Received: 2 March 2021 | Accepted: 8 September 2021

DOI: 10.1111/rssb.12481

ROYAL STATISTICAL SOCIETY
Series B Statistical Methodology

ORIGINAL ARTICLE

Nonparametric, tuning-free estimation of S-shaped functions

Oliver Y. Feng[1] | **Yining Chen**[2] | **Qiyang Han**[3] | **Raymond J. Carroll**[4,5] | **Richard J. Samworth**[1]

[1]Statistical Laboratory, University of Cambridge, Cambridge, UK

[2]Department of Statistics, London School of Economics and Political Science, London, UK

[3]Department of Statistics, Rutgers University, Piscataway, New Jersey, USA

[4]Department of Statistics, Texas A&M University, College Station, Texas, USA

[5]School of Mathematical and Physical Sciences, University of Technology Sydney, Sydney, Australia

Correspondence
Oliver Y. Feng, Statistical Laboratory, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, CB3 0WB, UK.
Email: o.feng@statslab.cam.ac.uk

Funding information
NSF, Grant/Award Number: DMS-1916221 and CCF-1934904; National Cancer Institute, Grant/Award Number: U01-CA057030; EPSRC, Grant/Award Number: EP/P031447/1 and EP/N031938

Abstract

We consider the nonparametric estimation of an S-shaped regression function. The least squares estimator provides a very natural, tuning-free approach, but results in a non-convex optimization problem, since the inflection point is unknown. We show that the estimator may nevertheless be regarded as a projection onto a finite union of convex cones, which allows us to propose a mixed primal-dual bases algorithm for its efficient, sequential computation. After developing a projection framework that demonstrates the consistency and robustness to misspecification of the estimator, our main theoretical results provide sharp oracle inequalities that yield worst-case and adaptive risk bounds for the estimation of the regression function, as well as a rate of convergence for the estimation of the inflection point. These results reveal not only that the estimator achieves the minimax optimal rate of convergence for both the estimation of the regression function and its inflection point (up to a logarithmic factor in the latter case), but also that it is able to achieve an almost-parametric rate when the true regression function is piecewise affine with not too many affine pieces. Simulations and a real data application to air pollution modelling also confirm

the desirable finite-sample properties of the estimator, and our algorithm is implemented in the `R` package `Sshaped`.

KEYWORDS

sequential algorithm, shape-constrained regression, S-shaped functions

1 | INTRODUCTION

We define a function $f : [0, 1] \to \mathbb{R}$ to be *S-shaped* if it is increasing, and if there exists $m_0 \in [0, 1]$ such that f is convex on $[0, m_0]$ and concave on $[m_0, 1]$. The point m_0 is called an *inflection point*, and we do not insist that f is continuous at m_0; the cases $m_0 = 0$ and $m_0 = 1$ correspond to increasing concave and increasing convex functions respectively. Various examples of S-shaped functions are shown in Figure 1. In many areas of applied science, there are domain-specific reasons to model the regression of a response variable on a covariate as an S-shaped function. For instance, development curves for individuals or populations often exhibit S-shaped behaviour in the context of biological growth (Archontoulis & Miguez, 2015; Cao et al., 2019; Zeidi, 1993) or skill proficiency (Gibbs, 2000). Further examples where time is the covariate can be found in audio signal processing (Smith, 2010) and sociology (Tarde, 1903). In agronomy, the van Genuchten–Gupta model (van Genuchten & Gupta, 1993) postulates an inverted S-shaped relationship between crop yield and soil salinity, and S-shaped trends are also observed for the production levels of commercial goods as labour or other resources are scaled up (Ginsberg, 1974). For the latter, economic principles such as the Regular Ultra Passum law (Frisch, 1964) have been formulated to describe scenarios where marginal gains (i.e. returns to scale) increase up to a point of maximal productivity and then taper off.

In some of the examples above, for instance when population or disease dynamics can be modelled by some governing differential equation, it may be natural to confine attention to certain parametric subclasses of S-shaped functions, such as those consisting of sigmoidal (i.e. logistic) functions of the form

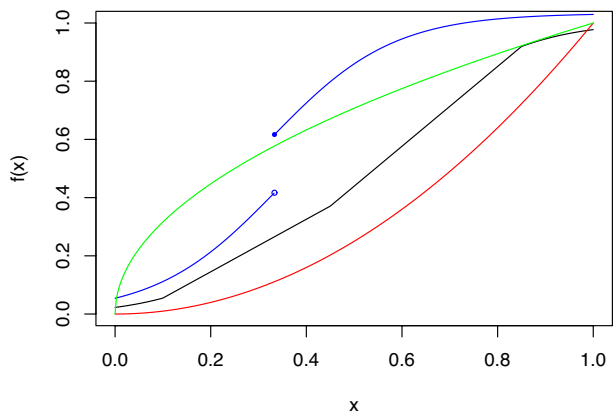

FIGURE 1 Some examples of S-shaped functions on $[0, 1]$

$$f(x; A, a, b) = \frac{A}{1 + e^{-ax+b}}, \tag{1}$$

with A, $a > 0$ and $b \in \mathbb{R}$; see also Jarne et al. (2007). However, in many other settings, such domain-specific knowledge is often lacking, and parametric assumptions may be excessively restrictive. To illustrate this effect, see Figure 2, where we compare two popular parametric fits of an S-shaped regression function with the estimator we propose in this paper. The first parametric method fits a logistic curve of the form (1) using nonlinear least squares. The second uses segmented linear regression with two kinks, fitted using least squares and a search over the locations of the kinks. Although these parametric fits appear to the naked eye to be satisfactory, it turns out that their estimation performance, as measured by the squared error loss on the training data, is roughly six times worse than that of our proposal (on average 0.38 and 0.43 compared with 0.067, over 100 repetitions). If the noise standard deviation is halved, then these parametric methods become 17 times and 19 times worse than our proposal respectively. Notice also that our S-shaped estimator is sufficiently flexible to be able to capture the discontinuity of the regression function, whereas the parametric methods struggle in this respect. The benefits of our nonparametric approach are also apparent in the analysis of real data: see Section 5.3, where we study the way that a quantity related to atmospheric mercury concentration varies with distance from an experimental device close to a geothermal power station.

Motivated by the limitations described in the previous paragraph, the goal of this paper is to introduce a flexible framework for nonparametric estimation of S-shaped functions. The

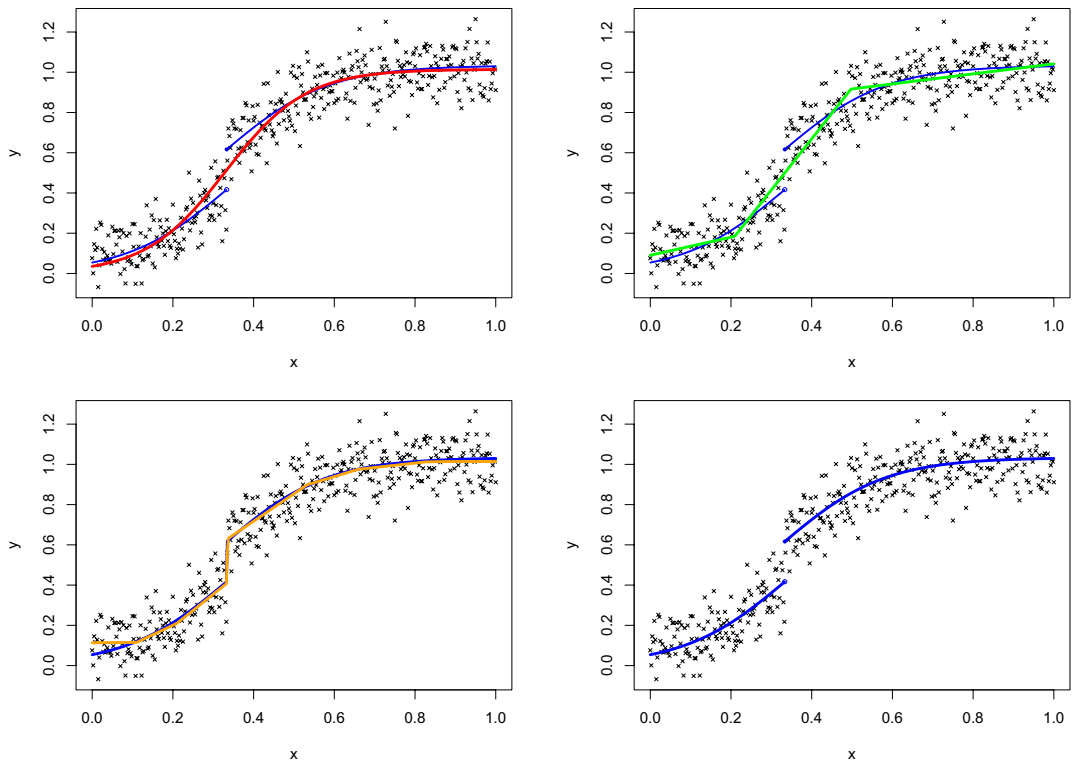

FIGURE 2 Logistic (red, top left), segmented linear regression (green, top right) and S-shaped (orange, bottom left) estimates of the true regression function $x \mapsto \frac{5}{6}(1+e^{-8(x-1/3)})^{-1} + \frac{1}{5}\mathbb{1}_{\{x>1/3\}}$ (blue, all plots)

main challenges in removing the parametric restrictions are two-fold: first, the class $\mathcal{F}$ of S-shaped functions on [0, 1] is infinite-dimensional; and second, since the inflection point is unknown, the family $\mathcal{F}$ is non-convex. Despite this non-convexity, we are able to develop methodology based on suitably defined L^2-'projections' of general distributions onto $\mathcal{F}$. The significant advantage of working in this additional generality is that, having established continuity properties of the projection, results on the consistency and robustness under misspecification of the estimator follow as simple corollaries of basic facts about convergence of empirical distributions. Nevertheless, since the fully general statements are fairly involved, we defer this formal presentation to Section S3 of the supplementary material (Feng et al., 2021b), and focus in Section 2 on the special case of projections of the empirical distribution of data of the form $(x_1, Y_1), \ldots, (x_n, Y_n) \in [0, 1] \times \mathbb{R}$ with $x_1 < \cdots < x_n$. This allows us to prove that an S-shaped least squares estimator always exists, and to study its uniqueness properties. Moreover, when the design is fixed and the errors are independent and identically distributed with mean zero and finite variance, we present a basic consistency result that follows from the general theory in Section S3.

In Section 3, we take up the challenge of computing the S-shaped least squares estimator. Since its inflection point occurs at one of the design points, a naive strategy would be to fit, for each choice of $m \in \{x_1, \ldots, x_n\}$, the least squares estimate over the class of S-shaped functions with inflection point m, before selecting a solution that minimizes the residual sum of squares. The individual constrained estimates are straightforward to compute using, for example, active set methods (Dümbgen et al., 2007; Nocedal & Wright, 2006, Chapters 12 and 16.5), but it can be time-consuming to run the active set method n times. We show how a simple refinement of the search strategy can improve the running time by a factor of around 4, but our major contribution here begins with the observation that the global S-shaped least squares estimate can be obtained as a concatenation of a convex increasing least squares estimate to the left of an estimated inflection point, with a concave increasing least squares estimate to the right. This enables us to pursue a sequential approach, where we reveal new observations one by one, and update the least squares fits using a mixed primal-dual bases algorithm (Fraser & Massam, 1989; Meyer, 1999). Our algorithm, which is available in the `R` package `Sshaped` (Feng et al., 2021a), is shown to be around 40 times faster than the naive strategy in examples; see Figure 5.

Our main theoretical contributions are presented in Section 4, under an independent and sub-Gaussian error assumption. Here, we derive worst-case and adaptive sharp oracle inequalities for the S-shaped least squares estimator. When combined with our corresponding minimax lower bounds, this theory reveals in particular that the S-shaped least squares estimator attains the optimal worst-case risk of order $n^{-2/5}$ with respect to L^2-loss, in the case where the design points are not too irregularly spaced. These results apply both when the S-shaped regression function hypothesis is correctly specified, and where it is misspecified, provided in the latter case that we interpret the loss as the distance to the projection of the signal onto $\mathcal{F}$. For adversarially chosen design configurations, we show that the risk bound can deteriorate to $n^{-1/3}$ in the worst case. Moreover, the S-shaped least squares estimator adaptively attains the parametric rate of order $n^{-1/2}$ (up to a logarithmic factor), when the projection of the signal is piecewise affine with a relatively small number of affine pieces. Finally, we study the delicate problem of estimating the true inflection point m_0, which represents the boundary between the convex and concave parts of the signal. Under an appropriate local smoothness assumption indexed by a parameter $\alpha > 0$, we show that the inflection point $\hat{m}_n$ of the least squares estimator converges to m_0 at rate $O_p((n^{-1} \log n)^{1/(2\alpha+1)})$, which matches our local asymptotic minimax

lower bound, up to the logarithmic factor. Interestingly, the combination of the monotonicity with the convexity/concavity means that our S-shaped estimator is sufficiently regularized to avoid boundary problems at the endpoints $\{0, 1\}$ of the covariate domain; other common shape-constrained methods are known to lead to boundary estimation inconsistency (Balabdaoui et al., 2011; Balász et al., 2015; Cule et al., 2010; Han & Kato, 2021; Kulikov & Lopuhaä, 2006; Samworth, 2018).

In Section 5, we study the empirical properties of our S-shaped least squares estimator, comparing both its running time and statistical performance with those of alternative approaches on simulated data. We also present a real data application of these techniques in air pollution modelling, which highlights the convenience and efficacy of our proposal. We conclude by discussing some possible directions for future research in Section 6. The appendix (Section A) provides further details of the mixed primal-dual bases algorithm that we use to compute our estimator. The proofs of our main results are deferred to the supplementary material (Feng et al., 2021b), in which the results and sections appear with an 'S' before the relevant label number.

Previous work on nonparametric estimation of S-shaped functions includes Yagi et al. (2019, 2020), who, in the context of production theory in economics, apply a method known as shape-constrained kernel least squares to estimate multivariate production functions that are S-shaped along one-dimensional rays. Kachouie and Schwartzman (2013) use local polynomial regression techniques to identify an inflection point of a smooth signal from corrupted observations. In both of these works, kernel bandwidths must be chosen carefully to control the bias-variance tradeoff and (for the approach of Kachouie and Schwartzman (2013) in particular) to ensure that the fitted curve does not have multiple inflection points. Liao and Meyer (2017) instead estimate univariate convex-concave functions using cubic splines defined with respect to a number of user-specified knots, and establish rates of convergence for the inflection points of the resulting estimators. Their method is implemented in the `R` package `ShapeChange` (Liao & Meyer, 2016), which Lee et al. (2020) subsequently used in combination with the `scam` (Shape Constrained Additive Models) package of Pya and Wood (2015) to estimate S-shaped disease trajectories of patients with Huntington's disease. We also mention the extremum distance estimator and extremum surface estimator proposed by Christopoulos (2016), with the aim of locating the inflection point of a smooth function based on its geometric properties. We provide a numerical comparison of our procedure with those of Liao and Meyer (2017), Yagi et al. (2019, 2020) and Christopoulos (2016) in Section 5.2.

