The Annals of Statistics
2016, Vol. 44, No. 6, 2401–2432
DOI: 10.1214/15-AOS1409
© Institute of Mathematical Statistics, 2016

NONPARAMETRIC ESTIMATION OF DYNAMICS OF MONOTONE TRAJECTORIES

BY DEBASHIS PAUL[1], JIE PENG[2] AND PRABIR BURMAN[3]

University of California, Davis

We study a class of nonlinear nonparametric inverse problems. Specifically, we propose a nonparametric estimator of the dynamics of a monotonically increasing trajectory defined on a finite time interval. Under suitable regularity conditions, we show that in terms of L^2-loss, the optimal rate of convergence for the proposed estimator is the same as that for the estimation of the derivative of a function. We conduct simulation studies to examine the finite sample behavior of the proposed estimator and apply it to the Berkeley growth data.

1. Introduction. Monotone trajectories representing the evolution of states over time appear widely in scientific studies, particularly, in the study of growth of organisms such as humans or plants [11, 23, 32]. There are many parametric models for modeling the growth trajectories or modeling their rate of change, that is, the derivative of the trajectories [16, 23]. Other examples of monotone trajectories appear in population dynamics under negligible resource constraints [35], in dose-response analysis in pharmacokinetics [19], in auction price dynamics in e-Commerce [17, 20, 38], and in analysis of trajectories of aircrafts after take-off [26].

Our goal in this paper is to estimate the functional relationship between the rate of change and the state, that is, the dynamics of the trajectory, through a nonparametric model. Many systems such as growth of organisms or economic activity of a country/region are intrinsically dynamic in nature. A dynamics model provides a mechanistic description of the system rather than a purely phenomenological one. Due to insufficient scientific knowledge, quite often there is a need for nonparametric modeling of the dynamical system, which can also be used to develop measures of goodness-of-fit for hypothesized parametric models.

A key observation is that for any smooth monotone trajectory $x(\cdot)$, its dynamics can be described by a first-order autonomous differential equation:

$$x'(t) = \big(x' o x^{-1}\big)\big(x(t)\big) = g\big(x(t)\big), \qquad t \in [0, 1], \tag{1}$$

Received January 2015; revised October 2015.

[1]Supported in part by NSF Grants DMR-1035468, DMS-11-06690 and DMS-14-07530.

[2]Supported in part by NSF Grants DMS-10-01256 and DMS-11-48643.

[3]Supported in part by NSF Grants DMS-09-07622 and DMS-11-48643.

MSC2010 subject classifications. Primary 62G08; secondary 62G20.

Key words and phrases. Autonomous differential equation, nonlinear inverse problem, monotone trajectory, nonparametric estimation, perturbation theory, spline.

where $g = x'ox^{-1}$ is the gradient function. In this paper, we propose to estimate the unknown gradient function g nonparametrically by representing it in a B-spline basis where the number of basis functions grows with the sample size. We adopt a nonlinear least squares approach for model fitting. We then carry out a detailed theoretical analysis and derive the rate of convergence of the proposed estimator.

We now highlight the major theoretical developments of this work. Here, we are dealing with a nonlinear nonparametric inverse problem. Although there is a large literature on linear nonparametric inverse problems [4, 5, 10, 18], especially on the nonparametric estimation of the derivative of a curve [12, 14, 25], there is very little theoretical development on nonlinear nonparametric inverse problems. Thus, our work makes an important contribution to this area. Specifically, we first quantify the degree of ill-posedness of the estimation of the gradient function g as the number of basis functions grows to infinity. We then use this result to show that if g is p times differentiable then the L^2-risk of the proposed estimator has the same optimal rate of convergence, namely, $O(n^{-2p/(2p+3)})$, as that of the estimation of the derivative of a trajectory, assuming that the latter is $p+1$ times differentiable. We also show that this optimal rate is indeed the minimax rate for the estimation of g under L^2 loss if the class of estimators is restricted to be uniformly Lipschitz.

There is an extensive literature on nonparametric estimation of monotone functions including [1, 22, 29, 30, 39]. However, most of these works are not concerned with the estimation of the gradient function of the trajectory, except for [30], which modeled the trajectory in terms of a second-order differential equation, and obtained an estimate of the gradient as a byproduct.

Although there are many works for fitting parametric differential equations [2, 3, 6, 7, 13, 31, 40, 41, 43], relatively few works exist for nonparametric ODE modeling. Among the latter, [44] dealt with estimating a parametric ODE with smooth time-varying parameters. Wu et al. [42] proposed a sparse additive model for describing the dynamics of a multivariate state vector and developed a combination of two-stage smoothing and sparse penalization for fitting the model. Their model can be seen as a multidimensional generalization of the autonomous ODE model studied here. For the theory in their paper, it is assumed that whenever the gradient function g is p times differentiable, the state x is at least $3p+1$ times differentiable. However, the representation $g = x'ox^{-1}$ [see (1)] implies that g is p times differentiable if and only if x is $p+1$ times differentiable. Therefore, the assumptions made in [42] are not satisfied if p is the maximal order of smoothness of g, and hence the rate of convergence derived there is suboptimal.

A possible alternative route for a nonparametric estimation of g is through the two-stage procedure where the trajectories and their derivatives are first estimated nonparametrically, and then the ODE is fitted by regressing the fitted derivatives to the fitted trajectories [6, 7, 36]. Cao and Zhao [3] gave a comprehensive theoretical analysis of such an approach, and [15] proposed a computationally tractable one-step estimation procedure that mitigates some statistical inefficiencies of two-stage

estimators. In spite of their simplicity, two-stage estimators are often unsatisfactory partly due to the difficulty of resolving the bias-variance trade-off in a data-dependent way. In all the numerical studies carried out in this paper, the proposed estimator performs better than the two-stage estimator.

The rest of the paper is organized as follows. In Section 2, we briefly describe the model and the estimation procedure. We present the main theoretical results in Section 3 and outline the proof in Section 4. We present a simulation study in Section 5 and an application to the Berkeley growth data in Section 6. We discuss some related issues in Section 7. Some proof details are provided in the Appendix. Additional derivations and more detailed summaries of simulation results are provided in a supplementary material (SM) [28].

2. Model. The class of models studied in this paper is of the form

$$x_g'(t) = g\big(x_g(t)\big), \qquad x_g(0) = x_0, \qquad t \in [0,1], \tag{2}$$

where g is an unknown smooth function which is assumed to be positive on the range of $\{x_g(t) : t \in [0,1]\}$. Therefore, the sample trajectory $x_g(t)$ is a strictly increasing function of time t. The observations are

$$Y_j = x_g(t_j) + \varepsilon_j, \qquad j = 1, \ldots, n, \tag{3}$$

where $0 \le t_1, \ldots, t_n \le 1$ are observation times. The noise terms ε_j's are assumed to be independent with mean 0 and variance $\sigma_\varepsilon^2 > 0$.

Our goal is to estimate the gradient function g based on the observed data Y_js. We propose to approximate g through a basis representation:

$$g(y) \approx g_{\boldsymbol{\beta}}(y) := \sum_{k=1}^{M} \beta_k \phi_{k,M}(y), \tag{4}$$

where $\{\phi_{k,M}(\cdot)\}_{k=1}^M$ is the rescaled B-spline basis, that is, a B-spline basis with equally spaced knots on an interval such that the functions are rescaled to have L^2-norm equal to 1. Henceforth, we use ϕ_k to denote $\phi_{k,M}$.

Since g is assumed to be either strictly positive or strictly negative throughout its domain, an alternative approach is to represent $\log g$ in a smooth basis. It can be shown that the corresponding estimate has the same rate of convergence. A brief sketch of the implementation of this approach is given in Section S6 of the SM [28].

We now describe the proposed estimator. For the time being, assume that we observe the two endpoints $x_0 = x_g(0)$ and $x_1 = x_g(1)$ noiselessly and so the combined support of $\{\phi_1, \ldots, \phi_M\}$ is the interval $[x_0, x_1]$. Given any $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_M)$ so that $g_{\boldsymbol{\beta}}$ is positive on the support of $\{\phi_k(\cdot)\}_{k=1}^M$, we can solve the initial value problem:

$$x'(t) = g_{\boldsymbol{\beta}}\big(x(t)\big), \qquad t \in [0,1], \qquad x(0) = x_0 \tag{5}$$

to obtain the corresponding trajectory $x(t) \equiv x(t; \boldsymbol{\beta})$. From now onward, we use $x(\cdot; \boldsymbol{\beta})$ to denote the solution of (5). Define the L^2 loss function:

$$L(\boldsymbol{\beta}) := \sum_{j=1}^{n} \big(Y_j - x(t_j; \boldsymbol{\beta})\big)^2. \tag{6}$$

Then the proposed *nonlinear least squares* estimator of g is defined as

$$\hat{g}(y) := g_{\hat{\boldsymbol{\beta}}}(y) = \sum_{k=1}^{M} \hat{\beta}_k \phi_{k,M}(y) \qquad \text{where } \hat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^M} L(\boldsymbol{\beta}). \tag{7}$$

Minimization of $L(\boldsymbol{\beta})$ is a nonlinear least squares problem. We use a Levenberg–Marquardt iterative updating scheme, which is known to be very stable for solving nonlinear regression problems (cf. [27]). This method requires evaluating the trajectory $x(t; \boldsymbol{\beta})$ and its derivative with respect to $\boldsymbol{\beta}$: Given the current estimate of $\boldsymbol{\beta}$, we solve the corresponding differential equations numerically by the fourth-order Runge–Kutta method. Details of the model fitting procedure are given in Appendix A.4.

