each other, with $\langle x - \gamma(t_i), \gamma'(t_i)\rangle \approx 0$. We therefore consider a distance function of the form

$$|\langle x - \mu_{l,h}, v_{l,h}\rangle| + c\,\|x - \mu_{l,h}\|$$

for some $c > 0$. The value of c cannot be too small, because the distance would then fail for highly self-entangled curves, or too large, because $\mathrm{dist}(x, h)$ should remain small when x is close to the slice $S_{l,h}$. The optimal choice of c depends on the curve and the distribution of data around it. We define the distance function $\mathrm{dist}(x, h)$ separately in the "thin" slice scenario and the "wide" slice scenario:

$$\mathrm{dist}(x,h) = \begin{cases} |\langle x - \mu_{l,h}, v_{l,h}\rangle|^2 + \frac{\lambda_d(\Sigma_{l,h})}{\lambda_1(\Sigma_{l,h})} \|x - \mu_{l,h}\|^2 & \text{if } G_{l,h} > 0 \\ \|x - \mu_{l,h}\|^2 + \frac{\lambda_d(\Sigma_{l,h})}{\lambda_1(\Sigma_{l,h})} |\langle x - \mu_{l,h}, v_{l,h}\rangle|^2 & \text{if } G_{l,h} < 0 \end{cases}, \tag{3}$$

and similarly for its empirical counterpart $\widehat{\mathrm{dist}}(x, h)$. In the "thin" slice scenario, this distance function focuses on measuring the displacement between x and $S_{l,h}$ along the direction $v_{l,h}$, and is less sensitive to the displacement orthogonal to $v_{l,h}$. In the "wide" scenario, $\mathrm{dist}(x, h)$ instead pays less attention to the displacement along $v_{l,h}$ and focuses on the displacement orthogonal to $v_{l,h}$. Note that when the shape of the slice $S_{l,h}$ (and $\widehat{S}_{l,h}$) is roughly isotropic, λ_d/λ_1 is roughly one, so the two cases in (3) are consistent with each other and the distance above varies regularly as $G_{l,h}$ changes sign. This distance function $\mathrm{dist}(x, h)$ has the following advantages: (i) it focuses on local slices while incorporating information about the geometry of each slice; (ii) inside the local neighborhood, it pays special attention to displacement along the tangential direction; (iii) it is distribution-adaptive, allowing, for example, to handle robustly heteroscedasticity in the distribution of X around the curve; (iv) its performance is amenable to mathematical analysis.

We use (3) in the convergence analysis. This distance is a simplification of the Mahalanobis distance from x to a slice $S_{l,h}$ $\mathrm{dist}(x, h) := ||\Sigma_{l,h}^{-1/2}(x-\mu_{l,h})||$, which also puts heavier weight on the displacement along the tangential direction in the "thin" slice scenario and lighter weight in the "wide" slice scenario. As it is a bit harder to analyze mathematically than the distance in (3), we use it only in numerical simulations.

Slices containing few observations yield high-variance estimates and are therefore discarded: let $n_{l,h} := \#\widehat{S}_{l,h}$ be the number of observations in the empirical slice $\widehat{S}_{l,h}$, and define the index set of "heavy" slices

$$\mathcal{H}_l := \{h \in \{1, \dots, l\} : n_{l,h} \geqslant n_{loc}\}.$$

Here n_{loc} is a threshold. Recall that ρ_t denotes the density of the pushforward of ρ_X under Π_γ. If $\rho_t \geqslant c > 0$ on a region of interest, then with n samples and l slices we expect $n_{loc} \asymp cn/l$ for slices whose inverse-image arcs lie in that region. For $x \in \Omega_\gamma$, we define the "nearest" index for the slice that x belongs to as

$$h_x := \arg\min_{h \in \mathcal{H}_l} \mathrm{dist}(x, h),$$

and the "correct" index for the slice that x belongs to as the unique index h'_x such that $F(x) \in R_{l,h'_x}$. For almost every $x \in \Omega_\gamma$, the minimizer h_x is unique because the set of points that cannot be uniquely assigned is a subset of $\{x \in \Omega_\gamma \subseteq \mathbb{R}^d : \exists h', h'' \text{ s.t. } \mathrm{dist}(x, h') = \mathrm{dist}(x, h'')\}$, which is a finite union of hypersurfaces in $\mathbb{R}^d$ and thus has measure zero. Under suitable assumptions, we show that in the thin-slice regime the nearest index h_x is almost correct, i.e., $|h'_x - h_x| \leqslant 1$ (Section 3). The possibility of an error $|h_x - h'_x| = 1$ stems

from the possibility that points near the boundary of a slice can be misclassified to one of its adjacent slices. Given sample data, the "nearest" slice is estimated using the empirical counterpart of the distance, yielding

$$\widehat{h}_x := \arg\min_{h \in \mathcal{H}_l} \widehat{\mathrm{dist}}(x, h)\,.$$

We prove that $|\widehat{h}_x - h_x| \leqslant 1$ with high probability. Both $|h'_x - h_x| = 1$ and $|\widehat{h}_x - h_x| = 1$ arise when points near a slice boundary are assigned to an adjacent slice. Nonetheless, the probability of adjacent misclassification (i.e., $\widehat{h}_x = h'_x \pm 1$) is relatively small but does not decrease to zero as the sample size increases: to avoid artifacts in the regression stage, we include data from adjacent slices to estimate the regression function in each local neighborhood. Combining the two one-slice bounds gives $|\widehat{h}_x - h'_x| \leqslant 2$ on the good event used in the regression analysis. The accepted regression neighborhood uses the ± 1 estimated slices, while the seven-slice projection window in Step 3.a contains the true response index of every accepted point on this event.

2.3 Local estimator of the link function f and global estimator of the regression function F

After assigning a point $x \in \mathbb{R}^d$ to the estimated nearest index $\widehat{h}_x$, we use a piecewise-polynomial estimator in that local neighborhood. For a fixed index h, an observation (X_i, Y_i) from the second half is used when $|\widehat{h}(X_i) - h| \leqslant 1$. This decision uses X_i and the geometry estimated from the first half, but does not use Y_i. We project these inputs onto the line with direction $\widehat{v}_{l,h}$, partition the projected interval $I^{(l,h)}$ into j subintervals, and fit a polynomial on each subinterval by least squares, obtaining $\widehat{f}_{j|\widehat{v}_{l,h}}$. The degree of the local polynomials required to attain estimation rates that are optimal up to logarithmic factors depends on the regularity of the function. A proper partition (or scale) is then chosen to minimize the expected mean squared error (MSE) using classical bias-variance trade-off arguments. Composing this local estimator of f with the nearest-slice assignment gives the untruncated local-polynomial value $\widehat{f}_{j|\widehat{v}_{l,\widehat{h}_x}}(\Pi_{I^{(l,\widehat{h}_x)}}\langle \widehat{v}_{l,\widehat{h}_x}, x\rangle)$, where the projected coordinate is first clipped to the interval $I^{(l,\widehat{h}_x)}$, and where the fit of the nearest retained bin is used, so that the value returned always comes from a bin carrying enough data; Algorithm 1 returns its truncation to $[-M, M]$.

For the theoretical guarantees, $M \geqslant |f|_{L^\infty}$, and n_{loc} is chosen at the scale specified in the proof of the corresponding theorem; in particular, $n_{loc} \geqslant 2d+1$. The interval endpoints are constructed from an independent quarter-sample; clipping and the nearest-retained-bin convention affect only boundary or low-count cases and are analyzed in Lemmas 22–24.

Algorithm 1: Significant Vector Regression

Input : Samples $\{(X_i, Y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \mathbb{R}$, numbers of partitions $l, j \in \mathbb{N}$, polynomial degree $m \in \{0, 1\}$, slice-retention threshold $n_{loc} \geqslant 2d+1$, and truncation level $M \in (0, \infty]$.

Output: $\widehat{F}$ estimate of F.

0. Randomly split and relabel the sample. For simplicity, assume that n is divisible by four. Use $i \leqslant n/4$ for the response slices and local geometry (Steps 1.a–2.b), $n/4 < i \leqslant n/2$ for the projected intervals in Step 3.a, and $i > n/2$ for regression (Steps 3.b–3.c).

1.a Using only $i \leqslant n/4$, compute the response range R. Construct $\{R_{l,h}\}_{h=1}^l$, the uniform partition of R into l intervals, and set $\widehat{S}_{l,h} = \{X_i : i \leqslant n/4,\ Y_i \in R_{l,h}\}$.

1.b Denote $n_{l,h} = \#\{i \leqslant n/4 : Y_i \in R_{l,h}\}$ and let $\mathcal{H}_l := \{h \in \{1, \dots, l\} : n_{l,h} \geqslant n_{loc}\}$. For each $h \in \mathcal{H}_l$, compute:
$\widehat{\mu}_{l,h} = \frac{1}{n_{l,h}} \sum_{i=1}^{n/4} X_i \mathbb{1}\{Y_i \in R_{l,h}\}$, the empirical mean;
$\widehat{\Sigma}_{l,h} = \frac{1}{n_{l,h}} \sum_{i=1}^{n/4} (X_i - \widehat{\mu}_{l,h})(X_i - \widehat{\mu}_{l,h})^\intercal \mathbb{1}\{Y_i \in R_{l,h}\}$, the empirical covariance matrix;
$\widehat{\lambda}_{l,h,r}$, the r-th eigenvalue of $\widehat{\Sigma}_{l,h}$, in descending order, for $r = 1, \dots, d$;
$\widehat{v}_{l,h,r}$, the corresponding r-th eigenvector of $\widehat{\Sigma}_{l,h}$;
$\widehat{G}_{l,h} = \log\left(\frac{(\widehat{\lambda}_{l,h,2} \times \cdots \times \widehat{\lambda}_{l,h,d-1})^{2/(d-2)}}{\widehat{\lambda}_{l,h,1} \widehat{\lambda}_{l,h,d}}\right)$.
Let the empirical significant vector $\widehat{v}_{l,h}$ equal $\widehat{v}_{l,h,d}$ if $\widehat{G}_{l,h} > 0$ and equal $\widehat{v}_{l,h,1}$ otherwise.

2.a Given $x \in \mathbb{R}^d$, for each $h \in \mathcal{H}_l$, compute the estimated distance between x and $\widehat{S}_{l,h}$:

$$\widehat{\operatorname{dist}}(x, h) = \begin{cases} |\langle x - \widehat{\mu}_{l,h}, \widehat{v}_{l,h} \rangle|^2 + \frac{\lambda_d(\widehat{\Sigma}_{l,h})}{\lambda_1(\widehat{\Sigma}_{l,h})} \|x - \widehat{\mu}_{l,h}\|^2 & \text{if } \widehat{G}_{l,h} > 0 \\ \|x - \widehat{\mu}_{l,h}\|^2 + \frac{\lambda_d(\widehat{\Sigma}_{l,h})}{\lambda_1(\widehat{\Sigma}_{l,h})} |\langle x - \widehat{\mu}_{l,h}, \widehat{v}_{l,h} \rangle|^2 & \text{if } \widehat{G}_{l,h} < 0 \end{cases} .$$

2.b Compute the estimated nearest slice index $\widehat{h}_x = \arg\min_{h \in \mathcal{H}_l} \widehat{\operatorname{dist}}(x, h)$.

3.a For each $h \in \mathcal{H}_l$, use the second quarter to compute the smallest interval $I^{(l,h)}$ containing the projected coordinates $\langle \widehat{v}_{l,h}, X_i \rangle$ of all X_i with $n/4 < i \leqslant n/2$ and $|\widehat{h}(X_i) - h| \leqslant 3$, and construct the uniform partition $\{I_{j,k}^{(l,h)}\}_{k=1}^j$.

3.b Let

$$n^{(l,h)} = \#\{i > n/2 : |\widehat{h}(X_i) - h| \leqslant 1\}, \quad h \in \mathcal{H}_l$$
$$n_{j,k}^{(l,h)} = \#\{i > n/2 : |\widehat{h}(X_i) - h| \leqslant 1,\ \langle \widehat{v}_{l,h}, X_i \rangle \in I_{j,k}^{(l,h)}\}, \quad k = 1, \dots, j .$$

Let $\mathcal{K}_j^{(l,h)} = \{k : n_{j,k}^{(l,h)} \geqslant \max\{m+1, c_0 n^{(l,h)}/j\}\}$, where $c_0 > 0$ is a sufficiently small fixed constant depending only on the fixed model-class constants. For each $k \in \mathcal{K}_j^{(l,h)}$, compute

$$\widehat{f}_{j,k|\widehat{v}_{l,h}} = \operatorname*{arg\,min}_{\deg(p) \leqslant m} \sum_{i=n/2+1}^{n} |Y_i - p(\langle \widehat{v}_{l,h}, X_i \rangle)|^2 \, \mathbb{1}\{|\widehat{h}(X_i) - h| \leqslant 1\} \mathbb{1}_{I_{j,k}^{(l,h)}}(\langle \widehat{v}_{l,h}, X_i \rangle) .$$

3.c For $h \in \mathcal{H}_l$, compute $\widehat{f}_{j|\widehat{v}_{l,h}}(r) = \widehat{f}_{j,k|\widehat{v}_{l,h}}(r)$ for $k = k^{(l,h)}(r) := \arg\min_{k \in \mathcal{K}_j^{(l,h)}} \operatorname{dist}(r, I_{j,k}^{(l,h)})$, the index of the retained bin nearest to r, and $\widehat{f}_{j|\widehat{v}_{l,h}} \equiv 0$ if $\mathcal{K}_j^{(l,h)} = \emptyset$; let Π_I be the metric projection onto the closed interval I; then return the estimator truncated at level M,

$$\widehat{F}(x) = \left(\widehat{f}_{j|\widehat{v}_{l,\widehat{h}_x}}(\Pi_{I^{(l,\widehat{h}_x)}} \langle \widehat{v}_{l,\widehat{h}_x}, x \rangle) \vee (-M)\right) \wedge M \ ,$$

2.4 The main algorithm and guarantees on the estimator it produces

Algorithm 1 summarizes the construction of the proposed estimator of the regression function F. The sample split makes the geometry independent of the responses used for regression. Since each half has order n observations, the split changes only numerical constants in the rates. We prove, under the assumptions detailed in Section 3, that the mean squared error of our estimator is the sum of an estimation term and a curve-approximation term. In the main thin-slice regime, which uses Assumption **(LCV)**, the estimation term decays at the one-dimensional minimax nonparametric rate up to logarithmic factors, down to a saturation level determined by quantities that are expected to be small, or even zero, as discussed below. At a schematic level, suppressing fixed model constants, the proof yields

$$\mathrm{MSE} \lesssim \underbrace{\text{saturation}}_{\text{geometry/noise}} + \underbrace{\text{slice-geometry error}}_{\text{PCA/assignment}} + \underbrace{(lj)^{-2s}}_{\text{regression bias}} + \underbrace{\sigma_\zeta^2\, lj \log j/n}_{\text{regression variance}} .$$

The formal theorem chooses l and j to balance the decaying terms while respecting the sample-size and noise-scale constraints.

Theorem 3 (MSE of the estimator constructed by Algorithm 1) *Assume that $\sigma_\zeta > 0$, that the conditions* **($\mathbf{X} \in \psi_2 \cap \mathcal{C}^2$)**, **($\mathbf{Y} \in \psi_2$)**, **($\boldsymbol{\zeta} \in \psi_2$)**, **($\boldsymbol{\gamma_1}$)**, **(LCV)**, *and* **($\boldsymbol{\omega_f}$)** *hold, and that $f \in \mathcal{C}^s(\mathbb{R}^1)$ with $s \in \left[\frac{1}{2}, 1\right]$. Let $C_{\gamma,f}, M^*, l_{\max}$ be the constants specified in (4) below, and C a universal constant. Then, for n such that $\frac{n}{\log^{3/2} n} \gtrsim C_{\gamma,f} \frac{C_f \mathrm{len}_\gamma}{C'_f \sigma_\gamma}$, choose, up to integer rounding,*

$$l^* := \min\left\{ C_{\gamma,f}^{-1} n \log^{-2} n,\ l_{\max},\ M^* n^{\frac{1}{2s+1}} \right\} , \qquad j^* := \left\lceil C \frac{M^* n^{\frac{1}{2s+1}}}{l^*} \right\rceil .$$

Then the estimator constructed by Algorithm 1, with running time $\mathcal{O}(d^2 n \log n)$, satisfies

$$\mathbb{E}\left[\left| \widehat{F}(X) - F(X) \right|^2 \right] \lesssim \underbrace{C_1(f, \gamma, \rho_X, \sigma_\zeta, d)\, n^{-\frac{2s}{2s+1}} \log n}_{\textit{one-dimensional estimation term}} + \underbrace{C_2(f, \gamma, \rho_X, \sigma_\zeta, d)}_{\textit{geometry- and noise-dependent saturation term}} .$$

Evaluating $\widehat{F}$ at a new point can be achieved in $O(dn^{\frac{1}{2s+1}})$ operations.

The constants in the preceding theorem are

$$C_1(f, \gamma, \rho_X, \sigma_\zeta, d) := \left(\frac{C_f}{C'_f} \mathrm{len}_\gamma \sigma_\zeta^2 \right)^{\frac{2s}{2s+1}} \left([f]_{\mathcal{C}^s} + |f|_{L^\infty} [\rho_X]_{\mathcal{C}^s} \right)^{\frac{2}{2s+1}} , \tag{4}$$

$$C_2(f, \gamma, \rho_X, \sigma_\zeta, d) := [f]_{\mathcal{C}^s}^2 \left(\frac{\bar{\sigma}_\gamma C_f}{\mathrm{reach}_\gamma} \max(\sigma_\zeta, \omega_f) \right)^{2s}, \ C_{\gamma,f} := \frac{R_0^3 C_f^2}{C'_f} \max\left(\frac{\mathrm{len}_\gamma d^{3/2}}{\sigma_\gamma^4}, \frac{R_0^5 C_f^2 d^4}{C_f'^3 \sigma_\gamma^8} \right) ,$$

$$M^* := \left(\sigma_\zeta^{-1} (C_f C_Y R_0)^s \left([f]_{\mathcal{C}^s} + |f|_{L^\infty} [\rho_X]_{\mathcal{C}^s} \right) \right)^{\frac{2}{2s+1}} \quad , \quad l_{\max} := \frac{C C_Y R_0}{\max(\sigma_\zeta, \omega_f)} .$$

Under the assumptions of Theorem 3, our estimator achieves the minimax-optimal one-dimensional nonparametric regression rate, up to logarithmic factors, together with an

additional approximation error of order $\mathcal{O}(\sigma_\zeta^2)$ for Lipschitz monotone links. In this precise sense, the estimator defeats the curse of dimensionality by exploiting the compositional structure of F, even though the inner function is nonlinear and its regularity does not scale with the ambient dimension. Theorem 3 should not be read as a guarantee for arbitrary nonlinear links: coarse monotonicity is used to make response slices local, and Assumption **(LCV)** is used to make the smallest principal component in a slice identify the tangential direction. The result of Theorem 3 is scaling invariant in X and Y.

The rate exponent is independent of the ambient dimension, while the sample-size thresholds and computational constants grow only polynomially in d and in the geometric complexity of the curve. The regularity assumptions on f and γ are also dimension independent. Moreover, the slice centers and significant vectors provide an interpretable piecewise-linear sketch of γ and local estimates of the nonlinear coordinate Π_γ, in addition to the final estimate of F.

Remark 4 *The regression bias depends on l and j through their product. We therefore take l^* as large as permitted by the sample-size and noise-scale constraints and allocate the remaining optimal product budget to j^*, so that*

$$l^* j^* \asymp M^* n^{1/(2s+1)}.$$

The proof verifies the lower slicing-scale requirement and the retained-bin count condition in all regimes. We keep the ratio C_f/C_f' explicit in the constants to display the dependence on the link function.

There are three important scope restrictions in our results:

(i) The link f is coarsely monotone, with the fixed ratio C_f/C_f' bounded by an absolute constant. Coarse monotonicity makes each response slice correspond to a localized portion of the curve; the bounded ratio ensures that equal response intervals correspond to comparable arclengths and that the fixed-width buffer in Step 3.a is sufficient. Without such localization, a response slice may contain many disconnected curve segments and inverse regression would require an additional clustering or disambiguation mechanism.

(ii) Assumption **(LCV)** requires enough variation normal to the curve relative to the observational noise and the coarse-monotonicity scale. It is central to the thin-slice theorem because it makes the smallest principal component of a slice align with the tangent. When **(LCV)** fails, Theorem 6 gives a wide-slice guarantee instead.

(iii) In the noisy theorem, the curve-approximation term does not vanish with n. It is small when the observational noise or curvature is small, vanishes for a straight line, and, by (6), decreases with the ambient dimension in the regime of the model. Removing this term would require higher-order local geometric approximations, which we leave to future work.

In the noisy regime of Theorem 3, the computational complexity of constructing $\widehat{F}$ is near-optimal. Its evaluation cost at a new point could potentially be reduced by building a

multiscale data structure (e.g. cover trees) of size $O(n \log n)$, similar to those used for fast nearest-neighbor search, that exploits the intrinsic one-dimensionality of the curve γ and the family of slices. Such a structure could reduce the evaluation cost to $O(d \log n)$ per query.

For a strictly monotone, Lipschitz function with zero observational noise, we obtain the rate $\mathcal{O}(n^{-2})$. Here, the curve-approximation term vanishes because there is no upper bound on the number of slices l, contrary to the noisy case in Theorem 3:

Theorem 5 (MSE of Algorithm 1 in the noiseless case) *Assume that* $(\mathbf{X} \in \psi_2 \cap \mathcal{C}^2)$, $(\mathbf{Y} \in \psi_2)$, $(\boldsymbol{\zeta} \in \psi_2)$, $(\boldsymbol{\gamma_1})$, **(LCV)**, *and* $(\boldsymbol{\omega_f})$ *hold. Assume that there is no observational noise, i.e.,* $\zeta \equiv 0$ *almost surely, that the link function* f *is perfectly monotone, i.e.,* $\omega_f = 0$, *and that* $f \in \mathcal{C}^s$ *with* $s \in \left[\frac{1}{2}, 1\right]$. *Let* $c_\rho > 0$ *be arbitrary, set* $\Omega_0(c_\rho) := \{x \in \Omega_\gamma : \rho_t(\Pi_\gamma x) > c_\rho\}$, *and write* $\Omega_0 = \Omega_0(c_\rho)$ *for brevity, where* ρ_t *is the density of the pushforward of* ρ_X *under* Π_γ. *With* $C_{\gamma,f}$ *in* (4), *when* $\frac{n}{\log^{3/2} n} \gtrsim \frac{C_f C_{\gamma,f}}{c_\rho C'_f \sigma_\gamma}$, *if we choose* $l^* = C \frac{C'_f c_\rho \mathrm{len}_\gamma}{C_f C_{\gamma,f}} \frac{n}{\log^{3/2} n}$ *and* $j^* = C$, *then the estimation error of Algorithm 1 satisfies*

$$\mathbb{E}\left[\left|\widehat{F}(X) - F(X)\right|^2\right] \lesssim ([f]_{\mathcal{C}^s} + |f|_{L^\infty} [\rho_X]_{\mathcal{C}^s})^2 \left(\frac{C_f^2 C_{\gamma,f}^2}{c_\rho^2 C_f'^2} \frac{\log^3 n}{n^2}\right)^s + \mathbb{P}(X \notin \Omega_0)\, |f|_{L^\infty}^2 \ .$$

Thus the theorem gives a family of localized bounds indexed by the arbitrary level $c_\rho > 0$. *If, in addition,* ρ_t *has a uniform positive lower bound, one may choose* c_ρ *below it, in which case* $\Omega_0 = \Omega_\gamma$ *and the complementary risk term vanishes; the estimation error is then* $\tilde{O}(n^{-2s})$, *and in particular* $\tilde{O}(n^{-2})$ *at the Lipschitz endpoint* $s = 1$. *Here no cap on* l^* *is needed:* $\sigma_\zeta = \omega_f = 0$ *gives* $l_{\max} = \infty$.

The set Ω_0 makes explicit the low-density qualification in the noiseless result: after removing a uniform lower bound on the pushforward density ρ_t from the main assumptions, the fast noiseless rate is asserted on regions where ρ_t is not too small, with the remaining contribution controlled by $\mathbb{P}(X \notin \Omega_0)\, |f|_{L^\infty}^2$.

The faster rate $\mathcal{O}((n^{-2} \log^3 n)^s)$ is a consequence of having zero observational noise. Bauer et al. (2017) proves that, in the case of the L^∞-norm, the minimax rate of non-parametric regression for functions in $\mathcal{C}^s(\mathbb{R}^d)$ with noiseless observations is $(\log n/n)^{s/d}$. The mean-squared-error rate in Theorem 5 is consistent with this benchmark throughout $s \in [\frac{1}{2}, 1]$ and $d = 1$, as if γ (and therefore Π_γ) were known. The restriction to the Hölder–Lipschitz range matches the first-order local approximation of the curve and its closest-point coordinate used by Algorithm 1; exploiting regularity beyond the Lipschitz level would require higher-order local geometric approximations. Our estimator avoids the curse of dimensionality by exploiting the compositional structure of F, even though the inner function is unknown, nonlinear, and only moderately smooth, with regularity that does not scale with the ambient dimension. A key difference between Theorem 3 and Theorem 5 is that there is no curve-approximation term in the latter: noiseless observations allow us to perform an unlimited amount of partitioning, obtaining "thin" slices, provided that enough samples are available (to control the variance of the objects we estimate in each slice).

When Assumption **(LCV)** is not satisfied, our estimator admits the following saturation bound:

Theorem 6 (NSVM without (LCV)) *Assume that* $(\mathbf{X}\in \psi_2 \cap \mathcal{C}^2)$, $(\mathbf{Y}\in \psi_2)$, $(\boldsymbol{\zeta} \in \psi_2)$, $(\boldsymbol{\gamma_1})$, $(\boldsymbol{\omega_f})$, *and* **(SC)** *hold. Assume also that* $f \in \mathcal{C}^s$ *with* $s \in [\frac{1}{2}, 1]$. *When* $\frac{n}{\log^{1/2} n} \gtrsim \frac{C_Y R_0^7 d^3}{C_f^6 \max(\sigma_\zeta, \omega_f)^7}$, *if we choose* $l^* = \frac{C_Y R_0}{\max(\sigma_\zeta, \omega_f)}$ *and* j^* *the smallest integer for which*

$$[f]_{\mathcal{C}^s} \left(\frac{C_f \max(\sigma_\zeta, \omega_f)}{j^*} \right)^s \leqslant [f]_{\mathcal{C}^s} \left(\frac{\bar{\sigma}_\gamma C_f}{\mathrm{reach}_\gamma} \max(\sigma_\zeta, \omega_f) \right)^s, \tag{5}$$

and if, in addition,

$$\frac{\sigma_\zeta^2 l^* j^* \log j^*}{n} \lesssim [f]_{\mathcal{C}^s}^2 \left(\frac{\bar{\sigma}_\gamma C_f}{\mathrm{reach}_\gamma} \max(\sigma_\zeta, \omega_f) \right)^{2s},$$

then the estimation error of Algorithm 1 satisfies

$$\mathbb{E}\left[\left| \widehat{F}(X) - F(X) \right|^2 \right] \lesssim [f]_{\mathcal{C}^s}^2 \left(\frac{\bar{\sigma}_\gamma C_f}{\mathrm{reach}_\gamma} \max(\sigma_\zeta, \omega_f) \right)^{2s}.$$

Condition (5) chooses enough local bins to make the polynomial approximation bias no larger than the wide-slice saturation level. The corresponding bias–variance calculation is given in the proof.

3. Analysis of the Estimator

We introduce several properties of the model that will be assumed in various combinations throughout the analysis. We start by collecting a few conditions on the distribution of X, Y, and ζ that are fairly standard:

$(\mathbf{X}\in \psi_2 \cap \mathcal{C}^2)$ X is sub-Gaussian with variance proxy R_0^2 and has a compactly supported density $\rho_X \in \mathcal{C}^2$ with $[\rho_X]_{\mathcal{C}^2} < \infty$.

$(\mathbf{Y}\in \psi_2)$ Y has sub-Gaussian distribution with variance proxy $C_Y^2 R_0^2$.

$(\boldsymbol{\zeta} \in \psi_2)$ ζ is centered and sub-Gaussian with variance proxy σ_ζ^2.

Recall that X can be decomposed as position along the underlying curve and deviation away from the curve γ as in (1). The following assumption on Z_{d-1} describes how X deviates from the curve γ:

$(\boldsymbol{\gamma_1})$ γ has a Lipschitz derivative $\gamma' : [0, \mathrm{len}_\gamma] \to \mathbb{R}^d$. For each $t_0 \in [0, \mathrm{len}_\gamma]$, the conditional random vector $Z_{d-1}|t = t_0$ is mean zero, isotropic with variance $\sigma_\gamma(t_0)^2 \mathbb{I}_{d-1}$, and supported in a Euclidean ball $B(0, c\,\mathrm{reach}_\gamma(t_0)) \subseteq \mathbb{R}^{d-1}$ for some $c < 1$.

The ball is centered at the origin of the normal space $\mathbb{R}^{d-1}$, not at $\gamma(t_0) \in \mathbb{R}^d$: the condition bounds the deviation from the curve, so that $\|X - \gamma(\Pi_\gamma X)\| \leqslant c\,\mathrm{reach}_\gamma$ almost surely. Combined with isotropy, it also forces the tube to be thin relative to the reach in every direction: $\sigma_\gamma \lesssim \mathrm{reach}_\gamma / \sqrt{d-1}$. This assumption is satisfied, for example, by the following

natural generative model: sample a point $\gamma(t)$ on the curve and, conditional on t, generate X according to (1), with Z_{d-1} having a sub-Gaussian distribution as in Assumption $(\boldsymbol{\gamma_1})$. This does not imply that, marginally, the points in the normal directions to the curve are uniformly or isotropically distributed.

Assumption $(\boldsymbol{\gamma_1})$ implies that for any $t_0 \in [0, \mathrm{len}_\gamma]$, the conditional mean is on the curve

$$\mu_{t_0} := \mathbb{E}[X \mid t = t_0] = \gamma(t_0)\,,$$

and the conditional covariance matrix has eigenvalue 0 on the eigenspace $\mathrm{span}\{\gamma'(t_0)\}$ and eigenvalue $\sigma_\gamma(t_0)^2$ on eigenspace $\mathrm{span}\{\gamma'(t_0)\}^\perp$, since

$$\Sigma_{t_0} := \mathbb{E}[(X - \mu_{t_0})(X - \mu_{t_0})^\intercal \mid t = t_0] = \sigma_\gamma(t_0)^2 \mathbb{I}_d - \sigma_\gamma(t_0)^2 \gamma'(t_0)\gamma'(t_0)^\intercal\,.$$

Because γ is unknown, we cannot condition on $t = \Pi_\gamma X$. Instead, we condition on the observed response Y_i and require a property that partially reveals the one-to-one correspondence between $t = \Pi_\gamma X$ and $F(X) = f(\Pi_\gamma X)$:

$(\boldsymbol{\omega_f})$ Let $\mathcal{R}_f := f([0, \mathrm{len}_\gamma])$ be the range of the link. There exist constants $\omega_f \geqslant 0$ and $C_f > C_f' > 0$, depending only on f, such that, for every interval T with $|T \cap \mathcal{R}_f| \geqslant \omega_f$,

$$C_f' |T \cap \mathcal{R}_f| \leqslant \left|\left[\min f^{-1}(T), \max f^{-1}(T)\right]\right| \leqslant C_f |T \cap \mathcal{R}_f|.$$

The fixed ratio C_f / C_f' is bounded by an absolute constant.