1.1 | Notation

For $n \in \mathbb{N}$, we write $[n] := \{1, ..., n\}$, and given $0 \leq x_1 < \cdots < x_n \leq 1$, define $\mathcal{G} \equiv \mathcal{G}[x_1, \ldots, x_n]$ to be the set of continuous, piecewise affine $f : [0, 1] \to \mathbb{R}$ with kinks in $\{x_2, \ldots, x_{n-1}\}$. If $\tilde{f}_n : [0, 1] \to \mathbb{R}$ minimizes[1] $f \mapsto \sum_{i=1}^n (Y_i - f(x_i))^2 =: S_n(f)$ over some class $\tilde{\mathcal{F}}$ of functions on $[0, 1]$, we say that $\tilde{f}_n$ is a *least squares estimator (LSE) over* $\tilde{\mathcal{F}}$ based on $\{(x_i, Y_i) : 1 \leq i \leq n\}$. We write $a_n \lesssim b_n$ to mean that there exists a universal constant $C > 0$ such that $a_n \leq C b_n$ for all n.

[1]Since there may be multiple minimizers, we will also assume throughout and without further comment that $\tilde{f}_n$ is chosen to depend measurably on $(x_1, Y_1), \ldots, (x_n, Y_n)$. Likewise, we will assume the same property for estimated inflection points.

2 | EXISTENCE, UNIQUENESS AND CONSISTENCY OF S-SHAPED LEAST SQUARES ESTIMATORS

The purpose of this section is to study the existence, uniqueness and consistency of S-shaped least squares estimators. We will see later that in a suitable sense, these estimators can be regarded as L^2-projections onto $\mathcal{F}$ of the empirical distribution of the data. As such, the results in this section turn out to be special cases of a much more general theory, presented in Section S3, concerning the existence and continuity of L^2-projections of arbitrary distributions on $[0, 1] \times \mathbb{R}$ having finite variance. The generality of this projection framework remains of importance to statisticians, particularly in terms of providing results on the robustness of S-shaped least squares estimators to model misspecification; however, the results are of a more technical nature, so to facilitate understanding of the main ideas, we focus on the well-specified case here.

Suppose we have observations $(x_1, Y_1), \ldots, (x_n, Y_n) \in [0, 1] \times \mathbb{R}$ with $x_1 < \cdots < x_n$. For each $m \in [0, 1]$, we denote by $\mathcal{F}^m$ the class of S-shaped functions with an inflection point at m, that is, the set of all $f : [0, 1] \to \mathbb{R}$ that are convex on $[0, m]$, concave on $[m, 1]$ and increasing (i.e. non-decreasing) on $[0, 1]$. Thus $\mathcal{F} := \bigcup_{m\in[0,1]} \mathcal{F}^m$ is the set of all S-shaped functions on $[0, 1]$, but this union of convex sets is not itself convex.

Proposition 1 *For each $m \in [0, 1]$, there exists an LSE $\tilde{f}_n^m$ over $\mathcal{F}^m$ that is uniquely determined at $x_1, \ldots, x_n$. Moreover, there exists an LSE $\tilde{f}_n$ over $\mathcal{F}$ with an inflection point in $\{x_2, \ldots, x_{n-1}\}$.*

A straightforward and direct proof of this result is given in Section S1. As part of the projection framework in Section S3, we obtain generalizations of Proposition 1 in Corollaries S10(d) and S14(a). Since our objective criterion only measures the error incurred at the design points, it is no surprise that any LSE $\tilde{f}_n^m$ over $\mathcal{F}^m$ can only be unique at $x_1, \ldots, x_n$. There is a canonical way to define $\tilde{f}_n^m$ on the whole of $[0, 1]$, namely by linear interpolation between its kinks. Thus, the slope remains constant on $[0, x_2]$, $[x_2, x_3]$, $\ldots$, $[x_{n-2}, x_{n-1}]$, $[x_{n-1}, 1]$, and we denote this interpolating function by $\widehat{f}_n^m \in \mathcal{G} \equiv \mathcal{G}[x_1, \ldots, x_n]$. A subtle issue, however, is that when m is not a design point, $\widehat{f}_n^m$ need not belong to $\mathcal{F}^m$; see the left panel of Figure 3. To finesse this point, for $m \in [0, 1]$, denote by $\mathcal{H}^m \equiv \mathcal{H}^m[x_1, \ldots, x_n]$ the class of all $f \in \mathcal{G}$ for which there exists $g \in \mathcal{F}^m$ with $f = g$ on

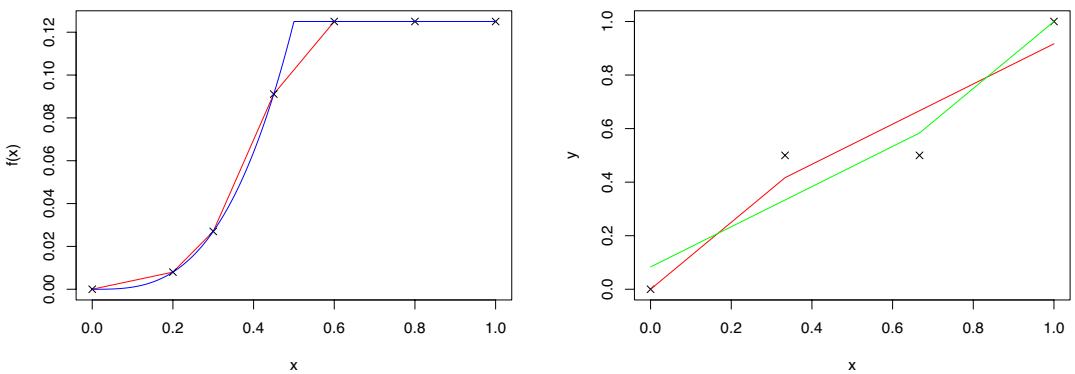

FIGURE 3 Left: For noiseless observations of the blue regression function at the black crosses, the red curve illustrates the linear interpolation $\widehat{f}_n^m$ of the least squares estimator (LSE), with $m = 0.5$; here, the segment of steepest slope does not contain $x = 0.5$, so $\widehat{f}_n^m$ does not belong to $\mathcal{F}^m$ with $m = 0.5$. Right: For the data given by the black crosses, both the red curve and the green curve are LSEs over $\mathcal{F}$

$\{x_1, \ldots, x_n\}$. Then $\mathcal{H}^m$ is a closed, convex cone, and the LSE over $\mathcal{H}^m$ based on $\{(x_i, Y_i) : 1 \le i \le n\}$ is precisely the function $\widehat{f}_n^m$. We refer to $\widehat{f}_n^0$ and $\widehat{f}_n^1$ as the *increasing concave* LSE and *increasing convex* LSE (based on $\{(x_i, Y_i) : 1 \le i \le n\}$) respectively.

It turns out, however, that in general an LSE $\tilde{f}_n$ over $\mathcal{F}$ is not even uniquely defined at the design points. For instance, if our data are (0, 0), (1/3, 1/2), (2/3, 1/2), (1, 1), then the linear interpolations of both (0, 0), (1/3, 5/12), (2/3, 2/3), (1, 11/12) and (0, 1/12), (1/3, 1/3), (2/3, 7/12), (1, 1) are LSEs over $\mathcal{F}$; see the right panel of Figure 3. We remark that this non-uniqueness is not related to the small number of data points, but rather to the symmetry of the data configuration.

In order to present a basic consistency result, we introduce a model where we regard our data $\{(x_1, Y_1), \ldots, (x_n, Y_n)\} \equiv \{(x_{n1}, Y_{n1}), \ldots, (x_{nn}, Y_{nn})\}$ as being realized from a triangular array sampling scheme

$$Y_{ni} = f_0(x_{ni}) + \xi_{ni}, \quad i = 1, \ldots, n, \tag{2}$$

where $f_0 : [0, 1] \to \mathbb{R}$ is a Borel measurable regression function, where $\xi_{n1}, \ldots, \xi_{nn}$ are independent noise variables with mean zero and finite variance for each n, and where $0 \le x_{n1} < \cdots < x_{nn} \le 1$ are fixed design points. We write $\mathbb{P}_n := n^{-1} \sum_{i=1}^n \delta_{(x_{ni}, Y_{ni})}$ and $\mathbb{P}_n^X := n^{-1} \sum_{i=1}^n \delta_{x_{ni}}$ for the joint and X-marginal empirical distributions respectively.

For a finite Borel measure ν on $[0, 1]$, we denote by $\operatorname{supp} \nu$ the *support* of ν, which is defined as the smallest closed set A such that $\nu(A^c) = 0$, or equivalently the set of all $x \in [0, 1]$ with the property that $\nu(U) > 0$ for any open neighbourhood U of x in $[0, 1]$.

Proposition 2 *In model (2), assume that $f_0 \in \mathcal{F}$ has unique inflection point $m_0 \in [0, 1]$ and that $\xi_{n1}, \ldots, \xi_{nn}$ are independent and identically distributed for each n. For each $n \in \mathbb{N}$, let $\widehat{f}_n^{m_0}$ and $\tilde{f}_n$ denote LSEs over $\mathcal{F}^{m_0}$ and $\mathcal{F}$ respectively. Suppose further that $(\mathbb{P}_n^X)$ converges weakly to a distribution P_0^X on $[0, 1]$ satisfying $\operatorname{supp} P_0^X = [0, 1]$ and $P_0^X(\{m\}) = 0$ for all $m \in [0, 1]$. Then, for $\tilde{g}_n \in \{\widehat{f}_n^{m_0}, \tilde{f}_n\}$ and with $\tilde{m}_n$ denoting any inflection point of $\tilde{g}_n$, we have*

a. $\tilde{m}_n \xrightarrow{p} m_0$;
b. $\sup_{x \in A} |(\tilde{g}_n - f_0)(x)| \xrightarrow{p} 0$ for any closed set $A \subseteq [0, 1] \setminus \{m_0\}$;
c. If $m_0 \in (0, 1)$, then $\int_0^1 |\tilde{g}_n - f_0|^q \, dP_0^X \xrightarrow{p} 0$ for all $q \in [1, \infty)$;
d. If $m_0 \in (0, 1)$ and in addition f_0 is continuous at m_0, then $\sup_{x \in [0,1]} |(\tilde{g}_n - f_0)(x)| \xrightarrow{p} 0$.

Proposition 2 follows from Proposition S16 in Section S3, which handles the more general case where f_0 need not belong to $\mathcal{F}$, and where it may have multiple inflection points. A proof of the latter result is given in Section S6.

3 | COMPUTATION OF S-SHAPED LEAST SQUARES ESTIMATORS

Returning to the setting of data $(x_1, Y_1), \ldots, (x_n, Y_n) \in [0, 1] \times \mathbb{R}$ with $x_1 < \cdots < x_n$, we now consider the problem of computing an S-shaped LSE over $\mathcal{F}$. In light of the non-uniqueness discussion in Section 2, we will take as our target the LSE $\widehat{f}_n := \widehat{f}_n^{\widehat{m}_n}$, where $\widehat{m}_n := x_{\widehat{\jmath}_n}$ and $\widehat{\jmath}_n := \operatorname{sargmin}_{1 \le j \le n} S_n(\widehat{f}_n^{x_j})$; here and below, sargmin denotes the smallest element of the argmin. One of the main challenges here is that in general the function $j \mapsto S_n(\widehat{f}_n^{x_j})$ has multiple local

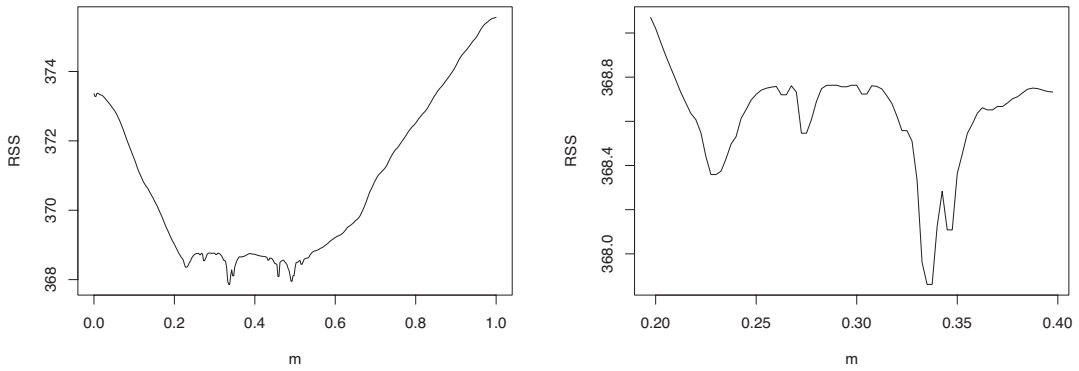

FIGURE 4 Plots of the residual sum of squares $S_n(\widehat{f}_n^m)$ of the least squares estimator with inflection point at m over $m \in [0, 1]$ (left) and $m \in [0.2, 0.4]$ (right), illustrating the multiple local minima of this function. Here, with $n = 400$, the data were generated according to $Y_i = f(x_i) + \xi_i$ for $i = 1, ..., n$, with f taken to be the blue regression function from Figure 1, $x_i = i/n$ for $i = 1,..., n$ and $\xi_1,..., \xi_n$ independent $N(0, 1)$ random errors

minima; see Figure 4. A 'brute-force' method that we call `ScanAll`, then, is to compute each of the LSEs $\widehat{f}_n^{x_1}, \ldots, \widehat{f}_n^{x_n}$ directly by solving n separate constrained least squares problems. In each instance, we can run the support reduction algorithm (Groeneboom et al., 2008) or a generic active set algorithm (Dümbgen et al., 2007; Nocedal & Wright, 2006, Chapters 12 and 16.5) on the whole dataset $\{(x_i, Y_i) : 1 \le i \le n\}$, but it is computationally expensive to repeat this n times, even when n is only moderately large; see Section 5.1.

To improve the overall efficiency of this procedure, it would therefore be desirable to both refine the initial search strategy as well as exploit any common structure underlying the individual minimization problems. For instance, we might hope to be able to obtain $\widehat{f}_n^{x_j}$ via a faster update step that takes as input the previous LSE $\widehat{f}_n^{x_{j-1}}$, but it is not immediately clear how this can be done.