In the following, we briefly discuss some key steps of the estimation procedure and the basic steps toward establishing the consistency and rates of convergence of the proposed estimator. Formal derivations and expositions are presented in the subsequent sections.

2.1. *Asymptotic results*. We establish consistency and derive rates of convergence in terms of L^2-loss of the estimator of g (Theorem 3.1). Specifically, we show that, for sufficiently smooth g and an appropriate range of values of M, there exists a local optimum of the objective function, which is also a global optimum in a neighborhood of the g^*, the point closest to g in L^2-norm within the model space. Moreover, this local optimum converges to the true g at a rate the same as the optimal rate of convergence of the estimation of the derivative function $x'(\cdot)$. The key steps and techniques are:

(i) using a linear approximation of the trajectory in terms of the parameter $\boldsymbol{\beta}$ near the optimal point $\boldsymbol{\beta}^*$ within the model space, accomplished by using the perturbation theory of differential equations;

(ii) quantifying the degree of ill-conditioning of the Hessian matrix of the loss function with respect to $\boldsymbol{\beta}$ as a function of M. This is achieved by using a novel technique involving a version of Halperin–Pitt inequalities [24];

(iii) using the result in (ii) to get a local quadratic approximation of $\hat{g}$ (equivalently $\hat{\boldsymbol{\beta}}$) in a neighborhood of the optimal point g^* ($\boldsymbol{\beta}^*$), which encompasses the bias and variance terms.

2.2. *Initial estimator of g.* Since we are dealing with a nonconvex optimization problem, it is important to specify a reasonable initial estimate for g (equivalently, $\boldsymbol{\beta}$). We consider a two-stage estimator, where, in the first stage we estimate $x(t)$ and its derivative $x'(t)$ using local polynomial smoothing and in the second stage we regress $\hat{x}'(t)$ on the basis $(\phi_k(\hat{x}(t)))_{k=1}^{\tilde{M}}$. We use $\tilde{M}$ here to distinguish it from the M used in the proposed nonlinear least squares estimator. The procedure is described in detail in Section 3.5, where we also show that with a proper choice $\tilde{M}$, the two-stage estimator is close enough to the optimal approximation $g^* = g_{\boldsymbol{\beta}^*}$ such that the proposed estimator, with this initial estimate, achieves the optimal rate.

2.3. *Boundary issue.* There is a fundamental limitation in estimation of g on the boundary, even when the value of x_g at the boundaries, that is, $x_0 = x(0), x_1 = x(1)$, are known exactly. Note the solution $x_g(t)$ of (5) can be expressed as $x_g(t) = x_0 + \int_0^t g(x_g(s))\,ds$. Thus, if t is close to zero, $x_g(t)$ and $x_{\hat{g}}(t)$, where $\hat{g}$ is an estimate of g based on the observed data, can be very close even when g and $\hat{g}$ are quite different. A similar phenomenon takes place when t is close to 1. Therefore, there is an intrinsic limitation (for any estimation procedure) in the accuracy that can be achieved in estimating $g(y)$ when y is in a neighborhood of x_0 and x_1.

For carrying out the theoretical analysis, we thus modify the loss function $L_\delta(\boldsymbol{\beta})$ (6) to include only those time points t_j that are within the interval $[\delta, 1-\delta]$ for some small positive number δ. In practice, we may choose δ to be the time point such that about 5% of the data fall in the intervals $[0, \delta]$ and $[1-\delta, 1]$. Throughout the paper, δ is treated as a fixed quantity.

Let $\hat{x}_{0,\delta}$ and $\hat{x}_{1,\delta}$ denote the estimates of $x_{0,\delta} := x(\delta)$, and $x_{1,\delta} := x(1-\delta)$, respectively. We then define

$$x_{0,M} = \hat{x}_{0,\delta} - \eta_M, \qquad x_{1,M} = \hat{x}_{1,\delta} + \eta_M, \tag{8}$$

where η_M is a small positive number satisfying $\eta_M = o(M^{-1})$. The estimation of $x_{0,\delta}, x_{1,\delta}$ and the choice of η_M are discussed in detail in Section 3.6.

We set the combined support of the basis functions $\{\phi_{k,M}\}_{k=1}^M$ as the interval $[x_{0,M}, x_{1,M}]$, and use the following modified loss function to derive an estimator for g:

$$\tilde{L}_\delta(\boldsymbol{\beta}) = \sum_{j=1}^{n} \big(Y_j - x(t_j; \boldsymbol{\beta}, \hat{x}_{0,\delta})\big)^2 \mathbf{1}_{[\delta,1-\delta]}(t_j), \tag{9}$$

where $x(t; \boldsymbol{\beta}, a)$ denotes the integral curve of the ODE:

$$x'(t) = g_{\boldsymbol{\beta}}\big(x(t)\big), \qquad t \in [\delta, 1-\delta], \qquad x(\delta) = a. \tag{10}$$

The estimated $\hat{g}$ is through minimizing the above loss function with respect to $\boldsymbol{\beta}$ [i.e., equation (7) with L replaced by $\tilde{L}_\delta$].

An alternative specification of the support of $g_{\boldsymbol{\beta}}$ is to treat the end points of the support as additional parameters in the loss function $L(\boldsymbol{\beta})$ defined in (6). Then $\hat{\boldsymbol{\beta}}$ is obtained by minimizing $L(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ as well as the end points.

2.4. *Model selection.* As in all nonparametric estimation problems, model selection, here the choice of the number of basis functions M, is an important issue. We propose to use the Mallow's C_p criterion for selecting M:

$$C_p := \frac{\sum_{j=1}^{n}(Y_j - x(t_j; \hat{\boldsymbol{\beta}}_M))^2}{\hat{\sigma}_\varepsilon^2} - (n - 2M),$$

where $\hat{\boldsymbol{\beta}}_M$ is the estimator under M basis functions and $\hat{\sigma}_\varepsilon^2$ is an unbiased estimator of the noise variance σ_ε^2. Utilizing the smoothness of the trajectories, we propose to use: $\hat{\sigma}_\varepsilon^2 = \frac{1}{2(n-1)} \sum_{j=2}^{n}(Y_j - Y_{j-1})^2$.

3. Consistency. In this section, we discuss the consistency and rates of convergence of the estimator $\hat{g}$ defined by the loss function (9). The asymptotic framework is that the number of basis functions M goes to infinity together with the number of measurements n. The consistency of the estimator $\hat{g}$ over $[x_{0,\delta}, x_{1,\delta}]$ is formulated in terms of the convergence of the L^2-loss:

$$\int_{x_{0,\delta}}^{x_{1,\delta}} |\hat{g}(u) - g(u)|^2 \, du \longrightarrow 0 \qquad \text{in probability as } n \to \infty.$$

In Theorem 3.1, we derive an upper bound on the rate of convergence of the L^2-loss as $n, M \to \infty$ that depends upon the degree of smoothness of g. Specifically, the optimal rate is $O_P(n^{-2p/(2p+3)})$ for $p > 3$ when g is p times continuously differentiable. Here, and henceforth, we use the phrase "p-times continuously differentiable", or the notation $f(\cdot) \in C^p$, to mean that p is the degree of smoothness of the function f, that is, $\lfloor p \rfloor$th derivative $f^{(\lfloor p \rfloor)}$ of f exists and is bounded on the domain of f and $|f^{(\lfloor p \rfloor)}(y) - f^{(\lfloor p \rfloor)}(z)| \le K|y - z|^{p - \lfloor p \rfloor}$ for all y, z in the domain of f, for some positive constant K. Here, for any $p \in \mathbb{R}$, $\lfloor p \rfloor$ denotes the largest integer less than or equal to p.

3.1. *Assumptions.* The following assumptions are being made.

A1. $g \in C^p(D)$, and $g > 0$ on D for some $p \ge 2$, where D is an open interval containing $[x_0, x_1]$.