Assumption $(\boldsymbol{\omega_f})$ is a large-scale bi-Lipschitz condition. If f is bi-Lipschitz, it holds with $\omega_f = 0$; for $\omega_f > 0$, it allows bounded oscillations below the scale ω_f, and we refer to this as coarse monotonicity. If an interval extends outside $\mathcal{R}_f$, its inverse image is unchanged after intersecting the interval with $\mathcal{R}_f$, so the upper inverse-image estimates used below are unaffected. Throughout the theoretical analysis, the fixed response scale $C_Y R_0$ is chosen comparable to $|\mathcal{R}_f|$; changing it by a fixed factor changes only fixed constants. The bounded ratio C_f / C_f' implies that equal response intervals correspond to curve segments of comparable arclength. The two elementary consequences needed below are summarized in Lemma 12.

The following assumption gives a lower bound on the conditional variance:

(LCV) Define $\sigma_\gamma := \min_{t_0 \in [0, \mathrm{len}_\gamma]} \sigma_\gamma(t_0)$ as the minimum value of $\sigma_\gamma(t_0)$. We also write $\bar{\sigma}_\gamma := \max_{t_0 \in [0, \mathrm{len}_\gamma]} \sigma_\gamma(t_0)$. We assume that $\sigma_\gamma \geqslant C_{\mathrm{LCV}} C_f \max(\sigma_\zeta, \omega_f)$, where $C_{\mathrm{LCV}} > 0$ is a sufficiently large universal constant.

The minimum σ_γ is used for lower eigengap and concentration bounds, whereas the maximum $\bar{\sigma}_\gamma$ is used when the off-curve variation must be bounded from above. The purpose of Assumption **(LCV)** is that for any interval $T \subseteq [0, \mathrm{len}_\gamma]$, it allows us to compute the conditional mean

$$\mu_T := \mathbb{E}[X \mid t \in T] = \mathbb{E}_t[\gamma(t) \mid t \in T]$$

and conditional covariance matrix

$$\begin{aligned}\Sigma_T &:= \mathbb{E}[(X - \mu_T)(X - \mu_T)^\intercal \mid t \in T] \\ &= \mathbb{E}\left[(\gamma(t) - \mu_T)(\gamma(t) - \mu_T)^\intercal \mid t \in T\right] + \mathbb{E}[\sigma_\gamma(t)^2 \mid t \in T]\mathbb{I}_d - \mathbb{E}[\sigma_\gamma(t)^2 \gamma'(t)\gamma'(t)^\intercal \mid t \in T]\ .\end{aligned}$$

This identity explains why Significant Vector Regression estimates the tangent direction when the slices are sufficiently thin. If T is small enough that the first term has negligible spectral norm relative to the remaining terms, then the second term is a multiple of the identity and the third is negative semidefinite. Consequently, the smallest principal component of Σ_T is approximately the largest principal component of $\mathbb{E}[\sigma_\gamma(t)^2\gamma'(t)\gamma'(t)^\intercal \mid t \in T]$, and hence estimates the local tangent direction.

Given the discussion above, thin slices are desirable. This requires both a small interval T and a lower bound on σ_γ, which motivates Assumption **(LCV)**: scales below the observational-noise level σ_ζ or the coarse-monotonicity scale ω_f are not informative for our inverse-regression approach. Assumption **(LCV)** is therefore not merely a proof convenience; it specifies the regime in which the smallest-vector slice geometry used in the main theorem is statistically stable. When this condition fails, the estimator is analyzed separately in the wide-slice setting of Theorem 6.

The assumptions above impose several restrictions on the parameters of γ: Assumption $(\boldsymbol{\gamma_1})$ implies that for $t_0 \in [0, \mathrm{len}_\gamma]$, $\sigma_\gamma(t_0) < \min(R_0, \mathrm{reach}_\gamma/\sqrt{d})$; Assumption **(LCV)** implies that $\sigma_\gamma \geqslant C_{\mathrm{LCV}} C_f \sigma_\zeta$; and Assumption $(\boldsymbol{\omega_f})$, together with the response-scale convention above, gives $C_f'|\mathcal{R}_f| \leqslant \mathrm{len}_\gamma \leqslant C_f|\mathcal{R}_f|$ with $|\mathcal{R}_f| \asymp C_Y R_0$.

Combining two of these gives a fact that is used twice below and is worth isolating. Assumption $(\boldsymbol{\gamma_1})$ places the conditional law of the normal component in a ball of radius $c\,\mathrm{reach}_\gamma(t_0)$ of the normal space $\mathbb{R}^{d-1}$ and requires it to be isotropic, which forces $\sqrt{d-1}\,\sigma_\gamma \leqslant \sqrt{d-1}\,\bar{\sigma}_\gamma \leqslant c\,\mathrm{reach}_\gamma$; combining this with Assumption **(LCV)** gives

$$C_f \max(\sigma_\zeta, \omega_f) \;\leqslant\; \frac{\sigma_\gamma}{C_{\mathrm{LCV}}} \;\leqslant\; \frac{c\,\mathrm{reach}_\gamma}{C_{\mathrm{LCV}}\sqrt{d-1}}\,. \tag{6}$$

Two consequences are worth recording. First, (6) is precisely the second condition of Assumption **(SC)**, with a constant better by a factor $\sqrt{d-1}$ than that assumption requires: the apparent asymmetry between **(LCV)**, which contains no upper bound on $C_f \max(\sigma_\zeta, \omega_f)$, and **(SC)**, which contains one, is therefore only apparent, since in the thin-slice regime the upper bound is automatic. Second, the saturation constant of Theorem 3 obeys

$$C_2 = [f]_{\mathcal{C}^s}^2 \left(\frac{\bar{\sigma}_\gamma C_f}{\mathrm{reach}_\gamma} \max(\sigma_\zeta, \omega_f)\right)^{2s} \;\lesssim\; [f]_{\mathcal{C}^s}^2 \left(\frac{\bar{\sigma}_\gamma}{\sqrt{d-1}}\right)^{2s},$$

so the curve-approximation term does not merely fail to vanish in n: it decays in the ambient dimension. The non-vanishing term discussed in the third scope restriction of Section 2.4 is thus harmless precisely in the high-dimensional regime that motivates the model.

3.1 Estimation of slice parameters with Assumption (LCV)

Consider the event of bounded data

$$\mathcal{B} := \left\{\|X\| \leqslant C_X\sqrt{d}R_0, |Y| \leqslant C_Y R_0\right\}$$

for some $C_X, C_Y \geqslant 1$ fixed constant from now on. We define the following bounded version of $\mu_{l,h}$ and $\Sigma_{l,h}$:

$$\mu_{l,h}^b := \mathbb{E}\left[X | Y \in R_{l,h}, \mathcal{B}\right], \quad \Sigma_{l,h}^b := \mathrm{Cov}\left[X | Y \in R_{l,h}, \mathcal{B}\right] \;.$$

Given the event

$$\mathcal{B}_i := \{\|X_i\| \leqslant C_X\sqrt{d}R_0, |Y_i| \leqslant C_Y R_0\} \ ,$$

we define

$$n^b := \sum_i \mathbb{1}(\mathcal{B}_i), \quad n^b_{l,h} := \sum_i \mathbb{1}(\{Y_i \in R_{l,h}\} \cap \mathcal{B}_i) \ .$$

The random variable n^b, assuming $(\mathbf{X} \in \psi_2 \cap \mathcal{C}^2)$ and $(\mathbf{Y} \in \psi_2)$, is larger than a constant fraction of n with high probability. The empirical counterparts of $\mu^b_{l,h}$ and $\Sigma^b_{l,h}$ are

$$\widehat{\mu}^b_{l,h} := \frac{1}{n^b_{l,h}} \sum_{i:\{Y_i \in R_{l,h}\} \cap \mathcal{B}_i} X_i \, , \, \widehat{\Sigma}^b_{l,h} := \frac{1}{n^b_{l,h}} \sum_{i:\{Y_i \in R_{l,h}\} \cap \mathcal{B}_i} (X_i - \widehat{\mu}^b_{l,h})(X_i - \widehat{\mu}^b_{l,h})^\intercal \, .$$

We denote by $v^b_{l,h}$ and $\widehat{v}^b_{l,h}$ the significant vector of $\Sigma^b_{l,h}$ and $\widehat{\Sigma}^b_{l,h}$, respectively. Given $\mathcal{B}$, it is natural to pick the interval R in Significant Vector Regression as

$$R := [-C_Y R_0, C_Y R_0] \ .$$

Taking slices $Y \in R_{l,h}$ is equivalent to conditioning on $Y \in R_{l,h}$. In this procedure, we obtain information on conditional random variables such as the slice center $\mu^b_{l,h} := \mathbb{E}[X|Y \in R_{l,h}, \mathcal{B}]$, the slice covariance matrix $\Sigma^b_{l,h}$, and the slice significant vector $v^b_{l,h}$. Moreover, the eigenvalues $\lambda_1(\Sigma^b_{l,h}), \ldots, \lambda_d(\Sigma^b_{l,h})$ describe the geometry of the conditional slice $S^b_{l,h}$; in particular, the smallest eigenvalue $\lambda_d(\Sigma^b_{l,h})$ determines its width. We next show that these parameters can be estimated accurately with a moderate sample size. Recall that $\Pi_\gamma : \Omega_\gamma \to [0, \mathrm{len}_\gamma]$ maps points to the one-dimensional interval $[0, \mathrm{len}_\gamma]$ that encodes position along the curve γ. We start with a proposition on estimating slice position on curve $\Pi_\gamma X|Y \in T, \mathcal{B}$.

Proposition 7 *Suppose* $(\boldsymbol{\zeta} \in \psi_2)$ *and* $(\boldsymbol{\omega_f})$ *hold. Let* $T \subseteq R$ *be a bounded interval with* $|T| \geqslant \max(\sigma_\zeta, \omega_f)$. *Then:*

(a) *For every* $i = 1, \ldots, n$ *and every* $\tau \geqslant 1$

$$\mathbb{P}\left\{|\Pi_\gamma X_i - \mathbb{E}\left[\Pi_\gamma X|Y \in T, \mathcal{B}\right]| \gtrsim C_f(|T| + \sqrt{\tau \log n}\sigma_\zeta) \mid Y_i \in T, \mathcal{B}_i\right\} \leqslant 2n^{-\tau} \, .$$

(b) $\mathrm{Var}\left[\Pi_\gamma X|Y \in T, \mathcal{B}\right] \lesssim C_f^2(|T|^2 + \sigma_\zeta^2)$.

See Appendix A.1 for the proof. We now bound the estimation error for the tangential direction and the smallest eigenvalue $\lambda_d(\Sigma^b_{l,h})$:

Proposition 8 *Suppose* $(\boldsymbol{\zeta} \in \psi_2)$, $(\boldsymbol{\gamma_1})$ *and* $(\boldsymbol{\omega_f})$ *hold. Let* $T \subseteq R$ *be a bounded interval with* $|T| \geqslant \max(\sigma_\zeta, \omega_f)$. *Then:*

(a) *For every* $i = 1, \ldots, n$ *and every* $\tau \geqslant 1$, *we have*

$$\mathbb{P}\left\{\left|\langle v^b_{l,h}, X_i\rangle - \mathbb{E}[\langle v^b_{l,h}, X\rangle|Y \in T, \mathcal{B}]\right| \gtrsim C_f(|T| + \sqrt{\tau \log n}\sigma_\zeta) \mid Y_i \in T, \mathcal{B}_i\right\} \leqslant 2n^{-\tau} \, ;$$

(b) $\lambda_d(\Sigma^b_{l,h}) = \mathrm{Var}\left[\langle v^b_{l,h}, X\rangle \mid Y \in T, \mathcal{B}\right] \lesssim C_f^2(|T|^2 + \sigma_\zeta^2)$.

We now show under which assumptions $\lambda_d(\Sigma^b_{l,h})$ is small compared to $\lambda_{d-1}(\Sigma^b_{l,h})$ and the other eigenvalues, yielding the "thin-slice scenario":

Corollary 9 *Suppose* $(\mathbf{X}\in \psi_2 \cap \mathcal{C}^2)$, $(\mathbf{Y}\in \psi_2)$, $(\boldsymbol{\zeta} \in \psi_2)$, $(\boldsymbol{\gamma_1})$, **(LCV)**, *and* $(\boldsymbol{\omega_f})$ *hold. Then, for every* l *such that* $l \gtrsim C_f C_Y R_0/\sigma_\gamma$, *and* $|R^b_{l,h}| \geqslant \max(\sigma_\zeta, \omega_f)$, *we have*

$$\lambda_{d-1}(\Sigma^b_{l,h}) - \lambda_d(\Sigma^b_{l,h}) \gtrsim \sigma^2_\gamma \,.$$

See Appendix A.2 for a proof of the proposition and its corollary.

In part 1. b) of Algorithm 1, we compute on each slice its sample mean $\widehat{\mu}^b_{l,h}$, sample covariance matrix $\widehat{\Sigma}^b_{l,h}$, eigenvalues $\widehat{\lambda}_{l,h,m}$ and eigenvectors $\widehat{v}_{l,h,m}$ of the sample covariance matrix. It is natural to ask how accurately these parameters can be estimated: we address this in Proposition 20 and Lemma 19. These technical results are postponed to the Appendix. Here we record the resulting near square-root consistency in terms of the local sample size within each slice.

Corollary 10 *Suppose* $(\mathbf{X}\in \psi_2 \cap \mathcal{C}^2)$, $(\mathbf{Y}\in \psi_2)$, $(\boldsymbol{\zeta} \in \psi_2)$, $(\boldsymbol{\gamma_1})$, **(LCV)**, *and* $(\boldsymbol{\omega_f})$ *hold. Then, for every* l *such that* $l \gtrsim C_f C_Y R_0/\sigma_\gamma$, *and* $|R^b_{l,h}| \geqslant \max(\sigma_\zeta, \omega_f)$ *for all* h, *for every* $t > 0$ *and* $\tau \geqslant 1$, *if* n *is sufficiently large so that* $\frac{n^b_{l,h}}{\sqrt{\tau \log n}} \gtrsim \left(\frac{C_f C_Y R_0}{\sigma_\gamma}\right)^2 d(t + \log d + \log l)$, *we have*

$$\mathbb{P}\left\{\exists h : \left|\langle v^b_{l,h}, \widehat{\mu}^b_{l,h} - \mu^b_{l,h}\rangle\right| \gtrsim C_Y C_f R_0 \sqrt{t + \log l + \log d}\sqrt{\frac{\sqrt{\tau \log n}}{n^b_{l,h} l^2}} \;\middle|\; n^b_{l,h}\right\} \lesssim e^{-t} + n^{-\tau} \;\; ;$$

$$\mathbb{P}\left\{\exists h : \left\|\widehat{v}^b_{l,h} - v^b_{l,h}\right\| \gtrsim C_Y C_f R_0^2 \sigma_\gamma^{-2} \sqrt{t + \log l + \log d}\sqrt{\frac{d\sqrt{\tau \log n}}{n^b_{l,h} l^2}} \;\middle|\; n^b_{l,h}\right\} \lesssim e^{-t} + n^{-\tau} \;\; .$$

Here and below, each $\widehat{v}^b_{l,h}$ *is oriented so that* $\langle \widehat{v}^b_{l,h}, v^b_{l,h}\rangle \geqslant 0$. *Moreover, if* $\frac{n^b_{l,h}}{\log^{3/2} n} \gtrsim C_Y^2 C_f^2 R_0^4 \sigma_\gamma^{-4} d(\log d + \log l)$, *then for any* h *and* $p \geq \frac{1}{2}$,

$$\mathbb{E}\left[\left\|\widehat{v}^b_{l,h} - v^b_{l,h}\right\|^{2p} \;\middle|\; n^b_{l,h}\right] \lesssim C(p)(C_Y C_f R_0^2 \sigma_\gamma^{-2})^{2p} (d \log d)^p \left(\frac{(\log n)^{1.5}}{n^b_{l,h} l^2}\right)^p \;\; .$$

This follows from Appendix A.3.

3.2 Estimation of the distance function and classification accuracy

Here we assume $(\mathbf{X}\in \psi_2 \cap \mathcal{C}^2)$, $(\mathbf{Y}\in \psi_2)$, $(\boldsymbol{\zeta} \in \psi_2)$, $(\boldsymbol{\gamma_1})$, **(LCV)**, and $(\boldsymbol{\omega_f})$: by Corollary 9 we are in "thin" slice scenario, i.e., $G_{l,h} > 0$, and the distance function simplifies to $\widehat{\mathrm{dist}}(x,h) = |\langle x - \widehat{\mu}^b_{l,h}, \widehat{v}^b_{l,h}\rangle|^2 + \frac{\lambda_d(\widehat{\Sigma}^b_{l,h})}{\lambda_1(\widehat{\Sigma}^b_{l,h})}\left\|x - \widehat{\mu}^b_{l,h}\right\|^2$. In Algorithm 1, we use the slice $\widehat{h}_x$ having the smallest estimated distance to x as an estimator of the correct index h'_x. The following proposition states that the population nearest index h_x is almost correct; its proof is deferred to Appendix A.5.

Proposition 11 (nearest index is almost correct) *Let $h_x = \arg\min_{h\in\mathcal{H}_l} \operatorname{dist}(x,h)$ be the nearest index and define the correct index h'_x to be the unique h such that $F(x) \in R_{l,h}$. Suppose* **($\boldsymbol{\zeta} \in \psi_2$)**, **($\boldsymbol{\omega_f}$)**, *and* **(LCV)** *hold. Suppose that $|R_{l,h}| \geqslant \max(\sigma_\zeta, \omega_f)$ for all $h \in \mathcal{H}_l$. Suppose in addition that at least one of the correct response slice and its adjacent slices is retained, i.e., $\{h'_x - 1, h'_x, h'_x + 1\} \cap \mathcal{H}_l \neq \varnothing$. Then $|h_x - h'_x| \leqslant 1$. Moreover, adjacent misclassification (i.e., $|h_x - h'_x| = 1$) only occurs for points near the boundary of some slices.*

The next lemma records two elementary consequences of Assumption **($\boldsymbol{\omega_f}$)**.

Lemma 12 (consequences of coarse monotonicity) *Assume* **($\mathbf{Y}\in\psi_2$)**, **($\boldsymbol{\omega_f}$)**, *and* **(LCV)**, *and let $l \geqslant 1$ satisfy $|R_{l,h}| \geqslant \max(\sigma_\zeta, \omega_f)$ for all h. Then:*

(i) *inverse images of equal-length response intervals have comparable arclengths, and*

$$C_f \max\left(\sigma_\zeta, \omega_f, \frac{\operatorname{len}_\gamma}{C'_f l}\right) \lesssim C'_f |R_{l,h}|;$$

(ii) *after fixing the universal constant in $l_{\max} \asymp C_Y R_0 / \max(\sigma_\zeta, \omega_f)$ sufficiently large, the requirements*

$$l \gtrsim \frac{C_f \operatorname{len}_\gamma}{C'_f \sigma_\gamma}, \qquad l \leqslant l_{\max}$$

are compatible.

In Algorithm 1, we use the sample slice with index $\widehat{h}_x$, which has the smallest estimated distance to x, to estimate the correct h'_x: this gives the classification with $|\widehat{h}_x - h'_x| \leq 1$ w.h.p.:

Proposition 13 (classification accuracy) *Assume* **($\mathbf{X}\in\psi_2\cap\mathcal{C}^2$)**, **($\mathbf{Y}\in\psi_2$)**, **($\boldsymbol{\zeta}\in\psi_2$)**, **($\boldsymbol{\gamma_1}$)**, **(LCV)**, *and* **($\boldsymbol{\omega_f}$)** *hold, and fix $\tau \geqslant 1$. Let $\mathcal{H}_l := \{h : n^b_{l,h} \geqslant n_{loc}\}$ denote the subset of "heavy" slices, whose sample counts are bounded below. If $l \gtrsim \frac{C_f \operatorname{len}_\gamma}{C'_f \sigma_\gamma}$ and $|R_{l,h}| \geqslant \max(\sigma_\zeta, \omega_f)$ for all $h \in \mathcal{H}_l$ (equivalently, $l \leqslant l_{\max}$), then the probability of misclassification by at least two slices in Step 2.b of Algorithm 1 satisfies*

$$\mathbb{P}\left(\left|\widehat{h}_x - h_x\right| \geqslant 2\right) \lesssim ld \exp\left(-c\frac{n_{loc}}{\sqrt{\tau \log n}} \min\left(\frac{C'_f \sigma_\gamma^4}{C_f^2 R_0^3 d^{3/2} \operatorname{len}_\gamma}, \frac{\sigma_\gamma^8 C'^4_f}{R_0^8 C_f^4 d^4}\right)\right) + ln^{-\tau}.$$

The preceding bound shows that the probability of misclassification decays exponentially, with an exponent proportional to $n_{loc}/\sqrt{\tau \log n}$, as long as n and l satisfy

$$\frac{n_{loc}}{\log^{3/2} n} \gtrsim C_{\gamma,f} := \frac{R_0^3 C_f^2}{C'_f} \max\left(\frac{d^{3/2}\operatorname{len}_\gamma}{\sigma_\gamma^4}, \frac{R_0^5 C_f^2 d^4}{C'^3_f \sigma_\gamma^8}\right) \quad \text{and} \quad l \gtrsim \frac{C_f \operatorname{len}_\gamma}{C'_f \sigma_\gamma}. \tag{7}$$

For large n, the error contributed by misclassification by at least two slices is therefore negligible relative to the error under classification with $|\widehat{h}_x - h'_x| \leq 1$. Similar to Proposition 13, misclassification to adjacent slices is unavoidable for points near the boundary of some slices. To prevent this effect from undermining the performance of the estimator $\widehat{F}$, we will include data from adjacent slices in the regression of the link function f in each local neighborhood after the projection onto the local tangent to γ (as in Step 3.b of Algorithm 1).

3.3 Function estimation error corresponding to almost correct classification

We study the regression error on the event of almost correct classification. Condition on the first half of the sample. Then the estimated geometry, the map $x \mapsto \widehat{h}(x)$, the intervals $I^{(l,h)}$, and their partitions are fixed. For each $h \in \mathcal{H}_l$, set

$$A_{l,h} := \{x \in \Omega_\gamma : |\widehat{h}(x) - h| \leqslant 1\}, \qquad Z := \langle \widehat{v}^b_{l,h}, X\rangle.$$

An observation (X, Y) from the second half is used for index h only when $X \in A_{l,h}$. This event depends on X and on the first half, but not on Y. Since ζ is independent of X and of the first half and has mean zero,

$$\mathbb{E}\left[\zeta \mid Z,\ X \in A_{l,h},\ \{(X_i, Y_i)\}_{i=1}^{n/2}\right] = 0.$$

In the rest of this subsection, conditioning on the first half is omitted from the notation, but the conditioning on $X \in A_{l,h}$ is kept explicitly. The use of seven neighboring empirical response slices in Step 3.a matches the accepted-region geometry used below. On the good classification event, if $X \in A_{l,h}$, then $|h'_x - h| \leqslant 3$, and if a test point is assigned to h, then $|h'_x - h| \leqslant 2$. Thus, except on the same classification and first-half slice-coverage exceptional set controlled in the theorem proofs, every accepted regression point and every retained test point for index h has projected coordinate in $I^{(l,h)}$. The factors $\mathbb{1}_{I^{(l,h)}}(Z)$ below only remove that exceptional set, whose contribution is absorbed into the stated high-probability remainders.

Throughout this subsection, we say that classification is *almost correct* at x when $|\widehat{h}_x - h'_x| \leqslant 2$, i.e., when the estimated index lies within two slices of the correct index. Proposition 11 gives $|h_x - h'_x| \leqslant 1$, and Proposition 13 gives $|\widehat{h}_x - h_x| \leqslant 1$ with high probability; hence $|\widehat{h}_x - h'_x| \leqslant 2$ with high probability. This is the event used in Proposition 14. The phrase "off by at most one slice" refers separately to the two comparisons $|h_x - h'_x| \leqslant 1$ and $|\widehat{h}_x - h_x| \leqslant 1$. Lemma 22, proved in the analysis of Theorem 3, makes the interval-coverage claim quantitative.

With this projected coordinate, we keep the slice-wise centering constant $c_{l,h|v^b_{l,h}} = \mathbb{E}[\Pi_\gamma X - \langle v^b_{l,h}, X\rangle \mid X \in S_{l,h}]$, where $v^b_{l,h}$ is the significant vector of the slice $S_{l,h}$. Define the random variable

$$\eta_h := f(\Pi_\gamma X) - f(\langle \widehat{v}^b_{l,h}, X\rangle + c_{l,h|v^b_{l,h}}).$$

The constant $c_{l,h|v^b_{l,h}}$ is defined on the single slice $S_{l,h}$, whereas η_h is measured over the accepted region $A_{l,h}$, which, on the almost-correct event, meets at most seven consecutive true response slices (Lemma 21 and the containment $|h'_x - h| \leqslant 3$). Applying the one-slice local-straightness estimate to this seven-slice arc enlarges the slice-width factor $\max(\sigma_\zeta, \omega_f, \mathrm{len}_\gamma/(C'_f l))$ only by a universal constant, which is absorbed in the $\lesssim$ bounds below.

For a measurable function $g : \mathbb{R} \to \mathbb{R}$, define its local regression error by

$$E_{l,h|\widehat{v}^b_{l,h}}(g) := \mathbb{E}\left[|Y - g(Z)|^2\, \mathbb{1}\{X \in A_{l,h}\}\mathbb{1}_{I^{(l,h)}}(Z)\right].$$

We define the regression function with respect to the one-dimensional projection along $\widehat{v}^b_{l,h}$ by

$$f^*_{l,h|\widehat{v}^b_{l,h}} := \operatorname*{arg\,min}_{g:\mathbb{R}\to\mathbb{R}} E_{l,h|\widehat{v}^b_{l,h}}(g).$$

Equivalently, for almost every $z \in I^{(l,h)}$ with respect to the conditional law of Z given $X \in A_{l,h}$,

$$\begin{aligned} f^*_{l,h|\widehat{v}^b_{l,h}}(z) &= \mathbb{E}\left[Y \mid Z = z,\ X \in A_{l,h}\right] \\ &= \mathbb{E}\left[f(\Pi_\gamma X) \mid Z = z,\ X \in A_{l,h}\right]. \end{aligned}$$

Therefore, on $X \in A_{l,h}$,

$$f^*_{l,h|\widehat{v}^b_{l,h}}(Z) = f(Z + c_{l,h|v^b_{l,h}}) + \mathbb{E}\left[\eta_h \mid Z,\ X \in A_{l,h}\right].$$

Using the definitions of η_h and $f^*_{l,h|\widehat{v}^b_{l,h}}$, we decompose Y on the event $X \in A_{l,h}$ as follows:

$$Y = f(\Pi_\gamma X) + \zeta = f^*_{l,h|\widehat{v}^b_{l,h}}(Z) + \zeta',$$

where

$$\zeta' = \eta_h - \mathbb{E}[\eta_h | Z,\ X \in A_{l,h}] + \zeta,$$

with $\mathbb{E}[\zeta' | Z,\ X \in A_{l,h}] = 0$. Consequently, after conditioning on the first half, the local regression problem on the accepted region has regression function $f^*_{l,h|\widehat{v}^b_{l,h}}$ and mean-zero noise ζ'.

Fix $j \in \mathbb{N}$ (the number of subintervals) and $m \in \{0, 1\}$ (the polynomial degree: $m = 0$ for piecewise-constant regression and $m = 1$ for piecewise-linear regression). Let $\{I^{(l,h)}_{j,k}\}_{k=1}^{j}$ be the uniform partition of $I^{(l,h)}$. Define $f^*_{j|\widehat{v}^b_{l,h}}$ to be any piecewise-polynomial function on $I^{(l,h)}$ such that, for every k with positive accepted probability,

$$f^*_{j|\widehat{v}^b_{l,h}}\Big|_{I^{(l,h)}_{j,k}} \in \operatorname*{arg\,min}_{g \in \mathcal{P}_m(I^{(l,h)}_{j,k})} \mathbb{E}\Big[|Y - g(Z)|^2 \,\Big|\, X \in A_{l,h},\ Z \in I^{(l,h)}_{j,k}\Big].$$

On a bin of zero accepted probability, choose the restriction of $f^*_{j|\widehat{v}^b_{l,h}}$ arbitrarily, since its value does not affect the conditional L^2 risk. Here $\mathcal{P}_m(I)$ denotes the polynomials of degree at most m on I. The empirical counterpart of $f^*_{j|\widehat{v}^b_{l,h}}$ is the piecewise-polynomial estimator $\widehat{f}_{j|\widehat{v}^b_{l,h}}$, assembled in Step 3.c from the local least-squares fits computed in Step 3.b of Algorithm 1.

Because $|F(x)| \leqslant M$ and clipping onto $[-M, M]$ cannot increase the squared error, it suffices to analyze the untruncated local-polynomial value. For a test point x assigned to the index $h = \widehat{h}(x)$, set $Z_x = \langle \widehat{v}^b_{l,h}, x \rangle$. The estimator of Step 3.c evaluates $\widehat{f}_{j|\widehat{v}^b_{l,h}}$ at $\Pi_{I^{(l,h)}} Z_x$ rather than at Z_x; the two coincide for every interior index on the interval-coverage event of Lemma 22, so the decomposition below, and Proposition 14 with it, are unaffected by the clipping. The two extreme indices are excluded from the decomposition and contribute additively; they are bounded in Lemma 24. We use the decomposition

$$\begin{aligned} &F(x) - \widehat{f}_{j|\widehat{v}^b_{l,h}}(Z_x) \\ &= \underbrace{F(x) - f^*_{l,h|\widehat{v}^b_{l,h}}(Z_x)}_{\text{(NCA)}} + \underbrace{f^*_{l,h|\widehat{v}^b_{l,h}}(Z_x) - f^*_{j|\widehat{v}^b_{l,h}}(Z_x)}_{\text{(B)}} + \underbrace{f^*_{j|\widehat{v}^b_{l,h}}(Z_x) - \widehat{f}_{j|\widehat{v}^b_{l,h}}(Z_x)}_{\text{(V)}}. \end{aligned}$$

Proposition 14 (Mean squared error conditional on almost-correct classification) *Assume* $(\mathbf{X}\in \psi_2 \cap \mathcal{C}^2)$, $(\mathbf{Y}\in \psi_2)$, $(\boldsymbol{\zeta} \in \psi_2)$, $(\boldsymbol{\gamma_1})$, $(\mathbf{LCV})$, *and* $(\boldsymbol{\omega_f})$. *Suppose* $f \in \mathcal{C}^s$ *with* $s \in \left[\frac{1}{2}, 1\right]$. *Conditional on almost-correct classification in Step 2.b of Algorithm 1 and on the retained-bin count event of Lemma 23 and the interval-coverage event at interior indices of Lemma 22 in Steps 3.a–3.b, for* n *sufficiently large,*

$$
\begin{aligned}
&\mathrm{MSE}_{(NCA)} + \mathrm{MSE}_{(B)} + \mathrm{MSE}_{(V)} \\
&\lesssim [f]_{\mathcal{C}^s}^2 \left(\frac{\bar{\sigma}_\gamma C_f}{\mathrm{reach}_\gamma}\right)^{2s} \max\left(\sigma_\zeta, \omega_f, \frac{\mathrm{len}_\gamma}{C_f' l}\right)^{2s} \\
&\quad + [f]_{\mathcal{C}^s}^2 \left(C_X R_0^2 \mathrm{len}_\gamma \sigma_\gamma^{-2} d \log d\right)^{2s} \left(\frac{(\log n)^{1.5}}{\max(n, n_{loc} l^2)}\right)^{s} \\
&\quad + \left([f]_{\mathcal{C}^s} + C_Y R_0 l^{-1} [\rho_X]_{\mathcal{C}^s}\right)^2 C_f^{2s} \max\left(\sigma_\zeta, \omega_f, \frac{\mathrm{len}_\gamma}{C_f' l}\right)^{2s} j^{-2s} + \sigma_\zeta^2 \frac{lj \log j}{n}
\end{aligned}
$$

Putting together these bounds and optimizing yields our main theorems.