We now describe and justify an alternative approach that achieves both of the above objectives. For $j \in [n]$, we write $\widehat{f}_{1,j} \in \mathcal{G}[x_1, \ldots, x_j]$ for the increasing convex LSE based on $\{(x_i, Y_i) : 1 \le i \le j\}$ and $\widehat{f}_{n,j} \in \mathcal{G}[x_j, \ldots, x_n]$ for the increasing concave LSE based on $\{(x_i, Y_i) : j \le i \le n\}$, recalling from, for example, Ghosal and Sen (2017, Lemma 2.2) that

$$\widehat{f}_{1,j}(x_j) \ge Y_j \ge \widehat{f}_{n,j}(x_j) \quad \text{for all } j \in [n]. \tag{3}$$

We then define $\widehat{h}_n^j \in \mathcal{G}[x_1, \ldots, x_n]$ for $j \in [n-1]$ by

$$\widehat{h}_n^j(x_i) := \begin{cases} \widehat{f}_{1,j}(x_i) & \text{for } i \in \{1, \ldots, j\} \\ \widehat{f}_{n,j+1}(x_i) & \text{for } i \in \{j+1, \ldots, n\}. \end{cases} \tag{4}$$

In other words, $\widehat{h}_n^j$ is obtained by partitioning the data into two disjoint subsets $\{(x_1, Y_1), \ldots, (x_j, Y_j)\}$ and $\{(x_{j+1}, Y_{j+1}), \ldots, (x_n, Y_n)\}$, and then fitting separate increasing convex and increasing concave LSEs on the left and right pieces respectively. In general, $\widehat{h}_n^j$ is not guaranteed to be S-shaped or even increasing on $[0, 1]$, in which case $\widehat{h}_n^j$ does not coincide with the LSE $\widehat{f}_n^{x_j}$ over $\mathcal{H}^{x_j} \equiv \mathcal{H}^{x_j}[x_1, \ldots, x_n] = \mathcal{F}^{x_j} \cap \mathcal{G}[x_1, \ldots, x_n]$. Nevertheless, observe that $\widehat{h}_n^j$ is the LSE over a larger subclass of $\mathcal{G}[x_1, \ldots, x_n]$ that contains $\mathcal{H}^{x_j}$. Together with Equation (3), this immediately implies Proposition 3 below, a key fact that we will exploit in our algorithm.

Proposition 3 *For $j \in [n-1]$, we have $\widehat{h}_n^j = \widehat{f}_n^{x_j}$ if and only if $\widehat{h}_n^j \in \mathcal{H}^{x_j}$, that is, if and only if*

$$\frac{\widehat{f}_{n,j+1}(x_{j+2}) - \widehat{f}_{n,j+1}(x_{j+1})}{x_{j+2} - x_{j+1}} \leq \frac{\widehat{f}_{n,j+1}(x_{j+1}) - \widehat{f}_{1,j}(x_j)}{x_{j+1} - x_j}. \tag{5}$$

If Equation (5) holds, then $Y_j \leq \widehat{h}_n^j(x_j) \leq \widehat{h}_n^j(x_{j+1}) \leq Y_{j+1}$.

In addition, we have the following crucial result for all global S-shaped LSEs over $\mathcal{H} \equiv \mathcal{H}[x_1, \ldots, x_n] := \mathcal{F} \cap \mathcal{G}$, namely those $\widehat{f}_n^{x_{j'}}$ for which $j' \in \operatorname{argmin}_{1 \leq j \leq n} S_n(\widehat{f}_n^{x_j})$.

Proposition 4 *Given any S-shaped LSE $\tilde{f}_n$ over $\mathcal{H}$, if $j \in [n-1]$ is such that either x_j is the smallest inflection point of $\tilde{f}_n$ or x_{j+1} is the largest inflection point of $\tilde{f}_n$, then $\widehat{h}_n^j = \tilde{f}_n$ and hence $Y_j \leq \tilde{f}_n(x_j) \leq \tilde{f}_n(x_{j+1}) \leq Y_{j+1}$.*

We explain in the final example of Section S1 that Proposition 4 is a consequence of Proposition S4(c, d, e), whose proof also reveals why $\widehat{h}_n^j = \widehat{f}_n^{x_j}$ is not guaranteed to hold for a pre-specified $j \in [n-1]$. A further remark is that the localization property for $\tilde{f}_n$ in Proposition 4 is only valid for particular choices of partition of our data into subintervals, namely where the split occurs at the smallest or largest inflection points of $\tilde{f}_n$. In other words, if for example x_j is chosen to be a kink of $\tilde{f}_n$ that is strictly to the left of the smallest inflection point, then $\tilde{f}_n$ is not guaranteed to agree with the increasing convex LSE $\widehat{f}_{1,j}$ on $[x_1, x_j]$. This presents a substantial additional difficulty for both computation and theory in comparison with the problem of unimodal regression (Shoung & Zhang, 2001; Stout, 2008), where, for every jump x_j of the unimodal LSE $\tilde{g}_n$ to the left of its mode, it is the case that $\tilde{g}_n$ agrees on $[x_1, x_j]$ with the increasing LSE based on $\{(x_i, Y_i) : 1 \leq i \leq j\}$. These issues are discussed in greater depth in Section S1.

Propositions 3 and 4 motivate the following generic procedure as an improvement on `ScanAll`:

Algorithm 1. Generic algorithm for computing $(\hat{m}_n, \hat{f}_n)$.

(I) Discard all $j \in [n-1]$ for which $Y_j > Y_{j+1}$.

(II) For each of the remaining indices $j \in [n-1]$, compute $\hat{f}_{1,j}$ based on $\{(x_i, Y_i) : 1 \leq i \leq j\}$ and $\hat{f}_{n,j+1}$ based on $\{(x_i, Y_i) : j+1 \leq i \leq n\}$, and concatenate these to obtain $\hat{h}_n^j$ via (4). Discard j if $\hat{h}_n^j \notin \mathcal{H}^{x_j}$, i.e.

$$\frac{\hat{f}_{n,j+1}(x_{j+2}) - \hat{f}_{n,j+1}(x_{j+1})}{x_{j+2} - x_{j+1}} > \frac{\hat{f}_{n,j+1}(x_{j+1}) - \hat{f}_{1,j}(x_j)}{x_{j+1} - x_j}.$$

(III) Let $\mathcal{J}$ be the set of indices $j \in [n-1]$ that are retained after Step II. Find $\tilde{j} := \operatorname{sargmin}_{j \in \mathcal{J}} S_n(\hat{h}_n^j)$ by computing $S_n(\hat{h}_n^j) = n^{-1} \sum_{i=1}^n \big(Y_i - \hat{h}_n^j(x_i)\big)^2$ for each $j \in \mathcal{J}$, and return $(x_{\tilde{j}}, \hat{h}_n^{\tilde{j}})$.

To see that the output $(x_{\tilde{j}}, \widehat{h}_n^{\tilde{j}})$ of Algorithm 1 is indeed $(\widehat{m}_n, \widehat{f}_n)$, note first that by Proposition 3, the set $\mathcal{J}$ in Step III consists precisely of those $j \in [n-1]$ for which $\widehat{h}_n^j = \widehat{f}_n^{x_j}$. In addition, by Proposition 4, $\widehat{\jmath}_n = \operatorname{sargmin}_{1\le j\le n} S_n(\widehat{f}_n^{x_j}) \in \mathcal{J}$ since $\widehat{m}_n = x_{\widehat{\jmath}_n}$ is the smallest inflection point of $\widehat{f}_n = \widehat{f}_n^{\widehat{m}_n}$. Thus, $\tilde{\jmath} = \operatorname{sargmin}_{j\in\mathcal{J}} S_n(\widehat{f}_n^{x_j}) = \widehat{\jmath}_n$, and hence $x_{\tilde{\jmath}} = \widehat{m}_n$ and $\widehat{h}_n^{\tilde{\jmath}} = \widehat{f}_n$, as desired.

The most obvious implementation of Step II of Algorithm 1 simply computes $\widehat{f}_{1,j}$ and $\widehat{f}_{n,j+1}$ from scratch for each different j; we refer to this as the `ScanSelected` algorithm. Even this naive modification has two significant advantages over `ScanAll`:

(i) In advance of carrying out any least squares minimization, we can restrict the set of candidates for $\widehat{\jmath}_n$ based on just $n-1$ pairwise comparisons. If $(x_1, Y_1), \ldots, (x_n, Y_n)$ are drawn according to a regression model (2) featuring a continuous f_0 and independent and identically distributed errors with zero mean, then Step I typically screens out about half of the indices in $[n]$ when n is reasonably large.

(ii) For the remaining indices j in Step II, we do not attempt to compute the S-shaped function $\widehat{f}_n^{x_j}$ based on all n data points, but instead fit the increasing convex LSE $\widehat{f}_{1,j}$ and the increasing concave LSE $\widehat{f}_{n,j+1}$ using j and $n-j$ observations respectively.

The main drawback of the `ScanSelected` algorithm, however, is that it fails to exploit the commonalities in the computation of $\widehat{f}_{1,j}$ for different j (and similarly of $\widehat{f}_{n,j+1}$ for different j). Our main computational contribution, then, is to show that for $k \in [j-1]$, it is possible to obtain $\widehat{f}_{1,j}$ by modifying $\widehat{f}_{1,k}$ appropriately when the observations $\{(x_i, Y_i) : k < i \le j\}$ are introduced. We can therefore proceed in a sequential manner and hence make significant computational gains.

Recall that for $j \in [n]$ and a closed, convex cone $\Lambda \subseteq \mathbb{R}^j$, there exists a unique L^2-projection $\Pi_\Lambda : \mathbb{R}^j \to \Lambda$, given by

$$\Pi_\Lambda(y) := \operatorname*{argmin}_{u\in\Lambda} \|u - y\|.$$

The key to our approach is to develop a mixed primal-dual bases algorithm (Fraser & Massam, 1989; Meyer, 1999) that allows us to compute $\Pi_\Lambda(L)$ when $L \subseteq \mathbb{R}^j$ is a line segment and Λ is a polyhedral convex cone. An important observation is that, given $v(0), v(1) \in \mathbb{R}^j$, the map $t \mapsto \Pi_\Lambda((1-t)v(0) + tv(1))$ is continuous and piecewise linear on $[0, 1]$, where the individual linear pieces correspond to projections onto different faces of Λ; see Remark 1 in Appendix A. This enables us to compute $\Pi_\Lambda(v(1))$ when $\Pi_\Lambda(v(0))$ is known. Indeed, we give a detailed description of a general procedure for this task in Algorithm 2 in Appendix A, and we focus here on its application to increasing convex regression (increasing concave regression for the right-hand end can be handled very similarly). In this case, the cones of particular interest to us are those of increasing convex sequences based on $x_1, \ldots, x_j$ for some $j \in [n]$, which we denote by

$$\Lambda^j := \{(g(x_1), \ldots, g(x_j)) : g \in \mathcal{F}^1\} = \left\{ (z_1, \ldots, z_j) \in \mathbb{R}^j : 0 \le \frac{z_2 - z_1}{x_2 - x_1} \le \cdots \le \frac{z_j - z_{j-1}}{x_j - x_{j-1}} \right\}. \tag{6}$$

Given $k \in [j-1]$ and supposing that we have already fitted the increasing convex LSE $\widehat{f}_{1,k}$ (which is linear on $[x_{k-1}, 1]$), an appropriate choice of $v(0), v(1)$ is

$$v(0) = \left(Y_1, \ldots, Y_k, \widehat{f}_{1,k}(x_{k+1}), \ldots, \widehat{f}_{1,k}(x_j)\right) \quad \text{and} \quad v(1) = (Y_1, \ldots, Y_j); \tag{7}$$

indeed, $\Pi_{\Lambda^j}(v(1)) = (\widehat{f}_{1,j}(x_1), \ldots, \widehat{f}_{1,j}(x_j))$ is what we seek to compute, and moreover we claim that $\Pi_{\Lambda^j}(v(0)) = (\widehat{f}_{1,k}(x_1), \ldots, \widehat{f}_{1,k}(x_j))$ (which is known). To establish this claim, observe that for any $u \equiv (u_1, \ldots, u_j) \in \Lambda^j$, we have

$$\|v(0)-u\|^2 \geq \sum_{i=1}^{k}(Y_i-u_i)^2 \geq \sum_{i=1}^{k}\left(Y_i-\widehat{f}_{1,k}(x_i)\right)^2 = \left\|v(0)-\left(\widehat{f}_{1,k}(x_1),\ldots,\widehat{f}_{1,k}(x_j)\right)\right\|^2, \tag{8}$$

and $(\widehat{f}_{1,k}(x_1), \ldots, \widehat{f}_{1,k}(x_j)) \in \Lambda^j$. In fact, we will apply this version of the mixed primal-dual bases algorithm with $k = j-1$, so that the observations $Y_1, \ldots, Y_n$ are introduced sequentially. Note that when $Y_j \geq \widehat{f}_{1,j-1}(x_j)$, we have by the same argument as in Equation (8) that $(\widehat{f}_{1,j}(x_1), \ldots, \widehat{f}_{1,j}(x_j)) = (\widehat{f}_{1,j-1}(x_1), \ldots, \widehat{f}_{1,j-1}(x_{j-1}), Y_j)$, so no calculations are required. We refer to this sequential implementation of Algorithm 1 as `SeqConReg`.

4 | THEORETICAL PROPERTIES OF S-SHAPED LEAST SQUARES ESTIMATORS

4.1 | Worst-case and adaptive sharp oracle inequalities

Our first main results of this section consist of worst-case and adaptive sharp oracle inequalities for S-shaped least squares estimators. These reveal not only risk bounds when our S-shaped regression function hypothesis is correctly specified, but also control the way in which the performance of the estimators deteriorate as the model becomes increasingly misspecified.

We will work in the setting of model (2), and now make the following assumption on the errors:

Assumption 1 $\{\xi_i \equiv \xi_{ni} : 1 \leq i \leq n\}$ is a collection of independent sub-Gaussian random variables with parameter 1, so that $\mathbb{E}(e^{t\xi_{ni}}) \leq e^{t^2/2}$ for all $t \in \mathbb{R}$ and $i \in [n]$.

For fixed $n \in \mathbb{N}$ and $f : [0, 1] \to \mathbb{R}$, we write $x_i \equiv x_{ni}$ for $i \in [n]$ and let $\|f\|_n := \|f\|_{L^2(\mathbb{P}_n^X)} = \left(\sum_{i=1}^n f^2(x_i)/n\right)^{1/2}$. Also, for $f \in \mathcal{H} \equiv \mathcal{H}[x_1, \ldots, x_n]$, let $V(f) := f(x_n) - f(x_1) = \max_{1\leq i\leq n} f(x_i) - \min_{1\leq i\leq n} f(x_i)$ and denote by $k(f)$ the number of affine pieces of f, so that $k(f)$ is the smallest $k \in [n]$ with the property that f is affine on each of k subintervals $I_1, \ldots, I_k$ that partition $[0, 1]$.