A2. The collection of basis functions $\Phi_M := \{\phi_{1,M}, \ldots, \phi_{M,M}\}$ satisfies:
   - (i) $\phi_{k,M}$'s have unit L^2 norm;
   - (ii) the combined support of Φ_M is $D_0 \equiv D_{0,M} := [x_{0,M}, x_{1,M}]$ and for every k, the length of the support of $\phi_{k,M}$ is $O(M^{-1})$;
   - (iii) $\phi_{k,M} \in C^2(D_0)$ for all k;
   - (iv) $\sup_{u \in D_0} \sum_{k=1}^{M} |\phi_{k,M}^{(j)}(u)|^2 = O(M^{1+2j})$, for $j = 0, 1, 2$;
   - (v) the Gram matrix $\mathbf{G}_{\Phi_M} := ((\int_{x_{0,M}}^{x_{1,M}} \phi_{k,M}(u)\phi_{l,M}(u)\, du))_{k,l=1}^{M}$ is such that there exist constants $0 < \underline{c} \le \overline{c} < \infty$, not depending on M such that $\underline{c} \le \lambda_{\min}(\mathbf{G}_{\Phi_M}) \le \lambda_{\max}(\mathbf{G}_{\Phi_M}) \le \overline{c}$ for all M;

(vi) for every M, there is a $\boldsymbol{\beta}^* \in \mathbb{R}^M$ (referred to as the optimal point) such that $\sup_{t\in[\delta,1-\delta]} |x_g(t) - x(t;\boldsymbol{\beta}^*)| = O(M^{-(p+1)})$ and $\sup_{u\in[x_{0,\delta},x_{1,\delta}]} |g^{(j)}(u) - g^{(j)}_{\boldsymbol{\beta}^*}(u)| = O(M^{-p+j})$ for $j = 0, 1, 2$, where $g_{\boldsymbol{\beta}} = \sum_{k=1}^{M} \beta_k \phi_{k,M}$ and $x(t;\boldsymbol{\beta}) \equiv x(t;\boldsymbol{\beta}, x_{0,\delta})$ with $x(t;\boldsymbol{\beta}, a)$ being the solution of (10).

A3. The observation times $\{t_j\}_{j=1}^n$ are realizations of $\{T_j\}_{j=1}^n$, where T_j's are i.i.d. from a continuous distribution F_T supported on $[0, 1]$ with a density f_T satisfying $\underline{c}' \le f_T \le \overline{c}'$ for some $0 < \underline{c}' \le \overline{c}' < \infty$.

A4. The noise ε_j's are i.i.d. sub-Gaussian random variables with mean 0 and variance $\sigma_\varepsilon^2 > 0$.

In A4, we follow the definition of sub-Gaussianity as given in [37]. A brief summary of properties of such random variables is given in Section S4 of the SM [28].

A1 ensures sufficient smoothness of the solution paths of the differential equation (2): By A1, $x_g(\cdot)$ has smoothness index $p + 1$. A2(i) to A2(vi) are satisfied by a normalized B-spline basis of order $\ge \max\{3, p\}$ with equally spaced knots on the interval $[x_{0,M}, x_{1,M}]$ where the basis functions are rescaled to have unit L^2 norm. To show that such a basis satisfiesA2(vi) is nontrivial and this is done in Appendix A.2. This result, which relies on the approximation property of splines [8, 21] is key to quantifying the estimation bias when using a spline basis. As shown in Section 3.6, A2(ii) ensures that the combined support of the basis functions covers the range of the data used in estimating g. A2(vi) ensures that a solution $x(t;\boldsymbol{\beta})$ of (10) on $t \in [\delta, 1 - \delta]$ exists for all $\boldsymbol{\beta}$ sufficiently close to the optimal point $\boldsymbol{\beta}^*$. This allows us to apply the perturbation theory of differential equations to bound the fluctuations of the sample paths when we perturb the parameter $\boldsymbol{\beta}$. A3 allows us to work with the random variables $\tilde{T}_j$ defined as T_j conditional on $T_j \in [\delta, 1 - \delta]$ with the conditional density $\tilde{f}_T(t) = f_T(t)/(F_T(1 - \delta) - F_T(\delta))$. The properties of f_T ensure that $\tilde{f}_T$ satisfies the same properties on $[\delta, 1 - \delta]$ with possibly modified values of the constants c_1 and c_2. It should be noted that the key derivations leading to the consistency of $\hat{g}$ are conditional on $\mathbf{T}$ and therefore A3 is only for mathematical convenience. The main asymptotic result (Theorem 3.1) holds if instead of being randomly distributed, the time points form a fixed regular grid, say, with equal spacing. A4 is a fairly standard assumption that allows calculation of the metric entropy used in proving Theorem 4.1, which shows the existence of an estimator $\hat{g}$ with a near-optimal convergence rate.

3.2. *Rate of convergence.* The estimation of $g(\cdot)$ is a nonlinear inverse problem since $x'(\cdot)$ is not directly observable. In addition, this is also an ill-posed estimation problem.

Let $x^{\boldsymbol{\beta}}(\cdot;\boldsymbol{\beta})$ be the partial derivative of $x(\cdot;\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, where $x(\cdot;\boldsymbol{\beta}) \equiv x(\cdot;\boldsymbol{\beta}, x_{0,\delta})$ is the solution of (10) with $x(0) = x_{0,\delta}$. Let $\boldsymbol{\beta}^* \in \mathbb{R}^M$ be the optimal point as in A2. Define

$$G_* := \mathbb{E}\big(x^{\boldsymbol{\beta}}(\tilde{T}_1;\boldsymbol{\beta}^*)\big(x^{\boldsymbol{\beta}}(\tilde{T}_1;\boldsymbol{\beta}^*)\big)^T\big), \tag{11}$$

where the expectation is with respect to the distribution of $\tilde{T}_1$. Clearly, G_* is a positive semi-definite matrix. The degree of ill-posedness of the estimation problem is determined by the size of the operator norm of the matrix G_*^{-1} as a function of M. The following proposition gives a precise quantification of the degree of ill-posedness. The situation here is in contrast with standard nonparametric function estimation problems where the corresponding matrix is well conditioned.

PROPOSITION 3.1. *Suppose that assumptions* A1–A3 *hold with* $p \geq 2$. *Assume further that* (a) $\max_{j=0,1} |x_{j,M} - x_{j,\delta}| = o(M^{-1})$ (*a.s.*) *and* (b) $\min\{x_{1,M} - x_{1,\delta}, x_{0,\delta} - x_{0,M}\} \gg M^{-3/2}$. *Then*

$$\|G_*^{-1}\| = O(M^2) \qquad a.s. \tag{12}$$

Under the condition of Theorem 3.1, (a) and (b) of Proposition 3.1 hold if the estimators $\hat{x}_j$ of $x_{j,\delta}$ as in Lemma 3.1 and η_M as in Section 3.6 are used in the definition of $x_{j,M}$ (8) ($j = 1, 2$). See Section 3.6 for details.

We now state the main result on the consistency and rate of convergence of the estimator $\hat{g}$.

THEOREM 3.1. *Suppose that the observed data* $\{(t_j, Y_j) : j = 1, \ldots, n\}$ *follow the model described by equations* (2) *and* (3) *and that* A1–A4 *are satisfied with* $p > 3$. *Further, suppose that* M *satisfies* $c'(n/\sigma_\varepsilon^2)^{1/(2p+3)} \leq M \leq c''(n/\sigma_\varepsilon^2 \log n)^{1/7}$ *for some* $c', c'' > 0$. *Then as* $n \to \infty$, *with probability tending to one, there exists a local minimum* $\hat{\boldsymbol{\beta}}$ *of the objective function* $\tilde{L}_\delta(\boldsymbol{\beta})$ [*defined by* (9)], *which is also a global minimum within radius* cM^{-2} (*for some* $c > 0$) *of* $\boldsymbol{\beta}^*$ [*defined in* A2(vi)] *such that, with* $\hat{g} := g_{\hat{\boldsymbol{\beta}}}$, *we have*

$$\begin{aligned} &\int_{x_{0,\delta}}^{x_{1,\delta}} \big(\hat{g}(u) - g(u)\big)^2 \, du \\ &\quad = O_P\left(\frac{\sigma_\varepsilon^2 M^3}{n}\right) + O_P\big(M^{-2p}\big) + O_P\big(M^2(\sigma_\varepsilon^2/n)^{2(p+1)/(2p+3)}\big), \end{aligned} \tag{13}$$

with the optimal rate given by $O_P((\sigma_\varepsilon^2/n)^{2p/(2p+3)})$, *which is obtained when* $M = c(n/\sigma_\varepsilon^2)^{1/(2p+3)}$ *for some* $c > 0$.