The proof follows five stages. First, response slices localize the latent coordinate. Second, population slice covariance identifies the local tangent direction. Third, concentration and eigenspace perturbation transfer this geometry to empirical slices. Fourth, stability of the local distance yields slice assignment with $|h - h'_x| \leq 1$. Finally, one-dimensional local regression gives the approximation, bias, and variance terms, while Lemmas 22–24 handle interval and bin boundary cases. Appendix A follows this order.

4. Numerical Experiments

We test a pooled implementation of Algorithm 1 on synthetic data, using 90% of each sample for training and 10% for testing. Unlike the theoretical estimator, the numerical implementation reuses the training observations for geometry and regression; it also uses Mahalanobis slice assignment and practical finite-sample bin-retention and integer-rounding conventions. These choices preserve the estimator's main structure and the l–j scaling principle studied in the theory. Full parameter traces, implementation details, and reproduction scripts are available in the accompanying repository.[2] All numerical experiments reported below use smooth link functions and degree-one local polynomial regression ($m = 1$). Since the links are smooth on their compact domains, they are Lipschitz; consistently with the theoretical range $s \in [\frac{1}{2}, 1]$, we apply the results throughout the numerical discussion at the endpoint $s = 1$. We vary n from 10^4 to 10^6 and use the theorem-prescribed values l^* and j^*, up to the practical conventions above. In practice, l and j may instead be selected by ten-fold cross-validation, which gave similar results in the experiments where it was used. We study the mean squared error $\mathbb{E}[|\widehat{F}_n(X) - F(X)|^2]$, the tangential error of the estimated slice center, and the difference between the significant vector and the tangential direction. Each experiment is repeated independently five times. Section 4.3 gives a stylized application to learning a reaction coordinate in a high-dimensional stochastic system.

2. `https://github.com/williamwuyantao/Nonlinear-Single-Variable-Model-scripts`

In each example in this section, we randomly generate $n_{\max} := 2 \times 10^6$ points from the underlying curve γ, for which we will have a complete parametrization. We use all these 2×10^6 data to compute an approximation to the center $\mu_{l,h} := \mathbb{E}[X \mid Y \in R_{l,h}]$ and the average tangential vector $\gamma'_{l,h} := \frac{\mathbb{E}[\gamma'(\Pi_\gamma X)|Y \in R_{l,h}]}{\|\mathbb{E}[\gamma'(\Pi_\gamma X)|Y \in R_{l,h}]\|}$ on each slice. Here, the unit vector $\gamma'_{l,h}$ is parallel to the average tangential direction $\mathbb{E}[\gamma'(\Pi_\gamma X) \mid X \in S_{l,h}]$ on the slice $S_{l,h}$.

To estimate accurately the nonlinear curve-approximation term $\mathrm{MSE}_{(NCA)}$, which accounts for the additive term that does not go to 0 as n increases in the bound in Theorem 3, we replace the estimated parameters $\{(\widehat{\mu}_{l,h}, \widehat{v}_{l,h})\}_{h=1}^{l}$ by the "oracle" parameters $\{(\mu_{l,h}, \gamma'_{l,h})\}_{h=1}^{l}$ (computed on the $n_{\max}$ points as described above) when performing the local linear projection and the local polynomial regression on each sample slice. For the oracle nonlinear-curve-approximation diagnostics described above, we retain the same values (l^*, j^*) as above and report the corresponding MSE at $n = n_{\max}$, denoted by "MSE at $n = 2 \times 10^6$" in Figs. 4, 5, 6, 7, 9, and 10. In this way, we reduce errors arising from the estimation of the parameters $(\widehat{\mu}_{l,h}, \widehat{v}_{l,h})_{h=1}^{l}$, so the reported "MSE at $n = 2 \times 10^6$" primarily reflects the nonlinear curve-approximation term $\mathrm{MSE}_{(NCA)}$.

Throughout this section, we use log–log plots to study how the following quantities decay with the number of samples n: the mean squared error $\mathbb{E}[|\widehat{F}_n(X) - F(X)|^2]$, the estimation error of the center along the tangential direction $\mathbb{E}[|\langle \hat{\mu}_{l^*,h} - \mu_{l^*,h}, \gamma'_{l^*,h} \rangle|]$, and the difference between the significant vector and the tangential direction $\mathbb{E}[||\widehat{v}_{l^*,h} - \gamma'_{l^*,h}||]$. On the interval $n \in [10^5, 10^6]$, we use linear least-squares regression to estimate the decay rates. This slope estimate becomes less reliable for large n because the additive term in the main theorem causes error saturation; consequently, the reported rates may be conservative. We add a dashed vertical line $n = 10^5$ for those figures to emphasize that the learning rate only corresponds to $n \in [10^5, 10^6]$. Because the MSE contains a curve-approximation term, the fitted slope reflects only the range $n \in [10^5, 10^6]$. Since the smooth links are analyzed at $s = 1$, the corresponding one-dimensional benchmark exponent for the MSE is $2/3$. The fitted slope may be smaller in magnitude once the curve-approximation term dominates and the overall error approaches its saturation level. This term also depends on curvature, which may itself vary with dimension.

4.1 Example 1: Circular Arcs and Verifying Theorem 3

We begin with circular arcs, the simplest nonlinear curves: they have constant curvature and admit a linear embedding in $\mathbb{R}^2$. We embed arcs of fixed length $\mathrm{len}_\gamma = 1$ in ambient dimension $d = 20$. Again, we fix $\sigma_\gamma = 0.5$ and the random vector Z_{d-1} follows the normal distribution $\mathcal{N}(0, \sigma_\gamma^2 \mathbb{I}_{d-1})$ with truncation at $\|Z_{d-1}\| < 0.9\,\mathrm{reach}_\gamma$. The curvature varies from 0.04 to 0.4. We set our upper bound for the curvature to be 0.4 because otherwise reach_γ becomes too small and the variance of Z_{d-1} will no longer be approximately σ_γ^2.

At the Lipschitz endpoint $s = 1$, Theorem 3 predicts that the curve-approximation term is proportional to $\|\gamma''\|^2$, which is smaller for curves with lower curvature. We examine this prediction numerically. We choose the smooth link $f(t) = \exp(t)$ on $[0, 1]$. Since it is Lipschitz on this compact interval, Theorem 3 applies with $s = 1$. The observational noise follows the normal distribution $\mathcal{N}(0, \sigma_\zeta^2)$ with $\sigma_\zeta = 0.03$. Figure 4 supports the following conclusions, consistent with Theorem 3 and our analysis: (i) When the observational noise is nontrivial, $\sigma_\zeta > 0$, the mean squared error has a nonzero asymptotic floor independent of n.

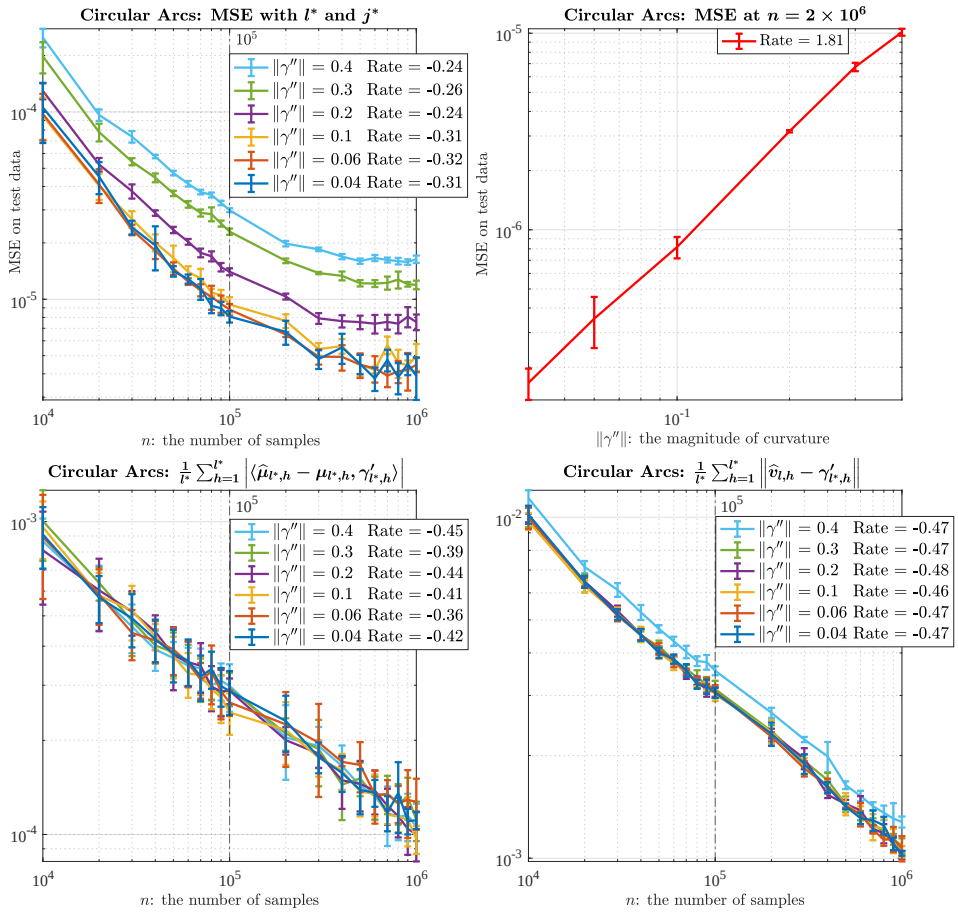

Figure 4: Numerical tests in Section 4.1: Circular arcs embedded in $\mathbb{R}^d$, $d = 20$, with unit length and curvature varying in $[0.04, 0.4]$. We fix the noise level $\sigma_\zeta = 0.03$. Upper row: MSE for $\widehat{F}$ (left) and MSE at $n = 2 \times 10^6$ as a function of the curvature (right); bottom row: estimation error for the center along tangential direction (left) and difference between estimated significant vector and the tangential direction (right) over 5 runs.

In the upper-left panel, the MSE decays more slowly as n increases and changes very little near $n \approx 10^6$, supporting the presence of a constant approximation error. The average decay rate over $n \in [10^5, 10^6]$ is estimated by least squares. (ii) At $s = 1$, the curve-approximation term is proportional to $\|\gamma''\|^2$. The upper-right plot shows that the MSE at $n = 2 \times 10^6$ is larger for curves with higher curvature and smaller for curves of lower curvature. A least-squares fit of the slope is consistent with the theoretical prediction. (iii) Corollary 10 states that the decay rate is $\mathcal{O}(n^{-\frac{1}{2}})$ for $\mathbb{E}[|\langle \hat{\mu}_{l^*,h} - \mu_{l^*,h}, \gamma'_{l^*,h} \rangle|]$ and $\mathbb{E}[\|\hat{v}_{l^*,h} - \gamma'_{l^*,h}\|]$.

4.2 Meyer helix

We perform numerical tests for a collection of curves called Meyer helices in $\mathbb{R}^d$, which we construct so that the curve complexity grows with the ambient dimension d. This is inspired by a curve called the Meyer staircase (named after Y. Meyer), defined by the

map $[0,1] \to L^p(\mathbb{R})$, for some $p \geq 1$, given by $t \mapsto \mathbf{1}_{[0,\frac{1}{2}]}(\cdot - t)$. We smooth out this original example by considering translations of a Gaussian, and we add oscillatory twists to increase the curve's complexity, thereby obtaining the Meyer helix. We measure the growing complexity of this family of curves as a function of d, and show that $\mathrm{len}_\gamma \asymp d^{3/2}, \mathrm{diam}_\gamma \asymp d^{1/2}, \|\gamma''\|_\infty \asymp d^{-1/2}, \mathrm{reach}_\gamma \asymp d^{1/2}$, and the effective linear dimensions defined below scale as d: see details in Appendix C. This collection of curves allows us to verify the performance of our algorithm when varying d: we fix $\sigma_\gamma = 0.5$ and let the random vector Z_{d-1} follow the normal distribution $\mathcal{N}(0, \sigma_\gamma^2 \mathbb{I}_{d-1})$ with truncation at $|Z_{d-1}| < 0.9\sqrt{d}$.

4.2.1 Example: Verifying Theorem 3 and Corollary 10

We let the underlying curve γ be the Meyer helix in dimension $d = 7$, with $\mathrm{len}_\gamma = 53.78$ and $\mathrm{reach}_\gamma = 2.65$. The link function $f(t) = \mathrm{len}_\gamma \exp(t/\mathrm{len}_\gamma)$ on $[0, \mathrm{len}_\gamma]$ is smooth and hence Lipschitz, so the corresponding theorem is applied with $s = 1$. The observational noise is distributed as $\zeta \sim \mathcal{N}(0, \sigma_\zeta^2)$, with σ_ζ varying from 0.05 to 0.2. Figure 5 supports the following theoretical conclusions: (i) When $\sigma_\zeta > 0$, the mean squared error has a nonzero error floor. (ii) At $s = 1$, the curve-approximation term is proportional to σ_ζ^2. (iii) Consistently with Corollary 10, the decay rate is $\mathcal{O}(n^{-\frac{1}{2}})$ for $\mathbb{E}[|\langle \hat{\mu}_{l^*,h} - \mu_{l^*,h}, \gamma'_{l^*,h} \rangle|]$ and $\mathbb{E}[||\widehat{v}_{l^*,h} - \gamma'_{l^*,h}||]$.

4.2.2 Example: Verifying Theorem 3 and Corollary 10

We consider a collection of Meyer helices with the following ambient dimensions and geometric parameters:

d	3	4	5	6	7	8	9
len_γ	20.86	33.39	35.35	43.89	53.78	90.20	96.65
reach_γ	1.73	2.00	2.24	2.45	2.65	2.83	3.00

We use the smooth link $f(t) = \mathrm{len}_\gamma \exp(t/\mathrm{len}_\gamma)$ on $[0, \mathrm{len}_\gamma]$. It is Lipschitz on this compact interval, so Theorem 3 is applied with $s = 1$. The observational noise ζ follows the normal distribution $\mathcal{N}(0, \sigma_\zeta^2)$ with $\sigma_\zeta = 0.2$.

In the left panel of Figure 6, we observe that, for fixed $n = 10^5$, the mean squared error is larger for Meyer helices with increasing ambient dimension d. Since the smooth link is analyzed as Lipschitz, Theorem 3 applies with $s = 1$ and its estimation term has rate $n^{-2/3} \log n$; the coefficient is larger for curves with larger len_γ. In the right panel, we observe that, in contrast, the mean squared error at $n = 2 \times 10^6$ is smaller for Meyer helices in larger ambient dimension d. This is consistent with Theorem 3 at $s = 1$, for which the curve-approximation term is proportional to $\mathrm{reach}_\gamma^{-2}$. For smaller d, the Meyer helix has smaller len_γ and reach_γ but larger curvature. Consequently: (i) the sample-size requirement in Theorem 3 is smaller; (ii) the first term in the mean squared error, which is $\mathcal{O}(n^{-2/3} \log n)$ at $s = 1$, has a smaller coefficient; (iii) the second term in the bound for the mean squared error in Theorem 3, which is the curve-approximation term, has larger magnitude; (iv) If n_1 denotes the sample size at which the two terms in the MSE upper bound balance, then n_1 increases with d. Consequently, over the interval $n \in [10^5, 10^6]$ used to estimate the

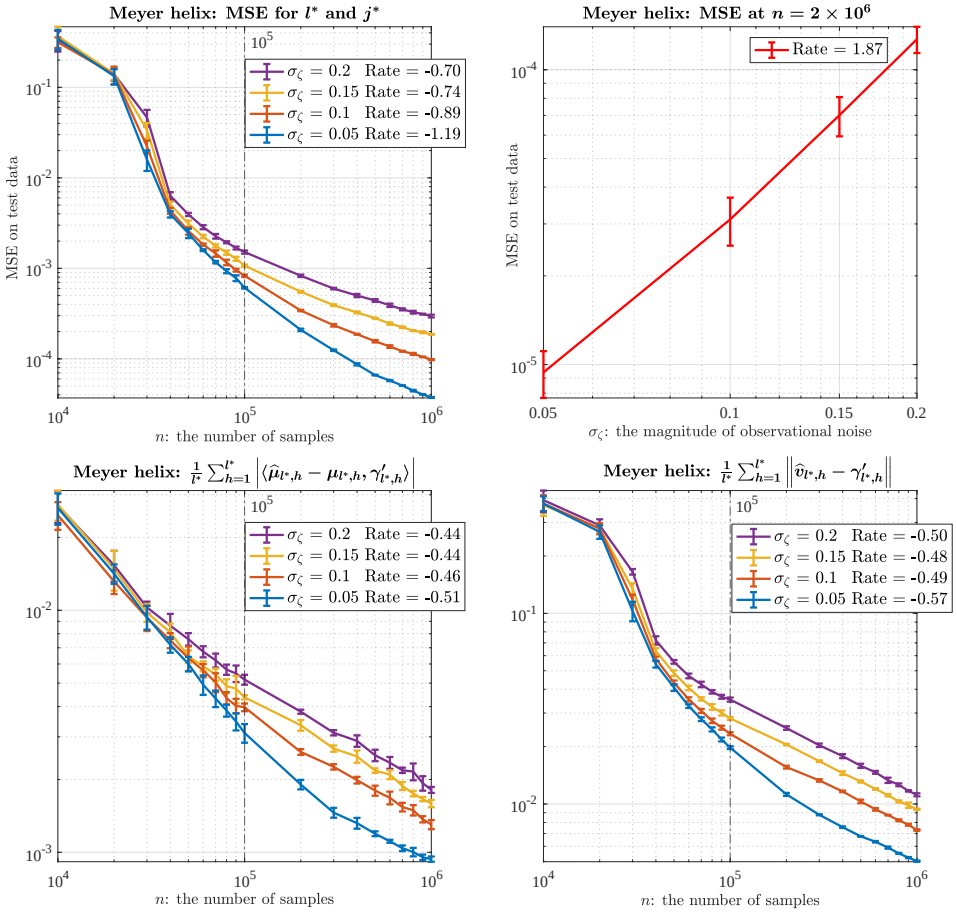

Figure 5: Numerical tests in Section 4.2.1 for a Meyer helix in dimension $d = 7$. The link is smooth and monotone; because it is Lipschitz on the compact parameter interval, we apply the theory with $s = 1$. The noise level σ_ζ varies over $[0.05, 0.2]$. Top row: MSE for $\widehat{F}$ (left) and MSE at $n = 2 \times 10^6$ as a function of σ_ζ (right). Bottom row: tangential error of the estimated slice center (left) and error of the estimated significant vector relative to the tangent direction (right), averaged over five independent runs.

decay rate, one must account for the value of n_1, which determines the dominant term in the mean squared error. In particular, the Meyer helix in $d = 3$ dimensions has $n_1 \approxeq 10^5$, with the MSE exhibiting good decay for $n \leq n_1$, and saturating for $n \geq n_1$. In contrast, the Meyer helix in dimension $d = 7$ has $n_1 \approxeq 10^6$, and thus we observe a good decay rate on the interval $n \in [10^5, 10^6]$. For Meyer helices in higher dimensions, the sample-size requirement in Theorem 3 is even larger. The numerical test also exhibits a transition for $d = 8, 9$: the error increases as n grows from 10^4 to 2×10^4 and decreases once $n \geqslant 4 \times 10^4$. First, this is not a contradiction with Theorem 3 because the minimum sample-size condition is not satisfied for the Meyer helix in dimension $d = 8, 9$ when $n \leqslant 3 \times 10^4$. Second, this increase of error for small n is due to the transition from the "wide" slice scenario to the "thin" slice scenario. Further investigation on the average empirical slice-shape score $\frac{1}{l^*}\sum_{h=1}^{l^*} \widehat{G}_{l^*,h}$

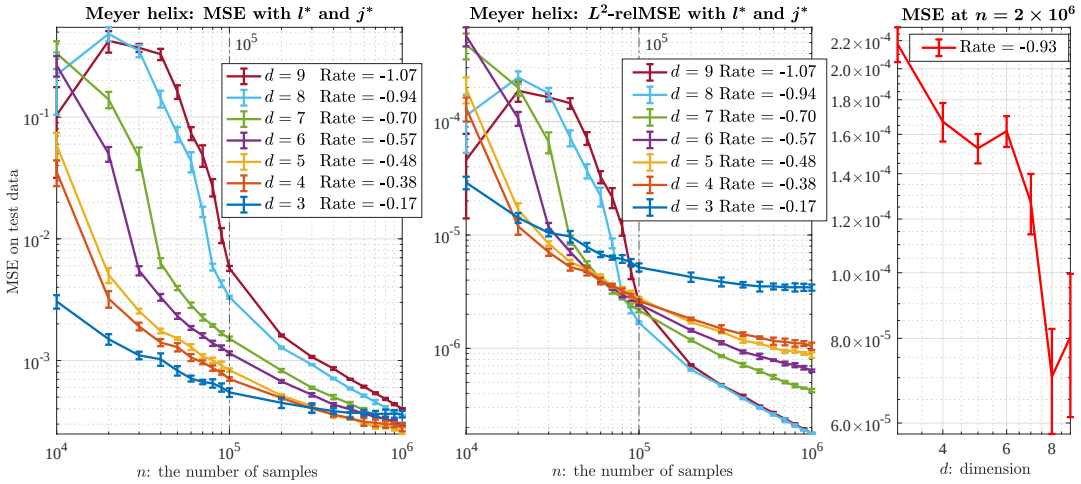

Figure 6: Numerical tests in Section 4.2.2 for Meyer helices in $\mathbb{R}^d$, $d \in \{3, \dots, 9\}$. The link is smooth and monotone and is analyzed at the Lipschitz endpoint $s = 1$; the noise level is $\sigma_\zeta = 0.2$. Left: MSE for $\widehat{F}$. Middle: relative MSE, $\mathbb{E}[|\widehat{F} - F|^2]/\mathbb{E}[|F(X) - \mathbb{E}(F(X))|^2]$, averaged over five independent runs. Right: MSE of $\widehat{F}$ at $n = 2 \times 10^6$ as a function of d.

shows that the slices are in the "wide" regime when $n \leqslant 3 \times 10^3$ and in the "thin" regime when $n \geqslant 4 \times 10^4$. For $n \in [3 \times 10^4, 4 \times 10^4]$, the average empirical slice-shape score $\frac{1}{l^*} \sum_{h=1}^{l^*} \widehat{G}_{l^*,h} \approx 0$ and thus the slices are roughly isotropic, which, as discussed in Section 2.1, prevents an accurate estimate of the significant vector.

Figure 6 supports the following conclusions, in line with our theoretical analysis: (i) the constants in $\mathcal{O}(n^{-2/3})$ at $s = 1$ are bigger for curves with greater len_γ; (ii) the sample-size requirement in Theorem 3 is larger for more complex curves; (iii) the significant vector cannot be estimated well when the geometric shape of a slice is roughly isotropic; (iv) the balancing sample size n_1 is larger for curves with greater length and reach; (v) at $s = 1$, the curve-approximation term is proportional to reach_γ^{-2};

4.2.3 Example: Verifying Theorem 5

We consider a collection of Meyer helices in the following ambient dimensions and geometric parameters:

d	5	6	7	8	9	10	12	24	36	48
len_γ	35.35	43.89	53.78	90.20	96.65	93.88	135.76	435.48	730.62	1306.78
reach_γ	2.24	2.45	2.65	2.83	3.00	3.16	3.46	4.90	6.00	6.93

Note that for Meyer helices with $d = 5, \dots, 12$, the sample size is sufficient when $n \geq 10^4$, while for $d = 24, 36, 48$, the sample-size requirement is substantially larger. Figure 7 supports the following theoretical conclusions: (i) when $\sigma_\zeta = 0$, the mean squared error has no positive asymptotic floor because the curve-approximation term vanishes; (ii) the

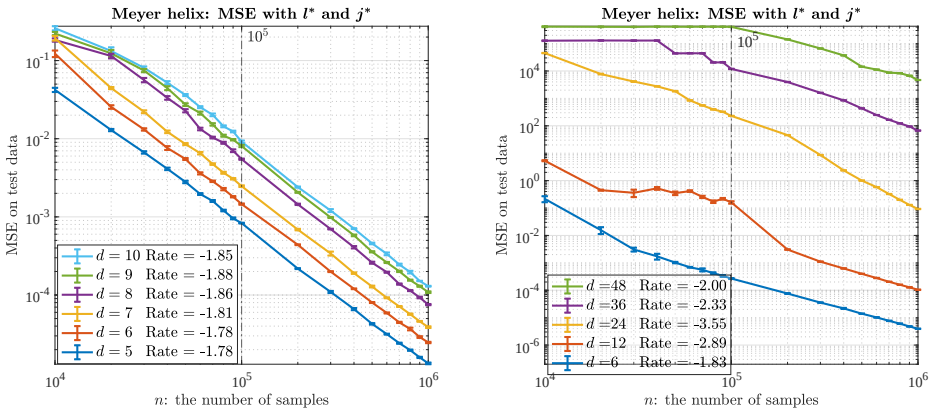

Figure 7: Numerical tests in Section 4.2.3 for Meyer helices of varying ambient dimension. Left: $d = 5, \ldots, 10$. Right: $d = 6, 12, 24, 36, 48$. In all tests, the observational noise is zero, $\zeta \equiv 0$, and the smooth link is analyzed at the Lipschitz endpoint $s = 1$. The reported errors are averages over five independent repetitions.

decay rate is $\mathcal{O}(n^{-2})$ if the link function f is monotone and Lipschitz; (iii) the sample-size requirement is larger for more complex curves.

4.3 Learning reaction paths and committor functions in overdamped Langevin dynamics

We consider a stylized application to the overdamped Langevin dynamics

$$dX_t = -\nabla U(X_t)\, dt + \sigma_{BM}\, dW_t, \tag{8}$$

where U is a potential and σ_{BM} is the noise amplitude. For two metastable sets A and B, let τ_A and τ_B be their first hitting times and define the committor

$$F_A(x) := \mathbb{P}_x(\tau_A < \tau_B).$$

In the interior outside $A \cup B$, F_A is harmonic with respect to the backward generator

$$\mathcal{L} = \frac{\sigma_{BM}^2}{2}\Delta - \nabla U \cdot \nabla, \qquad \mathcal{L}F_A = 0,$$

with boundary values $F_A|_A = 1$ and $F_A|_B = 0$. Computing this function directly is difficult in high dimension; see (Freidlin and Wentzell, 1998; E and Vanden-Eijnden, 2006; Metzner et al., 2006, 2009) and the references therein. One of the approaches to make this feasible is to use dimension reduction techniques that produce so-called reaction coordinates, that are such that the projected process is nearly Markovian, with transition rates and commitor functions that approximate the original ones (Coifman et al., 2008; Rohrdanz et al., 2011; Zheng et al., 2011).

Assume that the dominant transition geometry is described by a smooth reaction path γ and that the potential approximately separates into tangential and normal components,

$$U(x) = U_{tan}(\Pi_\gamma x) + U_{nor}(\mathrm{dist}(x, \gamma)). \tag{9}$$

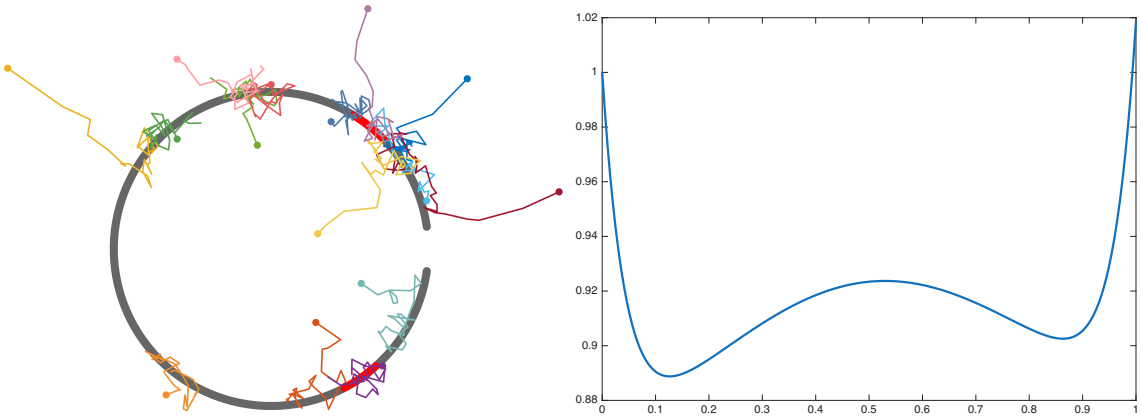

Figure 8: Left: short trajectories for (8) around a circular reaction path; the two metastable basins are highlighted in red. Right: the tangential double-well potential U_{tan}. Low-dimensional projections of the sampled configurations are shown in Figure 11 in Appendix B.

When the normal confinement is strong, trajectories rapidly relax toward γ and evolve primarily through the coordinate $s = \Pi_\gamma X_t$. In the limiting regime the committor therefore factors as

$$F_A(x) = f(\Pi_\gamma x), \qquad f(s) := F_A(\gamma(s)),$$

and for large but finite confinement this relation remains an approximation. This is precisely the NSVM structure: Π_γ is the nonlinear reaction coordinate and f describes progress along the transition path. As shown in Figure 11, the latent circular reaction path is not readily visible in either a random three-dimensional projection or the projection onto the first three principal components of the predictor cloud, illustrating the need to learn a nonlinear reaction coordinate.

Training responses can be obtained by Monte Carlo. From an initial condition x, simulate K independent trajectories and set

$$Y(x) := \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}\left\{ \tau_A^{(k)}(x) < \tau_B^{(k)}(x) \right\}. \tag{10}$$

Then $\mathbb{E}[Y(x) \mid x] = F_A(x)$ and $\mathrm{Var}(Y(x) \mid x) = K^{-1} F_A(x)(1 - F_A(x))$, so the Monte Carlo labels fit the regression setting considered in the paper.

For the numerical example, we take $d = 40$, a unit-length circular reaction path, a double-well tangential potential, and harmonic normal confinement. We generate $n = 4 \times 10^4$ configurations and estimate each response from $K = 4 \times 10^4$ independent trajectories. The full potential coefficients, basin tolerances, sampling variances, and time-discretization parameters are recorded in Appendix B and in the public repository.

We apply Algorithm 1 for several choices of l and j. Because the true committor is unavailable, performance is measured by ten-fold cross-validation against the Monte Carlo labels. Several nearby choices of (l, j) attain a cross-validation MSE of 1.7×10^{-3}, indicating that the practical performance is stable across a range of resolutions.

MSE l \ j	1	2	4	8	16	32
10	0.0022	0.0021	0.0021	0.0021	0.0021	0.0021
12	0.0020	0.0019	0.0020	0.0019	0.0019	0.0019
14	0.0018	0.0018	0.0018	0.0018	0.0018	0.0018
16	0.0018	0.0018	0.0018	0.0017	0.0017	0.0018
18	0.0018	0.0017	0.0017	0.0017	0.0017	0.0017
20	0.0018	0.0017	0.0017	0.0017	0.0017	0.0017
24	0.0021	0.0021	0.0021	0.0021	0.0021	0.0021

Table 2: Ten-fold cross-validation MSE $\mathbb{E}[|\widehat{F}(X) - Y|^2]$ for the committor experiment.