Theorem 1 *For fixed $n \geq 2$, suppose that Assumption 1 holds and let $\tilde{f}_n$ be any LSE over $\mathcal{F}$. Let $R := n^{-1}(x_n - x_1)/\min_{2\leq i\leq n}(x_i - x_{i-1})$. Then there exists a universal constant $C > 0$ such that for every $f_0 : [0, 1] \to \mathbb{R}$ and $t > 0$, we have*

$$\|\tilde{f}_n - f_0\|_n \leq \inf_{f\in\mathcal{H}}\left\{\|f-f_0\|_n + \frac{C(1+V(f))^{1/3}}{n^{1/3}} \wedge \frac{CR^{1/10}(1+V(f))^{1/5}}{n^{2/5}}\right\} + \sqrt{\frac{8t}{n}} \tag{9}$$

with probability at least $1 - e^{-t}$.

By integrating this tail bound, we obtain the worst-case risk bound

$$\mathbb{E}_{f_0}(\|\tilde{f}_n - f_0\|_n) \leq \inf_{f\in\mathcal{H}}\left\{\|f-f_0\|_n + \frac{C(1+V(f))^{1/3}}{n^{1/3}} \wedge \frac{CR^{1/10}(1+V(f))^{1/5}}{n^{2/5}}\right\} + \sqrt{\frac{2\pi}{n}}. \tag{10}$$

In the special case where $f_0 \in \mathcal{F}$, we may take $f = f_0$ in Theorem 1 to conclude that

$$\mathbb{E}_{f_0}(\|\tilde{f}_n - f_0\|_n) \lesssim \frac{(1+V(f_0))^{1/3}}{n^{1/3}} \wedge \frac{R^{1/10}(1+V(f_0))^{1/5}}{n^{2/5}};$$

thus, when R and $V(f_0)$ are of constant order, we obtain a worst-case risk bound of order $n^{-2/5}$. More generally, Equations (9) and (10) reveal the impact of both non-equispaced design and the range of the signal. In fact, an alternative, more complicated definition of R is possible, and this further refines our bounds for certain designs; see the discussion following the proof of Theorem 1 in Section S2.1. To see that the rate of order $n^{-2/5}$ cannot in general be attained for arbitrary configurations of design points, we appeal to Bellec (2018, Theorem 4.5) for a suitable minimax lower bound: for any $V \geq n^{-1/2}$, there exist design points $x_1 < \cdots < x_n$ that depend on V such that if $\xi_1, \ldots, \xi_n \overset{\text{iid}}{\sim} N(0, 1)$ in Equation (2), then

$$\inf_{\check{g}_n} \sup_{f_0 \in \mathcal{F}^1 : V(f_0) \leq 2V} \mathbb{P}_{f_0}(\|\check{g}_n - f_0\|_n \geq C(V/n)^{1/3}) \geq c,$$

where the infimum is taken over all estimators $\check{g}_n \equiv \check{g}_n(x_1, Y_1, \ldots, x_n, Y_n)$, and $c, C > 0$ are universal constants.

Another very attractive aspect of Theorem 1 is that, in cases where $f_0 \notin \mathcal{F}$, we can control the performance of an LSE $\tilde{f}_n$ over $\mathcal{F}$ via approximation error and estimation error terms. The fact that the approximation error term $\|f - f_0\|_n$ has leading constant 1 (which is the best possible) is the reason that Equations (9) and (13) are referred to as sharp oracle inequalities.

To complement the worst-case sharp oracle inequality in Equation (10), we now consider the more favourable situation where f_0 is well approximated by a piecewise affine function with not too many affine pieces. The fact that an LSE $\tilde{f}_n$ over $\mathcal{F}$ can approximate such a signal with a relatively small number of kinks suggests that we may be able to obtain improved sharp oracle inequalities in such cases.

Theorem 2 *For fixed $n \geq 2$, suppose that Assumption 1 holds, and let $\tilde{f}_n$ be any LSE over $\mathcal{F}$. Then for every $f_0 : [0, 1] \to \mathbb{R}$ and $t > 0$, we have*

$$\|\tilde{f}_n - f_0\|_n \leq \inf_{f \in \mathcal{H}} \left\{ \|f - f_0\|_n + \sqrt{\frac{32(k(f)+1)}{n} \log\left(\frac{en}{k(f)+1}\right)} \right\} + \sqrt{\frac{2(t + \log n)}{n}} \tag{11}$$

with probability at least $1 - e^{-t}$.

As with Theorem 1, we can integrate the tail bound from Equation (11) to obtain

$$\begin{aligned} \mathbb{E}_{f_0}(\|\tilde{f}_n - f_0\|_n) &\leq \inf_{f \in \mathcal{H}} \left\{ \|f - f_0\|_n + \sqrt{\frac{32(k(f)+1)}{n} \log\left(\frac{en}{k(f)+1}\right)} \right\} + \sqrt{\frac{2\log n}{n}} + \sqrt{\frac{\pi}{2n}} \\ &\leq \inf_{f \in \mathcal{H}} \left\{ \|f - f_0\|_n + 8\sqrt{\frac{k(f)+1}{n} \log\left(\frac{en}{k(f)+1}\right)} \right\}. \end{aligned} \tag{12}$$

In particular, we see from Equation (12) that if $f_0 \in \mathcal{F}$ has k affine pieces, then any LSE $\tilde{f}_n$ over $\mathcal{F}$ attains the parametric rate $k^{1/2}/n^{1/2}$, up to a logarithmic factor.

Adaptation to signals of low complexity is one of the particularly intriguing aspects of shape-constrained estimators (Guntuboyina & Sen, 2018; Samworth, 2018). For instance, Guntuboyina and Sen (2013), Chatterjee et al. (2015) and Chatterjee and Lafferty (2019) investigated the adaptive behaviour of univariate convex, isotonic and unimodal LSEs respectively when the truth is well approximated by a function with a small number of affine or constant pieces. For multivariate extensions of these results, see for example Han and Wellner (2016), Kur et al. (2020) and Han (2021) among others. Sharp oracle inequalities of a similar flavour to Theorem 2 have been obtained for a variety of LSEs (Bellec, 2018), including multivariate isotonic LSEs (Han et al., 2019; Pananjady & Samworth, 2021). In log-concave density estimation, adaptation results of this type were established for the log-concave maximum likelihood estimator by Kim et al. (2018) and Feng et al. (2021) in univariate and multivariate settings respectively. Finally, Baraud and Birgé (2016) introduced a ρ-estimation framework for univariate shape-constrained estimation and studied its adaptation properties.

4.2 | Inflection point estimation

A particular feature of S-shaped function estimation that differentiates it from other shape-constrained estimation problems is the existence of an inflection point m_0. In some respects, this is like a boundary point, because it represents the point of transition from convex to concave parts of the function, and the behaviour of the function is therefore less regulated there (in particular, the derivative of an S-shaped function may diverge to infinity as we approach the inflection point). When $m_0 \in (0, 1)$, we may well have design points on either side of m_0, and in that sense the inflection point may be regarded as an interior point. The distinguished nature of the inflection point means that its location is often of interest in applications such as the modelling of economic growth (e.g. Jarne et al., 2007) and disease progression in longitudinal studies (e.g. Lee et al., 2020). For instance, in the latter work, S-shaped functions were used to model the deterioration in motor function associated with Huntington's disease, and the estimated inflection points from a nonparametric procedure were seen to be clinically useful indicators of the onset of severe motor dysfunction, in the sense of having the potential to facilitate timely diagnosis and intervention.

In studying the inflection point estimation problem, we will assume that $f_0 \in \mathcal{F}$ and the following additional conditions hold:

Assumption 2 Suppose that $f_0 \in \mathcal{F}$ has a unique inflection point $m_0 \in (0, 1)$, and that there exist $B > 0$ and $\alpha \in (0, 1) \cup (1, \infty)$ such that as $x \to m_0$, we have

$$f_0(x) = \begin{cases} f_0(m_0) - B(1+o(1))\,\mathrm{sgn}(x-m_0)|x-m_0|^\alpha & \text{when } \alpha \in (0,1) \\ f_0(m_0) + f_0'(m_0)(x-m_0) + B(1+o(1))\,\mathrm{sgn}(x-m_0)|x-m_0|^\alpha & \text{when } \alpha > 1. \end{cases} \tag{13}$$

In the regression model (2), suppose also that $x_{ni} = i/n$ and $\xi_{ni} \overset{d}{=} \xi$ for all $n \in \mathbb{N}$ and $i \in [n]$, where ξ is a sub-Gaussian random variable with parameter 1.

When $\alpha \geq 3$ is an integer, Equation (13) holds if f_0 is α-times continuously differentiable in a neighbourhood of m_0 and $f_0^{(k)}(m_0) = 0 \neq f_0^{(\alpha)}(m_0)$ for $2 \leq k \leq \alpha - 1$. Under this stronger assumption, α must in fact be odd, and $f_0^{(\alpha)}(m_0) < 0$. Indeed, for all $x \in [0, 1]$ sufficiently close to the inflection point m_0, we have $f_0''(x) \geq 0$ if $x \leq m_0$ and $f_0''(x) \leq 0$ if $x \geq m_0$, and since $f_0^{(\alpha)}$ is continuous at m_0, a Taylor expansion reveals that $f_0''(x) = f_0^{(\alpha)}(m_0)(1 + o(1))(x - m_0)^{\alpha-2}/(\alpha - 2)!$ as $x \to m_0$.

Theorem 3 *Let $(\tilde{f}_n)$ be any sequence of LSEs over $\mathcal{F}$, and for each n, let $\tilde{m}_n$ be an inflection point of $\tilde{f}_n$. Under Assumption 2, we have $\tilde{m}_n - m_0 = O_p((n/\log n)^{-1/(2\alpha+1)})$.*

We mention that Liao and Meyer (2017) study a least squares estimator over a subclass of $\mathcal{F}$ consisting of cubic splines (where the number of knots is of order $n^{1/9}$); they show that its inflection point converges to the true m_0 at rate $O_p(n^{-8/63})$ in a random design setting where f_0 satisfies (a stronger version of) Equation (13) with $\alpha = 3$. The proof of their Theorem 2 relies on a quantitative result on the quality of local approximations to f_0 near m_0 by convex or concave functions (Liao & Meyer, 2017, Lemma 2), as well as a global rate of convergence for their spline-based estimator.

In our setting, Theorem 3 shows that the inflection point estimator $\tilde{m}_n$ (based on an LSE $\tilde{f}_n$ over the entire class $\mathcal{F}$) converges to m_0 at rate $O_p((n/\log n)^{-1/7})$ when $\alpha = 3$. The proof of Theorem 3, which is given in Section S2, is lengthy and broken up into several steps, each of which requires some delicate technical arguments; see Figure S1 for an illustration. The crucial Step 2a exploits the observation that if $\tilde{m}_n$ is a long way from m_0, then there is a long interval between the two on which one of $f_0, \tilde{f}_n$ is convex and the other is concave. On such an interval, we show that $\tilde{f}_n$ has a long affine piece, as would be intuitively expected, and thereby quantify the approximation error due to misspecification; see Lemma S6. Another important aspect of our proof strategy is that we find a suitable way to localize the analysis of $\tilde{f}_n$ to a neighbourhood of m_0, rather than rely on global considerations that would lead to a suboptimal bound. As we explain in Section S1, our localization technique for convex or S-shaped LSEs relies on non-trivial 'boundary adjustments' that are not needed for isotonic or unimodal LSEs. Nevertheless, a simpler version of the proof of Theorem 3 allows us to recover the result of Shoung and Zhang (2001) on the rate of convergence of the mode of the LSE of a unimodal regression function, at least under our sub-Gaussian assumption on the errors ξ_{ni} and their local smoothness condition (1.3).

The rate of convergence of $\tilde{m}_n$ to m_0 in Theorem 3 matches that in the following complementary local asymptotic minimax lower bound, up to a logarithmic factor. For $r > 0$, let $\mathcal{F}(f_0, r) := \{f \in \mathcal{F} : \int_0^1 (f - f_0)^2 < r^2\}$. Although f_0 has a unique inflection point m_0 under Assumption 2, not every function in $\mathcal{F}(f_0, r)$ has a unique inflection point, so for $f \in \mathcal{F}$, we denote by $\mathcal{I}_f$ the subinterval of inflection points of f and define $d(x, \mathcal{I}_f) := \inf_{z \in \mathcal{I}_f} |x - z|$ for $x \in [0, 1]$.

Proposition 5 *Under Assumption 2, and with $\xi_{n1}, \ldots, \xi_{nn} \overset{\text{iid}}{\sim} N(0, 1)$ for all n, we have*

$$\sup_{\tau > 0} \liminf_{n \to \infty} \inf_{\check{m}_n} \sup_{f \in \mathcal{F}(f_0, \tau/\sqrt{n})} n^{1/(2\alpha+1)} \, \mathbb{E}_f(d(\check{m}_n, \mathcal{I}_f)) > 0, \tag{14}$$

where the infimum is taken over all estimators $\check{m}_n \equiv \check{m}_n(x_1, Y_1, \ldots, x_n, Y_n)$ taking values in $[0, 1]$, and $\mathbb{E}_f$ is the expectation operator under the model (2) with f in place of f_0.

5 | SIMULATIONS AND REAL DATA EXAMPLE

In this section, we first investigate the computation time and empirical performance of our S-shaped estimator in some numerical experiments. We then demonstrate the use of our estimator in a real data application to air pollution modelling.

5.1 | Computation time

We compare the running time of our sequential cone projection Algorithm 2, denoted as `SeqConReg`, with two other possible approaches. The first, which we call `ScanAll`, relies on a brute-force search that scans through all possible inflection points $m \in \{x_1, \ldots, x_n\}$ as described in the introduction, performing least squares over $\mathcal{F}^m$, and determining the candidate that minimizes the residual sum of squares. Here the active set least squares procedure used for each m is based on a simple modification of the `R` package `scar` (Chen & Samworth, 2014, 2016). The second approach, which we call `ScanSelected`, is based on the observation in Step I of Algorithm 1 that there is no need to scan through all design points. Instead, we restrict attention to those indices j for which $Y_j \leq Y_{j+1}$, fitting an increasing convex function to $\{(x_i, Y_i) : 1 \leq i \leq j\}$, an increasing concave function to $\{(x_i, Y_i) : j+1 \leq i \leq n\}$ (both using `scar`), before finding the smallest j that minimizes the residual sum of squares.