REMARK 3.1. Assuming σ_ε to be a constant, the optimal rate of convergence of $\hat{g}$ is the same as the optimal rate in terms of the L^2-loss for estimating $x'(t)$ based on the data $\{Y_j : j = 1, \ldots, n\}$ given by (2) when $x \in C^{p+1}([0, 1])$. The fact that an estimator of g can attain this rate can be anticipated from the representation of g as $g = x_g' o x_g^{-1}$. It should be noted that when $2 \leq p \leq 3$, we can establish that the rate of convergence of $\hat{g}$ is of the order $O_P((\sigma_\varepsilon^2/n)^{2p/(2p+3)} \log n)$ (see Theorem 4.1). However, in this case, the presence of the factor $\log n$ is conjectured to be suboptimal. A formal argument, given in Section 3.3, shows that, at least if the

class of estimators is restricted to be uniformly Lipschitz, the rate of convergence $\hat{g}$ is the optimal rate when $p > 3$ and B-splines of sufficiently high order are used to model g.

REMARK 3.2. We give a brief explanation of the difference in the upper bounds of rates for $p \in [2, 3]$ and $p > 3$. After proving the existence of a local minimizer $\hat{g} \equiv g_{\hat{\boldsymbol{\beta}}}$ in Theorem 4.1, we obtain a refinement in its rate of convergence through a Taylor series expansion of the gradient of the loss function $\tilde{L}_\delta(\boldsymbol{\beta})$ (details are given in Section S2 of SM [28]). This expansion gives rise to a second-order representation of $\hat{\boldsymbol{\beta}}$. However, only when $p > 3$, we are able to control higher order terms in this expansion, so that their contributions to the L^2 loss for $\hat{g}$ are dominated by the first-order bias and variance terms. On the other hand, for $p \in [2, 3]$, we do not have such a representation and instead, we are only able to provide the rate of convergence given in Theorem 4.1.

The main steps of the proof of Theorem 3.1 are given in Section 4, with further details provided in the supplementary material [28].

3.3. *Lower bound on convergence rate.* In this subsection, we show that if the estimators of the gradient function g are restricted to a class of uniformly Lipschitz function (which includes the proposed estimator), then the minimax rate for estimation of g is of the order $n^{-2p/(2p+3)}$, and thus the proposed estimator is optimal within this class of estimators. Accordingly, we first specify the function class for g as

$$\mathcal{G} = \big\{g : D \to \mathbb{R}_+ : c_0 \le g \le c_1;\ |g'| \le c_2;\ g \in C^p(D)\big\}, \tag{14}$$

where $0 < c_0 < c_1 < \infty$ and $0 < c_2 < \infty$ are constants. Define the class of uniformly Lipschitz functions

$$\mathcal{L} = \big\{h : D \to \mathbb{R} : |h(y) - h(z)| \le c_4|y - z| \text{ for all } y, z \in D\big\},$$

where $c_4 \in (0, \infty)$ depends on (at least as large as) c_2 in (14). If $g \in \mathcal{G}$, then we have $x_g \in C^{p+1}([0, 1])$ and $x_g' \in C^p([0, 1])$. In addition, we assume the observation model (3) with the noise ε_i's are i.i.d. sub-Gaussian with mean zero and variance σ_ε^2.

Let δ be as in Section 2. By the condition $c_0 \le g \le c_1$, for $0 < c_0 < c_1$, we know that there exist $c_0(\delta) < c_1(\delta)$ such that $c_0(\delta) \le x_g(t) \le c_1(\delta)$ for all $t \in [\delta, 1 - \delta]$, for all $g \in \mathcal{G}$. Note that, we can take $c_0(\delta) = x_g(0)$. Define, $\|f\|_{2,\delta} = (\int_\delta^{1-\delta} (f(t))^2\,dt)^{1/2}$. Then there are constants $c_2(\delta), c_3(\delta) > 0$ such that for any given estimator $\hat{g} \in \mathcal{L}$ of g,

$$\begin{aligned} c_2(\delta)\|\hat{g}ox_g - gox_g\|_{2,\delta}^2 &\le \int_{x_g(\delta)}^{x_g(1-\delta)} \big|\hat{g}(u) - g(u)\big|^2\,du \\ &\le c_3(\delta)\|\hat{g}ox_g - gox_g\|_{2,\delta}^2. \end{aligned} \tag{15}$$

Recall that $gox_g = x'_g$.

On the other hand, since $x_g(\cdot) \in C^{p+1}([0,1])$, there exists (cf. [33]) an estimator $\hat{x}_{\mathrm{op}}(\cdot)$ with the property that, given $\epsilon > 0$, there exists a constant $K_1(\epsilon) > 0$ such that

$$\sup_{g\in\mathcal{G}} \mathbb{P}\big(\|\hat{x}_{\mathrm{op}} - x_g\|_{2,\delta}^2 > K_1(\epsilon) n^{-2(p+1)/(2p+3)}\big) < \epsilon \tag{16}$$

for all $n \geq N_1(\epsilon)$.

We define the estimator $\tilde{x}' := \hat{g}o\hat{x}_{\mathrm{op}}$ for x'_g. Then, by the triangle inequality,

$$\begin{aligned} \|\hat{g}ox_g - gox_g\|_{2,\delta} &= \|\hat{g}ox_g - x'_g\|_{2,\delta} \\ &\geq \|\tilde{x}' - x'_g\|_{2,\delta} - \|\hat{g}o\hat{x}_{\mathrm{op}} - \hat{g}ox_g\|_{2,\delta} \\ &\geq \|\tilde{x}' - x'_g\|_{2,\delta} - c_4\|\hat{x}_{\mathrm{op}} - x_g\|_{2,\delta}, \end{aligned} \tag{17}$$

where, in the last step we have used the fact that $\hat{g} \in \mathcal{L}$.

Since $x'_g \in C^p([0,1])$, the minimax rate of estimation of x'_g in terms of the L^2 loss $\|\cdot\|_{2,\delta}^2$ is of the order $n^{-2p/(2p+3)}$. This can be derived directly for g restricted to $\mathcal{G}$ by only slightly modifying the arguments in [33]. Combining this fact with (15), (16) and (17), we obtain that there exists $K_2 > 0$, such that

$$\lim_{n\to\infty} \inf_{\hat{g}\in\mathcal{L}} \sup_{g\in\mathcal{G}} \mathbb{P}\left(\int_{x_g(\delta)}^{x_g(1-\delta)} \big|\hat{g}(u) - g(u)\big|^2 du > K_2 n^{-2p/(2p+3)}\right) > 0.$$

In other words, as long as $\hat{g}$ is uniformly Lipschitz, the rate $n^{-2p/(2p+3)}$ is a lower bound on the rate for estimating g in terms of the L^2-loss. We note that, the requirement $\hat{g} \in \mathcal{L}$ can be relaxed by requiring that this holds with probability approaching one as $n \to \infty$. The latter is satisfied by the estimator we proposed. Thus, combining with Theorem 3.1, we deduce that the optimal rate of estimation of g is $n^{-2p/(2p+3)}$ for $p > 3$.

3.4. *Asymptotic variance of* $\hat{g}$. Using a consistent root $\hat{\boldsymbol{\beta}}$ and the equation $\nabla L(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0$, we can derive an approximate expression for the asymptotic variance of $\hat{g}(\cdot)$. Specifically, by the asymptotic representation of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ appeared in the proof of Theorem 3.1 (see Section S2 of SM [28]), ignoring higher order terms and the contribution of the model bias, and finally evaluating the expressions at $\hat{\boldsymbol{\beta}}$ instead of $\boldsymbol{\beta}^*$ (which is unknown), we have with M basis functions

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_M) \approx D(\hat{\boldsymbol{\beta}}_M) := \hat{\sigma}_{\varepsilon,M}^2 \left[\sum_{j=1}^n \left(\frac{\partial x(t_j;\hat{\boldsymbol{\beta}}_M)}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial x(t_j;\hat{\boldsymbol{\beta}}_M)}{\partial \boldsymbol{\beta}}\right)^T\right]^{-1}. \tag{18}$$

Here, the estimated noise variance $\hat{\sigma}_{\varepsilon,M}^2$ can be computed as the mean squared error $(n-M)^{-1}\sum_{j=1}^n (Y_j - x(t_j;\hat{\boldsymbol{\beta}}_M))^2$. Expression (18) allows us to obtain an

approximate asymptotic variance for $\hat{g}_M(y)$ by $V(y) := \boldsymbol{\phi}(y)^T D(\hat{\boldsymbol{\beta}}_M)\boldsymbol{\phi}(y)$, for any given y, where $\boldsymbol{\phi}(y) = (\phi_{1,M}(y), \ldots, \phi_{M,M}(y))^T$, $y \in \mathbb{R}$. Note that the formula appearing in equation (18) is the standard sampling variance estimator for Gauss–Newton optimization of a nonlinear least squares method.

3.5. *Initial estimator of g*. In Theorem 3.1, we prove the rate of convergence for a local minimizer, which is a global minimizer within a radius of $O(M^{-2})$ of the optimal point $\boldsymbol{\beta}^*$ for a suitable range of values of M. Therefore, we need an initial estimate which resides within this domain. In the following, we describe one way of obtaining such an initial estimate, through a two-stage approach, which is similar in spirit to the approaches by [6, 7].