5. Robustness: Performance of Algorithm 1 in a General Setting

Recall that Assumption **(LCV)** gives a quantitative requirement for the "thin" slice scenario. If we relax this assumption and consider instead the "wide" slice scenario, we expect that the largest principal component on each slice gives a proper approximation of the tangential direction under some assumption. Small curvature will do:

(SC) Let $\sigma_\gamma := \min_{t_0 \in [0, \mathrm{len}_\gamma]} \sigma_\gamma(t_0)$ and $\bar{\sigma}_\gamma := \max_{t_0 \in [0, \mathrm{len}_\gamma]} \sigma_\gamma(t_0)$. We assume that there exist $c_1, c_2 > 0$ such that $\bar{\sigma}_\gamma < c_1 C_f \max(\sigma_\zeta, \omega_f)$ and $C_f \max(\sigma_\zeta, \omega_f) \leqslant c_2 \mathrm{reach}_\gamma$. The maximum is the relevant quantity here because the normal covariance must be small on every retained wide slice.

In the thin-slice regime, the second condition follows automatically from (6), with a stronger dimension-dependent bound. The substantive distinction between the two regimes is therefore whether the normal variation $\bar{\sigma}_\gamma$ is large or small relative to $C_f \max(\sigma_\zeta, \omega_f)$.

Proposition 15 ((SC) implies "wide" slices) *Suppose* $(\mathbf{X} \in \psi_2 \cap \mathcal{C}^2)$, $(\mathbf{Y} \in \psi_2)$, $(\boldsymbol{\zeta} \in \psi_2)$, $(\boldsymbol{\gamma_1})$, *and* $(\boldsymbol{\omega_f})$ *hold, together with* **(SC)**, *with* c_1, c_2 *smaller than a small-enough universal constant. Then, for every* l *such that* $|R_{l,h}| \asymp \max(\sigma_\zeta, \omega_f), h \in \mathcal{H}_l$, *we have*

$$\lambda_1(\Sigma^b_{l,h}) - \lambda_2(\Sigma^b_{l,h}) \gtrsim C_f^2 |R_{l,h}|^2 \ .$$

Recall that we have the following distance function in this situation of "wide" slices, i.e., $G_{l,h} < 0$, $\mathrm{dist}(x,h) = \left\| x - \mu^b_{l,h} \right\|^2 + \frac{\lambda_d(\Sigma^b_{l,h})}{\lambda_1(\Sigma^b_{l,h})} |\langle x - \mu^b_{l,h}, v^b_{l,h} \rangle|^2$, and similarly for its empirical counterpart. With a proof as in Proposition 13, we conclude that

Proposition 16 (classification accuracy without (LCV)) *Assume that* $(\mathbf{X} \in \psi_2 \cap \mathcal{C}^2)$, $(\mathbf{Y} \in \psi_2)$, $(\boldsymbol{\zeta} \in \psi_2)$, $(\boldsymbol{\gamma_1})$, $(\boldsymbol{\omega_f})$, *and* **(SC)** *hold, and fix* $\tau \geqslant 1$. *Fix* $l = \frac{C_Y R_0}{\max(\sigma_\zeta, \omega_f)}$. *Then, the probability of misclassification by at least two slices in Step 2.b of Algorithm 1 can be bounded by*

$$\mathbb{P}\left(\left| \widehat{h}_x - h_x \right| \geqslant 2 \right) \lesssim ld \exp\left(-c \frac{C_f^6 \max(\sigma_\zeta, \omega_f)^7 n}{C_Y R_0^7 d^3 \sqrt{\log n}} \right) + l n^{-\tau} \ .$$

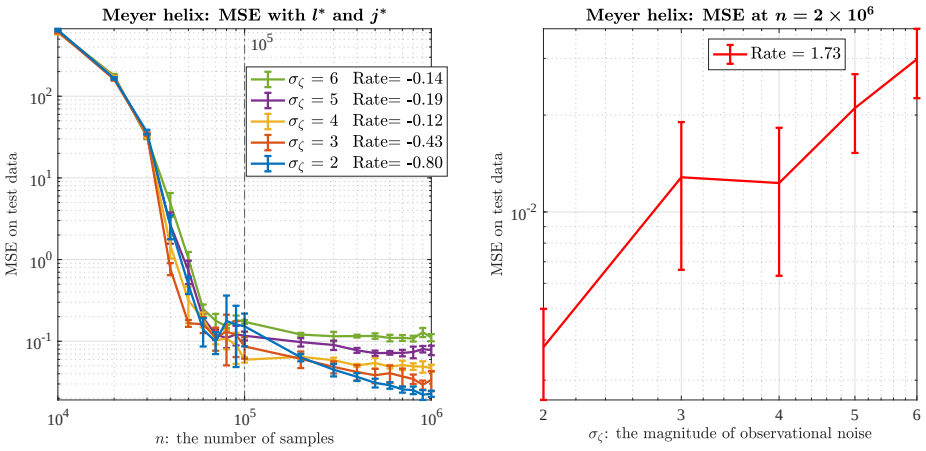

Figure 9: Numerical tests in Section 5.1: the Meyer helix in dimension $d = 21$ with $\sigma_\gamma = 0.5$. The link is smooth and hence Lipschitz, so the theorem is applied with $s = 1$; the noise level σ_ζ varies from 2 to 6. Left: mean squared error; right: mean squared error at $n = 2 \times 10^6$. The reported errors are averages over five independent repetitions.

5.1 Example: Verifying Theorem 6

Recall that in the "wide" slice scenario, the nonvanishing error is of the same order as the curve-approximation term. We verify this behavior in the following numerical test. Let the underlying curve γ be the Meyer helix in dimension $d = 21$, with $\mathrm{len}_\gamma = 370.63$ and $\mathrm{reach}_\gamma = 4.58$. The link function $f(t) = \mathrm{len}_\gamma \exp(t/\mathrm{len}_\gamma)$ on $[0, \mathrm{len}_\gamma]$ is smooth and hence Lipschitz, so the corresponding theorem is applied with $s = 1$. The observational noise is distributed as $\zeta \sim \mathcal{N}(0, \sigma_\zeta^2)$, with σ_ζ varying from 2 to 6. Assumption **(LCV)** is not satisfied in this case. We also vary σ_γ, which controls the magnitude of the deviation of X from the hidden curve γ. Smaller σ_γ means that the distribution of X is more concentrated near γ.

Figures 9 and 10 show that when Assumption **(LCV)** is not satisfied, the mean squared error saturates at the level of the curve-approximation term.

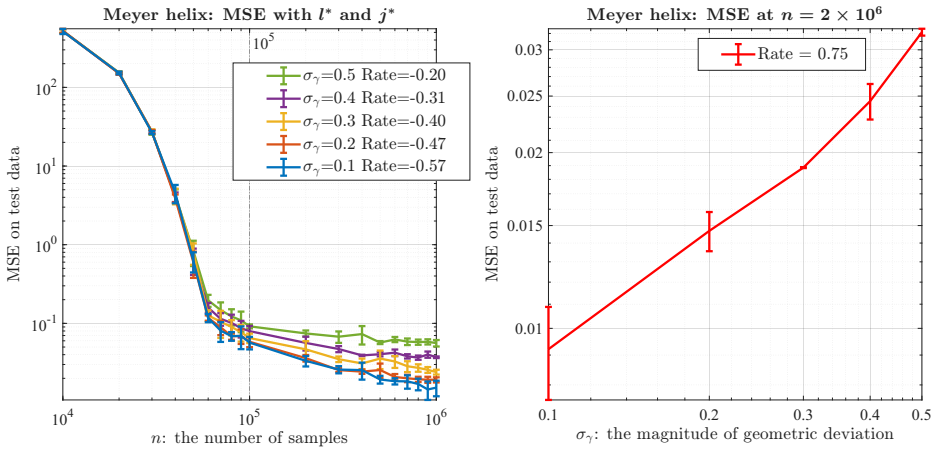

Figure 10: Numerical tests in Section 5.1: the Meyer helix in dimension $d = 21$ with $\sigma_\gamma = 0.1, 0.2, \ldots, 0.5$. The link is smooth and hence Lipschitz, so the theorem is applied with $s = 1$; the noise level is $\sigma_\zeta = 5$. Left: mean squared error. Right: mean squared error at $n = 2 \times 10^6$. The reported errors are averages over five independent repetitions.

6. Conclusion

We studied the Nonlinear Single-Variable Model, a compositional model $F = f \circ g$ for functions in high dimensions where both f and g are nonlinear, but g has a one-dimensional range and f is therefore a function of one variable. Using the geometric structure inherent in g, and under the coarse-monotonicity and lower conditional variance conditions stated above, inverse regression allows us to estimate local level-set geometry, estimate the nonlinear coordinate, and then learn f nonparametrically, with learning rates and sample requirements not cursed by the ambient dimension and with computationally efficient algorithms whose constants have only low-order polynomial dependence on the ambient dimension. The main novelty is the slice-geometry-based significant-vector construction, thin/wide-slice adaptation, data-adaptive local distance, and local polynomial regression step, which together convert the nonlinear curve-projection structure into the MSE guarantees proved here. The formal risk guarantees are stated for the Hölder–Lipschitz range $s \in [\frac{1}{2}, 1]$, in which the finite-slicing curve-approximation term is compatible with the one-dimensional nonparametric rate. The present estimator uses a first-order local approximation of the curve and its closest-point coordinate; exploiting regularity beyond the Lipschitz level is a natural direction for higher-order geometric estimators.

Extending the method to functions f that are not roughly monotone presents challenges for inverse regression. The pre-image of a set of response values can consist of several disconnected slices, which are difficult to cluster in high dimensions. Related obstructions appear, for example, in work on stochastic-gradient methods for single- and multi-index models (Arous et al., 2021; Bietti et al., 2025). The sharp statistical and computational tradeoffs in these settings remain incompletely understood. Another extension of interest is a Nonlinear Multi-Variable Model in which the curve γ is replaced by a higher-dimensional manifold $\mathcal{M}$. This setting presents additional challenges because the geometry of the level

sets of F becomes substantially more complicated. Both extensions are under investigation and are left to future work.

An important open direction is to understand how compositional structure affects estimator design and the behavior of general-purpose construction algorithms, such as SGD applied to a suitable loss. This question is especially compelling for compositions with more than two levels and for identifying when such structure can avoid the curse of dimensionality.

Acknowledgments

The authors are grateful to A. Lanteri, S. Vigogna and T. Klock (during their stays at Johns Hopkins) for early discussions and experiments on an initial version of this model, and to Fei Lu and Xiong Wang for their feedback on an early version of this manuscript. We thank the reviewers of the initial version of this manuscript for comments and feedback that improved the presentation of the work. This research was partially supported by AFOSR FA9550-21-1-0317 and FA9550-23-1-0445. Prisma Analytics Inc. provided and managed the computing resources for the numerical experiments.

Appendix A. Proof of Propositions

Lemma 17
Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be independent copies of a sub-Gaussian pair $(X, Y) \in \mathbb{R}^{d+1}$ with variance proxy R_0^2. Fix $a_X, a_Y > 0$ large enough that

$$p := \delta_X + \delta_Y - 1 > 0, \qquad \delta_X := 1 - 2\exp(-a_X^2/2), \qquad \delta_Y := 1 - 2\exp(-a_Y^2/2),$$

and write $E := B(0, a_X\sqrt{d}\, R_0) \times [-a_Y R_0, a_Y R_0]$. Then, for every $\gamma \in (0,1)$,

$$\mathbb{P}\big\{\#\{i : (X_i, Y_i) \in E\} < \gamma p\, n\big\} \leqslant 2\exp\Big(-p\frac{(1-\gamma)^2/2}{1 + (1-\gamma)/3}\, n\Big).$$

Proof By the sub-Gaussian norm bound of Lanteri et al. (2022, Lemma B.2), $\mathbb{P}(\|X\| > a_X\sqrt{d}\, R_0) \leq 2e^{-a_X^2/2}$ and $\mathbb{P}(|Y| > a_Y R_0) \leq 2e^{-a_Y^2/2}$, so a union bound yields $\mathbb{P}((X,Y) \in E) \geq p$. The claimed concentration is then the Bernstein inequality for the i.i.d. indicators $\mathbb{1}_E(X_i, Y_i)$, as in Lanteri et al. (2022, Lemma B.1) (cf. also their Lemma B.3, which is the same argument for the X-marginal alone). ■

Lemma 18 (Matrix Bernstein Inequality) *We recall the following matrix Bernstein bound; see Vershynin (2018, Thm. 5.4.1). Let $M_1, \dots, M_n$ be i.i.d. random matrices in $\mathbb{R}^{d_1 \times d_2}$, with $\mathbb{E}M_i = 0$ and $\|M_i\| \leqslant B$. Write $\widehat{M} = \frac{1}{n}\sum_{i \leqslant n} M_i$ for the sample mean and $\sigma^2 = \max\left(\|\mathbb{E}[M_i M_i^\intercal]\|, \|\mathbb{E}[M_i^\intercal M_i]\|\right)$ for the covariance proxy. Then*

$$\mathbb{P}\left(\left\|\widehat{M}\right\| \gtrsim t\right) \lesssim (d_1 + d_2)\exp\left(-\frac{cnt^2}{Bt + \sigma^2}\right).$$

Lemma 19 (Concentration Inequalities for means, covariances and eigenvalues) *Suppose that X_i are iid bounded by $\|X_i\| \leqslant R_0\sqrt{d}$, let $\mu_{l,h}^b = \mathbb{E}[X]$ and $\widehat{\mu}_{l,h}^b = \frac{1}{n_{l,h}^b}\sum_{i\leqslant n_{l,h}^b} X_i$ denote the mean and sample mean. Let $\Sigma_{l,h}^b = \mathbb{E}[(X-\mu_{l,h}^b)(X-\mu_{l,h}^b)^\intercal]$ denote the covariance matrix, $\widehat{\Sigma}_{l,h}^b = \frac{1}{n_{l,h}^b}\sum_{i\leqslant n_{l,h}^b}(X_i-\widehat{\mu}_{l,h}^b)(X_i-\widehat{\mu}_{l,h}^b)^\intercal$ the sample covariance matrix, and $\widetilde{\Sigma}_{l,h}^b = \frac{1}{n_{l,h}^b}\sum_{i\leqslant n_{l,h}^b}(X_i-\mu_{l,h}^b)(X_i-\mu_{l,h}^b)^\intercal$ be the augmented covariance matrix. Then, for every $\alpha > 0$, every $0 < \delta < \alpha d$, and every $0 < \beta < \alpha d$:*

$$\mathbb{P}\left(\left\|\widehat{\mu}_{l,h}^b - \mu_{l,h}^b\right\| \gtrsim R_0\sqrt{\frac{\delta}{\alpha}}\right) \lesssim d\exp\left(-\frac{cn_{l,h}^b\delta}{\alpha d}\right), \quad \text{where } \delta < \alpha d \ ;$$

$$\mathbb{P}\left(\left\|\widetilde{\Sigma}_{l,h}^b - \Sigma_{l,h}^b\right\| \gtrsim \beta\frac{R_0^2}{\alpha}\right) \lesssim d\exp\left(-\frac{c\beta^2 n_{l,h}^b}{\alpha^2 d^2}\right), \quad \text{where } \beta < \alpha d \ ;$$

$$\mathbb{P}\left(\left|\lambda_1(\widehat{\Sigma}_{l,h}^b) - \lambda_1(\Sigma_{l,h}^b)\right| \gtrsim \beta\frac{R_0^2}{\alpha}\right) \lesssim d\exp\left(-\frac{c\beta^2 n_{l,h}^b}{\alpha^2 d^2}\right), \quad \text{where } \beta < \alpha d \ .$$

Proof The first two inequalities can be shown directly by Lemma 18. The exact centering identity is

$$\widehat{\Sigma}_{l,h}^b = \widetilde{\Sigma}_{l,h}^b - (\widehat{\mu}_{l,h}^b - \mu_{l,h}^b)(\widehat{\mu}_{l,h}^b - \mu_{l,h}^b)^\intercal,$$

and hence

$$\left\|\widehat{\Sigma}_{l,h}^b - \Sigma_{l,h}^b\right\| \leqslant \left\|\widetilde{\Sigma}_{l,h}^b - \Sigma_{l,h}^b\right\| + \left\|\widehat{\mu}_{l,h}^b - \mu_{l,h}^b\right\|^2.$$

Weyl's inequality therefore gives

$$\left|\lambda_1(\widehat{\Sigma}_{l,h}^b) - \lambda_1(\Sigma_{l,h}^b)\right| \leqslant \left\|\widehat{\Sigma}_{l,h}^b - \Sigma_{l,h}^b\right\| \leqslant \left\|\widetilde{\Sigma}_{l,h}^b - \Sigma_{l,h}^b\right\| + \left\|\widehat{\mu}_{l,h}^b - \mu_{l,h}^b\right\|^2,$$

and the third concentration estimate in Lemma 19 follows from the first two after adjusting the universal constants. ∎

A.1 Proof of Proposition 7

Proof Define intervals $I_k := (-\sqrt{2(k+1)}\sigma_\zeta, \sqrt{2k}\sigma_\zeta] \cup [\sqrt{2k}\sigma_\zeta, \sqrt{2(k+1)}\sigma_\zeta)$ for $k = 0, 1, 2, \ldots$. We first note that, thanks to $(\boldsymbol{\zeta} \in \psi_2)$, we have $\zeta_i \in \bigcup_{k\leqslant\tau\log n} I_k$ with probability higher than $1 - 2n^{-\tau}$, for every $i = 1, \ldots, n$. Conditioned on this event and on $Y_i \in T$, $\Pi_\gamma X_i \in f^{-1}(T + \bigcup_{k\leqslant\tau\log n} I_k)$. Meanwhile, $\mathbb{E}\left[\Pi_\gamma X | Y \in T, \mathcal{B}, \zeta \in I_k\right] \in [\min f^{-1}(T + I_k), \max f^{-1}(T + I_k)]$. Since $f^{-1}(S) = f^{-1}(S \cap \mathcal{R}_f)$ for every interval S, Assumption $(\boldsymbol{\omega_f})$ gives the same upper inverse-image bound even when a sub-Gaussian annulus I_k extends outside the link range; hence, upon conditioning on $Y_i \in T, \mathcal{B}, \zeta_i \in \bigcup_{k\leqslant\tau\log n} I_k$,

$$|\Pi_\gamma X_i - \mathbb{E}\left[\Pi_\gamma X | Y \in T, \mathcal{B}, \zeta \in I_k\right]| \leqslant C_f(|T| + \sqrt{\max(k, \tau\log n)}\sigma_\zeta).$$

After conditioning on $Y \in T$ and $\mathcal{B}$, the weights of each sub-Gaussian annulus I_k must also be conditional. In the applications of the theorem, the following comparison is used on the

corresponding inverse-image arc in Ω_0 (i.e., with $\rho_t > c_\rho$); Bayes' rule, the independence of ζ and X, the local density comparison there, and Assumption $(\boldsymbol{\omega_f})$ show that conditioning on $Y \in T$ changes each probability of the corresponding sub-Gaussian annulus I_k by at most a fixed multiplicative factor. Consequently,

$$\mathbb{P}(\zeta \in I_k \mid Y \in T, \mathcal{B}) \lesssim \mathbb{P}(\zeta \in I_k) \lesssim e^{-k}.$$

On $\{\zeta_i \in I_r\}$, $r \leqslant \tau \log n$, the law of total expectation must use these conditional probabilities of the sub-Gaussian annuli I_k, and therefore

$$\begin{aligned}
&|\Pi_\gamma X_i - \mathbb{E}[\Pi_\gamma X \mid Y \in T, \mathcal{B}]| \\
&\quad\leqslant \sum_{k=0}^{\infty} |\Pi_\gamma X_i - \mathbb{E}[\Pi_\gamma X \mid Y \in T, \mathcal{B}, \zeta \in I_k]|\, \mathbb{P}(\zeta \in I_k \mid Y \in T, \mathcal{B}) \\
&\quad\lesssim C_f \sum_{k=0}^{\infty} \big(|T| + (\sqrt{r+1} + \sqrt{k+1})\sigma_\zeta\big) e^{-k} \\
&\quad\lesssim C_f \big(|T| + \sqrt{\tau \log n}\, \sigma_\zeta\big).
\end{aligned}$$

Together with the sub-Gaussian probability of the complementary noise event, this proves part (a). For part (b), the corresponding conditional variance decomposition is

$$\begin{aligned}
\mathrm{Var}[\Pi_\gamma X \mid Y \in T, \mathcal{B}] = \sum_{k=0}^{\infty} \mathbb{E}\Big[(\Pi_\gamma X - \mathbb{E}[\Pi_\gamma X \mid Y \in T, \mathcal{B}])^2 \Big| Y \in T, \mathcal{B}, \zeta \in I_k\Big] \\
\times \mathbb{P}(\zeta \in I_k \mid Y \in T, \mathcal{B}).
\end{aligned}$$

Assumption $(\boldsymbol{\omega_f})$ bounds the k-th conditional second moment by $CC_f^2(|T|^2 + (k+1)\sigma_\zeta^2)$. Hence

$$\mathrm{Var}[\Pi_\gamma X \mid Y \in T, \mathcal{B}] \lesssim C_f^2 \sum_{k=0}^{\infty} (|T|^2 + (k+1)\sigma_\zeta^2) e^{-k} \lesssim C_f^2 (|T|^2 + \sigma_\zeta^2),$$

which proves part (b). ■

A.2 Proof of Proposition 8 and Corollary 9

Proof Recall that $X_i = \gamma(t_i) + W_i$ where $t_i = \Pi_\gamma X_i$ denotes position of X_i along curve and $W_i = M_{\gamma'(t_i)} \begin{pmatrix} W_i' \\ 0 \end{pmatrix}$ denotes the deviation from the curve. Here each $M_v \in \mathbb{O}(d)$ is a rotation matrix on $\mathbb{R}^d$ that maps the d-th canonical unit vector $\hat{e}_d = (0, \ldots, 0, 1)$ to the unit vector $v \in \mathbb{S}^{d-1}$. Observe that $\langle v_{l,h}^b, X_i \rangle = \langle v_{l,h}^b, \gamma(\Pi_\gamma X_i) \rangle + \langle v_{l,h}^b, W_i \rangle$, so we will show a high probability bound for each term.

We will utilize the contraction property of γ for the first term. Notice that $\gamma : [0, \mathrm{len}_\gamma] \to \mathbb{R}^d$ is a contraction map, i.e., it has Lipschitz constant 1: $\|\gamma(t_1) - \gamma(t_2)\| \leqslant |t_1 - t_2|$. Recall that in Proposition 7 part (a) we show that conditioned on the event that $\zeta_i \in \bigcup_{k \leqslant \tau \log n} I_k$ and $Y_i \in T$, we have $\Pi_\gamma X_i \in f^{-1}(T + \bigcup_{k \leqslant \tau \log n} I_k)$. As a consequence, the contraction

property of γ shows that conditioned on the same event, we also have $\gamma(\Pi_\gamma X_i) \in \gamma(f^{-1}(T + \bigcup_{k\leqslant \tau\log n} I_k))$ whose diameter is bounded by diameter of $f^{-1}(T + \bigcup_{k\leqslant\tau\log n} I_k)$. Following the proof of Proposition 7, we have

$$\mathbb{P}\left\{\|\gamma(\Pi_\gamma X_i) - \mathbb{E}[\gamma(\Pi_\gamma X_i) \mid Y \in T, \mathcal{B}]\| \gtrsim C_f(|T| + \sqrt{\tau\log n}\sigma_\zeta) \mid Y_i \in T, \mathcal{B}_i\right\} \leqslant 2n^{-\tau} \ ;$$

$$\|\mathrm{Cov}[\gamma(\Pi_\gamma X)|Y \in T, \mathcal{B}]\| \lesssim C_f^2(|T|^2 + \sigma_\zeta^2) \ ,$$

and as a consequence,

$$\mathbb{P}\left\{\left|\langle v_{l,h}^b, \gamma(\Pi_\gamma X_i)\rangle - \mathbb{E}[\langle v_{l,h}^b, \gamma(\Pi_\gamma X)\rangle \mid Y \in T, \mathcal{B}]\right| \gtrsim C_f(|T| + \sqrt{\tau\log n}\sigma_\zeta) \mid Y_i \in T, \mathcal{B}_i\right\} \leqslant 2n^{-\tau} \ ;$$

$$\mathrm{Var}[\langle v_{l,h}^b, \gamma(\Pi_\gamma X)\rangle|Y \in T, \mathcal{B}] \lesssim C_f^2(|T|^2 + \sigma_\zeta^2) \ .$$

Because ρ_X is compactly supported, choose C_X once and for all so that $\mathrm{supp}(\rho_X) \subseteq B(0, C_X\sqrt{d}\,R_0)$; the X-part of $\mathcal{B}$ is then automatic. Moreover, $T \subseteq R = [-C_Y R_0, C_Y R_0]$, so $Y \in T$ already implies the Y-part of $\mathcal{B}$. Conditional on $t = \Pi_\gamma X$, the event $Y \in T$ depends only on the independent observational noise ζ and therefore does not alter the conditional law of W. Hence $\mathcal{L}(W \mid t, Y \in T, \mathcal{B}) = \mathcal{L}(W \mid t)$.

The observations are independent across i; conditional on t_i, the law of W_i' is centered and isotropic, with covariance and support specified in Assumption $(\boldsymbol{\gamma_1})$. Conditionally on $t_i = \Pi_\gamma X_i$, Assumption $(\boldsymbol{\gamma_1})$ gives

$$\mathbb{E}[W_i \mid t_i] = 0, \qquad \mathbb{E}[W_i W_i^\intercal \mid t_i] = \sigma_\gamma(t_i)^2\big(I - \gamma'(t_i)\gamma'(t_i)^\intercal\big).$$

We therefore estimate the projection of W_i by first identifying the population smallest-eigenvalue direction, rather than by using a pointwise comparison between W_i and t_i. Evaluate the Rayleigh quotient of $\Sigma_{l,h}^b$ in the direction $\gamma'(\mathbb{E}[\Pi_\gamma X \mid Y \in T, \mathcal{B}])$. Since γ is parameterized by arclength, Taylor's formula and Proposition 7 imply

$$\begin{aligned}
&\mathrm{Var}\big[\langle\gamma'(\mathbb{E}[\Pi_\gamma X \mid Y \in T, \mathcal{B}]), \gamma(\Pi_\gamma X)\rangle \mid Y \in T, \mathcal{B}\big] \lesssim C_f^2(|T|^2 + \sigma_\zeta^2),\\
&\mathbb{E}\big[\langle\gamma'(\mathbb{E}[\Pi_\gamma X \mid Y \in T, \mathcal{B}]), W\rangle^2 \mid Y \in T, \mathcal{B}\big] \lesssim \|\gamma''\|_\infty^2\,\mathrm{reach}_\gamma^2 C_f^2(|T|^2 + \sigma_\zeta^2).
\end{aligned}$$

The mixed term vanishes conditionally on $\Pi_\gamma X$. Hence the Rayleigh principle yields

$$\lambda_d(\Sigma_{l,h}^b) \lesssim C_f^2(|T|^2 + \sigma_\zeta^2),$$

which proves part (b). For every unit vector orthogonal to the same local tangent, the conditional normal covariance in Assumption $(\boldsymbol{\gamma_1})$ contributes at least σ_γ^2, while the variation of the tangent over the conditional arc contributes at most $CC_f^2(|T|^2 + \sigma_\zeta^2)$. Thus

$$\lambda_{d-1}(\Sigma_{l,h}^b) \geqslant \sigma_\gamma^2 - CC_f^2(|T|^2 + \sigma_\zeta^2).$$

Together with the preceding upper bound, the scale restriction on l, and Assumption **(LCV)**, this proves Corollary 9. Choose the sign of $v_{l,h}^b$ so that $\langle v_{l,h}^b, \gamma'(\mathbb{E}[\Pi_\gamma X \mid Y \in T, \mathcal{B}])\rangle \geqslant 0$. The same tangent–normal block calculation, divided by this eigengap, gives

$$\left\|v_{l,h}^b - \gamma'(\mathbb{E}[\Pi_\gamma X \mid Y \in T, \mathcal{B}])\right\| \lesssim \|\gamma''\|_\infty C_f(|T| + \sigma_\zeta).$$

Since $\langle W_i, \gamma'(t_i)\rangle = 0$,

$$\begin{aligned}\left|\langle W_i, v_{l,h}^b\rangle\right| \leqslant\ & \|W_i\| \left\|v_{l,h}^b - \gamma'(\mathbb{E}[\Pi_\gamma X \mid Y \in T, \mathcal{B}])\right\| \\ & + \|W_i\| \left\|\gamma'(\mathbb{E}[\Pi_\gamma X \mid Y \in T, \mathcal{B}]) - \gamma'(t_i)\right\|.\end{aligned}$$

The support bound in Assumption $(\boldsymbol{\gamma_1})$, the Lipschitz continuity of γ', and Proposition 7 therefore bound the normal term at the conditional tail scale $C_f(|T| + \sqrt{\tau \log n}\,\sigma_\zeta)$. Combining this bound with the curve-component estimate obtained above proves part (a). ■

A.3 Proof of Proposition 20 and Corollary 10

Proposition 20 (local NVM) *Suppose* $(\mathbf{X}\in \psi_2 \cap \mathcal{C}^2)$, $(\mathbf{Y}\in \psi_2)$, $(\boldsymbol{\zeta} \in \psi_2)$, $(\boldsymbol{\gamma_1})$, **(LCV)**, *and* $(\boldsymbol{\omega_f})$ *hold. Let* $\mu_{l,h}^b$ *be the mean of the h-th slice and* $v_{l,h}^b$ *be the significant vector of the h-th slice. Then, for every l such that* $l \gtrsim C_f C_Y R_0/\sigma_\gamma$, $|R_{l,h}^b| \geqslant \max\{\sigma_\zeta, \omega_f\}$ *for all* $h \in \mathcal{H}_l$, *for every* $\epsilon \in (0,1)$ *and* $\tau \geqslant 1$, *we have:*

(a) *for any* $h \in \mathcal{H}_l$ *and any* $\epsilon > 0$, *the estimation error of the slice mean along the tangential direction can be bounded as*

$$\mathbb{P}\left\{\left|\langle v_{l,h}^b, \widehat{\mu}_{l,h}^b - \mu_{l,h}^b\rangle\right| > \frac{\sigma_\gamma^2 \epsilon}{R_0\sqrt{d}} \;\middle|\; n_{l,h}^b\right\} \lesssim d\exp\left(-\frac{c\sigma_\gamma^4 \epsilon^2 n_{l,h}^b (\tau\log n)^{-\frac{1}{2}}}{C_Y^2 C_f^2 R_0^4 d(l^{-2} + l^{-1}\epsilon)}\right) + n^{-\tau};$$

(b) *for any* $h \in \mathcal{H}_l$ *and any* $\epsilon > 0$, *for* $l \gtrsim C_f C_Y R_0/\sigma_\gamma$, *the estimation error of the significant vector can be bounded as*

$$\mathbb{P}\left\{\left\|\widehat{v}_{l,h}^b - v_{l,h}^b\right\| > \epsilon \mid n_{l,h}^b\right\} \lesssim d\exp\left(-\frac{c\sigma_\gamma^4 \epsilon^2 n_{l,h}^b (\tau\log n)^{-\frac{1}{2}}}{C_Y^2 C_f^2 R_0^4 d(l^{-2} + l^{-1}\epsilon)}\right) + d\exp\left(-\frac{c\sigma_\gamma^4 n_{l,h}^b}{R_0^4 d^2}\right) + n^{-\tau};$$