For $n \in \{100, 200, 500, 1000, 2000\}$, we set $x_i = i/(n+1)$ and $Y_i = \sin(\pi(x_i - 0.5)) + \sigma\epsilon_i$ for $i = 1, \ldots, n$, where $\epsilon_1, \ldots, \epsilon_n$ are independent normal random variables with zero mean and unit variance. Here, to examine the impact of the signal-to-noise ratio on the running time, we also vary the value of $\sigma \in \{1, 0.1, 0.01\}$, and plot the average running time of the different approaches in Figure 5. We see that `SeqConReg` is the fastest among all three approaches, being approximately 10 times more efficient than `ScanSelected` and 40 times faster than `ScanAll`. The

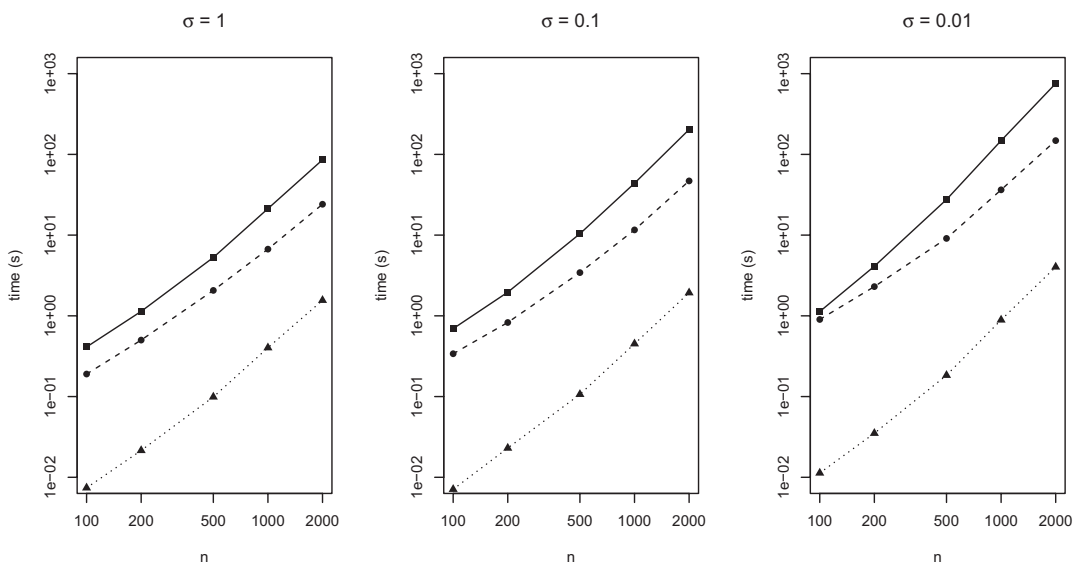

FIGURE 5 Log-log plots of the running time (in seconds) of the `SeqConReg` (▲), `ScanSelected` (•) and `ScanAll` (■) algorithms for least squares estimation of an S-shaped function, for sample sizes $n \in \{100, 200, 500, 1000, 2000\}$ and noise levels $\sigma \in \{1, 0.1, 0.01\}$

ratio of the timings becomes larger as the signal-to-noise ratio increases, because the resulting fitted function has more knots, which makes it more appealing to use algorithms of a sequential nature, such as `SeqConReg`.

5.2 | Statistical performance

We compare our estimator (denoted by LSE below) with the following alternatives:

- Spline: The method of Liao and Meyer (2017), based on cubic B-splines with shape constraints, which is implemented in the `R` package `ShapeChange` (Liao & Meyer, 2016);
- SCKLS: The shape-constrained kernel least squares method of Yagi et al. (2019, 2020) based on local linear kernels, with $M = 50$ evaluation points and kernel bandwidths selected according to the method of Ruppert et al. (1995);
- BEDE and BESE: The bisection extremum distance estimator and bisection extremum surface estimators of Christopoulos (2016), both developed based on the geometric properties of the inflection point for a smooth function and implemented in the `R` package `inflection` (Christopoulos, 2019).

For LSE, Spline and SCKLS, we assess their performance based on both the average $L^2(\mathbb{P}_n)$ loss and the mean absolute error of the estimated inflection point location, while for BEDE and BESE we compute only the mean absolute error of the estimated inflection point location. All results are based on numerical experiments over 1000 repetitions.

For $n \in \{100, 200, 500, 1000\}$, and design points $x_1, \ldots, x_n$, we set $Y_i = f_j(x_i) + 0.1\epsilon_i$ for $i = 1, ..., n$, where $\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} N(0, 1)$, for four different choices of signal function f_j:

$$f_1(x) = \begin{cases} 2(0.3 - \sqrt{0.09 - x^2}) & \text{for } x \in [0, 0.3) \\ 2\left\{0.3 + \sqrt{0.49 - (1-x)^2}\right\} & \text{for } x \in [0.3, 1] \end{cases}; \quad f_3(x) = x + \mathbb{1}_{\{x \geq 0.3\}};$$
$$f_2(x) = \sin((x - 0.3)\pi/1.4)\mathbb{1}_{\{x \geq 0.3\}}; \quad f_4(x) = 4/(1 + e^{-2(x-0.3)}). \tag{15}$$

These signals are plotted in Figure 6. The signals are designed in such a way that their ranges over [0,1] are roughly the same. Furthermore, they all belong to $\mathcal{F}$ and have a unique inflection point at $m_0 = 0.3$. Note that f_1 satisfies Assumption 2 with $\alpha = 1/2$, and f_2 and f_3 do not satisfy Assumption 2 for any $\alpha > 0$, while f_4 satisfies the assumption with $\alpha = 3$.

We consider two different designs by setting $x_i = F^{-1}(i/(n + 1))$ for $i = 1, ..., n$, where F is the distribution function of either the $U[0, 1]$ or Beta(4, 8) distributions. In the second setting, the design points are not equally spaced, and $m_0 = 0.3$ is the mode of the Beta(4, 8) distribution. The results are shown in Figures 7 and 8.

For the estimation of the regression function, the LSE performs well in all cases; in particular, it is able to adapt to inhomogeneous smoothness levels and asymmetric designs. The spline- and kernel-based approaches struggle in this regard, and perform much worse for signals f_1 and f_3 especially. In fact, the spline-based method appears to be inconsistent for signals f_1 and f_3, and the kernel-based approach seems to suffer the same problem for signal f_3 too. For the estimation of the inflection point, the story has some similarities, but also some differences: for signals f_1, f_2

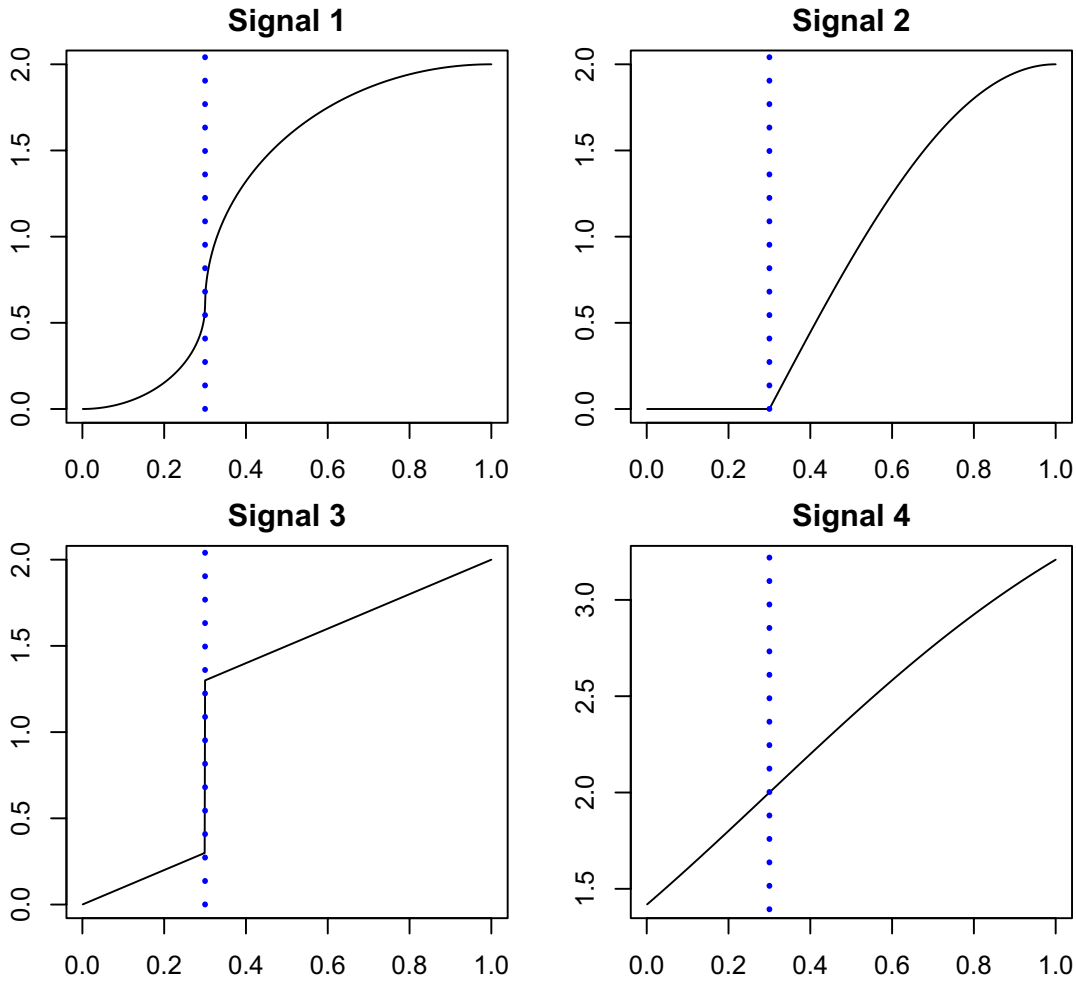

FIGURE 6 Plots of the signals f_1, f_2, f_3, f_4 defined in Equation (15), with the inflection points highlighted by dashed blue lines

and f_3, the least squares approach provides more reliable estimates, for two main reasons. First, it is able to adapt to a much wider range of local smoothnesses around m_0. Second, by carefully comparing Figure 8 to Figure 7, we see that the least squares approach is also able to take advantage of the additional design points near m_0 under the beta design to obtain improved estimation performance (relative to the uniform design). For signal f_4, the other methods are able to exploit the homogeneity of the signal across the entire domain (and the symmetry of the signal around the inflection point) and tend to have a smaller mean absolute error than the least squares approach. We recall Figure 2, which further illustrates the dangers of assuming smoothness of an S-shaped signal when it is not present.

5.3 | Real data example

In this subsection, we apply our nonparametric S-shaped procedure to $n = 221$ LIDAR (light detection and ranging) measurements for determining atmospheric concentrations of mercury

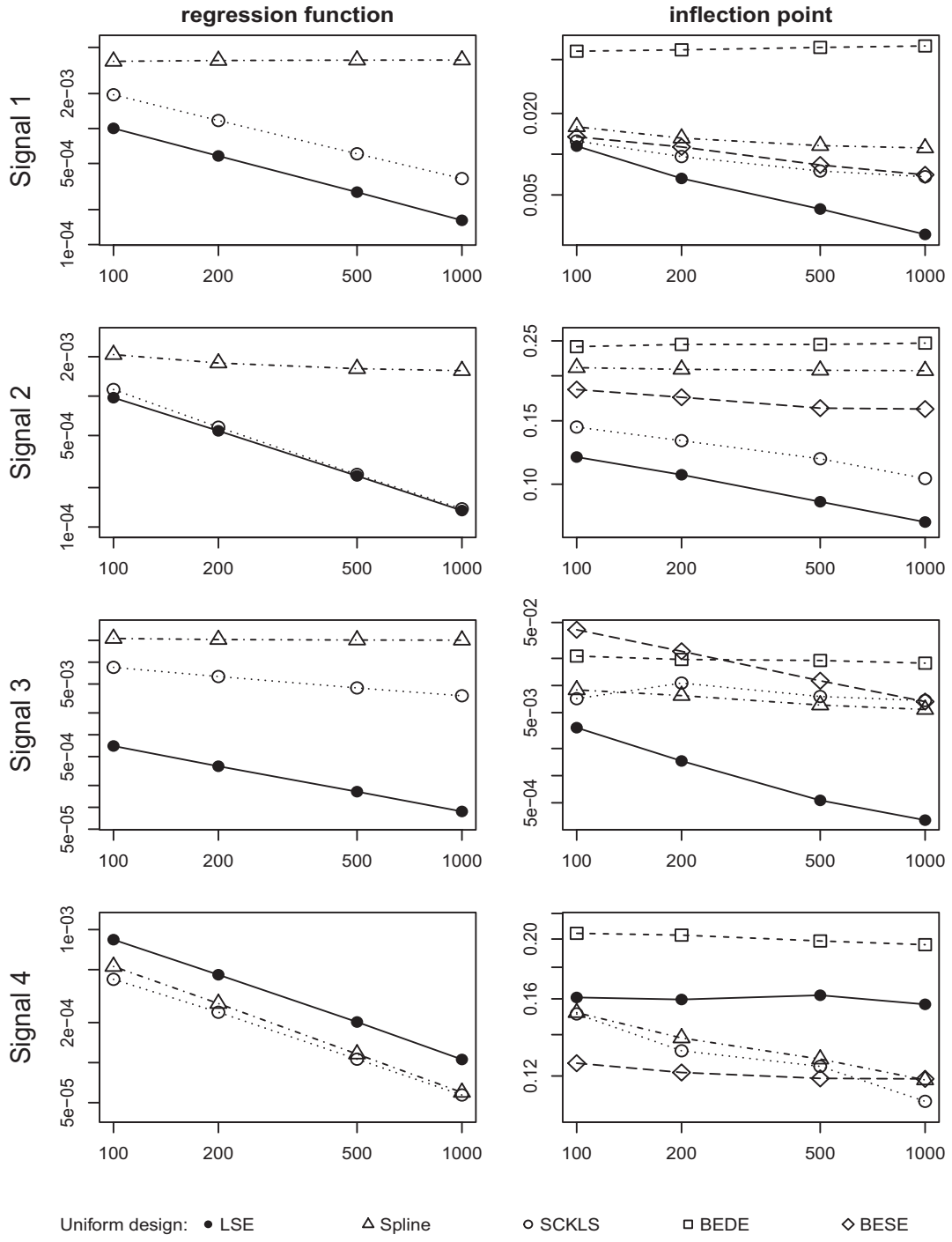

FIGURE 7 Log-log plots of the mean squared error of the fitted function on the design points, as well as the mean absolute distance between the estimated and true inflection points, based on $n = 100, 200, 500, 1000$ observations when the design points are equispaced and the signals are as in Figure 6