We first estimate $x(t)$ and $x'(t)$ by local polynomial smoothing [12] and denote these estimates by $\hat{x}(t)$ and $\hat{x}'(t)$, respectively. One may also use spline-based or other nonparametric approaches. Then we fit the regression model

$$\hat{x}'(T_j) = \boldsymbol{\phi}\big(\hat{x}(T_j)\big)^T \boldsymbol{\beta} + e_j, \qquad j = 1, \ldots, n \tag{19}$$

by ordinary least squares, where $\boldsymbol{\phi} = (\phi_{1,\tilde{M}}, \ldots, \phi_{\tilde{M},\tilde{M}})$. We refer to the resulting estimator $\tilde{\boldsymbol{\beta}}$ as the two-stage estimator of $\boldsymbol{\beta}$:

$$\begin{aligned}\tilde{\boldsymbol{\beta}} = {} & \left[\sum_{j=1}^{n} \boldsymbol{\phi}\big(\hat{x}(T_j)\big)\boldsymbol{\phi}\big(\hat{x}(T_j)\big)^T \mathbf{1}_{[\delta,1-\delta]}(T_j)\right]^{-1} \\ & \times \left(\sum_{j=1}^{n} \hat{x}'(T_j)\boldsymbol{\phi}\big(\hat{x}(T_j)\big)\mathbf{1}_{[\delta,1-\delta]}(T_j)\right).\end{aligned} \tag{20}$$

Since $x(\cdot) \in C^{p+1}$ and $x'(\cdot) \in C^p$ (by A1), and $\{\varepsilon_j\}$ is sub-Gaussian, with the optimal choice of bandwidths, we have

$$n^{-1}\sum_{j=1}^{n}\big(\hat{x}(T_j) - x_g(T_j)\big)^2 \mathbf{1}_{[\delta,1-\delta]}(T_j) = O_P\big((\sigma_\varepsilon^2/n)^{2(p+1)/(2p+3)}\big), \tag{21}$$

$$n^{-1}\sum_{j=1}^{n}\big(\hat{x}'(T_j) - x_g'(T_j)\big)^2 \mathbf{1}_{[\delta,1-\delta]}(T_j) = O_P\big((\sigma_\varepsilon^2/n)^{2p/(2p+3)}\big). \tag{22}$$

We state the following result about the rate of convergence of the two-stage estimator. The proof is given in Section S3 of the SM [28].

PROPOSITION 3.2. *Suppose that* $p \geq 2$ *and* A1–A4 *hold and that the two-stage estimator of g is given by* $\tilde{g} := g_{\tilde{\boldsymbol{\beta}}}$ *where* $\tilde{\boldsymbol{\beta}}$ *is defined in* (20). *Then, supposing that* $1 \leq \tilde{M} \ll n^{(p+1)/(2(2p+3))}$, *we have*

$$\int_{x_{0,\delta}}^{x_{1,\delta}} \big|\tilde{g}(u) - g(u)\big|^2 du = O_P(\tilde{\alpha}_n^2), \tag{23}$$

where

$$\tilde{\alpha}_n = \max\{(\sigma_\varepsilon^2/n)^{p/(2p+3)}, \tilde{M}^{-p}\}. \tag{24}$$

When $\sigma_\varepsilon \asymp 1$, the optimal $\tilde{\alpha}_n$ is of the order $n^{-p/(2p+3)}$, obtained by setting $\tilde{M} \asymp \tilde{M}_* = n^{1/(2p+3)}$. For all $p \geq 2$ this rate matches the lower bound on rate of convergence for estimating g reported in Section 3.3. This also shows that the rate of convergence of $\tilde{g}$ is faster than $O(\tilde{M}^{-2})$ if $\tilde{M} \ll n^{p/(2(2p+3))}$ if $p > 2$. So, for this range of $\tilde{M}$, which includes $\tilde{M}_*$, with a high probability, the two-stage estimator is within a ball of radius $O(\tilde{M}^{-2})$ around $\boldsymbol{\beta}^*$, over which $\hat{g}$ is a global optimizer of (9).

3.6. *Estimation of x_g at δ and $1-\delta$.* We estimate $x_{0,\delta} := x_g(\delta)$ and $x_{1,\delta} := x_g(1-\delta)$ by local polynomial smoothing and denote the estimators as $\hat{x}_{0,\delta}$ and $\hat{x}_{1,\delta}$, respectively. Recall that [equation (8)], $x_{0,M} = \hat{x}_{0,\delta} - \eta_M$ and $x_{1,M} = \hat{x}_{1,\delta} + \eta_M$, where η_M is a small positive number satisfying $\eta_M = o(M^{-1})$ which implies that $x_0 < x_{0,M} < x_{1,M} < x_1$ as n goes to infinity. At the same time, η_M should be large enough so that $\max_{j=0,1} |\hat{x}_{j,\delta} - x_{j,\delta}| = o_P(\eta_M)$ which ensures that $x_{0,M} < x_{0,\delta} < x_{1,\delta} < x_{1,M}$ and $\max_{j=0,1} |x_{j,\delta} - x_{j,M}| = O_P(\eta_M) = o_P(M^{-1})$ as n goes to infinity. For some technical considerations, to be utilized later, we also want $\eta_M \gg M^{-3/2}$. In practice, we may select η_M to be $\min\{M^{-3/2}\log n, s_M/\log n\}$ where s_M is the length of the smallest support among the basis functions $\{\phi_1, \ldots, \phi_M\}$. In addition, we also assume that $\hat{x}_{j,\delta}$, $j = 0, 1$ are estimated from a sample independent from that used in estimating $\boldsymbol{\beta}$. This can be easily achieved in practice by sub-sampling of the measurements. This assumption enables us to prove the consistency result (Theorem 3.1) conditionally on $\hat{x}_{j,\delta}$, $j = 0, 1$ and treating them as nonrandom sequences converging to $x_{j,\delta}$, $j = 0, 1$.

We have the following result with regard to the estimation of $x_{0,\delta}$ and $x_{1,\delta}$ by local polynomial smoothing (see, e.g., [12]). Define

$$\xi_n := n^{-(p+1)/(2p+3)}\sqrt{\log n}. \tag{25}$$

LEMMA 3.1. *Suppose that* A1, A3 *and* A4 *hold. Consider using a kernel of sufficient degree of smoothness to obtain estimates $\hat{x}_{j,\delta}$ for $x_{j,\delta}$, $j = 1, 2$, through local polynomial smoothing method with bandwidth of order $n^{-1/(2p+3)}$. Define $d_n := \max_{j=0,1} |\hat{x}_{j,\delta} - x_{j,\delta}|$. Then $d_n = O_P(n^{-(p+1)/(2p+3)})$ and given $\zeta > 0$, there exists $C(\zeta) > 0$ such that $d_n \leq C(\zeta)\xi_n$ with probability at least $1 - n^{-\zeta}$, where ξ_n is as in* (25).

If M is as in Theorem 3.1, then since $M^{-3/2} \ll \eta_M \ll M^{-1}$, it can be checked that $\xi_n = o(\eta_M)$ as $n \to \infty$. This ensures that $D_0 = [x_{0,M}, x_{1,M}]$ is within the interval $[x_0, x_1]$ a.s. for large enough n, and hence the properties of the function g hold on D_0. In addition, D_0 contains the interval $[x_{0,\delta}, x_{1,\delta}]$. Therefore, A2(ii) ensures that the combined support of the basis functions covers the range of the data used in estimating g.

4. Proofs. In this section, we outline the main steps of the proofs. Some technical details are deferred to the Appendix.

4.1. *Proof of Proposition* 3.1. For convenience of notation, we use $x_*(t)$ to denote the sample path $x(t;\boldsymbol{\beta}^*)$ and $x(t)$ to mean $x(t;\boldsymbol{\beta})$ Similarly, $x^{\beta_\ell}(t)$ is used to denote $x^{\beta_\ell}(t;\boldsymbol{\beta})$. Using the representation of $x^{\boldsymbol{\beta}}(\cdot;\boldsymbol{\beta})$ through (55) in Appendix A.1,

$$x^{\beta_\ell}(t) = g_{\boldsymbol{\beta}}\big(x(t)\big)\int_{x_0}^{x(t)} \frac{\phi_\ell(x)}{(g_{\boldsymbol{\beta}}(x))^2}\,dx, \qquad \ell = 1,\dots,M,$$

in order to prove Proposition 3.1, it suffices to find a lower bound on

$$\min_{\|\mathbf{b}\|=1}\int_\delta^{1-\delta}\bigg[\int_\delta^t g_{\mathbf{b}}\big(x_*(s)\big)/g_{\boldsymbol{\beta}^*}\big(x_*(s)\big)\,ds\bigg]^2 \tilde{f}_T(t)\,dt,$$

where $g_{\mathbf{b}}(u) = \mathbf{b}^T\boldsymbol{\phi}(u)$ with $\boldsymbol{\phi} = (\phi_1,\dots,\phi_M)^T$. By A3, without loss of generality, we can take the density $\tilde{f}_T(\cdot)$ to be uniform on $[\delta, 1-\delta]$.