(c) *for any* $h \in \mathcal{H}_l$, $u > 0$, *for* $l \gtrsim C_f C_Y R_0/\sigma_\gamma$, *the estimation error of the width of the slice can be bounded as*

$$\begin{aligned}\mathbb{P}\Big(\left|\lambda_d(\widehat{\Sigma}_{l,h}^b) - \lambda_d(\Sigma_{l,h}^b)\right| > & C_f^2 C_Y^2 R_0^2 u^2/l^2 \;\Big|\; n_{l,h}^b\Big) \lesssim d\exp\left(-\frac{c\sigma_\gamma^4 C_Y^2 u^2 n_{l,h}^b (\tau\log n)^{-\frac{1}{2}}}{R_0^4 d^{3/2}(\sqrt{d} + uC_f C_Y)}\right) \\ & + \exp\left(-\frac{cu^4 n_{l,h}^b}{1+u^2}\right) + d\exp\left(-\frac{c\sigma_\gamma^4 n_{l,h}^b}{R_0^4 d^2}\right) + n^{-\tau}.\end{aligned}$$

Proof of part (a): By Proposition 8 part (a), we know that conditioned on $Y_i \in R_{l,h}^b$ and $\mathcal{B}_i$, with probability higher than $1-2n^{-\tau}$, we have $\left|\langle v_{l,h}^b, X_i\rangle - \mathbb{E}\left[\langle v_{l,h}^b, X\rangle \mid Y \in R_{l,h}^b, \mathcal{B}\right]\right| \lesssim C_f C_Y R_0 \sqrt{\tau\log n} l^{-1}$. Also Proposition 8 part (b) implies that $\mathrm{Var}\left[\langle v_{l,h}^b, X\rangle \mid Y \in R_{l,h}^b, \mathcal{B}\right] \lesssim C_f^2 C_Y^2 R_0^2 l^{-2}$. Therefore, we have the following Bernstein-type inequality for any $\epsilon > 0$:

$$\mathbb{P}\left(\left|\langle v_{l,h}^b, \widehat{\mu}_{l,h}^b - \mu_{l,h}^b\rangle\right| > \frac{\sigma_\gamma^2 \epsilon}{R_0\sqrt{d}} \;\middle|\; n_{l,h}^b\right) \lesssim d\exp\left(-\frac{c\sigma_\gamma^4 \epsilon^2 n_{l,h}^b (\tau\log n)^{-\frac{1}{2}}}{C_f^2 C_Y^2 R_0^4 d(l^{-2} + l^{-1}\epsilon)}\right) + n^{-\tau}\ .$$

Proof of part (b) The main tool is the following Davis-Kahan type inequality. Choose the sign of $\widehat{v}^b_{l,h}$ so that $\langle \widehat{v}^b_{l,h}, v^b_{l,h}\rangle \geqslant 0$. The residual Davis–Kahan bound (Bhatia, 1997, Theorem VII.3.1), together with (Stewart and Sun, 1990, Ch. 1, Sec. 5.3, Theorem 5.5), gives, on the event $\left\|\widehat{\Sigma}^b_{l,h} - \Sigma^b_{l,h}\right\| \leqslant \frac{1}{2}(\lambda_{d-1}(\Sigma^b_{l,h}) - \lambda_d(\Sigma^b_{l,h}))$,

$$\left\|\widehat{v}^b_{l,h} - v^b_{l,h}\right\| \leqslant \frac{2\left\|(\widehat{\Sigma}^b_{l,h} - \Sigma^b_{l,h})v^b_{l,h}\right\|}{\lambda_{d-1}(\Sigma^b_{l,h}) - \lambda_d(\Sigma^b_{l,h})}. \tag{11}$$

Step 1: We want the denominator on the right-hand side of (11) to be large. Corollary 9 supplies the population gap. Moreover,

$$\widehat{\Sigma}^b_{l,h} - \Sigma^b_{l,h} = \widetilde{\Sigma}^b_{l,h} - \Sigma^b_{l,h} - (\widehat{\mu}^b_{l,h} - \mu^b_{l,h})(\widehat{\mu}^b_{l,h} - \mu^b_{l,h})^\intercal,$$

so

$$\left\|\widehat{\Sigma}^b_{l,h} - \Sigma^b_{l,h}\right\| \leqslant \left\|\widetilde{\Sigma}^b_{l,h} - \Sigma^b_{l,h}\right\| + \left\|\widehat{\mu}^b_{l,h} - \mu^b_{l,h}\right\|^2.$$

Lemma 19, with $\alpha = R_0^2/\sigma_\gamma^2$ and a sufficiently small fixed β, gives

$$\mathbb{P}\left(\max\{\left\|\widetilde{\Sigma}^b_{l,h} - \Sigma^b_{l,h}\right\|, \left\|\widehat{\mu}^b_{l,h} - \mu^b_{l,h}\right\|^2\} > C\beta\sigma_\gamma^2\right) \lesssim d\exp\left(-\frac{c\beta^2\sigma_\gamma^4 n^b_{l,h}}{R_0^4 d^2}\right).$$

Outside this event, the Davis–Kahan condition holds and the denominator in (11) is a constant multiple of σ_γ^2.

Step 2: We apply Bernstein's inequality to bound the numerator $\left\|(\widehat{\Sigma}^b_{l,h} - \Sigma^b_{l,h})v^b_{l,h}\right\|$ in the right-hand side of (11). Consider the following decomposition:

$$\begin{aligned}\left\|(\widehat{\Sigma}^b_{l,h} - \Sigma^b_{l,h})v^b_{l,h}\right\| &\leqslant \left\|(\widetilde{\Sigma}^b_{l,h} - \Sigma^b_{l,h})v^b_{l,h}\right\| + \left|\langle v^b_{l,h}, \widehat{\mu}^b_{l,h} - \mu^b_{l,h}\rangle\right| \left\|\widehat{\mu}^b_{l,h} - \mu^b_{l,h}\right\| \\ &\lesssim \left\|(\widetilde{\Sigma}^b_{l,h} - \Sigma^b_{l,h})v^b_{l,h}\right\| + R_0\sqrt{d}\left|\langle v^b_{l,h}, \widehat{\mu}^b_{l,h} - \mu^b_{l,h}\rangle\right| \quad .\end{aligned}$$

Recall that part (a) already gives desired bound for $\left|\langle v^b_{l,h}, \widehat{\mu}^b_{l,h} - \mu^b_{l,h}\rangle\right|$, so we only need to bound $\left\|(\widetilde{\Sigma}^b_{l,h} - \Sigma^b_{l,h})v^b_{l,h}\right\|$. Each centered summand in $(\widetilde{\Sigma}^b_{l,h} - \Sigma^b_{l,h})v^b_{l,h}$ has norm at most

$$\left|{v^b_{l,h}}^\intercal(X_i - \mu^b_{l,h})\right| \left\|X_i - \mu^b_{l,h}\right\| + \left\|\Sigma^b_{l,h}v^b_{l,h}\right\|.$$

By Proposition 8 part (a), conditional on $Y_i \in R^b_{l,h}$ and $\mathcal{B}_i$, the first term is at most $C_f C_Y R_0^2\sqrt{d\tau\log n}\, l^{-1}$ with probability at least $1 - 2n^{-\tau}$. Proposition 8 part (b) gives $\left\|\Sigma^b_{l,h}v^b_{l,h}\right\| = \lambda_d(\Sigma^b_{l,h}) \lesssim C_f^2 C_Y^2 R_0^2 l^{-2}$, so the same scale provides an ℓ^∞ bound for the centered summands.

Next, we consider the ℓ^2 bound (i.e., variance). Considering the following decomposition:

$$\mathbb{E}\left[\left\|(\widetilde{\Sigma}^b_{l,h} - \Sigma^b_{l,h})v^b_{l,h}\right\|^2 \mid n^b_{l,h}\right] = \mathbb{E}\left[\left\|\widetilde{\Sigma}^b_{l,h}v^b_{l,h}\right\|^2 \mid n^b_{l,h}\right] - \left\|\Sigma^b_{l,h}v^b_{l,h}\right\|^2,$$

where

$$\mathbb{E}\left[\left\|\widetilde{\Sigma}^b_{l,h}v^b_{l,h}\right\|^2 \mid n^b_{l,h}\right] = \frac{1}{(n^b_{l,h})^2} {v^b_{l,h}}^{\intercal}\mathbb{E}\left[\left(\sum_i (X_i-\mu^b_{l,h})(X_i-\mu^b_{l,h})^{\intercal}\right)^2 \mid n^b_{l,h}\right] v^b_{l,h}$$
$$\leqslant \frac{R_0^2 d}{n^b_{l,h}}\mathrm{Var}\left[\langle X_i-\mu^b_{l,h}, v^b_{l,h}\rangle \mid n^b_{l,h}\right] + \left\|\Sigma^b_{l,h}v^b_{l,h}\right\|^2 .$$

The preceding inequality, together with Proposition 8 part (b), gives us the following ℓ^2 bound:

$$\mathbb{E}\left[\left\|(\widetilde{\Sigma}^b_{l,h}-\Sigma^b_{l,h})v^b_{l,h}\right\|^2 \mid n^b_{l,h}\right] \lesssim \frac{C_f^2 C_Y^2 R_0^4 d}{n^b_{l,h} l^2} .$$

We use the ℓ^∞ and ℓ^2 bounds above and apply Bernstein inequality 18 to obtain: for any $\epsilon > 0$,

$$\mathbb{P}\left(\left\|(\widetilde{\Sigma}^b_{l,h}-\Sigma^b_{l,h})v^b_{l,h}\right\| > \sigma_\gamma^2\epsilon \;\middle|\; n^b_{l,h}\right) \lesssim d\exp\left(-\frac{c\sigma_\gamma^4\epsilon^2 n^b_{l,h}(\tau\log n)^{-\frac{1}{2}}}{C_f^2C_Y^2R_0^4 d(\epsilon l^{-1}+l^{-2})}\right) + n^{-\tau} .$$

Combining part (a) and the estimates in Step 1,2 finishes the proof.

Proof of part (c) Let $V_i = \langle v^b_{l,h}, X_i-\mu^b_{l,h}\rangle^2$; then $\mathbb{E}[V_i \mid Y_i \in R^b_{l,h}, \mathcal{B}_i] = \lambda_d(\Sigma^b_{l,h}) \lesssim C_f^2C_Y^2R_0^2 l^{-2}$. Moreover, we can follow the same argument as in Proposition 7 and Proposition 8 to show that $\mathbb{E}[V_i^2 \mid Y_i \in R^b_{l,h}, \mathcal{B}_i] \lesssim C_f^4C_Y^4R_0^4 l^{-4}$. Denote $\widehat{V}^b_{l,h} = \frac{1}{n^b_{l,h}}\sum_i \langle v^b_{l,h}, X_i - \mu^b_{l,h}\rangle^2 \mathbb{1}\left\{Y_i \in R^b_{l,h}\right\} \cap \mathcal{B}_i$. Thus, we use Bernstein's inequality to show that

$$\mathbb{P}\left(|\widehat{V}^b_{l,h} - \lambda_d(\Sigma^b_{l,h})| > \beta\sigma_\gamma^2 \;\middle|\; n^b_{l,h}\right) \lesssim \exp\left(-\frac{c\beta^2\sigma_\gamma^2 n^b_{l,h}}{C_f^2C_Y^2R_0^4\sigma_\gamma^{-2}l^{-4}(\beta l^2 + C_f^2C_Y^2R_0^2\sigma_\gamma^{-2})}\right) .$$

We use $\mathfrak{S}_h(\epsilon,\beta)$ to denote the event that

$$\mathfrak{S}_h = \left\{ \begin{array}{l} \left\|\widehat{\mu}^b_{l,h}-\mu^b_{l,h}\right\| < \dfrac{R_0\sqrt{d}}{2}, \left|\langle\widehat{\mu}^b_{l,h}-\mu^b_{l,h}, v^b_{l,h}\rangle\right| \lesssim \dfrac{\epsilon\sigma_\gamma^2}{R_0\sqrt{d}}, \left\|v^b_{l,h}-\widehat{v}^b_{l,h}\right\| \lesssim \epsilon, \\ \left\|(\widetilde{\Sigma}^b_{l,h}-\Sigma^b_{l,h})v^b_{l,h}\right\| < \epsilon\sigma_\gamma^2, \left|\widehat{V}^b_{l,h}-\lambda_d(\Sigma^b_{l,h})\right| \lesssim \beta\sigma_\gamma^2 \end{array} \right\} .$$

We know from part (a), part (b), and Lemma 19 that, $\mathfrak{S}_h(\epsilon,\beta)$ satisfies, for any $\epsilon,\beta>0$,

$$\mathbb{P}\left(\mathfrak{S}_h(\epsilon,\beta)^c\right) \lesssim d\exp\left(-\frac{c\sigma_\gamma^4\epsilon^2 n^b_{l,h}(\tau\log n)^{-\frac{1}{2}}}{C_f^2C_Y^2R_0^4 d(l^{-2}+\epsilon l^{-1})}\right)$$
$$+ \exp\left(-\frac{c\sigma_\gamma^2\beta^2 n^b_{l,h}}{C_f^2C_Y^2R_0^2 l^{-4}(\beta l^2 + C_f^2C_Y^2R_0^2\sigma_\gamma^{-2})}\right) + d\exp\left(-\frac{c\sigma_\gamma^4 n^b_{l,h}}{R_0^4 d^2}\right) + n^{-\tau} .$$

On the event $\mathfrak{S}_h(\epsilon,\beta)$, we have the following estimate: Using $\widehat{\Sigma}^b_{l,h} = \widetilde{\Sigma}^b_{l,h} - (\widehat{\mu}^b_{l,h} - \mu^b_{l,h})(\widehat{\mu}^b_{l,h} - \mu^b_{l,h})^\intercal$ and the Rayleigh characterization of the smallest eigenvalue, we have

$$\begin{aligned}\left|\lambda_d(\widehat{\Sigma}^b_{l,h}) - \lambda_d(\Sigma^b_{l,h})\right| &\leqslant \left|\langle \widehat{v}^b_{l,h}, (\widetilde{\Sigma}^b_{l,h} - \Sigma^b_{l,h})\widehat{v}^b_{l,h}\rangle\right| + \left|\langle \widehat{v}^b_{l,h}, \widehat{\mu}^b_{l,h} - \mu^b_{l,h}\rangle\right|^2 \\ &\quad + \left|\langle \widehat{v}^b_{l,h}, \Sigma^b_{l,h}\widehat{v}^b_{l,h}\rangle - \langle v^b_{l,h}, \Sigma^b_{l,h} v^b_{l,h}\rangle\right| \\ &\lesssim R_0^2 d\epsilon^2 + \frac{\epsilon^2\sigma_\gamma^4}{R_0^2 d} + \beta\sigma_\gamma^2 + \epsilon^2\sigma_\gamma^2 + \lambda_d(\Sigma^b_{l,h})\epsilon^2.\end{aligned}$$

which is bounded by $\frac{u^2 C_f^2 C_Y^2 R_0^2}{l^2}$ if one takes $\epsilon' = c\frac{uC_fC_Y}{l\sqrt{d}}$ and $\beta' = \frac{u^2C_f^2C_Y^2R_0^2}{\sigma_\gamma^2 l^2}$, for $l > C_fC_YR_0/\sigma_\gamma$. This means that we have for any $u > 0$,

$$\begin{aligned}&\mathbb{P}\left(\left|\lambda_d(\widehat{\Sigma}^b_{l,h}) - \lambda_d(\Sigma^b_{l,h})\right| \gtrsim \frac{u^2C_f^2C_Y^2R_0^2}{l^2} \;\middle|\; n^b_{l,h}\right) \leqslant \mathbb{P}\left(\mathfrak{S}_h(\epsilon',\beta')^c\right) \\ \lesssim& d\exp\left(-\frac{cu^2C_Y^2\sigma_\gamma^4 n^b_{l,h}(\tau\log n)^{-\frac{1}{2}}}{R_0^4 d^{3/2}(\sqrt{d} + uC_fC_Y)}\right) + \exp\left(-\frac{cn^b_{l,h}u^4}{1+u^2}\right) + d\exp\left(-\frac{c\sigma_\gamma^4 n^b_{l,h}}{R_0^4 d^2}\right) + n^{-\tau}.\end{aligned}$$

A.3.1 Proof of Corollary 10

Proof Take $\epsilon = C_fC_YR_0^2\sigma_\gamma^{-2}\left[d(t+\log l+\log d)\sqrt{\tau\log n}/(n^b_{l,h}l^2)\right]^{1/2}$. The term $\log d$ absorbs the prefactor d in the exponential bound, while $\log l$ accounts for the union bound over the slices. The requirement $\epsilon < 1/l$ gives $n^b_{l,h}/\sqrt{\tau\log n} \gtrsim C_f^2C_Y^2R_0^4\sigma_\gamma^{-4}d(t+\log d+\log l)$. The moment estimate follows by integrating the conditional tail bound, equivalently by choosing e^{-t} at the scale $(C_fC_YR_0^2\sigma_\gamma^{-2})^{2p}(d\log d)^p[(\log n)^{1.5}/(n^b_{l,h}l^2)]^p$. ■

A.4 Proof of Lemma 12

Proof For the equal-length response intervals used in the localized interior analysis, Assumption $(\boldsymbol{\omega_f})$ gives

$$\frac{C'_f}{C_f} \leqslant \frac{|f^{-1}(T)|}{|f^{-1}(T')|} \leqslant \frac{C_f}{C'_f},$$

so the pull-backs of the uniform response partition have comparable arclengths. Since $|R_{l,h}| \asymp C_YR_0/l$, the condition $|R_{l,h}| \geqslant \max(\sigma_\zeta,\omega_f)$ controls the first two terms in the maximum in part (i). Applying Assumption $(\boldsymbol{\omega_f})$ to $\mathcal{R}_f$ and using $|\mathcal{R}_f| \asymp C_YR_0$ gives $\mathrm{len}_\gamma \lesssim C_fC_YR_0$, and the bounded ratio C_f/C'_f yields

$$\frac{C_f\mathrm{len}_\gamma}{C'_f l} \lesssim C'_f|R_{l,h}|,$$

which proves (i).

For (ii), the same bound on len_γ and Assumption **(LCV)** imply

$$\frac{C_f\mathrm{len}_\gamma}{C'_f\sigma_\gamma} \lesssim \frac{C_YR_0}{\max(\sigma_\zeta,\omega_f)}.$$

Thus the lower slicing-scale requirement is below $l_{\max}$ after the universal constant in the definition of $l_{\max}$ is chosen sufficiently large. If $\max(\sigma_\zeta, \omega_f) = 0$, then $l_{\max} = \infty$ and there is nothing to prove. ■

A.5 Proof of Proposition 11

Proof Fix $x_0 \in \Omega_\gamma$ with $y_0 := F(x_0) \in R$, set $t_0 := \Pi_\gamma x_0$, and let h'_x be the unique index for which $y_0 \in R_{l,h}$. For every retained slice, Bayes' formula gives

$$\rho_{t|Y\in R_{l,h}}(t) = \frac{\mathbb{P}(\zeta \in R_{l,h} - f(t))\rho_t(t)}{\mathbb{P}(Y \in R_{l,h})}, \qquad \mathbb{E}(\Pi_\gamma X \mid Y \in R_{l,h}) = \frac{\int t\,\mathbb{P}(\zeta \in R_{l,h} - f(t))\rho_t(t)\,dt}{\int \mathbb{P}(\zeta \in R_{l,h} - f(t))\rho_t(t)\,dt}.$$

On the retained inverse-image arc in Ω_0 used in the theorem proofs, the local density comparison and Assumption $(\boldsymbol{\omega_f})$ show directly from these two integrals that

$$\min_{h\in\{h'_x-1,h'_x,h'_x+1\}\cap\mathcal{H}_l} |t_0 - \mathbb{E}(\Pi_\gamma X \mid Y \in R_{l,h})| \lesssim C_f(|R_{l,h}| + \sigma_\zeta).$$

The hypothesis that a slice h with $|h - h'_x| \le 1$ is retained in $\mathcal{H}_l$ guarantees that the set over which the minimum is taken is nonempty. Indeed, the contribution from the relevant inverse-image interval has the same order as its length, while the contribution from its complement is bounded by the sub-Gaussian tail of ζ. If $|h - h'_x| \geqslant 2$, the response interval $R_{l,h}$ is separated from y_0 by at least one complete response bin. Applying the lower bound in Assumption $(\boldsymbol{\omega_f})$ in the preceding Bayes formula, and estimating the remaining noise-tail integral by Assumption $(\boldsymbol{\zeta} \in \psi_2)$, gives

$$|t_0 - \mathbb{E}(\Pi_\gamma X \mid Y \in R_{l,h})| \gtrsim C'_f \operatorname{dist}(y_0, R_{l,h}) - C_f\sigma_\zeta.$$

Since $|R_{l,h}| \geqslant \max(\sigma_\zeta, \omega_f)$, the conditional center for an index h with $|h - h'_x| \le 1$ is strictly closer to t_0 than every nonadjacent conditional center, with constants depending only on the fixed model constants. Finally, Taylor's formula for γ and the tangent alignment in the proof of Corollary 9 transfer the same ordering to the population distance. On consecutive local slices,

$$||\langle x_0 - \mu_{l,h}, v_{l,h}\rangle| - |t_0 - \mathbb{E}(\Pi_\gamma X \mid Y \in R_{l,h})|| \lesssim \left\|\gamma''\right\|_\infty C_f^2(|R_{l,h}| + \sigma_\zeta)^2.$$

For a nonlocal portion of the curve, the reach gives a lower bound for $\|x_0 - \mu_{l,h}\|$, and the nonnegative second term in $\operatorname{dist}(x_0, h)$ prevents that portion from being selected. Hence

$$\operatorname{dist}(x_0, h) > \min_{\substack{r\in\mathcal{H}_l \\ |r-h'_x|\leqslant 1}} \operatorname{dist}(x_0, r), \qquad h \in \mathcal{H}_l, \quad |h - h'_x| \geqslant 2,$$

which proves $|h_x - h'_x| \leqslant 1$. The two neighboring values can be comparable only when y_0 is near their common response boundary, proving the final assertion. ■

A.6 Proof of Proposition 13

Proof We first make the population separation explicit. By Proposition 11, consecutive response slices correspond to consecutive arclength pieces. Taylor's formula for γ and the local tangential direction give a constant $K_0 > 0$, depending only on the fixed curvature constants, such that, for $2 \leqslant |k| \leqslant K_0 l$ with $h_x + k \in \mathcal{H}_l$, summing the separation of consecutive slice centers along the index path from h_x to $h_x + k$ yields

$$\sqrt{\mathrm{dist}(x, h_x + k)} - \sqrt{\mathrm{dist}(x, h_x)} \gtrsim |k| \frac{\mathrm{len}_\gamma}{l}.$$

The restriction $|k| \geqslant 2$ is necessary because an adjacent slice may have comparable distance near a response boundary. For $|k| \geqslant K_0 l$ with $h_x + k \in \mathcal{H}_l$, the reach separates the two curve pieces in Euclidean distance. The local density lower bound on the retained inverse-image interval in Ω_0, together with Assumption $(\boldsymbol{\omega_f})$, gives the tangential spread $\lambda_d(\Sigma^b_{l,h}) \gtrsim (C'_f |R_{l,h}|)^2$, while Assumption $(\boldsymbol{\gamma_1})$ and Proposition 8 give $\lambda_1(\Sigma^b_{l,h}) \lesssim \sigma_\gamma^2$. Therefore

$$\sqrt{\mathrm{dist}(x, h_x + k)} - \sqrt{\mathrm{dist}(x, h_x)} \gtrsim \sqrt{\frac{\lambda_d(\Sigma^b_{l,h})}{\lambda_1(\Sigma^b_{l,h})}}\, \mathrm{reach}_\gamma \gtrsim \frac{\mathrm{len}_\gamma}{l}.$$

Combining the local and remote cases gives

$$\sqrt{\mathrm{dist}(x, h)} - \sqrt{\mathrm{dist}(x, h_x)} \gtrsim \frac{\mathrm{len}_\gamma}{l}, \qquad |h - h_x| \geqslant 2.$$

To obtain correct classification, we require the distance-estimation error to be small enough that for all $|h' - h_x| \geqslant 2$,

$$\left|\widehat{\mathrm{dist}}(x, h') - \mathrm{dist}(x, h')\right| + \left|\widehat{\mathrm{dist}}(x, h_x) - \mathrm{dist}(x, h_x)\right| < \left|\mathrm{dist}(x, h') - \mathrm{dist}(x, h_x)\right| \quad . \tag{12}$$

Indeed, this will imply that $\widehat{h}_x = \arg\min_{h' \in \mathcal{H}^b_l} \widehat{\mathrm{dist}}(x, h')$ satisfies $|h_x - \widehat{h}_x| \leq 1$.

Consider the event $\mathfrak{S}$ that we have a small estimation error for information in all slices:

$$\begin{aligned}
\mathfrak{S}(\epsilon, \delta, \beta, u): \quad & \text{for all } h \in \mathcal{H}^b_l, \\
& \left|\langle \widehat{\mu}^b_{l,h} - \mu^b_{l,h}, v^b_{l,h} \rangle\right| \lesssim \frac{\epsilon \sigma_\gamma^2}{R_0 \sqrt{d}}, \ \left\|\widehat{\mu}^b_{l,h} - \mu^b_{l,h}\right\| \lesssim \sigma_\gamma \sqrt{\delta}, \\
& \left\|v^b_{l,h} - \widehat{v}^b_{l,h}\right\| \lesssim \epsilon, \ \left|\lambda_1(\widehat{\Sigma}^b_{l,h}) - \lambda_1(\Sigma^b_{l,h})\right| \lesssim \beta \sigma_\gamma^2, \\
& \left|\lambda_d(\widehat{\Sigma}^b_{l,h}) - \lambda_d(\Sigma^b_{l,h})\right| \lesssim \frac{u^2 C_f^2 C_Y^2 R_0^2}{l^2}.
\end{aligned}$$

Orient $\widehat{v}^b_{l,h}$ so that $\langle \widehat{v}^b_{l,h}, v^b_{l,h} \rangle \geqslant 0$. Proposition 20 and Lemma 19, used with $\alpha = R_0^2/\sigma_\gamma^2$, show that, when $l > C_f C_Y R_0/\sigma_\gamma$, for any $\delta, \beta < R_0^2 d \sigma_\gamma^{-2}$ and any $\epsilon, u > 0$, the event $\mathfrak{S}$ has

the following high-probability bound:

$$\begin{aligned}\mathbb{P}\left(\mathfrak{S}^c\right) \lesssim & ld\exp\left(-\frac{c\sigma_\gamma^4\epsilon^2 n_{l,h}^b(\tau\log n)^{-\frac{1}{2}}}{C_f^2C_Y^2R_0^4d(l^{-2}+l^{-1}\epsilon)}\right)+ld\exp\left(-cn_{l,h}^b\min\left(\frac{\sigma_\gamma^2\delta}{R_0^2d},\frac{\sigma_\gamma^4\beta^2}{R_0^4d^2},\frac{\sigma_\gamma^4}{R_0^4d^2}\right)\right)\\ & +ld\exp\left(-\frac{c\sigma_\gamma^4C_Y^2u^2n_{l,h}^b(\tau\log n)^{-\frac{1}{2}}}{R_0^4d^{3/2}(\sqrt{d}+uC_fC_Y)}\right)+l\exp\left(-\frac{cn_{l,h}^bu^4}{1+u^2}\right)+ln^{-\tau}\ .\end{aligned} \tag{13}$$

We now investigate how small these parameters should be. Conditioned on event $\mathfrak{S}(\epsilon,\delta,\beta,u)$, we can expand the estimation error in the distance function and estimate its upper-bound by the following calculation:

$$\begin{aligned}&\left|\widehat{\mathrm{dist}}(x,h)-\mathrm{dist}(x,h)\right|\\ \lesssim&\left(R_0\sqrt{d}\epsilon+\sigma_\gamma\sqrt{\delta}\epsilon+\frac{\sigma_\gamma^2\epsilon}{R_0\sqrt{d}}\right)^2+\left(R_0\sqrt{d}\epsilon+\sigma_\gamma\sqrt{\delta}\epsilon+\frac{\sigma_\gamma^2\epsilon}{R_0\sqrt{d}}\right)\sqrt{\mathrm{dist}(x,h)}\\ &+\frac{\lambda_d(\Sigma_{l,h}^b)}{\lambda_1(\Sigma_{l,h}^b)}\left(\sigma_\gamma^2\delta+R_0\sigma_\gamma\sqrt{d}\sqrt{\delta}\right)+\frac{u^2\mathrm{len}_\gamma^2}{l^2}\frac{R_0^2d}{\lambda_1(\Sigma_{l,h}^b)}+\beta\sigma_\gamma^2\frac{\lambda_d(\Sigma_{l,h}^b)R_0^2d}{\lambda_1(\Sigma_{l,h}^b)^2}\ .\end{aligned}$$

Thus, in the preceding inequality, we require the coefficient of $\sqrt{\mathrm{dist}(x,h)}$ to be smaller than $c\frac{\mathrm{len}_\gamma}{l}$, and all other terms to be smaller than $c\frac{\mathrm{len}_\gamma^2}{l^2}$, so that (12) can be guaranteed. To achieve this, choose the small scales ϵ,δ,β,u as follows. With $l\gtrsim C_fC_YR_0/\sigma_\gamma$,

$$\epsilon'=c\frac{\mathrm{len}_\gamma}{lR_0\sqrt{d}},\quad \delta'=c\frac{\sigma_\gamma^2}{R_0^2d},\quad \beta'=c\frac{\sigma_\gamma^2C_f'^2}{R_0^2C_f^2d},\ u'=c\frac{\sigma_\gamma}{R_0\sqrt{d}}\,.$$

Here $\delta',\beta'<R_0^2d\sigma_\gamma^{-2}$ automatically holds because of $\sigma_\gamma\leqslant R_0\leqslant R_0\sqrt{d}$ given by Assumption **(LCV)**. Therefore, we substitute these quantities in (13) and obtain

$$\mathbb{P}\left(\mathfrak{S}(\epsilon',\delta',\beta',u')^c\right)\lesssim ld\exp\left(-c\frac{n_{loc}}{\sqrt{\tau\log n}}\min\left(\frac{C_f'\sigma_\gamma^4}{C_f^2R_0^3d^{3/2}\mathrm{len}_\gamma},\frac{\sigma_\gamma^8C_f'^4}{R_0^8C_f^4d^4}\right)\right)+ln^{-\tau}$$

Here Lemma 19 is used with $\alpha=R_0^2/\sigma_\gamma^2$, and we also used $C_f'|\mathcal{R}_f|\leqslant\mathrm{len}_\gamma\leqslant C_f|\mathcal{R}_f|$ together with $|\mathcal{R}_f|\asymp C_YR_0$. ■

A.7 Proof of Proposition 14, Theorem 3, and Theorem 5

Proof Throughout this proof, condition on the first half. For each $h\in\mathcal{H}_l$, the set $A_{l,h}$ and the intervals $I_{j,k}^{(l,h)}$ are then fixed, while $Z=\langle\widehat{v}_{l,h}^b,X\rangle$ is a fixed function of X. The estimates below are first proved on the accepted regions $A_{l,h}$ and are transferred to the test regions by Lemma 21.