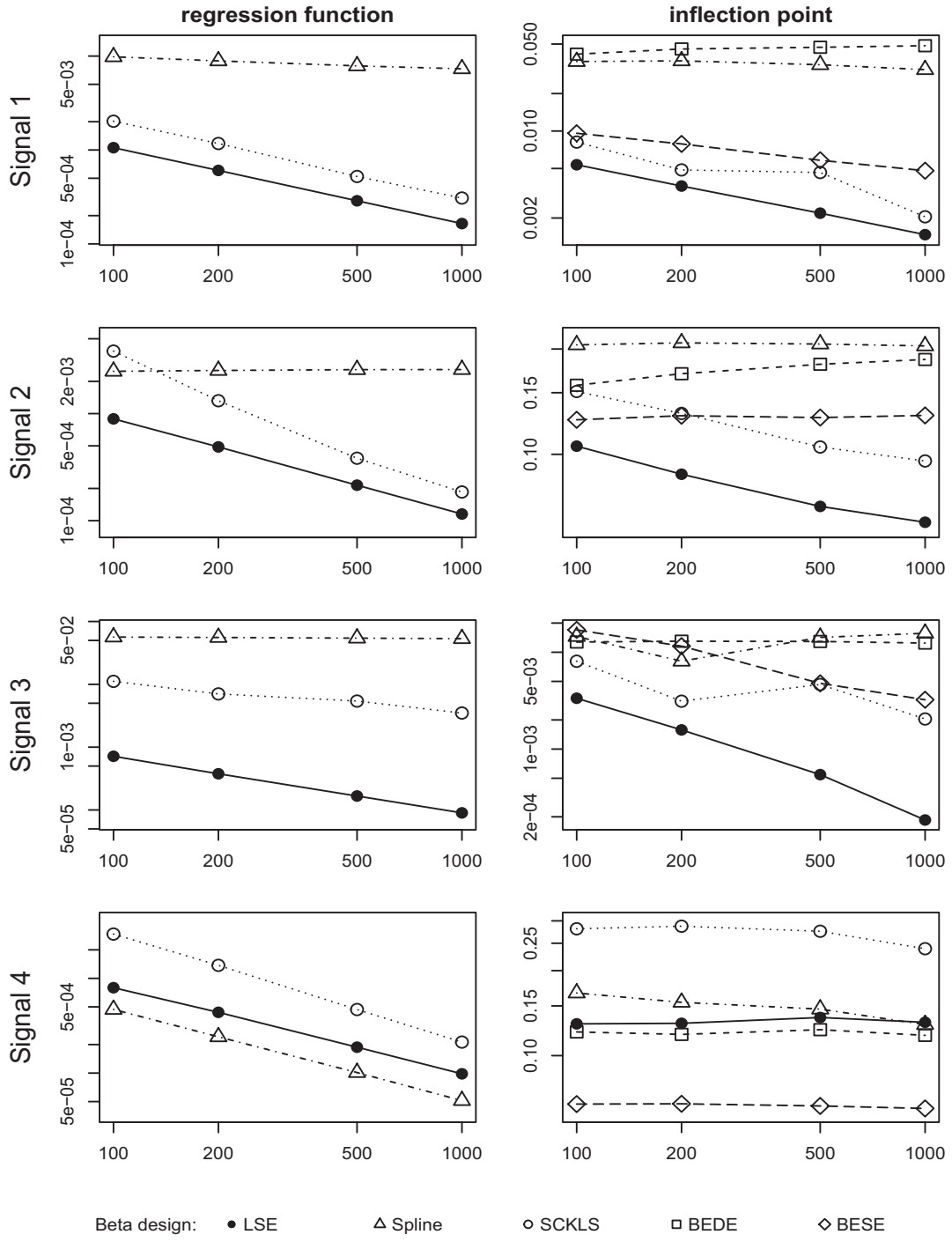

FIGURE 8 Log-log plots of the mean squared error of the fitted function on the design points, as well as the mean absolute distance between the estimated and true inflection points, based on $n = 100, 200, 500, 1000$ observations when the design points are quantiles of a Beta(4, 8) distribution and the signals are as in Figure 6

emissions from the Bella Vista geothermal power station in Italy. This dataset, which is of interest from an air pollution modelling perspective, is discussed at length by Ruppert et al. (2003) and included in the `R` package `SemiPar` (Wand, 2018).

To explain the rationale behind the use of the S-shaped regression model (2) in this context, we begin by briefly outlining the physical background and experimental setup; see Edner et al. (1989, 1992) and Holst et al. (1996, Section 2) for further details.[2] In this instance, the LIDAR equipment was set up at a fixed location downwind of the power station, at a distance of 390–720 m from the bulk of the mercury plume. The DIAL (differential absorption LIDAR) technique involves firing two laser beams in quick succession in the same direction towards the plume, where the first beam contains light at the resonant wavelength $\lambda_{\text{on}} = 253.6\,\text{nm}$ of mercury while the second 'reference' beam is set to a slightly different 'off-resonant' wavelength λ_{off}. The light in both beams is scattered (or reflected back) to roughly the same extent by particles and aerosols in the atmosphere, but the light at wavelength λ_{on} is absorbed much more strongly by atoms of mercury, the pollutant of interest. The LIDAR apparatus records the intensity (i.e. power) of the reflected signals from both incident beams as a function of time elapsed, which is proportional to the distance travelled by the light before it is reflected back towards the source. The latter is the independent variable `range` in the dataset. The intensity curves from 100 pairs of laser shots in the same direction were then averaged to produce power estimates $P(r_i;\, \lambda_{\text{on}})$ and $P(r_i;\, \lambda_{\text{off}})$ for $n = 221$ equispaced values r_i of `range` between 390 and 720 m (at intervals of 1.5 m). In view of the physical reasons outlined above, the relative sizes of these two quantities for different r_i can be used to estimate how the atmospheric concentration $g_0(r)$ of mercury (in ng/m^3) varies with distance r (in metres) along the path of the laser beams.

More precisely, based on an approximation of the governing equation for LIDAR scattering, Holst et al. (1996, Section 3) consider a regression model for the `logratio` values

$$\log \frac{P(r_i;\, \lambda_{\text{on}})}{P(r_i;\, \lambda_{\text{off}})} = f_0(r_i) + \xi_i, \quad i = 1, \ldots, n,$$

where on physical grounds, $f_0(r) = -C \int_0^r g_0(s)\, ds$ is defined for $r \geq 0$ as the integral of the concentration function g_0 over $[0,\, r]$ multiplied by $-C \equiv -C(\lambda_{\text{on}},\, \lambda_{\text{off}}) = -1.6 \times 10^{-5}\,\text{ng}^{-1}\,\text{m}^2$. Since mercury concentration is always non-negative and would generally be expected to decrease with distance from the interior of the plume, g_0 can reasonably be modelled as a non-negative unimodal function, in which case its antiderivative satisfies our definition of an S-shaped function. The data, shown in Figure 9, do indeed appear to support f_0 as an inverted S-shaped regression function. Moreover, Holst et al. (1996, Figure 4) present plots of suitably normalized residuals against `range` as well as the sample autocorrelations at different lags, which provide some empirical justification for the assumption that the errors $\xi_1, \ldots, \xi_n$ are independent.

The different panels of Figure 9 illustrate least squares fits over different classes of regression functions. In the top-left panel, we plot a fit of a logistic function

$$x \mapsto -\frac{A}{1 + e^{-ax+b}};$$

here we see the limitations of the parametric model in terms of its inability to capture the behaviour of the regression function in the range 390–550 m. The segmented linear regression fits shown in the

[2]For additional graphical illustrations, see for example http://www.nist.gov/programs-projects/differential-absorption-lidar-detection-and-quantification-greenhouse-gases as well as http://dialtechnology.info/history.html.

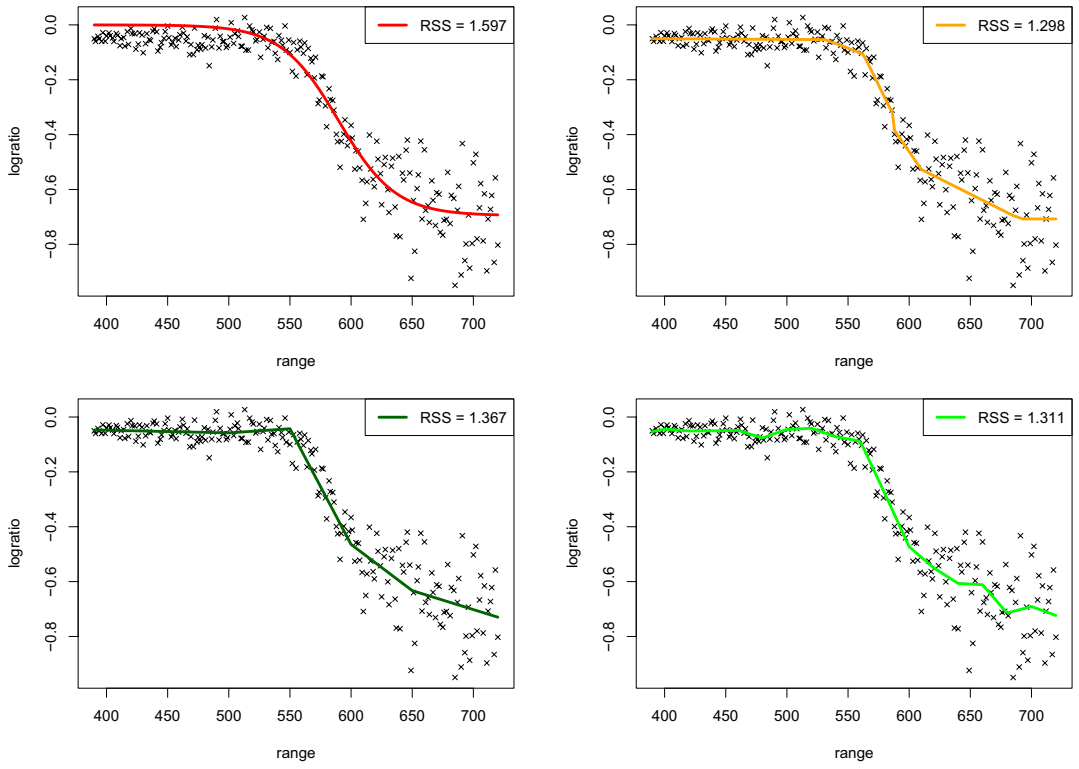

FIGURE 9 Least squares fits to the light detection and ranging dataset ($n = 221$) from Holst et al. (1996): logistic (top left), segmented linear with knots at `range` = 500, 550, 600, 650 (bottom left), segmented linear with knots at `range` = 400, 420, ..., 680, 700 (bottom right) and S-shaped (top right), along with their respective residual sums of squares

two bottom panels require the choice of a set of knots, and the left and right panels use 4 and 16 knots respectively. We see that the selection of the set of knots can have quite a significant influence, and moreover, the fits are not guaranteed to be S-shaped or even monotone. Interestingly, despite the overfitting that is apparent in the bottom-right plot of the figure, the residual sum of squares remains higher than that of the S-shaped LSE[3] illustrated in the top-right panel. Moreover, the S-shaped LSE selects the number and location of its knots adaptively, with no input required from the practitioner. Another attractive feature of the S-shaped LSE is that its theoretical guarantees presented in Theorems 1 and 2 allow for heteroscedasticity, which is clearly present in this dataset. Finally, we note that the inflection point of this LSE at `range` = 586 m yields an estimate of the distance from the LIDAR equipment to the central part of the plume, where the mercury concentration is highest.

6 | DISCUSSION

In this paper, we have developed a framework for the estimation of S-shaped regression functions and their inflection points via nonparametric least squares. In spite of the challenges

[3]Note that all the algorithms in Section 3 can be used without further modifications to compute S-shaped LSEs on any other interval $[a, b]$ besides $[0, 1]$.

of working with a non-convex shape-constrained function class, we have proposed and implemented an efficient sequential algorithm for the computation of S-shaped least squares estimators, and also established theoretical guarantees on the consistency, robustness and rates of convergence of our estimators. We will conclude by discussing some variations and possible extensions of our S-shaped regression problem that may prove to be interesting avenues for future research.

First, while our monotonicity requirement for S-shaped functions is natural in many practical applications, and useful for regulating the boundary behaviour of the least squares estimator at the endpoints of the covariate domain, much of our methodology and theory can be adapted straightforwardly to handle functions that are convex on $[0, m_0]$ and concave on $[m_0, 1]$, but not necessarily increasing on $[0, 1]$. On the computational side, our sequential strategy `SeqConReg` would still be applicable after the obvious small modifications to Step II of Algorithm 1. This modified algorithm would be justified by analogues of Propositions 3 and 4, and we could still use the mixed primal-dual bases algorithm (Algorithm 2) to sequentially compute convex LSEs on $\{(x_i, Y_i) : 1 \le i \le j\}$ and concave LSEs on $\{(x_i, Y_i) : j \le i \le n\}$ for $j \in [n]$. The theoretical results in Section 4 would also go through with some minor alterations (e.g. to the smoothness condition (13) in Assumption 2). The proofs of the oracle inequalities would be broadly the same, and the current localization argument for the inflection point result does not rely in any essential way on monotonicity near m_0. Some properties of our projection framework may need more significant adjustment, however, in order to handle potential boundary issues.

In another direction, one could consider the estimation of 'symmetric' S-shaped regression functions, by which we mean S-shaped functions f_0 with inflection point $m_0 \in (0, 1)$ such that $f_0(x) = 2f_0(m_0) - f_0(2m_0 - x)$ for $x \in [0 \vee (2m_0 - 1), (2m_0) \wedge 1]$. We believe that this additional symmetry constraint is likely to bring about considerable challenges when it comes to developing theory and algorithms for the LSE that minimizes the residual sum of squares over all symmetric S-shaped functions. In particular, unlike in our Proposition 1, it is not clear if the global minimizer in the least squares procedure can be attained at some symmetric S-shaped function with inflection point in $\{x_1, \ldots, x_n\}$. Moreover, the sequential strategy that underpins our current algorithm may no longer be valid, because in contrast to the conclusion of Proposition 4, the symmetric S-shaped LSE may not coincide with increasing convex or increasing concave LSEs on any subinterval. Theoretically, although the global risk bounds in Section 4.1 are likely to carry over even with the additional symmetry constraint, the rate of convergence of the inflection point estimator $\tilde{m}_n$ may be very different to that in Theorem 3, and may even be (nearly) parametric.

A further topic for future research could be to seek quantitative versions of the continuity result (Proposition S12) for our L^2-projection onto the class of S-shaped functions, in the spirit of the recent work of Barber and Samworth (2021) on the log-concave projection. Such a result could, for instance, provide insight into the rate at which the estimated inflection point converges to the inflection point of the projected regression function under model misspecification.