We make use of the following result known as Halperin–Pitt inequality [24].

LEMMA 4.1. *If f is locally absolutely continuous and f'' is in $L_2([0,A])$, then for any $\epsilon > 0$ the following inequality holds with $K(\epsilon) = 1/\epsilon + 12/A^2$:*

$$\int_0^A \big(f'(t)\big)^2\,dt \le K(\epsilon)\int_0^A f^2(t)\,dt + \epsilon\int_0^A \big(f''(t)\big)^2\,dt. \tag{26}$$

Now defining

$$R(t) := \int_\delta^t \frac{g_{\mathbf{b}}(x_*(s))}{g_{\boldsymbol{\beta}^*}(x_*(s))}\,ds,$$

we have

$$\begin{aligned} R'(t) &:= \frac{dR(t)}{dt} = \frac{g_{\mathbf{b}}(x_*(t))}{g_{\boldsymbol{\beta}^*}(x_*(t))},\\ R''(t) &:= \frac{d^2R(t)}{dt^2} = \bigg[\frac{g'_{\mathbf{b}}(x_*(t))}{g_{\boldsymbol{\beta}^*}(x_*(t))} - \frac{g_{\mathbf{b}}(x_*(t))g'_{\boldsymbol{\beta}^*}(x_*(t))}{g^2_{\boldsymbol{\beta}^*}(x_*(t))}\bigg]x'_*(t)\\ &= \bigg[\frac{g'_{\mathbf{b}}(x_*(t))}{g_{\boldsymbol{\beta}^*}(x_*(t))} - \frac{g_{\mathbf{b}}(x_*(t))g'_{\boldsymbol{\beta}^*}(x_*(t))}{g^2_{\boldsymbol{\beta}^*}(x_*(t))}\bigg]g_{\boldsymbol{\beta}^*}\big(x_*(t)\big). \end{aligned}$$

By A2(vi), we have $\sup_{t\in[\delta,1-\delta]}|x_g(t) - x_*(t)| = O(M^{-(p+1)})$, and hence

$$\begin{aligned} x_*(1-\delta) &\le x_{1,\delta} + \big|x_*(1-\delta) - x_{1,\delta}\big| < x_{1,M}, x_*(\delta)\\ &\ge x_{0,\delta} - \big|x_*(\delta) - x_{0,\delta}\big| > x_{0,M}. \end{aligned} \tag{27}$$

Hence, using the fact that $\phi_\ell(u)$'s are $O(M^{1/2})$ and $\phi'_\ell(u)$'s are $O(M^{3/2})$, and these functions are supported on intervals of length $O(M^{-1})$, we deduce that

$$\int_\delta^{1-\delta} \big(R''(t)\big)^2\,dt = O\big(M^2\big). \tag{28}$$

An application of Lemma 4.1 with $f(t) = R(t-\delta)$ and $A = 1-2\delta$ yields

$$\begin{aligned}&\int_\delta^{1-\delta} \big(R'(t)\big)^2\,dt\\ &\quad\le \big(1/\epsilon + 12/(1-2\delta)^2\big)\int_\delta^{1-\delta} \big(R(t)\big)^2\,dt + \epsilon\int_\delta^{1-\delta} \big(R''(t)\big)^2\,dt.\end{aligned} \tag{29}$$

Take $\epsilon = k_0 M^{-2}$ for some $k_0 > 0$, then by (28),

$$\int_\delta^{1-\delta} \big(R(t)\big)^2\,dt \ge k_1 M^{-2}\int_\delta^{1-\delta} \big(R'(t)\big)^2\,dt - k_2 M^{-2},$$

for constants $k_1, k_2 > 0$ dependent on k_0. Next, we write

$$\int_\delta^{1-\delta} \big(R'(t)\big)^2\,dt = \int_{x_*(\delta)}^{x_*(1-\delta)} \frac{g^2_{\mathbf{b}}(v)}{g^3_{\boldsymbol{\beta}^*}(v)}\,dv = \int_{x_*(\delta)}^{x_*(1-\delta)} g^2_{\mathbf{b}}(v)h(v)\,dv, \tag{30}$$

where $h(v) = g^{-3}_{\boldsymbol{\beta}^*}(v)$ which is bounded below by a positive constant on the interval $[x_*(\delta), x_*(1-\delta)]$.

Observe that by (27), the combined support of $\{\phi_{k,M}\}_{k=1}^M$, viz., $[x_{0,M}, x_{1,M}]$, contains (for sufficiently large M) the interval $[x_*(\delta), x_*(1-\delta)]$. Also, $|x_{1,M} - x_*(1-\delta)| \le |x_{1,M} - x_{1,\delta}| + |x_{1,\delta} - x_*(1-\delta)| = o(M^{-1})$ and $|x_{0,M} - x_*(\delta)| \le |x_{0,M} - x_{0,\delta}| + |x_{0,\delta} - x_*(\delta)| = o(M^{-1})$. These two facts and A2(v) imply that

$$\begin{aligned}&\int_{x_*(\delta)}^{x_*(1-\delta)} g^2_{\mathbf{b}}(v)h(v)\,dv\\ &\quad\ge \Big(\inf_{v\in[x_*(\delta),x_*(1-\delta)]} h(v)\Big)\mathbf{b}^T\bigg[\int_{x_{0,M}}^{x_{1,M}} \boldsymbol{\phi}(v)\big(\boldsymbol{\phi}(v)\big)^T\,dv - o(1)\bigg]\mathbf{b}\\ &\quad\ge k_3,\end{aligned}$$

for some constant $k_3 > 0$, for sufficiently large M. Thus, by appropriate choice of ϵ, we have $\int_\delta^{1-\delta}(R(t))^2\,dt \ge k_4 M^{-2}$ for some constant $k_4 > 0$, which yields (12).

4.2. *Proof of Theorem* 3.1. The main step toward the proof of Theorem 3.1 is the following slightly weaker version which is valid for wider range of the smooth index p.

THEOREM 4.1. *Suppose that the observed data* $\{Y_j : j = 1, \ldots, n\}$ *follow the model described by equations* (2) *and* (3) *and that assumptions* A1–A4 *are satisfied with* $p \geq 2$. *Suppose further that the sequence* M *is such that*

$$c_1' \left(\frac{n}{\sigma_\varepsilon^2 \log n} \right)^{1/(2p+3)} \leq M \leq c_1'' \left(\frac{n}{\sigma_\varepsilon^2 \log n} \right)^{1/7} \tag{31}$$

for some $c_1', c_1'' > 0$, $M^{-3/2} \ll \eta_M \ll M^{-1}$, *and* ξ_n *be as defined in Lemma* 3.1. *Let* $\bar{\alpha}_n := c_2' M^{-2}$ *for some* $c_2' > 0$ (*sufficiently small*) *and*

$$\alpha_n := C_0 M \max \left\{ \sigma_\varepsilon \sqrt{\frac{M \log n}{n}}, M^{-(p+1)}, \xi_n \right\}, \tag{32}$$

for some $C_0 > 0$. *Then as* $n \to \infty$, *with probability tending to one, there exists a local minimum* $\hat{\boldsymbol{\beta}}$ *of the objective function* $\tilde{L}_\delta(\boldsymbol{\beta})$ [*defined through* (9)], *which is also a global minimum within radius* $\bar{\alpha}_n$ *of* $\boldsymbol{\beta}^*$ [*note that*, $\alpha_n \leq \bar{\alpha}_n$ *by* (31)] *such that, with* $\hat{g} := g_{\hat{\boldsymbol{\beta}}}$,

$$\int_{x_{0,\delta}}^{x_{1,\delta}} \left| \hat{g}(u) - g(u) \right|^2 du = O\left(\alpha_n^2\right). \tag{33}$$

REMARK 4.1. Assuming σ_ε to be a constant, if M is chosen to be of the order $n^{1/(2p+3)}$ if $p > 2$ [and $(n/\log n)^{1/7}$ for $p = 2$], then α_n^2 in (33) simplifies to $n^{-2p/(2p+3)} \log n$ (correspondingly, bounded by the same for $p = 2$), which is within a factor of $\log n$ of the optimal rate in terms of the L^2-loss for estimating $x'(t)$ based on the data $\{Y_j : j = 1, \ldots, n\}$ given by (2) when $x \in C^{p+1}([0, 1])$. For $p > 3$, we have the improved rate of convergence of $\hat{g}$, that is, without the $\log n$ factor, as stated in Theorem 3.1.