Lemma 21 (accepted-region domination) *Condition on the first half. For every collection of nonnegative measurable functions G_h, $h \in \mathcal{H}_l$,*

$$\sum_{h\in\mathcal{H}_l} \mathbb{E}\left[G_h(X)\mathbb{1}\{\widehat{h}(X)=h\}\right] \leqslant \sum_{h\in\mathcal{H}_l} \mathbb{E}\left[G_h(X)\mathbb{1}\{X\in A_{l,h}\}\right].$$

Moreover, $\sum_{h\in\mathcal{H}_l} \mathbb{1}\{X\in A_{l,h}\} \leqslant 3$ pointwise.

The proof of Lemma 21 is straightforward: if $\widehat{h}(x)=h$, then $|\widehat{h}(x)-h|=0$, hence $x\in A_{l,h}$. The displayed inequality follows pointwise from $\mathbb{1}\{\widehat{h}(X)=h\} \leqslant \mathbb{1}\{X\in A_{l,h}\}$. For fixed x, the condition $x\in A_{l,h}$ is equivalent to $h\in\{\widehat{h}(x)-1,\widehat{h}(x),\widehat{h}(x)+1\}$, so at most three indices contribute.

Recall the definition of the random variable

$$\eta_h := f(\Pi_\gamma X) - f(\langle \widehat{v}^b_{l,h}, X\rangle + c_{l,h|v^b_{l,h}}).$$

On the retained-slice count event and on the event on which the conclusions of Propositions 11 and 13 both hold, if $X\in A_{l,h}$, then $|\widehat{h}(X)-h| \leqslant 1$, $|\widehat{h}(X)-h_x| \leqslant 1$, and $|h_x - h'_x| \leqslant 1$. Hence $|h'_x - h| \leqslant 3$. Thus the accepted region for the index h is contained, up to the complement already controlled by Proposition 13, in the union of at most seven consecutive true response slices. Applying the same local curve estimate used for one slice to this enlarged interval changes only the universal constant in the slice-width term. Hölder continuity and the support bound in Assumption $(\boldsymbol{\gamma_1})$ give the honest pointwise estimate

$$\begin{aligned}|\eta_h| \lesssim\ & [f]_{\mathcal{C}^s}\left(C_f \max\left\{\sigma_\zeta, \omega_f, \frac{\mathrm{len}_\gamma}{C'_f l}\right\}\right)^s \\ &+ [f]_{\mathcal{C}^s}\left(C_X\sqrt{d}\,R_0\right)^s \left\|\widehat{v}^b_{l,h} - v^b_{l,h}\right\|^s =: U_{\eta,h}.\end{aligned}$$

The sharper factor $\bar{\sigma}_\gamma/\mathrm{reach}_\gamma$ is used only after taking second moments of the normal displacement; it is not asserted as an almost-sure improvement. Consequently, for $\zeta' = \eta_h - \mathbb{E}[\eta_h \mid Z, X\in A_{l,h}] + \zeta$,

$$\mathrm{Var}(\zeta' \mid Z, X\in A_{l,h}) \leqslant \sigma^2_\zeta + \mathbb{E}[\eta_h^2 \mid Z, X\in A_{l,h}].$$

Using the seven-slice containment, the conditional second-moment identity for W, the seven-slice overlap bound, and Corollary 10, and summing only after taking these second moments, yields

$$\begin{aligned}\sum_{h\in\mathcal{H}_l} \mathbb{E}\left[\eta_h^2\mathbb{1}\{X\in A_{l,h}\}\right] \lesssim\ & [f]^2_{\mathcal{C}^s}\left(\frac{\bar{\sigma}_\gamma C_f}{\mathrm{reach}_\gamma}\right)^{2s} \max\left(\sigma_\zeta, \omega_f, \frac{\mathrm{len}_\gamma}{C'_f l}\right)^{2s} \\ &+ [f]^2_{\mathcal{C}^s}\left(C_X R_0^2 \mathrm{len}_\gamma \sigma_\gamma^{-2} d\log d\right)^{2s}\left(\frac{(\log n)^{1.5}}{n}\right)^s.\end{aligned}$$

We use this averaged bound in the proof of Theorem 3. Taking instead the uniform local-sample-size estimate in Corollary 10 gives

$$\sum_{h\in\mathcal{H}_l}\mathbb{E}\left[\eta_h^2\mathbb{1}\{X\in A_{l,h}\}\right]\lesssim [f]_{\mathcal{C}^s}^2\left(\frac{\bar\sigma_\gamma C_f}{\mathrm{reach}_\gamma}\right)^{2s}\max\left(\sigma_\zeta,\omega_f,\frac{\mathrm{len}_\gamma}{C_f' l}\right)^{2s} + [f]_{\mathcal{C}^s}^2\left(C_X R_0^2\mathrm{len}_\gamma\sigma_\gamma^{-2}d\log d\right)^{2s}\left(\frac{(\log n)^{1.5}}{n_{loc}l^2}\right)^s.$$

We use this version in the proof of Theorem 5.

We next control the three terms in the decomposition. By Lemma 21 and $(a+b+c)^2\leqslant 3(a^2+b^2+c^2)$, it is enough to control the corresponding errors on $A_{l,h}$ and then sum over $h\in\mathcal{H}_l$. For the nonlinear curve-approximation term, the definition of $f^*_{l,h|\widehat{v}^b_{l,h}}$ as the conditional mean on $A_{l,h}$ gives

$$\begin{aligned}\mathrm{MSE}_{(\mathrm{NCA})} &:= \sum_{h\in\mathcal{H}_l}\mathbb{E}\left[\left|f(\Pi_\gamma X)-f^*_{l,h|\widehat{v}^b_{l,h}}(Z)\right|^2\mathbb{1}\{X\in A_{l,h}\}\mathbb{1}_{I^{(l,h)}}(Z)\right]\\ &= \sum_{h\in\mathcal{H}_l}\mathbb{E}\left[\mathrm{Var}(f(\Pi_\gamma X)\mid Z,\ X\in A_{l,h})\mathbb{1}\{X\in A_{l,h}\}\mathbb{1}_{I^{(l,h)}}(Z)\right].\end{aligned}$$

Write $X=\gamma(t)+W$, with $t=\Pi_\gamma X$ and $W\perp\gamma'(t)$, and set $\phi_h(t)=\langle\widehat{v}^b_{l,h},\gamma(t)\rangle$ on the seven-slice arc of γ corresponding to $A_{l,h}$. On the geometry event of Proposition 13, the monotonicity estimate used in Lemma 22 makes ϕ_h bi-Lipschitz, so ϕ_h^{-1} is uniformly Lipschitz on its image. Extend ϕ_h^{-1} to $I^{(l,h)}$ without increasing its Lipschitz constant; no higher-order regularity of the inverse is used below. Since $Z=\phi_h(t)+\langle\widehat{v}^b_{l,h},W\rangle$,

$$\left|t-\phi_h^{-1}(Z)\right|\lesssim\left|\langle\widehat{v}^b_{l,h},W\rangle\right|.$$

The conditional mean minimizes squared error. Hence Hölder continuity of f, followed by Jensen's inequality, gives

$$\mathrm{Var}(f(t)\mid Z,X\in A_{l,h})\lesssim[f]_{\mathcal{C}^s}^2\ \mathbb{E}\left[\left|\langle\widehat{v}^b_{l,h},W\rangle\right|^{2s}\ \middle|\ Z,X\in A_{l,h}\right].$$

By Assumption $(\boldsymbol{\gamma_1})$,

$$\mathbb{E}\left[\langle\widehat{v}^b_{l,h},W\rangle^2\mid t\right]=\sigma_\gamma(t)^2\left\|P_{\gamma'(t)^\perp}\widehat{v}^b_{l,h}\right\|^2\leqslant\bar\sigma_\gamma^2\left\|\widehat{v}^b_{l,h}-\gamma'(t)\right\|^2.$$

Combining the population tangent alignment from Corollary 9, the empirical-vector moment bound in Corollary 10, and the seven-slice overlap bound yields

$$\begin{aligned}\mathrm{MSE}_{(\mathrm{NCA})}\lesssim\ &[f]_{\mathcal{C}^s}^2\left(\frac{\bar\sigma_\gamma C_f}{\mathrm{reach}_\gamma}\max\left\{\sigma_\zeta,\omega_f,\frac{\mathrm{len}_\gamma}{C_f' l}\right\}\right)^{2s}\\ &+[f]_{\mathcal{C}^s}^2\left(C_X R_0^2\mathrm{len}_\gamma\sigma_\gamma^{-2}d\log d\right)^{2s}\left(\frac{(\log n)^{1.5}}{\max\{n,n_{loc}l^2\}}\right)^s.\end{aligned}\tag{14}$$

For the bias term, let q be any degree-m piecewise polynomial on the partition of $I^{(l,h)}$. Since $f^*_{l,h|\widehat{v}^b_{l,h}}$ is a conditional expectation,

$$\mathbb{E}\left[\left|f^*_{l,h|\widehat{v}^b_{l,h}}(Z)-q(Z)\right|^2 \mathbb{1}\{X\in A_{l,h}\}\right] \leqslant \mathbb{E}\left[|f(t)-q(Z)|^2\,\mathbb{1}\{X\in A_{l,h}\}\right].$$

Choose q to be the usual degree-m piecewise-polynomial approximation of $z\mapsto f(\phi_h^{-1}(z))$. Because $0<s\leqslant 1$ and ϕ_h^{-1} is uniformly Lipschitz on the arc of γ corresponding to $A_{l,h}$, the composition $f\circ\phi_h^{-1}$ is s-Hölder with seminorm controlled by $[f]_{\mathcal{C}^s}$ up to fixed geometric constants. Splitting $f(t)-q(Z)$ through $f(\phi_h^{-1}(Z))$ and using the estimate that led to (14) gives

$$\begin{aligned}\mathrm{MSE}_{(\mathrm{B})} &:= \sum_{h\in\mathcal{H}_l}\mathbb{E}\left[\left|f^*_{l,h|\widehat{v}^b_{l,h}}(Z)-f^*_{j|\widehat{v}^b_{l,h}}(Z)\right|^2 \mathbb{1}\{X\in A_{l,h}\}\mathbb{1}_{I^{(l,h)}}(Z)\right]\\ &\lesssim \mathrm{MSE}_{(\mathrm{NCA})}+([f]_{\mathcal{C}^s}+C_Y R_0 l^{-1}\,[\rho_X]_{\mathcal{C}^s})^2 C_f^{2s}\max\left(\sigma_\zeta,\omega_f,\frac{\mathrm{len}_\gamma}{C'_f l}\right)^{2s} j^{-2s}.\end{aligned} \tag{15}$$

Conditional on the first half, the observations used for regression are independent, and their observational noises remain centered. The conditional response variance equals $\sigma_\zeta^2+\mathrm{Var}(f(t)\mid Z, X\in A_{l,h})$. On the retained-bin count and interval-coverage events, the calculations in (Liao et al., 2022, Proposition 2 and Lemma 5) therefore give

$$\begin{aligned}\mathrm{MSE}_{(\mathrm{V})} &:= \sum_{h\in\mathcal{H}_l}\mathbb{E}\left[\left|f^*_{j|\widehat{v}^b_{l,h}}(Z)-\widehat{f}_{j|\widehat{v}^b_{l,h}}(Z)\right|^2 \mathbb{1}\{X\in A_{l,h}\}\right]\\ &\lesssim \left(\mathrm{MSE}_{(\mathrm{NCA})}+\sigma_\zeta^2 l\right)\frac{j\log j}{n}.\end{aligned} \tag{16}$$

In the admissible sample-size regime, $j\log j/n\leqslant 1$, so the first contribution in (16) is absorbed into (14). Collecting the three bounds proves Proposition 14. Because Algorithm 1 clips its output and $|F|\leqslant M$, every exceptional event E contributes at most $4M^2\mathbb{P}(E)$ to the unconditional risk. Propositions 13 and Lemmas 22–23 show that the classification, interval-coverage, and retained-bin exceptional contributions are of smaller order than the displayed theorem rates.

Two situations are not exceptional events of this kind, and are accounted for separately and additively. The two extreme response indices are bounded in Lemma 24 by Cn^{-1} times the risk just bounded, plus $Cn^{-1}\,[f]^2_{\mathcal{C}^s}\,w^{2s}$. The bins (at most two) for each index that straddle an endpoint of $J^{(l,h)}$ need not be retained, and a query landing in one of them is evaluated, by Step 3.c, using the fit of the adjacent bin, which is covered and hence retained with high probability; the evaluation point is then at most one bin width outside that bin, so by (20) the squared error on these random boundary bins is controlled by the same conditional-variance and polynomial-stability argument as the bin-averaged risk, and the total contribution is again $O(\mathrm{MSE}_{(\mathrm{NCA})}+\mathrm{MSE}_{(\mathrm{B})}+\mathrm{MSE}_{(\mathrm{V})})$; the pointwise support estimate is needed only for the extreme endpoints treated in Lemma 24. The nearest-retained-bin convention therefore gives the same order of risk on these partial boundary bins as on the retained interior bins.

Theorems 3 and 5 follow by optimizing the parameters (l, j). For the noiseless case $\sigma_\zeta = 0$, we first trim the low-density region. Recall that ρ_t is the density of the pushforward of ρ_X under Π_γ. In Theorem 5, let $\Omega_0 := \{x \in \Omega_\gamma : \rho_t(\Pi_\gamma x) > c_\rho\}$. On its complement, the squared error is controlled by the L^∞ norm, producing the additional term in $\mathbb{E}[|\widehat{F}(X) - F(X)|^2]$. We take the largest l satisfying (7), namely

$$l^* = C\frac{C'_f c_\rho \mathrm{len}_\gamma}{C_f C_{\gamma,f}}\frac{n}{\log^{3/2} n}, \quad C_{\gamma,f} := \frac{R_0^3 C_f^2}{C'_f}\max\left(\frac{d^{3/2}\mathrm{len}_\gamma}{\sigma_\gamma^4}, \frac{R_0^5 C_f^2 d^4}{C_f'^3 \sigma_\gamma^8}\right).$$

This choice gives

$$n_{loc} := \frac{C'_f \mathrm{len}_\gamma c_\rho}{2C_f}\frac{n}{l^*} = C_{\gamma,f}\log^{3/2} n,$$

which satisfies (7). A union bound over the response slices whose relevant inverse-image arcs lie in Ω_0 gives

$$\mathbb{P}\Big(n_{l,h}^b \leqslant n_{loc} \text{ for some such } h\Big) \leqslant l^* \exp\Big(-\frac{n_{loc}}{4}\Big) \lesssim \frac{\log^3 n}{n^2}.$$

Thus every slice whose inverse-image arc lies in Ω_0 and is used in the localized analysis is retained with high probability. On this event, each test point in Ω_0 has a retained response slice h with $|h - h'_x| \leq 1$, so the hypothesis of Proposition 11 that a slice h with $|h - h'_x| \leq 1$ is retained in $\mathcal{H}_l$ is satisfied. Substituting l^* yields the desired bound for $\mathbb{E}[|\widehat{F}(X) - F(X)|^2 \mathbb{1}(X \in \Omega_0)]$. Finally, $\mathbb{E}[|\widehat{F}(X) - F(X)|^2 \mathbb{1}(X \notin \Omega_0)] \leqslant |f|_{L^\infty}^2 \mathbb{P}(X \notin \Omega_0)$, which gives the stated bound for the full mean squared error.

For the noisy case $\sigma_\zeta > 0$, we balance the bias term $\mathrm{MSE}_{(B)}$ and the variance term $\mathrm{MSE}_{(V)}$. Using the definition of M^* in (4), the optimizer satisfies

$$l^* j^* \asymp n^{\frac{1}{2s+1}} M^*. \tag{17}$$

We record why the selection rule of Theorem 3 realizes this product whichever constraint is active. Assumption $(\boldsymbol{\omega_f})$, applied to $\mathcal{R}_f$, together with $|\mathcal{R}_f| \asymp C_Y R_0$ and the bounded ratio C_f/C'_f, gives $\mathrm{len}_\gamma \asymp C'_f C_Y R_0$ up to fixed constants. Hence, for every admissible $l \leqslant l_{\max} \asymp C_Y R_0/\max(\sigma_\zeta, \omega_f)$,

$$\max\left(\sigma_\zeta, \omega_f, \frac{\mathrm{len}_\gamma}{C'_f l}\right) \asymp \frac{\mathrm{len}_\gamma}{C'_f l},$$

so by (15) the bias depends on (l, j) only through the product lj and decreases in it, while by (16) the variance $\lesssim (\mathrm{MSE}_{(\mathrm{NCA})} + \sigma_\zeta^2 l) j \log j/n$ increases in it. Any allocation achieving the product (17) is therefore as good as any other, and only the constraints on l decide the split. Second, $M^* n^{1/(2s+1)}$ is that product up to universal constants and a logarithmic factor: writing $A := [f]_{\mathcal{C}^s} + |f|_{L^\infty}[\rho_X]_{\mathcal{C}^s}$, balancing $A^2(C_f \mathrm{len}_\gamma/C'_f)^{2s}(lj)^{-2s}$ against $\sigma_\zeta^2 lj \log(lj)/n$ gives $lj \asymp \big(A^2(C_f C_Y R_0)^{2s} n/(\sigma_\zeta^2 \log n)\big)^{1/(2s+1)}$. The rule of Theorem

3 takes l^* as large as the three constraints on l allow and gives the remaining budget to j^*, so that $l^* j^* \asymp M^* n^{1/(2s+1)}$ in every case. Substituting,

$$\begin{aligned}
\mathrm{MSE}_{(\mathrm{B})} &\lesssim \mathrm{MSE}_{(\mathrm{NCA})} + C_1 n^{-\frac{2s}{2s+1}}, \\
\mathrm{MSE}_{(\mathrm{V})} &\lesssim \sigma_\zeta^2 \, l^* j^* \, \frac{\log n}{n} \asymp \sigma_\zeta^2 M^* n^{-\frac{2s}{2s+1}} \log n \lesssim C_1 n^{-\frac{2s}{2s+1}} \log n,
\end{aligned}$$

the last step because $C_1/(\sigma_\zeta^2 M^*) = \big(\mathrm{len}_\gamma/(C_f' C_Y R_0)\big)^{2s/(2s+1)} \gtrsim 1$. This is the bound claimed in Theorem 3, uniformly over which constraint binds.

It remains only to account for the finite-slicing entry $\mathrm{len}_\gamma/(C_f' l^*)$ in (14) when $l^* < l_{\max}$. Because $s \in [\frac{1}{2}, 1]$, the exponent in the nonlinear curve-approximation term is $2s$. If the sample-count constraint $l^* = C_{\gamma,f}^{-1} n \log^{-2} n$ is active, its contribution is $O(n^{-2s} \log^{4s} n)$ and therefore decays faster than the first term of Theorem 3. If $l^* = M^* n^{1/(2s+1)}$, the finite-slicing contribution has the same rate $n^{-2s/(2s+1)}$. Indeed, writing $A := [f]_{\mathcal{C}^s} + |f|_{L^\infty} [\rho_X]_{\mathcal{C}^s}$ and suppressing the arguments of C_1, direct substitution of the definition of M^* gives

$$\frac{[f]_{\mathcal{C}^s}^2 \left(\dfrac{\bar{\sigma}_\gamma C_f \mathrm{len}_\gamma}{\mathrm{reach}_\gamma C_f' M^*} \right)^{2s}}{C_1} = \left(\frac{[f]_{\mathcal{C}^s}}{A} \right)^2 \left(\frac{\bar{\sigma}_\gamma}{\mathrm{reach}_\gamma} \right)^{2s} \left(\frac{\mathrm{len}_\gamma}{C_f' C_Y R_0} \right)^{\frac{4s^2}{2s+1}} \lesssim 1.$$

Here the first factor is at most one, Assumption $(\boldsymbol{\gamma_1})$ controls the second factor, and Assumption $(\boldsymbol{\omega_f})$ together with the response-scale convention and the fixed ratio C_f/C_f' controls the third. Thus the product-constraint contribution is absorbed into the existing coefficient C_1. Finally, when $l^* = l_{\max}$, the finite-slicing scale is comparable to $\max(\sigma_\zeta, \omega_f)$ and the nonlinear curve-approximation term is precisely the n-independent saturation term C_2 printed in the theorem.

For completeness, a direct implementation has construction cost

$$O\big(d^2 n + n \log n + dn|\mathcal{H}_l| + d^3 |\mathcal{H}_l|\big).$$

In the noisy regime, $|\mathcal{H}_l| \leqslant l^* \leqslant l_{\max}$ and $l_{\max}$ is a fixed model-class parameter, so the assignment term is absorbed into the model-dependent constant. Moreover, $n_{loc} \geqslant 2d+1$ and $|\mathcal{H}_l| n_{loc} \leqslant n/4$ imply $d^3 |\mathcal{H}_l| = O(d^2 n)$. This yields the stated $O(d^2 n \log n)$ construction bound.

Recall that ρ_t is the density of the pushforward of ρ_X under Π_γ. Define

$$\begin{aligned}
c_{\rho,n} &:= (C_Y R_0)^{-2} \mathrm{len}_\gamma^{-1} C_1(f, \gamma, \rho_X, \sigma_\zeta, d) n^{-\frac{2s}{2s+1}} \log n, \\
\Omega_{0,n} &:= \{x \in \Omega_\gamma : \rho_t(\Pi_\gamma x) > c_{\rho,n}\}.
\end{aligned}$$

By direct integration over the set $\{\rho_t \leqslant c_{\rho,n}\}$,

$$\mathbb{E}[|\widehat{F}(X) - F(X)|^2 \mathbb{1}\{X \notin \Omega_{0,n}\}] \lesssim (C_Y R_0)^2 \mathbb{P}(X \notin \Omega_{0,n}) \lesssim C_1(f, \gamma, \rho_X, \sigma_\zeta, d) n^{-\frac{2s}{2s+1}} \log n,$$

which is the desired one-dimensional nonparametric rate. A union bound over the response slices whose relevant inverse-image arcs lie in $\Omega_{0,n}$ gives

$$\mathbb{P}\Big(n_{l,h}^b \leqslant n_{loc} \text{ for some such } h\Big) \leqslant l \exp\Big(-\frac{n_{loc}}{4}\Big),$$
$$n_{loc} := \frac{C_f' \max(\sigma_\zeta, \omega_f)}{2\mathrm{len}_\gamma C_Y^2 R_0^2} C_1(f, \gamma, \rho_X, \sigma_\zeta, d) n^{\frac{1}{2s+1}} \log n.$$

Thus every slice whose inverse-image arc lies in Ω_0 and is used in the noisy localized analysis is retained with high probability. On this event, each test point in Ω_0 has a retained response slice h with $|h - h_x'| \leq 1$, so the hypothesis of Proposition 11 that a slice h with $|h - h_x'| \leq 1$ is retained in $\mathcal{H}_l$ is satisfied. The last probability vanishes faster than $\mathcal{O}(n^{-\frac{2s}{2s+1}} \log n)$ and is therefore negligible in the final mean-squared-error bound. For Theorem 3, Lemmas 22–23 are applied with $c_\rho = c_{\rho,n}$ and $\Omega_0 = \Omega_{0,n}$. It remains to verify the interval-coverage and retained-bin events used in Proposition 14. These events are established in Lemmas 22 and 23, respectively; conditional on the first half, the required second-half counts are sums of independent Bernoulli variables. Lemma 24 then treats the two extreme response indices, which the interval-coverage event excludes.

We now make the two events invoked above quantitative. Both rely on the following elementary monotonicity fact, which follows directly from the geometry estimates. On the event $\mathfrak{S}(\epsilon', \delta', \beta', u')$ of Proposition 13, we have $\left\|\widehat{v}_{l,h}^b - v_{l,h}^b\right\| \lesssim \epsilon'$ with $\epsilon' = c\mathrm{len}_\gamma/(lR_0\sqrt{d})$; together with the population alignment from Corollary 9, this implies that

$$\langle \widehat{v}_{l,h}^b, \gamma'(t) \rangle \geqslant 1 - c \tag{18}$$

on every arc of length $\lesssim \mathrm{len}_\gamma/l$. Hence $t \mapsto \langle \widehat{v}_{l,h}^b, \gamma(t) \rangle$ is an increasing bi-Lipschitz bijection from such an arc onto an interval of projected coordinates, with constants in $[1-c, 1]$. Moreover, writing $t = \Pi_\gamma X$ and using $X - \gamma(t) \perp \gamma'(t)$,

$$\langle \widehat{v}_{l,h}^b, X - \gamma(t) \rangle = \langle \widehat{v}_{l,h}^b - \gamma'(t), X - \gamma(t) \rangle.$$

The three sources of error in $\widehat{v}_{l,h}^b - \gamma'(t)$ must all be accounted for here: the empirical-vector error, the population misalignment of $v_{l,h}^b$, and the variation of γ' over the local arc. By the triangle inequality,

$$\left\|\widehat{v}_{l,h}^b - \gamma'(t)\right\| \leqslant \underbrace{\left\|\widehat{v}_{l,h}^b - v_{l,h}^b\right\|}_{\lesssim \epsilon'} + \underbrace{\left\|v_{l,h}^b - \gamma'(\bar{t}_h)\right\|}_{\lesssim \|\gamma''\|_\infty C_f(|R_{l,h}^b| + \sigma_\zeta)} + \underbrace{\left\|\gamma'(\bar{t}_h) - \gamma'(t)\right\|}_{\leqslant \|\gamma''\|_\infty |t - \bar{t}_h|},$$

where $\bar{t}_h = \mathbb{E}[\Pi_\gamma X \mid Y \in R_{l,h}^b, \mathcal{B}]$ and the middle bound is the tangent alignment established in the proof of Corollary 9. The second and third terms are *population* quantities: they are not reduced by shrinking the universal constant in ϵ'. Assumption $(\boldsymbol{\gamma_1})$ bounds the off-curve deviation by $\|X - \gamma(t)\| \leqslant c\,\mathrm{reach}_\gamma$, which is sharper than $R_0\sqrt{d}$; combining this with $\|\gamma''\|_\infty \leqslant 1/\mathrm{reach}_\gamma$ gives the transversal fluctuation

$$\left|\langle \widehat{v}_{l,h}^b, X - \gamma(t) \rangle\right| \;\lesssim\; \epsilon' \mathrm{reach}_\gamma + C_f \max\left(\sigma_\zeta, \omega_f, \frac{\mathrm{len}_\gamma}{C_f' l}\right). \tag{19}$$

The first term is at most $c\text{len}_\gamma/(lR_0\sqrt{d})\cdot\text{reach}_\gamma \ll \text{len}_\gamma/l$ and is made smaller than any fixed fraction of the projected slice width by choosing the universal constant in ϵ' small. The second term does not shrink with n, and by Lemma 12(i) it is at most a fixed multiple of one projected slice width. By the bounded-ratio clause in Assumption $(\boldsymbol{\omega_f})$, this fixed multiple is absorbed by the one-slice buffer in Step 3.a; hence the total transversal fluctuation is less than one half of a local projected slice width on the stated event.

The same decomposition justifies the alignment (18). Indeed the second and third terms above are at most $\|\gamma''\|_\infty$ times the length of one slice arc, which by Lemma 12(i) is at most a fixed multiple of len_γ/l; the requirement $l \gtrsim C_f\text{len}_\gamma/(C'_f\sigma_\gamma)$ of Proposition 13 therefore bounds it by $C\sigma_\gamma\|\gamma''\|_\infty \leqslant C\sigma_\gamma/\text{reach}_\gamma$. Finally, the support condition in Assumption $(\boldsymbol{\gamma_1})$ combined with isotropy forces $\sqrt{d-1}\,\sigma_\gamma \lesssim \text{reach}_\gamma$, so this contribution is $O(d^{-1/2})$, and (18) holds with c as small as desired for d large, and after adjusting the universal constants for every d.

Lemma 22 (interval coverage) *Suppose that the event $\mathfrak{S}(\epsilon',\delta',\beta',u')$ of Proposition 13 holds and that every available empirical slice used in Step 3.a for an index $h \in \mathcal{H}_l$ contains at least n_{loc} bounded second-quarter observations. In this lemma, all probabilities are restricted to test points in $\Omega_0 = \{x : \rho_t(\Pi_\gamma x) > c_\rho\}$, as used in the theorem proofs. Except on an additional first-half event of probability at most $le^{-cn_{loc}}$, every test point x with $\widehat{h}(x) = h$ that is almost correctly classified, i.e., $|\widehat{h}(x) - h'_x| \leqslant 2$, and whose index h is not one of the two extreme response indices, satisfies*

$$Z_x = \langle \widehat{v}^b_{l,h}, x\rangle \in I^{(l,h)}.$$

Consequently,

$$\sum_{h\in\mathcal{H}_l \text{ interior}} \mathbb{P}\big(\widehat{h}(X) = h,\ Z \notin I^{(l,h)}\big) \lesssim l\,e^{-cn_{loc}} + l\,n^{-\tau},$$

and the contribution of this event to $\mathbb{E}\left|\widehat{F}(X) - F(X)\right|^2$ is at most $4M^2(le^{-cn_{loc}} + ln^{-\tau})$, which is negligible compared with the rate of Theorem 3. The two extreme response indices, for which no outer buffer exists, are treated separately in Lemma 24.

Proof An almost-correctly classified test point with $\widehat{h}(x) = h$ has its true response index in $\{h-2,\dots,h+2\}$. Step 3.a uses the seven estimated slices $\{x : \widehat{h}(x) = r\}$, $r = h-3,\dots,h+3$, which on the classification event of Proposition 13 differ from the response slices $h-3,\dots,h+3$ only within one slice of their boundaries. On the event $\mathfrak{S}(\epsilon',\delta',\beta',u')$, the projection along $\widehat{v}^b_{l,h}$ is monotone on this local arc, with derivative bounded above and below by fixed positive constants. The two outer slices therefore give a one-slice buffer on each available side of the five slices that may contain the test point. By (19) and the bounded ratio C_f/C'_f in Assumption $(\boldsymbol{\omega_f})$, the transversal fluctuation of both the test point and the buffer observations is at most one half of a projected slice width, so this one-slice buffer is sufficient. On the region Ω_0 used in the theorem proofs, the local density lower bound on each retained outer arc therein implies that a fixed subinterval of each available outer slice has conditional probability bounded below. Since each such outer slice contains at least n_{loc} second-quarter observations, Bernstein's inequality (Vershynin, 2018, Theorem 2.8.4)

gives probability at most $e^{-cn_{loc}}$ that the corresponding buffer contains no observation. A union bound over the retained indices and the two sides gives $le^{-cn_{loc}}$. This covers every interior index, that is, every h for which both outer slices $h \pm 3$ (or, at worst, the two slices immediately outside the five that may contain the test point) are available. Outside these count events and the classification exceptional event, $Z_x \in I^{(l,h)}$ for every interior index h. Proposition 13 bounds the latter event by $ln^{-\tau}$. Since $|\widehat{F}| \leqslant M$ and $|F| \leqslant M$, the excluded contribution to the squared risk is at most $4M^2$ times the displayed probability.