Finally, under local curvature conditions on an S-shaped function f_0 similar to those in Assumption 2, it would be of methodological and theoretical interest to be able to carry out (uniformly) asymptotically valid inference for $f_0(x)$ at fixed $x \in [0, 1]$, as well as for the inflection point m_0. For $x \neq m_0$, defining $[\tilde{u}_n(x), \tilde{v}_n(x)]$ to be the largest interval containing x on which the LSE $\tilde{f}_n$ is linear, we anticipate that the techniques of Deng et al. (2020) can be applied to obtain a limiting distribution for

$$\sqrt{n(\tilde{v}_n(x) - \tilde{u}_n(x))}(\tilde{f}_n(x) - f_0(x))$$

that does not depend on f_0, and hence construct asymptotically valid confidence intervals for $f_0(x)$. On the other hand, since m_0 marks the boundary between the convex and concave parts of f_0, we expect the problem of uncertainty quantification for m_0 and $f_0(m_0)$ to be more challenging and of a qualitatively different character. With this end in view, it is natural to seek tractable asymptotic distributions for $\tilde{m}_n$ and $\tilde{f}_n(\tilde{m}_n)$. As an initial step, one would need to refine the results in Section 4.2 by closing the logarithmic gap between the upper and lower bounds therein on the rate of convergence of $\tilde{m}_n$ to m_0. A satisfactory solution to this problem would ideally also settle the analogous problem for the plug-in mode estimator based on the unimodal LSE (Shoung & Zhang, 2001), and is likely to require significant further technical developments.

ACKNOWLEDGEMENTS

We thank the editor and three anonymous reviewers for their constructive comments and suggestions. QH was supported by the NSF grant DMS-1916221. RJC was supported by the National Cancer Institute grant U01-CA057030 and the TRIPODS grant NSF CCF-1934904 from the U.S. National Science Foundation. RJS was supported by EPSRC grants EP/P031447/1 and EP/N031938.

ORCID

Oliver Y. Feng https://orcid.org/0000-0003-0039-7039
Yining Chen https://orcid.org/0000-0003-1697-1920
Raymond J. Carroll https://orcid.org/0000-0002-5465-9682
Richard J. Samworth https://orcid.org/0000-0003-2426-4679

REFERENCES

Archontoulis, S.V. & Miguez, F.E. (2015) Nonlinear regression models and applications in agricultural research. *Agronomy Journal*, 107, 786–798.

Balabdaoui, F., Jankowski, H., Pavlides, M., Seregin, A. & Wellner, J.A. (2011) On the Grenander estimator at zero. *Statistica Sinica*, 21, 873–899.

Balász, G., Gyögy, A. & Szepesvári, C. (2015) Near-optimal max-affine estimators for convex regression. *Proceedings of Machine Learning Research*, 38, 56–64.

Baraud, Y. & Birgé, L. (2016) Rates of convergence of rho-estimators for sets of densities satisfying shape constraints. *Stochastic Processes and their Applications*, 12, 3888–3912.

Barber, R.F. & Samworth, R.J. (2021) Local continuity of log-concave projection, with applications to estimation under model misspecification. *Bernoulli*, 27, 2437–2472.

Barlow, R.E., Bartholomew, D.J., Bremner, J.M. & Brunk, H.D. (1972) *Statistical inference under order restrictions.*. New York: Wiley.

Bellec, P.C. (2018) Sharp oracle inequalities for least squares estimators in shape restricted regression. *Annals of Statistics*, 46, 745–780.

Cao, L., Shi, P.-J., Li, L. & Chen, G. (2019) A new flexible sigmoidal growth model. *Symmetry*, 11, 204.

Chatterjee, S. & Lafferty, J. (2019) Adaptive risk bounds in unimodal regression. *Bernoulli*, 25, 1–25.

Chatterjee, S., Guntuboyina, A. & Sen, B. (2015) On risk bounds in isotonic and other shape restricted regression problems. *Annals of Statistics*, 43, 1774–1800.

Chen, Y. & Samworth, R.J. (2014) scar: shape-constrained additive regression: a maximum likelihood approach. R package version 0.2-1. Available from: https://CRAN.R-project.org/package=scar.

Chen, Y. & Samworth, R.J. (2016) Generalized additive and index models with shape constraints. *Journal of the Royal Statistical Society, Series B*, 78, 729–754.

Christopoulos, D.T. (2016) On the efficient identification of an inflection point. *International Journal of Mathematics and Scientific Computing*, 6, 13–20.

Christopoulos, D.T. (2019) inflection: finds the inflection point of a curve. R package version 1.3.5. Available from: https://cran.r-project.org/web/packages/inflection.

Cule, M., Samworth, R. & Stewart, M. (2010) Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society, Series B (with discussion)*, 72, 545–607.

Deng, H., Han, Q. & Sen, B. (2020) Inference for local parameters in convexity constrained models. Available from: https://arxiv.org/abs/2006.10264.

Dümbgen, L., Hüsler, A. & Rufibach, K. (2007) Active set and EM algorithms for log-concave densities based on complete and censored data. Available from: https://arxiv.org/abs/0707.4643v4.

Edner, H., Faris, G.W., Sunesson, A. & Svanberg, S. (1989) Atmospheric atomic mercury monitoring using differential absorption lidar techniques. *Applied Optics*, 28, 921–930.

Edner, H., Ragnarson, P., Svanberg, S., Wallinder, E., Deliso, A., Ferrara, R. & Maserti B.E. (1992) Differential absorption lidar mapping of atmospheric atomic mercury in Italian geothermal fields. *Journal of Geophysical Research*, 97, 3779–3786.

Feng, O.Y., Guntuboyina, A., Kim, A.K.H. & Samworth, R.J. (2021) Adaptation in multivariate log-concave density estimation. *Annals of Statistics*, 49, 129–153.

Feng, O.Y., Chen, Y., Han, Q., Carroll, R.J. & Samworth, R.J. (2021a) Sshaped: estimation of an S-shaped function. R package version 0.99. Available from: https://CRAN.R-project.org/package=Sshaped.

Feng, O.Y., Chen, Y., Han, Q., Carroll, R.J. & Samworth, R.J. (2021b) Supplementary material to 'Nonparametric, tuning-free estimation of S-shaped functions'. Available from: https://arxiv.org/pdf/2107.07257.pdf.

Fraser, D.A.S. & Massam, H. (1989) A mixed primal-dual bases algorithm for regression under inequality constraints. Application to concave regression. *Scandinavian Journal of Statistics*, 16, 65–74.

Frisch, R. (1964) *Theory of production*. Berlin: Springer Science & Business Media.

van Genuchten, M.Th. & Gupta, S.K. (1993) A reassessment of the crop tolerance response function. *Journal of the Indian Society of Soil Science*, 41, 730–737.

Ghosal, P. & Sen, B. (2017) On univariate convex regression. *Sankhya Series A*, 79, 215–253.

Gibbs, M.N. (2000) Variational Gaussian process classifiers. *IEEE Transactions on Neural Networks*, 11, 1458–1464.

Ginsberg, W. (1974) The multiplant firm with increasing returns to scale. *Journal Economic Theory*, 9, 283–292.

Groeneboom, P. (1996) Inverse problems in statistics. In: Proceedings of the St. Flour Summer School in Probability. Lecture Notes in Mathematics, vol. 1648, pp. 67–164.

Groeneboom, P., Jongbloed, G. & Wellner, J.A. (2008) The support reduction algorithm for computing nonparametric function estimates in mixture models. *Scandinavian Journal of Statistics*, 35, 385–399.

Guntuboyina, A. & Sen, B. (2013) Global risk bounds and adaptation in univariate convex regression. *Probability Theory and Related Fields*, 163, 379–411.

Guntuboyina, A. & Sen, B. (2018) Nonparametric shape-restricted regression. *Statistical Science*, 33, 568–594.

Han, Q. (2021) Set structured global empirical risk minimizers are rate optimal in general dimensions. *Annals of Statistics*, 49, 2642–2671.

Han, Q. & Kato, K. (2021) Berry–Esseen bounds for Chernoff-type non-standard asymptotics in isotonic regression. *Annals of Applied Probability*, to appear.

Han, Q. & Wellner, J.A. (2016) Multivariate convex regression: global risk bounds and adaptation. Available from: https://arxiv.org/abs/1601.06844.

Han, Q., Wang, T., Chatterjee, S. & Samworth, R.J. (2019) Isotonic regression in general dimensions. *Annals of Statistics*, 47, 2440–2471.

Holst, U., Hössjer, O., Björklund, C., Ragnarson, P. & Edner, H. (1996) Locally weighted least squares kernel regression and statistical evaluation of LIDAR measurements. *Environmetrics*, 7, 401–416.

Jarne, G., Sanchez-Choliz, J. & Fatas-Villafranca, F. (2007) "S-shaped" curves in economic growth. A theoretical contribution and an application. *Evolutionary and Institutional Economics Review*, 3, 239–259.

Kachouie, N.N. & Schwartzman, A. (2013) Non-parametric estimation of a single inflection point in noisy observed signal. *Journal of Electrical and Electronic Systems*, 2, 1–9.

Kim, A.K.H., Guntuboyina, A. & Samworth, R.J. (2018) Adaptation in log-concave density estimation. *Annals of Statistics*, 46, 2279–2306.

Kulikov, V.N. & Lopuhaä, H.P. (2006) The behavior of the NPMLE of a decreasing density near the boundaries of the support. *Annals of Statistics*, 34, 742–768.

Kur, G., Gao, F., Guntuboyina, A. & Sen, B. (2020) Convex regression in multidimensions: suboptimality of least squares estimators. Available from: https://arxiv.org/abs/2006.02044.
Lee, U., Carroll, R.J., Marder, K., Wang, Y. & Garcia, T.P. (2020) Estimating disease onset from change points of markers measured with error. *Biostatistics*, 22, 819–835.
Liao, X. & Meyer, M. (2016) ShapeChange: change-point estimation using shape-restricted splines. R package version 1.4. Available from: https://CRAN.R-project.org/package=ShapeChange.
Liao, X. & Meyer, M. (2017) Change-point estimation using shape-restricted regression splines. *Journal of Statistical Planning and Inference*, 188, 8–21.
Meyer, M. (1999) An extension of the mixed primal-dual bases algorithm to the case of more constraints than dimensions. *Journal of Statistical Planning and Inference*, 81, 13–31.
Moreau, J.J. (1962) Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires. *Comptes rendus de l'Académie des Sciences*, 255, 238–240.
Nocedal, J. & Wright, S.J. (2006) *Numerical optimization*, 2nd edn. New York: Springer–Verlag.
Pananjady, A. & Samworth, R.J. (2021) Isotonic regression with unknown permutations: statistics, computation, and adaptation. *Annals of Statistics*, to appear.
Pya, N. & Wood, S.N. (2015) scam: Shape Constrained Additive Models. R package version 1.2–11. Available from: https://CRAN.R-project.org/package=scam.
Rockafellar, R.T. (1997) *Convex analysis*. Princeton, NJ: Princeton University Press.
Ruppert, D., Sheather, S.J. & Wand, M.P. (1995) An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90, 1257–1270.
Ruppert, D., Wand, M.P. & Carroll, R.J. (2003) *Semiparametric regression*. Cambridge: Cambridge University Press.
Samworth, R.J. (2018) Recent progress in log-concave density estimation. *Statistical Science*, 33, 493–509.
Shoung, J.-M. & Zhang, C.H. (2001) Least squares estimators of the mode of a unimodal regression function. *Annals of Statistics*, 29, 648–665.
Smith, J.O. (2010) *Physical audio signal processing*. Stanford: W3K Publishing.
Stout, Q.F. (2008) Unimodal regression via prefix isotonic regression. *Computational Statistics & Data Analysis*, 53, 289–297.
Tarde, G. (1903) *The laws of imitation*. New York: H. Holt & Co.
Wand, M. (2018) SemiPar: Semiparametic Regression. R package version 1.0–4.2. Available from: https://CRAN.R-project.org/package=SemiPar.
Yagi, D., Chen, Y., Johnson, A.L. & Morita, H. (2019) An axiomatic nonparametric production function estimator: modeling production in Japan's cardboard industry. Available from: https://arxiv.org/abs/1906.08359.
Yagi, D., Chen, Y., Johnson, A.L. & Kuosmanen, T. (2020) Shape-constrained kernel-weighted least squares: estimating production functions for Chilean manufacturing industries. *Journal of Business & Economic Statistics*, 38, 43–54.
Zeidi, B. (1993) Analysis of growth equations. *Forest Science*, 39, 594–616.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Feng, O.Y., Chen, Y., Han, Q., Carroll, R.J. & Samworth, R.J. (2022) Nonparametric, tuning-free estimation of S-shaped functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 00, 1–29. https://doi.org/10.1111/rssb.12481

APPENDIX A

A MIXED PRIMAL-DUAL BASES ALGORITHM

In this section, we describe a mixed primal-dual bases algorithm to compute the L^2-projection of a line segment onto the polyhedral convex cone of increasing convex sequences. This underpins our `SeqConReg` algorithm in Section 3. Our starting point is the following standard characterization of projections onto general closed, convex cones (e.g. Groeneboom, 1996; Moreau, 1962, Corollary 2.1). Here and below, we write $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ for the standard Euclidean norm and inner product on $\mathbb{R}^n$ for some $n \in \mathbb{N}$.

Lemma 1 *Let $\Lambda \subseteq \mathbb{R}^n$ be a closed, convex cone. For each $y \in \mathbb{R}^n$, there exists a unique projection of y onto Λ, given by $\Pi_\Lambda(y) = \mathrm{argmin}_{u \in \Lambda} \|u - y\|$, and we have the following:*

a. $\Pi_\Lambda(y)$ is the unique $\widehat{y} \in \Lambda$ for which $\langle v, y - \widehat{y} \rangle \leq 0$ for all $v \in \Lambda$ and $\langle \widehat{y}, y - \widehat{y} \rangle = 0$.
b. Suppose in addition that Λ is *finitely generated*, that is $\Lambda = \{\sum_{\ell=1}^r \lambda_\ell v^\ell : \lambda_1, \ldots, \lambda_r \geq 0\}$ for some generators $v^1, \ldots, v^r \in \Lambda$. Then $\widehat{y} = \Pi_\Lambda(y)$ if and only if $\widehat{y} = \sum_{\ell=1}^r \widehat{\lambda}_\ell v^\ell$ for some $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_r \geq 0$, and $\langle v^\ell, y - \widehat{y} \rangle \leq 0$ for all ℓ, with $\langle v^\ell, y - \widehat{y} \rangle = 0$ for any ℓ such that $\widehat{\lambda}_\ell > 0$.