The main idea behind the proof of Theorem 4.1 is to obtain a lower bound on $n^{-1}(L_\delta(\boldsymbol{\beta}) - L_\delta(\boldsymbol{\beta}^*))$ which is proportional to $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2$ when $\boldsymbol{\beta}$ lies in an annular region around $\boldsymbol{\beta}^*$. The outer radius of the annular region depends on the degree of ill-conditioning of the problem, as quantified by Proposition 3.1, and the smoothness of the function g and the approximating bases, as indicated in condition A2. This lower bound then naturally leads to the conclusion about the existence and rate of convergence of a local minimizer $\hat{g}$.

Define

$$\Gamma_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \frac{1}{n} \sum_{j=1}^{n} \left(x(T_j; \boldsymbol{\beta}) - x\left(T_j; \boldsymbol{\beta}^*\right) \right)^2 \mathbf{1}_{[\delta, 1-\delta]}(T_j), \tag{34}$$

$\mathcal{A}_M(\alpha_n, \bar{\alpha}_n) = \{\boldsymbol{\beta} \in \mathbb{R}^M : \alpha_n \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq \bar{\alpha}_n\}$, and

$$D_n^* = \frac{1}{n} \sum_{j=1}^{n} \left(x_g(T_j) - x\left(T_j; \boldsymbol{\beta}^*\right) \right)^2 \mathbf{1}_{[\delta, 1-\delta]}(T_j).$$

Suppose that $\boldsymbol{\beta} \in \mathcal{A}_M(\alpha_n, \bar{\alpha}_n)$. Henceforth, we use $x(t; \boldsymbol{\beta})$ to denote $x(t; \boldsymbol{\beta}; x_{0,\delta})$ and $x_g(t)$ to denote $x_g(t; x_{0,\delta})$. Then

$$
\begin{aligned}
&\frac{1}{n} L_\delta(\boldsymbol{\beta}) - \frac{1}{n} L_\delta\big(\boldsymbol{\beta}^*\big) \\
&\quad = \frac{1}{n} \sum_{j=1}^{n} \big(Y_j - x(T_j; \boldsymbol{\beta})\big)^2 \mathbf{1}_{[\delta, 1-\delta]}(T_j) \\
&\qquad - \frac{1}{n} \sum_{j=1}^{n} \big(Y_j - x\big(T_j; \boldsymbol{\beta}^*\big)\big)^2 \mathbf{1}_{[\delta, 1-\delta]}(T_j) \\
&\quad = \frac{1}{n} \sum_{j=1}^{n} \big(x(T_j; \boldsymbol{\beta}) - x\big(T_j; \boldsymbol{\beta}^*\big)\big)^2 \mathbf{1}_{[\delta, 1-\delta]}(T_j) \\
&\qquad - \frac{2}{n} \sum_{j=1}^{n} \varepsilon_j \big(x(T_j; \boldsymbol{\beta}) - x\big(T_j; \boldsymbol{\beta}^*\big)\big) \mathbf{1}_{[\delta, 1-\delta]}(T_j) \\
&\qquad - \frac{2}{n} \sum_{j=1}^{n} \big(x_g(T_j) - x\big(T_j; \boldsymbol{\beta}^*\big)\big)\big(x(T_j; \boldsymbol{\beta}) - x\big(T_j; \boldsymbol{\beta}^*\big)\big) \mathbf{1}_{[\delta, 1-\delta]}(T_j),
\end{aligned} \tag{35}
$$

where $U_{1n}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ and $U_{2n}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$, are the second and third summations in the above expression, respectively. Next, we write

$$
\begin{aligned}
&\frac{1}{n} \tilde{L}_\delta(\boldsymbol{\beta}) - \frac{1}{n} L_\delta(\boldsymbol{\beta}) \\
&\quad = \frac{1}{n} \sum_{j=1}^{n} \big[\big(Y_j - x(T_j; \boldsymbol{\beta}, \hat{x}_{0,\delta})\big)^2 \\
&\qquad - \big(Y_j - x(T_j; \boldsymbol{\beta})\big)^2\big] \mathbf{1}_{[\delta, 1-\delta]}(T_j) \\
&\quad = \frac{1}{n} \sum_{j=1}^{n} \big(x(T_j; \boldsymbol{\beta}; \hat{x}_{0,\delta}) - x(T_j; \boldsymbol{\beta})\big)^2 \mathbf{1}_{[\delta, 1-\delta]}(T_j) \\
&\qquad - \frac{2}{n} \sum_{j=1}^{n} \varepsilon_j \big(x(T_j; \boldsymbol{\beta}; \hat{x}_{0,\delta}) - x(T_j; \boldsymbol{\beta})\big) \mathbf{1}_{[\delta, 1-\delta]}(T_j) \\
&\qquad - \frac{2}{n} \sum_{j=1}^{n} \big(x(T_j; \boldsymbol{\beta}; \hat{x}_{0,\delta}) - x(T_j; \boldsymbol{\beta})\big)\big(x_g(T_j) - x(T_j; \boldsymbol{\beta})\big) \\
&\qquad \times \mathbf{1}_{[\delta, 1-\delta]}(T_j),
\end{aligned} \tag{36}
$$

where $V_{1n}(\boldsymbol{\beta})$, $V_{2n}(\boldsymbol{\beta})$, $V_{3n}(\boldsymbol{\beta})$ are the three summations in the last expression.

From (35) and (36), we deduce that

$$
\begin{aligned}
&\frac{1}{n}\tilde{L}_\delta(\boldsymbol{\beta}) - \frac{1}{n}\tilde{L}_\delta(\boldsymbol{\beta}^*) \\
&\quad = \frac{1}{n}\bigl(L_\delta(\boldsymbol{\beta}) - L_\delta(\boldsymbol{\beta}^*)\bigr) + \frac{1}{n}\bigl(\tilde{L}_\delta(\boldsymbol{\beta}) - L_\delta(\boldsymbol{\beta})\bigr) - \frac{1}{n}\bigl(\tilde{L}_\delta(\boldsymbol{\beta}^*) - L_\delta(\boldsymbol{\beta}^*)\bigr) \\
&\quad = \Gamma_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*) - U_{1n}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) - U_{2n}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + \bigl(V_{1n}(\boldsymbol{\beta}) - V_{1n}(\boldsymbol{\beta}^*)\bigr) \\
&\qquad - \bigl(V_{2n}(\boldsymbol{\beta}) - V_{2n}(\boldsymbol{\beta}^*)\bigr) + U_{3n}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) - U_{4n}(\boldsymbol{\beta}, \boldsymbol{\beta}^*),
\end{aligned} \tag{37}
$$

where

$$
\begin{aligned}
&U_{3n}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \\
&\quad = \frac{2}{n}\sum_{j=1}^{n}\bigl(x_g(T_j) - x(T_j; \boldsymbol{\beta}^*)\bigr)\bigl(x(T_j; \boldsymbol{\beta}; \hat{x}_{0,\delta}) - x(T_j; \boldsymbol{\beta})\bigr)\mathbf{1}_{[\delta, 1-\delta]}(T_j) \\
&\qquad - \frac{2}{n}\sum_{j=1}^{n}\bigl(x_g(T_j) - x(T_j; \boldsymbol{\beta}^*)\bigr)\bigl(x(T_j; \boldsymbol{\beta}^*; \hat{x}_{0,\delta}) - x(T_j; \boldsymbol{\beta}^*)\bigr)\mathbf{1}_{[\delta, 1-\delta]}(T_j),
\end{aligned}
$$

$$
\begin{aligned}
&U_{4n}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \\
&\quad = \frac{2}{n}\sum_{j=1}^{n}\bigl(x(T_j; \boldsymbol{\beta}) - x(T_j; \boldsymbol{\beta}^*)\bigr)\bigl(x(T_j; \boldsymbol{\beta}; \hat{x}_{0,\delta}) - x(T_j; \boldsymbol{\beta})\bigr)\mathbf{1}_{[\delta, 1-\delta]}(T_j).
\end{aligned}
$$