■

Lemma 23 (retained bins) *Condition on the first half of the data and suppose that the event $\mathfrak{S}(\epsilon', \delta', \beta', u')$ of Proposition 13 holds. Call a pair (h,k) with $h \in \mathcal{H}_l$ and $1 \leqslant k \leqslant j$* query-relevant *if the bin $I^{(l,h)}_{j,k}$ can actually be hit by a query, that is, if*

$$\exists\, x \in \Omega_0 : \quad \widehat{h}(x) = h, \quad |\widehat{h}(x) - h'_x| \leqslant 2, \quad \Pi_{I^{(l,h)}} \langle \widehat{v}^b_{l,h}, x \rangle \in I^{(l,h)}_{j,k}\,.$$

Only these bins are used when evaluating $\widehat{F}$, so only these need to be retained; the remaining bins of $I^{(l,h)}$ lie in the part of the interval contributed by the outer slices, carry little or no accepted mass, and are legitimately discarded by the threshold in Step 3.b. Write $J^{(l,h)}$ for the support of the conditional law of Z given $X \in A_{l,h}$, an interval, and call a bin covered *if it is contained in $J^{(l,h)}$. Since $J^{(l,h)}$ is an interval and the bins tile $I^{(l,h)}$, all query-relevant bins are covered except at most two per index, namely those containing an endpoint of $J^{(l,h)}$ in their interior. At an extreme response index the outermost bin of $I^{(l,h)}$, which is where the clipping of Step 3.c sends a query falling outside $I^{(l,h)}$, is covered. For every covered query-relevant bin $I^{(l,h)}_{j,k}$ for which the arc of γ corresponding to $A_{l,h}$ lies in $\Omega_0 = \{x : \rho_t(\Pi_\gamma x) > c_\rho\}$,*

$$\mathbb{P}\big(Z \in I^{(l,h)}_{j,k} \mid X \in A_{l,h}\big) > \frac{2c_0}{j}\,,$$

provided the fixed constant c_0 in Step 3.b is chosen sufficiently small. Conditional on the accepted counts $\{n^{(l,h)}\}_{h \in \mathcal{H}_l}$ and on the event $c_0 n^{(l,h)}/j \geqslant m+1$ for all retained h, Bernstein's inequality (Vershynin, 2018, Theorem 2.8.4) and a union bound over the covered query-relevant pairs (h,k) give

$$\mathbb{P}\left(\begin{array}{c} \exists\,(h,k) \text{ query-relevant and covered}: \\ n^{(l,h)}_{j,k} < \max\left\{ m+1, \dfrac{c_0 n^{(l,h)}}{j} \right\} \end{array} \;\middle|\; \{n^{(l,h)}\}_{h \in \mathcal{H}_l} \right) \lesssim \sum_{h \in \mathcal{H}_l} j \exp\left(-c \frac{n^{(l,h)}}{j} \right),$$

which is negligible compared with the rate of Theorem 3. Finally, the local fit is stable up to the edge of a retained bin and slightly beyond it: for $m \in \{0,1\}$, on the above event, and for any point z_0 that is measurable with respect to the first half of the data and lies within one bin width of a retained bin $I^{(l,h)}_{j,k}$,

$$\mathbb{E}\left[\left| f(z_0 + c_{l,h|v^b_{l,h}}) - \widehat{f}_{j|\widehat{v}^b_{l,h}}(z_0) \right|^2 \;\middle|\; \textit{first half} \right] \lesssim \frac{(\sigma_\zeta^2 + U^2_{\eta,h})\, j}{n^{(l,h)}} + [f]^2_{\mathcal{C}^s} \left(\frac{|I^{(l,h)}|}{j} \right)^{2s} + U^2_{\eta,h}\,, \tag{20}$$

which is, up to universal constants, the same bound as for the bin-averaged risk of $\widehat{f}_{j|\widehat{v}^b_{l,h}}$.

Proof On the arc of γ corresponding to $A_{l,h}$ contained in Ω_0, one has $\rho_t > c_\rho$ by definition. On this event, the map

$$t \longmapsto \langle \widehat{v}_{l,h}^{b}, \gamma(t) \rangle$$

is bi-Lipschitz on the arc of γ corresponding to $A_{l,h}$. The transversal perturbation $\langle \widehat{v}_{l,h}^{b}, X - \gamma(\Pi_\gamma X) \rangle$ is controlled by the preceding geometry estimate (19), which, by Assumption $(\boldsymbol{\omega_f})$, is at most half a projected slice width. The two outer slices in Step 3.a therefore keep every bin reachable by an almost-correctly classified test point inside the projected arc. On this arc of γ corresponding to $A_{l,h}$ in Ω_0, pushing the local tubular density forward through the bi-Lipschitz projection gives the required local upper and lower bounds for the conditional density of Z on $J^{(l,h)}$, hence on every covered query-relevant bin.

We first justify the structural claim about the outermost bin at an extreme response index h; the claim about the at most two uncovered bins is immediate from $J^{(l,h)}$ being an interval and the bins tiling $I^{(l,h)}$. At an extreme index the outer side of $I^{(l,h)}$ is not buffered, since the slices beyond h do not exist: the outer endpoint z_0 of $I^{(l,h)}$ is the extreme projected coordinate of a second-quarter observation with $\widehat{h}(X_i) = h$. Such an observation is accepted, because $A_{l,h} = \{|\widehat{h} - h| \leqslant 1\}$, so $z_0 \in J^{(l,h)}$; and $J^{(l,h)}$ reaches inward from z_0 across the two accepted slices $h-1, h$, a distance $\geqslant 2w(1-o(1))$, whereas the outermost bin has length $|I^{(l,h)}|/j \leqslant 4w(1+o(1))/j$. For $j \geqslant 2$ the outermost bin is therefore contained in $J^{(l,h)}$, that is, covered. Since a query falling outside $I^{(l,h)}$ is clipped to z_0, the value returned by Step 3.c at such a query comes from a retained bin. By contrast, at an interior index the outer endpoints of $I^{(l,h)}$ are contributed by the buffer slices $h \pm 3$, the outermost bins are not query-relevant, and nothing needs to be proved about them.

By Lemma 12(i), the projected widths of the response slices meeting the arc of γ corresponding to $A_{l,h}$ are comparable. The local conditional density of Z on this arc of γ corresponding to $A_{l,h}$ in Ω_0 is therefore bounded below by a fixed multiple of the inverse length of its support. Since a covered query-relevant bin has length comparable to $|I^{(l,h)}|/j$, there is a constant $c > 0$, depending only on the fixed model constants, such that

$$\mathbb{P}\big(Z \in I_{j,k}^{(l,h)} \mid X \in A_{l,h}\big) \geqslant \frac{c}{j}$$

uniformly over covered query-relevant bins. Choose the fixed threshold constant $c_0 < c/2$. This yields the required uniform gap above the retention threshold.

Conditional on the first half and on $n^{(l,h)}$, the accepted projected coordinates are independent draws from the conditional accepted design, and $n_{j,k}^{(l,h)}$ is binomial. The preceding uniform gap and Bernstein's inequality (Vershynin, 2018, Theorem 2.8.4) therefore give

$$\mathbb{P}\left(n_{j,k}^{(l,h)} < \frac{c_0\, n^{(l,h)}}{j} \,\middle|\, n^{(l,h)} \right) \lesssim \exp\left(-c \frac{n^{(l,h)}}{j} \right).$$

On the accepted-count event, $n^{(l,h)} \asymp n/l$. For the choices in Theorems 3, 5, and 6, one has $n/(lj) \to \infty$ (linearly in the wide-slice case), and therefore $c_0 n^{(l,h)}/j \geqslant m+1$ for n sufficiently large. A union bound over the covered query-relevant pairs proves the claim. For $m = 1$, after rescaling each interval to $[-1,1]$, the same density lower bound gives a uniformly positive lower bound for the population 2×2 polynomial Gram matrix. A

standard matrix Bernstein bound transfers this lower bound to the empirical Gram matrix on the same high-probability event, so the retained least-squares fit is well defined.

It remains to prove the pointwise estimate (20). Fix a retained bin $I_{j,k}^{(l,h)}$, rescale it to $[-1,1]$ by the affine map ψ, and let z_0 be a point with $\psi(z_0) \in [-3,3]$, which covers every point within one bin width of $I_{j,k}^{(l,h)}$. Write the least-squares fit as $\widehat{f}_{j,k|\widehat{v}_{l,h}^b}(z) = e(\psi(z))^\intercal \widehat{G}^{-1} \frac{1}{N} \sum_i e(\psi(Z_i)) Y_i$, where $\widehat{G} = \frac{1}{N} \sum_i e(\psi(Z_i)) e(\psi(Z_i))^\intercal$ is the empirical Gram matrix of the bin, $e(u) = (1, u, \dots, u^m)^\intercal$, and $N = n_{j,k}^{(l,h)}$. On the event above, $\widehat{G} \succeq c\,\mathrm{Id}$ with c universal, by the matrix Bernstein argument just given, and $\|e(u)\| \leqslant C$ for $|u| \leqslant 3$ and $m \leqslant 1$. Consequently the leverage $\left\| e(\psi(z_0))^\intercal \widehat{G}^{-1} \right\| \leqslant C$ is bounded by a universal constant for every such z_0, that is, the fit does not blow up when evaluated at the edge of its bin, or up to one bin width outside it. Decomposing $Y_i = f(Z_i + c_{l,h|v_{l,h}^b}) + \eta_h(X_i) + \zeta_i$ as in the proof of Proposition 14 and splitting the resulting error into the approximation error of the best degree-m polynomial on a set of diameter $\leqslant 3\left|I^{(l,h)}\right|/j$, which is $\lesssim [f]_{\mathcal{C}^s} \left(\left|I^{(l,h)}\right|/j\right)^s + U_{\eta,h}$, and the stochastic part, whose conditional variance is at most $C(\sigma_\zeta^2 + U_{\eta,h}^2)/N \leqslant C(\sigma_\zeta^2 + U_{\eta,h}^2)\, j/(c_0 n^{(l,h)})$ by the bounded leverage and the retention threshold $N \geqslant c_0 n^{(l,h)}/j$, gives (20), the factor c_0^{-1} being absorbed into the implicit constant. The bound is uniform in z_0 and therefore applies to first-half-measurable choices such as an endpoint of $I^{(l,h)}$. ∎

Lemma 24 (the two extreme response indices) *In the setting of Lemma 22, let h be one of the two extreme response indices and write $w \asymp \max(\sigma_\zeta, \omega_f, \mathrm{len}_\gamma/(C_f' l))$ for one projected slice width. Then, for each such h, the probability that an almost correctly classified test point x with $\widehat{h}(x) = h$ has $Z_x \notin I^{(l,h)}$ is at most $4/n$, and the total contribution of the two extreme indices to $\mathbb{E}\left|\widehat{F}(X) - F(X)\right|^2$ is at most*

$$\frac{C}{n}\left([f]_{\mathcal{C}^s}^2\, w^{2s} + \mathrm{MSE}_{(\mathrm{NCA})} + \mathrm{MSE}_{(\mathrm{B})} + \mathrm{MSE}_{(\mathrm{V})}\right).$$

This is negligible compared with the rates of both Theorem 3 and Theorem 5: the second group of terms is $1/n$ times the risk bounded in Proposition 14, and in the noiseless case $w \asymp \mathrm{len}_\gamma/(C_f' l^*) \asymp \log^{3/2} n/n$, so the first term is $O(n^{-1-2s} \log^{3s} n)$, smaller by a factor n^{-1} than the stated rate. Without the clipping of Step 3.c the same event would contribute $|f|_{L^\infty}^2/n$, which exceeds the rate of Theorem 5 for every $s > \frac{1}{2}$.

Proof For such an h the outer side of $I^{(l,h)}$ is not buffered: $I^{(l,h)}$ is the smallest interval containing finitely many projections, so its outer endpoint z_0 is the extreme order statistic of the projections of the N_h observations of the second quarter with $\widehat{h}(X_i) = h$, and lies strictly inside the support of the projected design. No choice of buffer inside the response range removes this, and the resulting failure probability is polynomially, not exponentially, small. We bound it without any assumption on how the mass is distributed among slices.

Condition on the first quarter of the sample. This fixes the slices, the direction $\widehat{v}_{l,h}^b$ and the classification rule $\widehat{h}(\cdot)$, and leaves the observations of the second quarter independent and

identically distributed; this is the reason for the second split in Step 0. Set $p_h := \mathbb{P}(\widehat{h}(X) = h)$, computed with $\widehat{h}$ frozen. The N_h second-quarter observations with $\widehat{h}(X_i) = h$ and an independent test point X with $\widehat{h}(X) = h$ are then exchangeable, and their projections along the frozen direction $\widehat{v}^b_{l,h}$ are exchangeable real random variables, so the test point exceeds all N_h of them with conditional probability $1/(N_h + 1)$. Since $N_h \sim \text{Binomial}(n/4, p_h)$ and $\mathbb{E}[(N+1)^{-1}] \leqslant ((M+1)p)^{-1}$ for $N \sim \text{Binomial}(M, p)$,

$$\mathbb{P}\left(\widehat{h}(X) = h,\ Z_X \notin I^{(l,h)}\right) \ \leqslant\ p_h\, \mathbb{E}\left[\frac{1}{N_h + 1}\right] \ \leqslant\ p_h \cdot \frac{4}{n\, p_h} \ =\ \frac{4}{n}\,,$$

whatever the mass p_h of the slice. The order in which the two restrictions are imposed matters here, and is the following. Lemma 22 restricts all its probabilities to test points in Ω_0, but the N_h second-quarter observations that determine z_0 are not so restricted, and could not be, since Ω_0 is not observable. The display above is therefore carried out with the test point untrimmed as well, which is what makes it exchangeable with those observations; dropping the restriction only enlarges the event being bounded, so the same inequality holds a fortiori for a test point restricted to Ω_0. The restriction is reinstated only at the risk step below, where it is needed for the density lower bound of Lemma 23. Two features of the construction are used here and are the reason for it. The observations defining z_0 are independent of the statistics conditioned upon, which is what makes them exchangeable with the test point; and they are selected by the same rule $\widehat{h}(\cdot) = h$ as the test point, so that the two are drawn from the same conditional law. The estimate also replaces the cruder $p_h \asymp 1/l$ and $n^{(l,h)} \asymp n/l$, which is not available at the extreme indices: under sub-Gaussian Y these are precisely the light slices, which is why $\mathcal{H}_l$ trims by count in the first place.

On the geometry event such a test point still lies in the slice h up to the transversal fluctuation (19), hence within one projected slice width w of z_0. Step 3.c evaluates it at z_0, which by Lemma 23 belongs to a retained bin. Using $\left|f(z + c_{l,h|v^b_{l,h}}) - f(z_0 + c_{l,h|v^b_{l,h}})\right| \leqslant [f]_{\mathcal{C}^s}\, |z - z_0|^s$,

$$\left|F(x) - \widehat{F}(x)\right|^2 \ \leqslant\ 2\,[f]^2_{\mathcal{C}^s}\, w^{2s} + 2\left|f(z_0 + c_{l,h|v^b_{l,h}}) - \widehat{f}_{j|\widehat{v}^b_{l,h}}(z_0)\right|^2\,,$$

and the second term is a pointwise, first-half-measurable quantity, not an average against the law of X; it is therefore not controlled by $\text{MSE}_{(\text{NCA})}, \text{MSE}_{(\text{B})}, \text{MSE}_{(\text{V})}$ directly, and bounding it by $4M^2$ would reinstate the $|f|^2_{L^\infty}/n$ term that the clipping was introduced to remove. Instead, its conditional expectation is bounded by (20): the stochastic and polynomial-approximation terms are controlled by $\text{MSE}_{(\text{V})}$ and $\text{MSE}_{(\text{B})}$, while the pointwise term $U^2_{\eta,h}$ contributes the explicit support-based term $[f]^2_{\mathcal{C}^s}\, w^{2s}$ together with the already controlled empirical-vector term. Multiplying by the probability $4/n$ obtained above and summing over the two extreme indices gives the stated bound. ∎

Note that

$$\sum_{h\in\mathcal{H}_l}\mathbb{E}\left[\eta_h^2\mathbb{1}\{X\in A_{l,h}\}\right] \lesssim [f]_{\mathcal{C}^s}^2\left(\frac{\bar\sigma_\gamma C_f}{\mathrm{reach}_\gamma}\right)^{2s}\max\left(\sigma_\zeta,\omega_f,\frac{\mathrm{len}_\gamma}{C_f' l}\right)^{2s} + [f]_{\mathcal{C}^s}^2\left(C_X R_0^2 \mathrm{len}_\gamma \sigma_\gamma^{-2} d\log d\right)^{2s}\left(\frac{(\log n)^{1.5}}{n}\right)^s.$$

The second term is negligible compared with $\mathcal{O}(n^{-\frac{2s}{2s+1}}\log n)$. The display is a useful rough bound for η_h; the sharper conditional-variance estimate (14) gives the saturation level

$$[f]_{\mathcal{C}^s}^2\left(\frac{\bar\sigma_\gamma C_f}{\mathrm{reach}_\gamma}\max(\sigma_\zeta,\omega_f)\right)^{2s}.$$

Finally, the optimal $\mathrm{MSE}_{(B)}+\mathrm{MSE}_{(V)}$ is

$$\left(\frac{C_f}{C_f'}\mathrm{len}_\gamma\sigma_\zeta^2\right)^{\frac{2s}{2s+1}}\left([f]_{\mathcal{C}^s}+C_Y R_0 l^{-1}\left[\rho_X\right]_{\mathcal{C}^s}\right)^{\frac{2}{2s+1}} n^{-\frac{2s}{2s+1}}\log n\,. \tag{21}$$

■

A.8 Proof of Proposition 15

Proof Proposition 7 and Assumption $(\boldsymbol{\omega_f})$ show that conditioning on $Y\in R_{l,h}$ localizes $t=\Pi_\gamma X$ to an arclength interval of length comparable to $C_f\max(\sigma_\zeta,\omega_f)$; summing over the sub-Gaussian annuli I_k changes only fixed constants. As in the noisy thin-slice proof, the lower-density estimates below are invoked only on retained intervals in Ω_0, while the complementary low-density contribution is controlled by the same direct-integration trimming argument. It therefore suffices to perform the following covariance calculation on an interval T of this length. Fix any interval $T\subseteq[0,\mathrm{len}_\gamma]$ with $|T|=C_f\max(\sigma_\zeta,\omega_f)$, and let $\mu_T=\mathbb{E}[X\mid t\in T]=\mathbb{E}[\gamma(t)\mid t\in T]$ and $t_1=\Pi_\gamma\mu_T$. Since γ is parameterized by arclength and $\gamma(t_1)$ is the closest point to μ_T, $\gamma'(t_1)$ is perpendicular to $\mu_T-\gamma(t_1)$. Assumption **(SC)** gives

$$\left\|\gamma(t)-\gamma(t_1)-\gamma'(t_1)(t-t_1)\right\|\leqslant\frac{1}{2}\left\|\gamma''\right\|_\infty|t-t_1|^2\leqslant\frac{1}{2}c_2|t-t_1|. \tag{22}$$

Hence $t_1\in T$ and $\|\gamma(t_1)-\mu_T\|\leqslant\frac{1}{2}c_2|T|$. The local density bounds on this retained interval in Ω_0 imply

$$\mathbb{E}[|t-t_1|^2\mid t\in T]\gtrsim|T|^2.$$

The covariance separates as

$$\Sigma_T=\mathrm{Cov}(\gamma(t)\mid t\in T)+\mathbb{E}[WW^\intercal\mid t\in T].$$

Expanding only the curve covariance around $\gamma(t_1)$ and using (22) gives

$$\left\|\mathrm{Cov}(\gamma(t)\mid t\in T)-\gamma'(t_1)\gamma'(t_1)^\intercal\mathrm{Var}(t\mid t\in T)\right\|\leqslant 2c_2^2|T|^2.$$

Thus the leading curve-covariance term is tangential, while the sum of all curve-Taylor remainder terms has operator norm at most $2c_2^2|T|^2$. Therefore

$$\gamma'(t_1)^\intercal \Sigma_T \gamma'(t_1) \gtrsim (1/4 - 2c_2^2)|T|^2.$$

For every unit vector orthogonal to $\gamma'(t_1)$, the curve contribution is at most $3c_2^2|T|^2$. The conditional normal covariance is bounded above by $\bar{\sigma}_\gamma^2 \leqslant c_1^2 C_f^2 \max(\sigma_\zeta, \omega_f)^2 = c_1^2|T|^2$ under Assumption **(SC)**. Thus

$$\lambda_1(\Sigma_T) - \lambda_2(\Sigma_T) \gtrsim |T|^2$$

for sufficiently small c_1, c_2. The tangential–normal block is $O((c_1+c_2)|T|^2)$, so the largest eigenvector is tangential up to the same order. Returning from $t \in T$ to $Y \in R_{l,h}$ by the initial decomposition over the sub-Gaussian annuli I_k proves $\lambda_1(\Sigma_{l,h}^b) - \lambda_2(\Sigma_{l,h}^b) \gtrsim C_f^2|R_{l,h}|^2$. ■

A.9 Proof of Proposition 16 and Theorem 6

Proof Similar to Proposition 20, we have the following high probability bound on the estimation error of parameters such as $\langle v_{l,h}^b, \widehat{\mu}_{l,h}^b - \mu_{l,h}^b \rangle$, $\left\| \widehat{v}_{l,h}^b - v_{l,h}^b \right\|$, $\left| \lambda_1(\Sigma_{l,h}^b) - \lambda_1(\widehat{\Sigma}_{l,h}^b) \right|$. The argument is the same, so the proof is omitted.

Proposition 25 (local NVM for a "wide" slice) *Suppose* **(X**$\in \psi_2 \cap \mathcal{C}^2$**)**, **(Y**$\in \psi_2$**)**, **(**$\boldsymbol{\zeta} \in \psi_2$**)**, **(**$\boldsymbol{\gamma_1}$**)**, **(SC)**, *and* **(**$\boldsymbol{\omega_f}$**)** *hold. Let* $\mu_{l,h}^b$ *be the mean of* h*-th slice and* $v_{l,h}^b$ *be the significant vector of* h*-th slice. Then, for every* l *such that* $|R_{l,h}^b| \asymp \max(\sigma_\zeta, \omega_f)$ *for all* h*, for every* $\epsilon \in (0,1)$ *and* $\tau \geqslant 1$*, on each slice*

(a) *For any* h *and any* $\epsilon > 0$*, the estimation error of the slice mean along the tangential direction can be bounded as*

$$\mathbb{P}\left\{ \left| \langle v_{l,h}^b, \widehat{\mu}_{l,h}^b - \mu_{l,h}^b \rangle \right| > \frac{C_f^2 |R_{l,h}^b|^2 \epsilon}{R_0 \sqrt{d}} \;\middle|\; n_{l,h}^b \right\} \lesssim d \exp\left(- \frac{c C_f^2 |R_{l,h}^b|^2 \epsilon^2 n_{l,h}^b (\tau \log n)^{-\frac{1}{2}}}{R_0^2 d + \epsilon C_f |R_{l,h}^b| R_0 \sqrt{d}} \right) + n^{-\tau} \ .$$

(b) *For any* h*, the estimation error of the significant vector can be bounded as*

$$\begin{aligned} &\mathbb{P}\left\{ \left\| \widehat{v}_{l,h}^b - v_{l,h}^b \right\| > \epsilon \mid n_{l,h}^b \right\} \\ &\qquad \lesssim d \exp\left(- \frac{c C_f^2 |R_{l,h}^b|^2 \epsilon^2 n_{l,h}^b (\tau \log n)^{-\frac{1}{2}}}{R_0^2 d + \epsilon C_f |R_{l,h}^b| R_0 \sqrt{d}} \right) + d \exp\left(- \frac{c C_f^4 |R_{l,h}|^4 n_{l,h}^b}{R_0^4 d^2} \right) + n^{-\tau} \ . \end{aligned}$$

(c) *For any* h *and any* $0 < u < \frac{R_0\sqrt{d}}{C_f |R_{l,h}^b|} < 1$*, the estimation error of the first principal value of the slice can be bounded as*

$$\begin{aligned} &\mathbb{P}\left(\left| \lambda_1(\widehat{\Sigma}_{l,h}^b) - \lambda_1(\Sigma_{l,h}^b) \right| > u^2 C_f^2 |R_{l,h}^b|^2 \;\middle|\; n_{l,h}^b \right) \\ &\qquad \lesssim d \exp\left(- c \frac{u^2 C_f^4 |R_{l,h}^b|^4 n_{l,h}^b (\tau \log n)^{-\frac{1}{2}}}{R_0^4 d^2 + u C_f^2 |R_{l,h}^b|^2 R_0^2 d} \right) + d \exp\left(- \frac{c C_f^4 |R_{l,h}|^4 n_{l,h}^b}{R_0^4 d^2} \right) + n^{-\tau} \ . \end{aligned}$$

Condition on $n^b_{l,h}$ and repeat the three steps of Proposition 20 with the largest rather than the smallest population eigenvector. Proposition 15 supplies the gap $\lambda_1(\Sigma^b_{l,h}) - \lambda_2(\Sigma^b_{l,h}) \gtrsim C_f^2|R^b_{l,h}|^2$. Scalar Bernstein applied to $\langle v^b_{l,h}, X_i - \mu^b_{l,h}\rangle$ gives the first bound. The exact identity $\widehat{\Sigma}^b_{l,h} = \widetilde{\Sigma}^b_{l,h} - (\widehat{\mu}^b_{l,h} - \mu^b_{l,h})(\widehat{\mu}^b_{l,h} - \mu^b_{l,h})^\intercal$, followed by residual Davis–Kahan, gives the second. Weyl's inequality and the same directional covariance concentration give the third. Replacing the thin-slice eigengap $\asymp \sigma_\gamma^2$ by the wide-slice eigengap $\asymp C_f^2|R^b_{l,h}|^2$ in the calculations of parts (a)–(c) yields the three displayed exponents.

Moreover, for any fixed constant $\beta < \frac{R_0^2 d}{C_f^2|R_{l,h}|^2}$, we have

$$\mathbb{P}\left(\max\left(\left\|\widetilde{\Sigma}^b_{l,h} - \Sigma^b_{l,h}\right\|, \left\|\widehat{\mu}^b_{l,h} - \mu^b_{l,h}\right\|^2\right) \gtrsim \beta C_f^2|R^b_{l,h}|^2\right) \lesssim d\exp\left(-\frac{c\beta^2 C_f^4|R^b_{l,h}|^4 n^b_{l,h}}{R_0^4 d^2}\right) \ .$$

Consider the event $\mathfrak{S}$ that we have a small estimation error for information in all slices:

$$\begin{aligned}\mathfrak{S}(\epsilon,\beta,u): \quad & \text{for all } h,\\ & \left|\langle \widehat{\mu}^b_{l,h} - \mu^b_{l,h}, v^b_{l,h}\rangle\right| \lesssim \frac{C_f^2|R^b_{l,h}|^2\epsilon}{R_0\sqrt{d}},\ \left\|\widehat{\mu}^b_{l,h} - \mu^b_{l,h}\right\| \lesssim \frac{R_0\sqrt{d}}{2},\\ & \left\|v^b_{l,h} - \widehat{v}^b_{l,h}\right\| \lesssim \epsilon,\ \left|\lambda_1(\widehat{\Sigma}^b_{l,h}) - \lambda_1(\Sigma^b_{l,h})\right| \lesssim u^2 C_f^2|R^b_{l,h}|^2,\\ & \max\left(\left\|\widehat{\mu}^b_{l,h} - \mu^b_{l,h}\right\|^2, \left\|\widetilde{\Sigma}^b_{l,h} - \Sigma^b_{l,h}\right\|\right) \lesssim \beta^2 C_f^2|R^b_{l,h}|^2.\end{aligned}$$

We claim that the conditions $\epsilon' \asymp \beta' \asymp \frac{C_f^2|R^b_{l,h}|^2}{R_0^2 d}$ and $u \asymp \frac{C_f|R^b_{l,h}|}{R_0\sqrt{d}}$ are sufficient to perform almost correct classification.

As in the thin-slice regime, correct classification requires the distance-estimation error to be small enough that for all $|h' - h_x| \geqslant 2$,

$$\left|\widehat{\mathrm{dist}}(x,h') - \mathrm{dist}(x,h')\right| + \left|\widehat{\mathrm{dist}}(x,h_x) - \mathrm{dist}(x,h_x)\right| < \left|\mathrm{dist}(x,h') - \mathrm{dist}(x,h_x)\right| \ .$$

Indeed, this will imply that $\widehat{h}_x = \arg\min_{h' \in \mathcal{H}^b_l} \widehat{\mathrm{dist}}(x,h')$ satisfies $|h_x - \widehat{h}_x| \leq 1$.

We now analyze the distance function. Notice that in the "wide" slice scenario, the difference in distance can be bounded as follows:

$$\sqrt{\mathrm{dist}(x,h_x+k)} - \sqrt{\mathrm{dist}(x,h_x)} \gtrsim C_f|R^b_{l,h}| \quad \text{for any } |k| \geqslant 2 \ .$$

(Recall that for large k, we can bound this by $\mathrm{reach}_\gamma \gtrsim C_f|R^b_{l,h}|$ by assumption **(SC)**.) This gives a lower bound for the right-hand side. We next control the estimation error on the left-hand side. We use the same argument in the proof of Proposition 13. Given event $\mathfrak{S}(\epsilon,\beta,u)$,

$$\begin{aligned}&\left|\widehat{\mathrm{dist}}(x,h) - \mathrm{dist}(x,h)\right|\\ &\lesssim C_f^2|R^b_{l,h}|^2\beta^2 + C_f|R^b_{l,h}|\beta\sqrt{\mathrm{dist}(x,h)} + u^2R_0^2 d + \epsilon^2 R_0^2 d + C_f^2|R^b_{l,h}|^2\beta^2\epsilon^2 + \frac{C_f^4|R^b_{l,h}|^4}{R_0^2 d}\epsilon^2\\ &\lesssim C_f^2|R^b_{l,h}|^2 \ ,\end{aligned}$$

when $\epsilon' \asymp \beta' \asymp \frac{C_f^2 |R_{l,h}^b|^2}{R_0^2 d}$ and $u \asymp \frac{C_f |R_{l,h}^b|}{R_0 \sqrt{d}}$. Therefore, we derive the conclusion that the event of misclassification by at least two slices has a small probability:

$$\mathbb{P}\left(\left|\widehat{h}_x - h_x\right| \geqslant 2\right) \lesssim ld \exp\left(-c \frac{C_f^6 \max(\sigma_\zeta, \omega_f)^7 n}{C_Y R_0^7 d^3 \sqrt{\log n}}\right) + ln^{-\tau} \ .$$

This completes the proof of Proposition 16. For Theorem 6, repeat the decomposition in the proof of Proposition 14, now with the largest significant vector. The saturation term is obtained by repeating the conditional-variance argument leading to (14), now using the largest-vector alignment above rather than a pointwise bound on the normal displacement. After taking the conditional second moment and integrating, this gives

$$\mathrm{MSE}_{(\mathrm{NCA})} \lesssim [f]_{\mathcal{C}^s}^2 \left(\frac{\bar{\sigma}_\gamma C_f}{\mathrm{reach}_\gamma} \max(\sigma_\zeta, \omega_f)\right)^{2s}.$$

The local polynomial bias and variance estimates are the same as in Proposition 14. With $l^* = C_Y R_0 / \max(\sigma_\zeta, \omega_f)$ the maximum in (15) is $\asymp \max(\sigma_\zeta, \omega_f)$, so the bias term reads $\lesssim ([f]_{\mathcal{C}^s} + C_Y R_0 (l^*)^{-1} [\rho_X]_{\mathcal{C}^s})^2 (C_f \max(\sigma_\zeta, \omega_f)/j^*)^{2s}$, and (5) is exactly the requirement that it not exceed the saturation term; unlike in Theorem 3, the slice width cannot be reduced to achieve this, since l^* is pinned at $l_{\max}$ by the wide-slice geometry, so the refinement must come from j^*. The variance term $\sigma_\zeta^2 l^* j^* \log j^*/n$ is below the saturation term by the second condition on n. For the fixed wide-slice choice of l and the sample-size range in the theorem, the remaining finite-sample regression term and the classification exceptional-event term are bounded by the displayed saturation term after enlarging its constant. This proves Theorem 6. ■

Appendix B. Parameters for the committor experiment

The experiment in Section 4.3 uses ambient dimension $d = 40$ and a unit-length circular arc of radius 1/6, with angular parameter ranging from 0 to 6. The tangential potential is

$$\begin{aligned} U_{tan}(s) = {} & 512s^{12} - 3072s^{11} + 8448s^{10} - 14080s^9 + 15840s^8 - 12672s^7 \\ & + 7392s^6 - 3168s^5 + 990.625s^4 - 221.25s^3 + 33.625s^2 - 2.9813s + 1. \end{aligned}$$

with minima at $s_A = 0.126$ and $s_B = 0.863$. The basin tolerances are $\delta_{tan} = 0.02$ and $\delta_{nor} = 0.05$, and the normal confinement is $U_{nor}(r) = 5r^2$. The initial distribution has normal standard deviation 0.1 in the first two coordinates, 0.12 in coordinates $3, \ldots, 13$, 0.15 in coordinates $14, \ldots, 26$, and 0.08 in coordinates $27, \ldots, 40$. We use $\sigma_{BM} = 0.2$, Euler–Maruyama time step $\Delta t = 2 \times 10^{-3}$, $n = 4 \times 10^4$ initial configurations, and $K = 4 \times 10^4$ trajectories per configuration. The public repository contains the complete simulation and reproduction workflow.