In Lemma 1(b), the vector $(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_r)$ is the minimizer of the quadratic function $(\lambda_1, \ldots, \lambda_r) \mapsto \|y - \sum_{\ell=1}^r \lambda_\ell v^\ell\|^2$ over the convex set $[0, \infty)^r$. When this constrained minimization problem is written in Lagrangian form, the associated KKT optimality conditions (e.g. Rockafellar, 1997, Theorem 28.3) correspond precisely to the three conditions in (a) that uniquely define $\Pi_\Lambda(y)$, namely (i) $\widehat{y} \in \Lambda$ (*primal feasibility*); (ii) $y - \widehat{y} \in \{u \in \mathbb{R}^n : \langle u, v \rangle \leq 0 \text{ for all } v \in \Lambda\}$, the *polar cone* of Λ (*dual feasibility*); and (iii) $\langle \widehat{y}, y - \widehat{y} \rangle = 0$ (*complementary slackness*).

Given $(x_1, Y_1), \ldots, (x_n, Y_n) \in [0, 1] \times \mathbb{R}$ with $x_1 < \cdots < x_n$, we now fix $j \in [n]$ and work with the cone Λ^j of increasing convex sequences based on $x_1, \ldots, x_j$, as defined in Equation (6). The projection of $(Y_1, \ldots, Y_j)$ onto Λ^j is $(\widehat{f}_{1,j}(x_1), \ldots, \widehat{f}_{1,j}(x_j))$, where $\widehat{f}_{1,j}$ is the increasing convex LSE based on $\{(x_i, Y_i) : i \in [j]\}$. The generators of Λ^j are $\pm u^0, u^1, \ldots, u^{j-1} \in \mathbb{R}^j$, where $u^0 = \mathbf{1}$ and $u_i^\ell = (x_i - x_\ell)^+$ for all $i \in [j]$ and $\ell \in [j-1]$. Since $u^0, u^1, \ldots, u^{j-1}$ are linearly independent, every $v \equiv (v_1, \ldots, v_j) \in \mathbb{R}^j$ can be represented uniquely in the form $v = \sum_{\ell=0}^{j-1} \lambda_\ell u^\ell$, where

$$\lambda_0 \equiv \lambda_0(v) = v_1; \quad \lambda_1 \equiv \lambda_1(v) = \frac{v_2 - v_1}{x_2 - x_1}; \quad \lambda_\ell \equiv \lambda_\ell(v) = \frac{v_{\ell+1} - v_\ell}{x_{\ell+1} - x_\ell} - \frac{v_\ell - v_{\ell-1}}{x_\ell - x_{\ell-1}}, \quad 2 \leq \ell \leq j-1, \tag{16}$$

so that $v \in \Lambda^j$ if and only if $\lambda_\ell(v) \geq 0$ for all $\ell \in [j-1]$; this is the *primal feasibility* condition from Lemma 1. For each $v = \sum_{\ell=0}^{j-1} \lambda_\ell u^\ell \in \mathbb{R}^j$, the unique $g_v \in \mathcal{G}[x_1, \ldots, x_j]$ satisfying $v = (g_v(x_1), \ldots, g_v(x_j))$ has a knot at x_ℓ if and only if $\lambda_\ell \neq 0$, so we refer to $A(v) := \{1 \leq \ell \leq j-1 : \lambda_\ell \neq 0\}$ as the set of *knots* of v (or 'active indices').

The following useful property of the projection map $\Pi_{\Lambda^j} : \mathbb{R}^j \to \Lambda^j$ can be derived easily from Lemma 1. A general version of this result for arbitrary closed, convex sets is stated as Lemma S17.

Lemma 2 *Let $A \subseteq [j-1]$ and $v', v'' \in \mathbb{R}^j$ be such that $A(\Pi_{\Lambda^j}(v)) = A$ for each $v \in \{v', v''\}$. Then for all $v \in [v', v''] := \{(1-t)v' + tv'' : t \in [0, 1]\}$, we have $A(\Pi_{\Lambda^j}(v)) = A$ and, defining the linear subspace $\mathcal{L}_A := \mathrm{span}\{u^\ell : \ell \in A \cup \{0\}\} = \{v \in \mathbb{R}^j : A(v) \subseteq A\}$, we have $\Pi_{\Lambda^j}(v) = \Pi_{\mathcal{L}_A}(v)$.*

Remark 1 For $A \subseteq [j-1]$, the orthogonal projection onto the linear subspace $\mathcal{L}_A$ is represented by $P_A := U_A(U_A^\top U_A)^{-1} U_A^\top \in \mathbb{R}^{j \times j}$, where $U_A \in \mathbb{R}^{j \times (|A|+1)}$ is the matrix obtained by

extracting the columns of $U := (u^0\ u^1 \cdots\ u^{j-1}) \in \mathbb{R}^{j\times j}$ indexed by $A \cup \{0\}$. By taking $v' = v''$ in Lemma 2, we recover a version of Ghosal and Sen (2017) Proposition 2.1): suppose that we are given $v \in \mathbb{R}^j$ and have oracle knowledge of $A \equiv A(\Pi_{\Lambda^j}(v))$, that is, the locations of the knots of $\Pi_{\Lambda^j}(v)$. Then to compute $\Pi_{\Lambda^j}(v)$, we can note that $\Pi_{\Lambda^j}(v) = P_A v = \sum_{\ell=0}^{j-1} \widehat{\lambda}_\ell u^\ell$, where $\widehat{\lambda}_\ell \equiv \widehat{\lambda}_\ell^A(v) := \lambda_\ell(P_A v)$ for $0 \le \ell \le j-1$, so that $\widehat{\lambda}_\ell = 0$ for all $\ell \notin A$ and

$$(\widehat{\lambda}_\ell : \ell \in A \cup \{0\}) = (U_A^\top U_A)^{-1} U_A^\top v = \operatorname*{argmin}_{(\lambda_\ell : \ell \in A\cup\{0\})} \sum_{i=1}^n \left(v_i - \lambda_0 - \sum_{\ell \in A} \lambda_\ell (x_i - x_\ell)^+ \right)^2 \tag{17}$$

solves an ordinary (*unconstrained*) least squares problem.

Observe now that if $v(0), v(1) \in \mathbb{R}^j$ are arbitrary and $v(t) := (1-t)v(0) + tv(1)$ for all $t \in (0,1)$, then $t \mapsto \Pi_{\Lambda^j}(v(t))$ is a continuous, piecewise affine function from $[0,1]$ to Λ^j. Indeed, by Lemma 2 (and the continuity of projections onto closed, convex cones), there exist $0 = t_0' < t_1' < \cdots < t_{s+1}' = 1$ and distinct subsets $A_0', A_1', \ldots, A_s' \subseteq [j-1]$ such that for each $0 \le r \le s$, we have $\Pi_{\Lambda^j}(v(t)) = \Pi_{\mathcal{L}_{A_r'}}(v(t)) = P_{A_r'} v(t)$ for all $t \in [t_r', t_{r+1}']$.

Suppose that we are given $v(0), v(1) \in \mathbb{R}^j$ and the projection $\Pi_{\Lambda^j}(v(0)) \in \Lambda^j$, and now seek to compute $\Pi_{\Lambda^j}(v(1))$. The reasoning in the previous paragraph suggests that we can proceed as in Algorithm 2 below.

Algorithm 2. Mixed primal-dual bases algorithm to compute projections onto the cone Λ^j.

(I) Starting at $t = t_0 := 0$, define $\hat{v}_0(t_0) := \Pi_{\Lambda^j}\big(v(0)\big)$ and let the initial active set be $A_0 := A\big(\hat{v}_0(t_0)\big)$, so that $\hat{v}_0(t_0) = \Pi_{\Lambda^j}\big(v(0)\big) = P_{A_0} v(0)$.

(II) For $r \in \mathbb{N}_0$, suppose inductively that at $t = t_r$, we are given that $\hat{v}_r(t_r) := \Pi_{\Lambda^j}\big(v(t_r)\big) = P_{A_r} v(t_r)$ for some $A_r \subseteq [j-1]$. Let $\hat{v}_r(t) := P_{A_r} v(t) = \hat{v}_r(t_r) - (t - t_r) P_{A_r} u$ for $t \in [t_r, 1]$, where $u := v(0) - v(1)$, and

$$t_{r+1} := \sup\big\{ t \ge t_r : \lambda_\ell\big(\hat{v}_r(s)\big) \ge 0,\ \langle u^\ell, v(s) - \hat{v}_r(s)\rangle \le 0 \text{ for all } s \in [t_r, t] \text{ and } \ell \in [j-1] \big\}. \tag{18}$$

By Lemma 9 and the fact that $\hat{v}_r(t_r) = \Pi_{\Lambda^j}\big(v(t_r)\big)$, the set on the right-hand side always contains t_r. In order to compute t_{r+1} explicitly, observe that for all $t \in [t_r, t_{r+1}]$, we have

(i) *Primal feasibility*: $\beta_\ell(t) := \lambda_\ell\big(\hat{v}_r(t)\big) = \lambda_\ell\big(\hat{v}_r(t_r) - (t - t_r) P_{A_r} u\big) = \beta_\ell(t_r) - (t - t_r)\, \lambda_\ell(P_{A_r} u) \ge 0$ for every $\ell \in [j-1]$, where equality holds if $\ell \in A_r^c$;

(ii) *Dual feasibility*: $\gamma_\ell(t) := \langle u^\ell, v(t) - \hat{v}_r(t)\rangle = \langle u^\ell, (I - P_{A_r}) v(t)\rangle = \gamma_\ell(t_r) - (t - t_r)\, \hat{\zeta}_\ell^{A_r}(u) \le 0$ for every $0 \le \ell \le j-1$, where equality holds if $\ell \in A_r \cup \{0\}$, and $\hat{\zeta}_\ell^{A_r}(u) := \langle u^\ell, (I - P_{A_r}) u\rangle$.

In particular, $\beta_\ell(t), \gamma_\ell(t)$ depend linearly on $t \in [t_r, t_{r+1}]$, so

$$t_{r+1} = t_r + \left(\min \left\{ \frac{\beta_\ell(t_r)}{\hat{\lambda}_\ell^{A_r}(u)} : \ell \in [j-1],\ \hat{\lambda}_\ell^{A_r}(u) > 0 \right\} \wedge \min \left\{ \frac{\gamma_\ell(t_r)}{\hat{\zeta}_\ell^{A_r}(u)} : \ell \in A_r^c,\ \hat{\zeta}_\ell^{A_r}(u) < 0 \right\} \right), \tag{19}$$

observing that $\hat{\zeta}_\ell^{A_r}(u) = \langle (I - P_{A_r})u^\ell, u \rangle = 0$ for $\ell \in A_r \cup \{0\}$.

(iii) *Complementary slackness* is maintained throughout this step: $\langle \hat{v}_r(t), v(t) - \hat{v}_r(t) \rangle = \langle P_{A_r} v(t), (I - P_{A_r}) v(t) \rangle = 0$, so $\Pi_{\Lambda^j}\big(v(t)\big) = \hat{v}_r(t) = P_{A_r} v(t)$ for all $t \in [t_r, t_{r+1}]$ by Lemma 9.

(III) If $t_{r+1} \geq 1$, then return $\hat{v}_r(1) = \hat{v}_r(t_r) - (1 - t_r) P_{A_r} u$ and terminate the algorithm. Otherwise, go to (IV), noting that when t approaches t_{r+1} from below, either

- A *primal variable* $\beta_\ell(t)$ with $\ell \in A_r$ is about to hit 0 and turn negative, or
- A *dual variable* $\gamma_\ell(t)$ with $\ell \in A_r^c$ is about to hit 0 and turn positive.

(IV) *Changing the 'active set'*: Define $A_r^- := \{\ell \in A_r : \beta_\ell(t_{r+1}) = 0\}$ and $A_r^+ := \{\ell \in A_r^c : \gamma_\ell(t_{r+1}) = 0\}$.

(i) If $|A_r^- \cup A_r^+| = 1$, then repeat (II) and (III) with $r+1$ in place of r and $A_{r+1} := (A_r \setminus A_r^-) \cup A_r^+$, observing that $\Pi_{\Lambda^j}\big(v(t_{r+1})\big) = P_{A_r} v(t_{r+1}) = P_{A_{r+1}} v(t_{r+1})$.

(ii) If $|A_r^- \cup A_r^+| > 1$, i.e. there is a *degeneracy* at t_{r+1}, then choose $A^\pm \subseteq A_r^\pm$ and carry out (II) with $r+1$ in place of r and $A_{r+1} = (A_r \setminus A^-) \cup A^+$. In doing so, if (19) yields a strict increase in t, then let the algorithm continue from there and pass to (III). Otherwise, retry this for different pairs of subsets $A^\pm \subseteq A_r^\pm$ until we can move a strictly positive distance in the next iteration of (II).

When defining the primal variables $\beta_\ell(t)$ in (II), it is convenient here that every $v \in \Lambda^j$ has a unique primal representation, which in this case is given by Equation (16). The same is true of any cone in $\mathbb{R}^j$ generated by $\pm\tilde{u}^0, \ldots, \pm\tilde{u}^{q-1}, \tilde{u}^q, \ldots, \tilde{u}^{j-1}$, for some linearly independent $\tilde{u}^0, \tilde{u}^1, \ldots, \tilde{u}^{j-1}$. Thus, Algorithm 2 is applicable to all such cones, provided that the 'active sets' are taken to be subsets of $\{q, q{+}1, \ldots, j{-}1\}$ (Fraser & Massam, 1989), so in particular, it can also be used to compute isotonic and convex LSEs (in a sequential manner, as described in Section 3). Indeed, the sequential application of this mixed primal-dual bases algorithm to the monotone cone $\Theta^\uparrow$ from the proof of Corollary S2 yields the widely-used, linear time 'pool adjacent violators' algorithm (PAVA) (Barlow et al., 1972). Moreover, with appropriate modifications, Algorithm 2 can be extended to general polyhedral cones (Meyer, 1999) and polyhedral convex sets.

Lemma 3 *Algorithm 2 always terminates after finitely many steps with the correct solution* $\Pi_{\Lambda^j}(v(1))$.

This follows from (i)–(iii) in Stage (II) and the following two observations:

(iv) The algorithm does not get stuck at any of the thresholds t_r; that is, when $t = t_r$ for some r, there is always a subsequent iteration of (II) that strictly increases t;
(v) At distinct thresholds t_r, the corresponding 'active sets' A_r are distinct subsets of $[j-1]$.

We will justify (iv) and (v) in Section S4 in the supplementary material, where we also exploit the specific structure of Λ^j to handle the degeneracies mentioned in Stage (IV)(b); see in particular modification (IV') and Proposition S18.