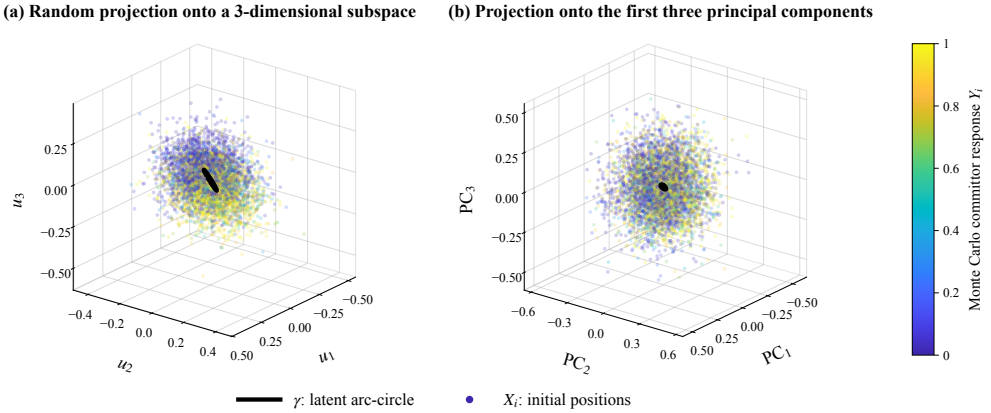

Figure 11: Low-dimensional projections of the predictors X_i in the committor experiment, colored by the Monte Carlo responses Y_i defined in (10). Left: a random projection onto a three-dimensional subspace. Right: projection onto the first three principal components. The latent circular reaction path is shown in black. In this example, the latent geometric structure is not readily revealed by either standard projection.

Appendix C. Case Analysis: Meyer helix

C.0.1 Background: standard & modified Meyer staircase

We begin with the standard Meyer staircase. Fix constant $\delta \geqslant 1$. Consider the unit interval $I = [0, 1]$ and the set of Gaussians $\mathcal{N}(t; \mu, \delta^2)$ where the mean μ takes values in I, and the density function is truncated to accept arguments $t \in I$ only. Varying $\mu \in I$ in this manner induces a smooth embedding of the interval I into the infinite-dimensional Hilbert space $L^2(I)$, i.e., a curve. Explicitly, we take the square root of the Gaussian density centered at $\mu \in I$ and truncate it to $t \in I$:

$$I \to L^2(I) \quad : \quad \mu \mapsto g_\mu(t) := \frac{1}{\sqrt[4]{2\pi}\sqrt{\delta}} \exp\left(-\frac{|t-\mu|^2}{4\delta^2}\right) . \tag{23}$$

By discretizing I, we may sample this manifold and project it into a finite-dimensional space. In particular, for any $d \in \mathbb{N}$, a grid $\Gamma_d \subseteq I$ of d points may be generated. It is obtained by subdividing I in d equal parts and thus $\Gamma_d(k) = k/d$ for $k = 1, \ldots, d$. Explicitly, the evaluation function is

$$L^2(I) \to \mathbb{R}^d \quad : \quad g_\mu(t) \mapsto (g_\mu(1/d), \ldots, g_\mu(1))^\intercal . \tag{24}$$

Thus, combining the preceding two maps (23) and (24) produces an embedding of interval I into $\mathbb{R}^d$ (which is equivalent to a curve in $\mathbb{R}^d$). We write it explicitly as $x(t) = (x_1(t), \ldots, x_d(t))^\intercal$ where $t \in I$ and for each $k = 1, \ldots, d$. For the standard Meyer staircase, the expression for $x_k(t)$ is

$$x_k(t) = \frac{1}{\sqrt[4]{2\pi}\sqrt{\delta}} \exp\left(-\frac{|k/d - t|^2}{4\delta^2}\right) . \tag{25}$$

This expression differs from Definition 1 because it does not use the unit-speed parameterization. It has two advantages: it is defined uniformly across dimensions, avoiding dimension-dependent parameter intervals, and it emphasizes that these curves are finite-dimensional approximations of $g_\mu(t) \in L^2(I)$. For the numerical experiments, we reparameterize the curve by arclength.

Notice that the map (23) describes a curve in $L^2(I)$. Here $\mu \in I$ parameterizes this curve while $t \in I$ is merely the argument for the function g_μ. On the other hand, Equation (25) describes a curve in $\mathbb{R}^d$. Here $t \in I$ parameterizes the curve while μ is replaced by a discrete grid Γ_d.

The standard Meyer staircase provides a curve in high-dimensional Euclidean space $\mathbb{R}^d$. However, because of the construction, as dimension $d \to \infty$, the standard Meyer staircase defined in Equation (25) will converge to a limit that corresponds to the function $g_\mu \in L^2(I)$. This implies that the complexity of the curve is bounded as $d \to \infty$. Because we are focusing on the regression problem for general curve classes, we need to consider various curves with different complexity and different ambient dimensions. Our strategy is to consider analogies of (25).

A direct modification of the standard Meyer staircase is to let $\delta = \frac{1}{d}$ in equation 25. This modified Meyer staircase is an example of a collection of curves whose complexity grows with dimension d. Besides parameters such as length, diameter, curvature, and reach, we also consider the effective linear dimension. One way to measure the effective linear dimension is to study the singular values of the curve. Suppose we perform singular value decomposition on a curve in $\mathbb{R}^d$ and obtain its singular value $\lambda_\gamma(k), k = 1, \ldots, d$ in descending order. Then, we can consider the sum of the singular values divided by the largest singular value, $\|\lambda_\gamma\|_1 := \sum_{k=1}^d \lambda_\gamma(k)/\lambda_\gamma(1)$, or count the number of singular values that are greater than 0.05 times the largest singular value, $\|\lambda_\gamma\|_0 := \#\{\lambda_\gamma(k) : \lambda_\gamma(k) > 0.05\lambda_\gamma(1)\}$. Both $\|\lambda_\gamma\|_1$ and $\|\lambda_\gamma\|_0$ are scaling invariant and measure the minimum number of linear dimensions needed to capture (in the mean squared sense) the underlying curve γ up to a given relative error. These quantities are commonly used as stable notions of matrix rank (often called numerical or stable rank).

Figure 12 illustrates the relationship between the modified Meyer staircase parameters and the dimension d. We observe that the length is roughly proportional to $d^{1.5}$, the diameter to $d^{0.5}$, the curvature to $d^{-0.5}$, and the reach to $d^{0.5}$. Both $\|\lambda_\gamma\|_1$ and $\|\lambda_\gamma\|_0$ are roughly proportional to d^1. Thus, the complexity of the modified Meyer staircase grows with d.

However, the modified Meyer staircase is still special in the following two respects: (i) It remains approximately on the sphere $\sqrt{d}\mathbb{S}^{d-1}$: For $t \in (0, 1)$,

$$\frac{1}{d}\|x(t)\|^2 := \frac{1}{d}\sum_{k=1}^{d} x_k(t)^2 \approx \int_0^1 \frac{1}{\sqrt{2\pi}\delta} \exp\left(-\frac{|s-t|^2}{2\delta^2}\right) ds \asymp 1 \quad \text{when } d \text{ is large } ;$$

(ii) the local reach of the modified Meyer staircase is almost the reciprocal of the magnitude of local curvature. These two properties indicate that the curve traverses the space with weak self-entanglement. This suggests the possibility of finding a linear projection $P : \mathbb{R}^d \to \mathbb{R}^{d'}$ with d' much smaller than d such that the projected image $P\gamma$ is a much simpler curve. For example, suppose we project the modified Meyer staircase linearly onto its first

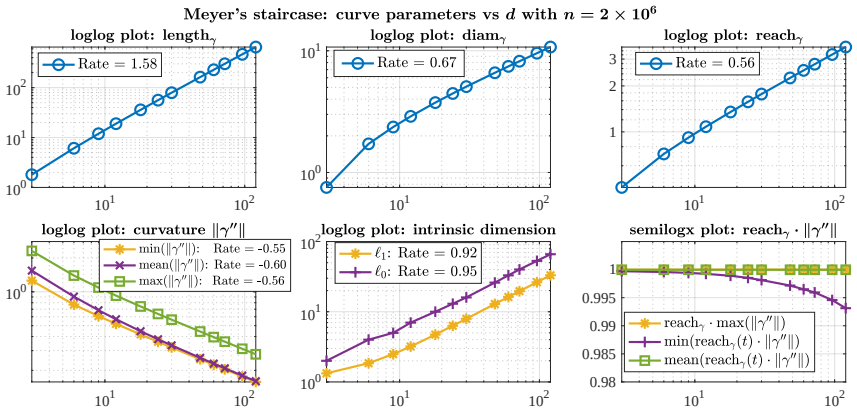

Figure 12: Behavior of geometric features of the modified Meyer staircase in $\mathbb{R}^d$ as a function of d.

few principal components: the projected image is a simple curve, and the learning problem can be significantly simplified if we study the regression problem on the projected curve. Hence, to test the performance of the regression algorithm, we aim to test curves that are so complex that there is no trivial dimension reduction, e.g., via standard techniques such as principal component analysis.

C.0.2 Meyer helix

For the numerical experiments, we seek sufficiently complex curves and use the following indicators of complexity: (i) geometric quantities—length, diameter, reach, and the effective linear dimensions $\|\lambda\|_1$ and $\|\lambda\|_0$—that grow with d; (ii) the curve γ admits no trivial dimension reduction. In particular, consider a linear projection $P_{d'} : \mathbb{R}^d \to \mathbb{R}^{d'}$, such as projection onto the first $d' \leqslant d$ principal components. Define the "regression complexity"

$$C_\gamma := \frac{\text{len}_\gamma}{\text{reach}_\gamma} \tag{26}$$

of a curve as its length divided by its reach. We seek curves sufficiently complex that, whenever the projected curve $P_{d'}\gamma$ has *regression complexity* $C_{P_{d'}\gamma} \lesssim C_\gamma$, the target dimension cannot be small; for example, $d' \gtrsim d$.

Because the regression problem is scale invariant, we can freely rescale the curve, and we choose the normalization such that the reach equals $\sqrt{d}$. This is consistent with Assumption $(\boldsymbol{\gamma_1})$ and allows us to take σ_γ, the deviation of data X away from the curve, of order one. In particular, we do not let curvature grow with d here.

We introduce the following curve, called the Meyer helix, as an analogue of the standard and modified Meyer staircases:

$$x_k(t) = \frac{1}{\sqrt[4]{2\pi}\sqrt{\delta_d}} \cos\left(ak + \frac{t - k/d}{\delta'_d}\right) G\left(\frac{|k/d - t|}{\delta_d}\right) \quad , \tag{27}$$

where $\delta_d = (1 + 0.3\cos(ak))/d$, $\delta'_d = (1 + 0.3\sin(ak))/d$, $a = 10$, and G is chosen to have the Bernstein-type decay $G(z) = \exp\left(-\frac{z^2}{1+z}\right)$. Compared with the standard and modified

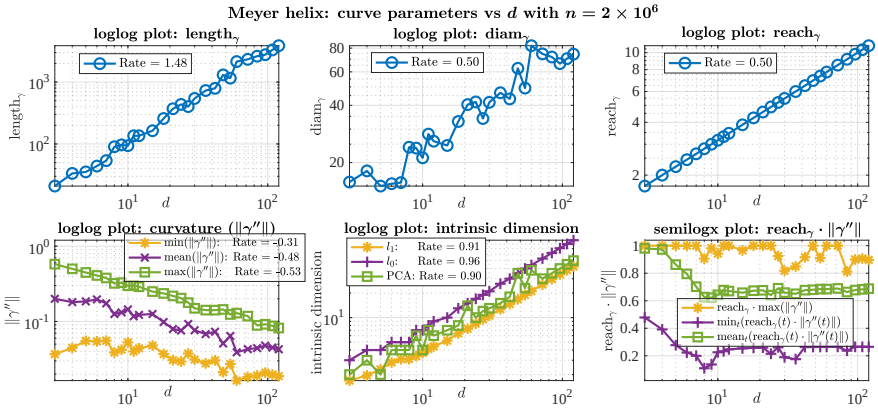

Figure 13: Behavior of geometric features of the Meyer helix in $\mathbb{R}^d$ as a function of d.

Meyer staircases, this curve contains an additional cosine factor that makes $x(t)$ traverse $\mathbb{R}^d$ more broadly and introduces greater self-entanglement. Moreover, the following quantities vary across coordinates: the frequency $1/\delta_d$ in function G, the frequency $1/\delta'_d$ in the cosine term, and the phase ak in the cosine term. These variations make the curve less structured while preserving the desired complexity.

In Figure 13, we plot the geometric parameters of the Meyer helix after scaling its reach to $\sqrt{d}$. As for the modified Meyer staircase, the length is roughly proportional to $d^{1.5}$, the diameter is roughly proportional to $d^{0.5}$, $\|\gamma''\|_\infty$ is roughly proportional to $d^{-1/2}$, the reach is roughly proportional to $d^{1/2}$, and the effective linear dimensions $\|\lambda_\gamma\|_1$ and $\|\lambda_\gamma\|_0$ are roughly proportional to d. Moreover, we can adopt the regression complexity (26) to measure the effective linear dimension: define d_{SVD} to be the smallest d' such that $C_{P_{d'}\gamma} \leqslant 1.2C_\gamma$, then this effective linear dimension is roughly proportional to d^1.

These complexity properties are robust: we can also choose Gaussian-type decay $G(z) = \exp(-z^2)$ or choose different constants in the construction (e.g., a, δ_d, and δ'_d) and these properties persist. This indicates that the complexity of this collection of Meyer helices grows with the ambient dimension d. Numerical tests suggest that this collection of curves does not admit a simple random-projection-based dimension reduction that preserves geometric properties such as length and reach. Recall the following form of the Johnson–Lindenstrauss random-projection lemma for manifolds:

Lemma 26 ((Baraniuk and Wakin, 2009, Theorem 3.1)) *Let $\mathcal{M}$ be a compact K-dimensional submanifold of $\mathbb{R}^N$ having condition number $1/\tau$, volume V, and geodesic covering regularity R. Fix $0 < \epsilon < 1$ and $0 < \rho < 1$. Let Φ be a random orthoprojector from $\mathbb{R}^N$ to $\mathbb{R}^M$ with $M \gtrsim \epsilon^{-2} K \log(NVR\tau^{-1}\epsilon^{-1}) \log(1/\rho)$. If $M \leqslant N$, then with probability at least $1-\rho$, the following statement holds: for every pair of points $x, y \in \mathcal{M}$,*

$$(1-\epsilon)\sqrt{\frac{M}{N}} \leqslant \frac{\|\Phi x - \Phi y\|_2}{\|x-y\|_2} \leqslant (1+\epsilon)\sqrt{\frac{M}{N}} \quad .$$

Applying this Johnson–Lindenstrauss lemma to a curve γ on $\mathbb{R}^d$ with length len_γ and reach reach_γ, then we can take ambient dimension $N = d$, the condition number $1/\tau =$

$1/\text{reach}_\gamma$, volume $V = \text{len}_\gamma$, and geodesic covering regularity $R = \mathcal{O}(1)$, and thus $M = \mathcal{O}\left(\epsilon^{-2}\log(1/\rho)\log\left(\frac{d}{\epsilon}\frac{\text{len}_\gamma}{\text{reach}_\gamma}\right)\right)$, which suggests the possibility of dimension reduction for the curve through linear projection. In our setting, however, the more relevant question is whether the projection reduces geometric complexity, for example $\text{len}_\gamma/\text{reach}_\gamma$; furthermore, our samples are not distributed on the curve, but in a tube around the curve of radius as large as a fraction of the reach. We perform numerical tests in the setting of Figure 1. For the Meyer helix in dimension $d = 36$, we project onto a 12-dimensional subspace using PCA or a random projection and compute the length and reach of the image. We consider ten independent repetitions and, for comparison, we also include the original curve as well as the projection onto the first 12 principal components. To be consistent with the scaling in the Johnson-Lindenstrauss lemma, for the PCA and random projection, we rescale points on the projected image of the curve by the factor $\sqrt{36/12} = \sqrt{3}$.

projection P	original γ	PCA	1	2	3	4	5	6	7	8	9	10
$\text{len}_{P\gamma}$	731	1136	703	720	709	717	639	748	766	709	737	706
$\text{reach}_{P\gamma}$	6.0	5.0	2.2	4.4	2.4	3.9	2.9	4.4	2.0	4.3	2.5	2.3
$\frac{\text{len}_{P\gamma}}{\text{reach}_{P\gamma}}$	122	226	322	164	295	187	223	171	383	166	291	313

These numerical results support our argument that the Meyer helix cannot be easily embedded in a lower-dimensional space without significantly affecting its complexity, which involves pointwise curvature/reach that are beyond the scope of random projections.

References

Giacomo Alberti, Alessandro Ingrosso, and Radu Toader. Gradient flow of the population risk in deep learning of single-index models. *Journal of Statistical Physics*, 182(3):1–33, 2021.

Gennady I. Arkhipov, Vladimir N. Chubarikov, and Anatoly A. Karatsuba. *Trigonometric sums in number theory and analysis*, volume 39 of *De Gruyter Expositions in Mathematics*. Walter de Gruyter, Berlin, 2004. ISBN 3-11-016266-0. Translated from the 1987 Russian original.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021. URL `https://www.jmlr.org/papers/v22/20-1288.html`.

Richard G. Baraniuk and Michael B. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, 2009. doi: 10.1007/s10208-007-9011-z. URL `https://doi.org/10.1007/s10208-007-9011-z`.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993. doi: 10.1109/18.256500.

Benedikt Bauer, Luc Devroye, Michael Kohler, Adam Krzyżak, and Harro Walk. Nonparametric estimation of a function from noiseless observations at random points. *Journal of Multivariate Analysis*, 160:93–104, 2017. ISSN 0047-259X. doi: 10.1016/j.jmva.2017.05.010. URL https://doi.org/10.1016/j.jmva.2017.05.010.

Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, New York, 1997. doi: 10.1007/978-1-4612-0653-8. URL https://doi.org/10.1007/978-1-4612-0653-8.

Peter J. Bickel and Bo Li. Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series*, 54:177–186, 2007. ISSN 07492170. URL http://www.jstor.org/stable/20461468.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *Advances in Neural Information Processing Systems*, 35:9768–9783, 2022.

Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow part i: General properties and two-timescale learning. *Communications on Pure and Applied Mathematics*, 78(12):2354–2435, 2025. doi: 10.1002/cpa.70006.

P. Binev, A. Cohen, W. Dahmen, R.A. DeVore, and V. Temlyakov. Universal algorithms for learning theory part i: piecewise constant functions. *J. Mach. Learn. Res.*, 6:1297–1321, 2005.

P. Binev, A. Cohen, W. Dahmen, R.A. DeVore, and V. Temlyakov. Universal algorithms for learning theory part ii: piecewise polynomial functions. *Constr. Approx.*, 26(2):127–152, 2007.

Peter Binev, Andrea Bonito, Ronald DeVore, and Guergana Petrova. Optimal learning. *Calcolo*, 61(1):15, 2024. doi: 10.1007/s10092-023-00564-y. URL https://doi.org/10.1007/s10092-023-00564-y.

Luca Brandolini, Giacomo Gigante, Allan Greenleaf, Alexander Iosevich, Andreas Seeger, and Giancarlo Travaglini. Average decay estimates for fourier transforms of measures supported on curves. *The Journal of Geometric Analysis*, 17(1):15–40, 2007. doi: 10.1007/BF02922080.

Leo Breiman and Jerome H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985. doi: 10.1080/01621459.1985.10478157. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1985.10478157.

Raymond J Carroll, Jianqing Fan, Irène Gijbels, and Matt P Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997.

Ronald R. Coifman, Ioannis G. Kevrekidis, Stéphane Lafon, Mauro Maggioni, and Boaz Nadler. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Modeling & Simulation*, 7(2):842–864, 2008. doi: 10.1137/070696325.

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. doi: 10.1080/03610920008832598. URL https://doi.org/10.1080/03610920008832598.

Raphaël Coudret, Benoit Liquet, and Jérôme Saracco. Comparison of sliced inverse regression approaches for underdetermined cases. *Journal de la Société Française de Statistique*, 155(2):72–96, 2014. URL https://www.numdam.org/item/JSFS_2014__155_2_72_0/.

Alex Damian, Loucas Pillaud-Vivien, Jason Lee, and Joan Bruna. Computational-statistical gaps in gaussian single-index models (extended abstract). In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1262–1262. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/damian24a.html.

Michel Delecroix and Marian Hristache. M-estimateurs semi-paramétriques dans les modèles à direction révélatrice unique. *Bulletin of the Belgian Mathematical Society Simon Stevin*, 6(2):161–186, 1999.

Michel Delecroix, Wolfgang Härdle, and Marian Hristache. Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis*, 86(2):213–226, 2003. doi: 10.1016/S0047-259X(02)00046-5. URL https://doi.org/10.1016/S0047-259X(02)00046-5.

Michel Delecroix, Marian Hristache, and Valentin Patilea. On semiparametric m-estimation in single-index regression. *Journal of Statistical Planning and Inference*, 136(3):730–769, 2006. doi: 10.1016/j.jspi.2004.09.006. URL https://doi.org/10.1016/j.jspi.2004.09.006.

Naihua Duan and Ker-Chau Li. Slicing regression: A link-free regression method. *The Annals of Statistics*, 19(2):505–530, 1991. ISSN 00905364. URL http://www.jstor.org/stable/2242072.

Weinan E and Eric Vanden-Eijnden. Towards a theory of transition paths. *Journal of Statistical Physics*, 123:503–523, 2006.

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

Mark I. Freidlin and Alexander D. Wentzell. *Random Perturbations of Dynamical Systems*, volume 260 of *Grundlehren der mathematischen Wissenschaften*. Springer, New York, second edition, 1998. doi: 10.1007/978-1-4612-0611-8. URL https://doi.org/10.1007/978-1-4612-0611-8. Translated from the 1979 Russian original by Joseph Szücs.

Stéphane Gaïffas and Guillaume Lecué. Optimal rates and adaptation in the single-index model using aggregation. *Electronic Journal of Statistics*, 1:538–573, 2007. doi: 10.1214/07-EJS077. URL https://doi.org/10.1214/07-EJS077.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002. doi: 10.1007/b97848. URL https://doi.org/10.1007/b97848.

Wolfgang Härdle and Thomas M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408): 986–995, 1989. ISSN 0162-1459. URL https://www.jstor.org/stable/2290074.

Wolfgang Härdle, Peter Hall, and Hidehiko Ichimura. Optimal smoothing in single-index models. *The Annals of Statistics*, 21(1):157–178, 1993. ISSN 0090-5364. URL https://www.jstor.org/stable/3035585.

Trevor J. Hastie and Robert J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990. ISBN 9780412343902.

Joel L. Horowitz. *Semiparametric Methods in Econometrics*. Springer, New York, 1998. doi: 10.1007/978-1-4612-0621-7. URL https://doi.org/10.1007/978-1-4612-0621-7.

Joel L. Horowitz and Enno Mammen. Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *The Annals of Statistics*, 35 (6):2589–2619, 2007.

Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *The Annals of Statistics*, 29(3):595–623, 2001. ISSN 00905364. URL http://www.jstor.org/stable/2673964.

Shuo Huang, Hippolyte Labarrière, Ernesto De Vito, Tomaso Poggio, and Lorenzo Rosasco. Learning multi-index models with hyper-kernel ridge regression. *arXiv preprint arXiv:2510.02532*, 2025. URL https://arxiv.org/abs/2510.02532.

Hidehiko Ichimura. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58(1–2):71–120, 1993. ISSN 0304-4076. doi: 10.1016/0304-4076(93)90114-K. URL https://doi.org/10.1016/0304-4076(93)90114-K.

Anatoli B. Juditsky, Oleg V. Lepski, and Alexandre B. Tsybakov. Nonparametric estimation of composite functions. *The Annals of Statistics*, 37(3):1360–1404, 2009. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/30243671.

Željko Kereta, Timo Klock, and Valeriya Naumova. Nonlinear Generalization of the Monotone Single Index Model. *Information and Inference: A Journal of the IMA*, 10(3): 987–1029, 2021. doi: 10.1093/imaiai/iaaa013. URL https://doi.org/10.1093/imaiai/iaaa013.

Samory Kpotufe. k-NN regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems 24*, pages 729–737. Curran Associates, Inc., 2011. ISBN 9781618395993. URL https://proceedings.neurips.cc/paper/2011/hash/05f971b5ec196b8c65b75d2ef8267331-Abstract.html.

Samory Kpotufe and Vikas Garg. Adaptivity to local smoothness and dimension in kernel regression. In *Advances in Neural Information Processing Systems 26*, pages 3075–3083. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/hash/28fc2782ea7ef51c1104ccf7b9bea13d-Abstract.html.

Ming-Jun Lai and Zhaiming Shen. The kolmogorov superposition theorem can break the curse of dimensionality when approximating high dimensional functions, 2021. URL https://arxiv.org/abs/2112.09963.

Alessandro Lanteri, Željko Kereta, Timo Klock, Mauro Maggioni, Valeriya Naumova, and Stefano Vigogna. Thinking geometrically in regression with structured models: Linear and nonlinear single-index models. Unpublished manuscript, 2018.

Alessandro Lanteri, Mauro Maggioni, and Stefano Vigogna. Conditional regression for single-index models. *Bernoulli*, 28(4):3051–3078, 2022. doi: 10.3150/22-BEJ1482. URL https://doi.org/10.3150/22-BEJ1482.

Jason D. Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with SGD near the information-theoretic limit. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 58716–58756. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6bd5fca2074dcd9ede9de50f71f7ec28-Paper-Conference.pdf.

Bing Li and Shaoli Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007. ISSN 01621459. URL http://www.jstor.org/stable/27639941.

Bing Li, Hongyuan Zha, and Francesca Chiaromonte. Contour regression: A general approach to dimension reduction. *The Annals of Statistics*, 33(4):1580–1616, 2005. ISSN 00905364. URL http://www.jstor.org/stable/3448618.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. ISSN 01621459. URL http://www.jstor.org/stable/2290563.

Wenjing Liao, Mauro Maggioni, and Stefano Vigogna. Learning adaptive multiscale approximations to data and functions near low-dimensional sets. In *2016 IEEE Information Theory Workshop (ITW)*, pages 226–230, Cambridge, United Kingdom, 2016. IEEE. doi: 10.1109/ITW.2016.7606829. URL https://doi.org/10.1109/ITW.2016.7606829.

Wenjing Liao, Mauro Maggioni, and Stefano Vigogna. Multiscale regression on unknown manifolds. *Mathematics in Engineering*, 4(4):1–25, 2022. ISSN 2640-3501. doi: 10.3934/mine.2022028. URL https://doi.org/10.3934/mine.2022028.

Hao Liu and Wenjing Liao. Learning functions varying along a central subspace. *SIAM Journal on Mathematics of Data Science*, 6(2):343–371, 2024. doi: 10.1137/23M1557751. URL https://doi.org/10.1137/23M1557751.

Hao Liu, Jiahui Cheng, and Wenjing Liao. Deep neural networks are adaptive to function regularity and data distribution in approximation and estimation. *Journal of Machine Learning Research*, 26(213):1–56, 2025a. URL https://www.jmlr.org/papers/v26/24-1148.html.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov–arnold networks. In *International Conference on Learning Representations (ICLR)*, 2025b. URL https://openreview.net/forum?id=Ozo7qJ5vZi.

Philipp Metzner, Christof Schütte, and Eric Vanden-Eijnden. Illustration of transition path theory on a collection of simple examples. *The Journal of Chemical Physics*, 125 (8):084110, 2006.

Philipp Metzner, Christof Schütte, and Eric Vanden-Eijnden. Transition path theory for markov jump processes. *Multiscale Modeling & Simulation*, 7(3):1192–1219, 2009.

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science*, 383(6690):1461–1467, 2024. doi: 10.1126/science.adi5639. URL https://www.science.org/doi/abs/10.1126/science.adi5639.

Mary A. Rohrdanz, Wenwei Zheng, Mauro Maggioni, and Cecilia Clementi. Determination of reaction coordinates via locally scaled diffusion map. *The Journal of Chemical Physics*, 134(12):124116, 2011. doi: 10.1063/1.3569857. URL https://doi.org/10.1063/1.3569857.

Johannes Schmidt-Hieber. Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020. doi: 10.1214/19-AOS1875. URL https://doi.org/10.1214/19-AOS1875.

Ohad Shamir. Discussion of: "nonparametric regression using deep neural networks with ReLU activation function". *The Annals of Statistics*, 48(4):1911–1915, 2020. URL https://doi.org/10.1214/19-AOS1915.

Guohao Shen, Yuling Jiao, Yuanyuan Lin, Joel L. Horowitz, and Jian Huang. Deep quantile regression: Mitigating the curse of dimensionality through composition. *arXiv preprint arXiv:2107.04907*, 2021. URL https://arxiv.org/abs/2107.04907.

Zhaiming Shen, Alex Havrilla, Rongjie Lai, Alexander Cloninger, and Wenjing Liao. Transformers for learning on noisy and task-level manifolds: Approximation and generalization insights. *arXiv preprint arXiv:2505.03205*, 2025. URL https://arxiv.org/abs/2505.03205.

G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Academic Press, Boston, 1990. ISBN 978-0-12-670230-9.

Thomas M. Stoker. Consistent estimation of scaled coefficients. *Econometrica*, 54(6):1461–1481, 1986. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1914309.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982. ISSN 00905364. URL http://www.jstor.org/stable/2240707.

Jan-Olov Strömberg. Computation with wavelets in higher dimensions. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 523–532, 1998.

Taiji Suzuki and Atsushi Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. In *Advances in Neural Information Processing Systems*, volume 34, pages 3609–3621, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/1dacb10f0623c67cb7dbb37587d8b38a-Abstract.html.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2018. doi: 10.1017/9781108231596. URL https://doi.org/10.1017/9781108231596.

Zihao Wang, Eshaan Nichani, and Jason D. Lee. Learning hierarchical polynomials with three-layer neural networks. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=QgwAYFrh9t.

Yingcun Xia. Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, 22(6):1112–1137, 2006. ISSN 02664666, 14694360. URL http://www.jstor.org/stable/4093215.

Wenwei Zheng, Mary A. Rohrdanz, Mauro Maggioni, and Cecilia Clementi. Polymer reversal rate calculated via locally scaled diffusion map. *The Journal of Chemical Physics*, 134(14):144109, 2011. doi: 10.1063/1.3575245. URL https://doi.org/10.1063/1.3575245.