Using the fact that $x^a(t; \boldsymbol{\beta}, a_0) := (\partial/\partial a)x(t; \boldsymbol{\beta}, a)|_{a=a_0}$ satisfies

$$
x^a(t; \boldsymbol{\beta}, a_0) = \frac{g_{\boldsymbol{\beta}}(x(t; \boldsymbol{\beta}, a_0))}{g_{\boldsymbol{\beta}}(a_0)}, \qquad t \in [\delta, 1-\delta],
$$

provided $g_{\boldsymbol{\beta}}(y) > 0$ for $y \in [a_0, x(1-\delta; \boldsymbol{\beta}, a_0)]$, we have, for all $\boldsymbol{\beta} \in \mathcal{A}_M(\alpha_n, \bar{\alpha}_n)$,

$$
\sup_{a_0 \in [x_{0,\delta} - \xi_n, x_{0,\delta} + \xi_n]} \sup_{t \in [\delta, 1-\delta]} \bigl|x(t; \boldsymbol{\beta}, a_0) - x(t; \boldsymbol{\beta}, x_{0,\delta})\bigr| \le C_1 \xi_n \tag{38}
$$

for some $C_1 > 0$. Here, we have used the fact that for $t \in [\delta, 1-\delta]$, and $a_0 \in [x_{0,\delta} - \xi_n, x_{0,\delta} + \xi_n]$,

$$
x(t; \boldsymbol{\beta}, a_0) = \tilde{G}_{\boldsymbol{\beta}}^{-1}\bigl(t - \delta + G_{\boldsymbol{\beta}}(a_0)\bigr) \qquad \text{where } \tilde{G}_{\boldsymbol{\beta}}(y) := \int_{x_{0,M}}^{y} \frac{du}{g_{\boldsymbol{\beta}}(u)}, \tag{39}
$$

and that

$$
\sup_{\boldsymbol{\beta} \in \mathcal{A}(\alpha_n, \bar{\alpha}_n)} \sup_{y \in [x_{0,M}, x_{1,M}]} \bigl|g_{\boldsymbol{\beta}}(y) - g_{\boldsymbol{\beta}^*}(y)\bigr| = O\bigl(\bar{\alpha}_n M^{1/2}\bigr) = O\bigl(M^{-3/2}\bigr),
$$

so that, by using (61), and the fact that $M^{-3/2} \ll \eta_M \ll M^{-1}$,

$$
\Bigl[x_{0,\delta} - \xi_n, \sup_{\boldsymbol{\beta} \in \mathcal{A}(\alpha_n, \bar{\alpha}_n)} \sup_{a_0 \in [x_{0,\delta} - \xi_n, x_{0,\delta} + \xi_n]} x(1-\delta; \boldsymbol{\beta}, a_0)\Bigr] \subset [x_{0,M}, x_{1,M}]
$$

for large enough M and n.

We now bound individual terms in the expansion (37). First, we have the following lower bound on $\Gamma_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$, the proof of which is given in Appendix A.3.

LEMMA 4.2. *Let $\Gamma_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ be as defined in (34). Then given $\eta > 0$, there exist constants $d_1(\eta) > 0$ and $d_2, d_3, d_4 > 0$ independent of η such that*

$$\begin{aligned}&\Gamma_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*)\\ &\quad\geq d_1(\eta)\frac{1}{M^2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2 - d_2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^4 M^2\big(1 + d_3\bar{\alpha}_n^2 M^2 + d_4 M^{-2}\big)\end{aligned} \tag{40}$$

uniformly in $\boldsymbol{\beta} \in \mathcal{A}_M(\alpha_n, \bar{\alpha}_n)$ with probability at least $1 - n^{-\eta}$.

Since $\bar{\alpha}_n M = c_2' M^{-1} = o(1)$, and the constant c_2' can be chosen to be small enough so that we can conclude from (40) that given $\eta > 0$, there exists $d_5(\eta) > 0$ such that

$$\mathbb{P}\bigg(\Gamma_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \geq \frac{d_5(\eta)}{M^2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2 \text{ for all } \boldsymbol{\beta} \in \mathcal{A}_M(\alpha_n, \bar{\alpha}_n)\bigg) \geq 1 - n^{-\eta}. \tag{41}$$

Next, by the Cauchy–Schwarz inequality, we have

$$\big|U_{2n}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)\big| \leq 2\sqrt{D_n^*}\sqrt{\Gamma_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*)}. \tag{42}$$

Next, by (38), we have

$$\max\Big\{V_{1n}(\boldsymbol{\beta}^*), \sup_{\boldsymbol{\beta} \in \mathcal{A}(\alpha_n, \bar{\alpha}_n)} V_{1n}(\boldsymbol{\beta})\Big\} \leq C_1^2 \xi_n^2, \tag{43}$$

and hence

$$\sup_{\boldsymbol{\beta} \in \mathcal{A}(\alpha_n, \bar{\alpha}_n)} \big|U_{3n}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)\big| \leq 4C_1 \xi_n \sqrt{D_n^*} \tag{44}$$

and

$$\big|U_{4n}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)\big| \leq 2C_1 \xi_n \sqrt{\Gamma_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*)}. \tag{45}$$

Next, defining

$$Z(\boldsymbol{\beta}) = \frac{\sum_{j=1}^n \varepsilon_j (x(T_j; \boldsymbol{\beta}) - x(T_j; \boldsymbol{\beta}^*)) \mathbf{1}_{[\delta, 1-\delta]}(T_j)}{\sigma_\varepsilon \sqrt{\sum_{j=1}^n (x(T_j; \boldsymbol{\beta}) - x(T_j; \boldsymbol{\beta}^*))^2 \mathbf{1}_{[\delta, 1-\delta]}(T_j)}},$$

and setting $Z(\boldsymbol{\beta})$ being zero if the denominator is zero, we have

$$\big|U_{1n}(\boldsymbol{\beta})\big| \leq \frac{2\sigma_\varepsilon}{\sqrt{n}}\sqrt{\Gamma_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*)}\big|Z(\boldsymbol{\beta})\big|. \tag{46}$$

Let $\mathcal{B}_M(\Delta; \alpha_n, \bar{\alpha}_n)$ be a Δ-net for $\mathcal{A}_M(\alpha_n, \bar{\alpha}_n)$. Then $|\mathcal{B}_M(\Delta; \alpha_n, \bar{\alpha}_n)| \leq 3(\bar{\alpha}_n/\Delta)^M$. Then, by using Lemma S.3 in the SM [28], and (41), we conclude

that given $\eta > 0$, there exist constants $c_1(\eta) > 0, C'(\eta) > 0$, and a set $A_{1\eta}$ with $\mathbb{P}(\mathbf{T} \in A_{1\eta}) \geq 1 - n^{-\eta}$, such that for all $\mathbf{T} \in A_{1\eta}$,

$$\mathbb{P}\Big(\max_{\boldsymbol{\beta} \in \mathcal{B}_M(\delta;\alpha_n,\bar{\alpha}_n)} \big|Z(\boldsymbol{\beta})\big| > c_1(\eta)\sqrt{M \log(\bar{\alpha}_n/\delta)} \big| \mathbf{T}\Big) \leq C'(\eta)\Big(\frac{\Delta}{\bar{\alpha}_n}\Big)^{\eta M}$$

for some constant $C' > 0$. Thus, taking δ to be sufficiently small, say, $\delta = n^{-c}$ for c large enough, and using the smoothness of the process $Z(\boldsymbol{\beta})$ as a function of $\boldsymbol{\beta}$, we can show that given any $\eta > 0$, there exists $c_2(\eta) > 0$, such that for all $\mathbf{T} \in A_{1\eta}$,

$$\mathbb{P}\Big(\sup_{\boldsymbol{\beta} \in \mathcal{A}_M(\alpha_n,\bar{\alpha}_n)} \big|Z(\boldsymbol{\beta})\big| \leq c_2(\eta)\sqrt{M \log n} \big| \mathbf{T}\Big) > 1 - n^{-\eta}. \tag{47}$$

Very similarly, defining

$$\tilde{Z}(\boldsymbol{\beta}) = \frac{\sum_{j=1}^n \varepsilon_j (x(T_j; \boldsymbol{\beta}; \hat{x}_{0,\delta}) - x(T_j; \boldsymbol{\beta})) \mathbf{1}_{[\delta,1-\delta]}(T_j)}{\sigma_\varepsilon \sqrt{\sum_{j=1}^n (x(T_j; \boldsymbol{\beta}; \hat{x}_{0,\delta}) - x(T_j; \boldsymbol{\beta}))^2 \mathbf{1}_{[\delta,1-\delta]}(T_j)}},$$

expressing $V_{2n}(\boldsymbol{\beta}) = 2\sigma_\varepsilon n^{-1/2} \sqrt{V_{1n}(\boldsymbol{\beta})} \tilde{Z}(\boldsymbol{\beta})$ and using (43), we have, for any given $\eta > 0$, there exists $c_3(\eta) > 0$ and a set $A_{2\eta}$ with $\mathbb{P}(\mathbf{T} \in A_{2\eta}) \geq 1 - n^{-\eta}$, such that for all $\mathbf{T} \in A_{2\eta}$,

$$\mathbb{P}\Big(\sup_{\boldsymbol{\beta} \in \mathcal{A}_M(\alpha_n,\bar{\alpha}_n) \cup \{\boldsymbol{\beta}^*\}} \big|V_{2n}(\boldsymbol{\beta})\big| \leq c_3(\eta)\sigma_\varepsilon \xi_n \sqrt{\frac{M \log n}{n}} \Big| \mathbf{T}\Big) > 1 - n^{-\eta}. \tag{48}$$

Finally, by A2(vi) we have the bound

$$D_n^* \leq \sup_{t \in [\delta,1-\delta]} \big|x_g(t) - x\big(t; \boldsymbol{\beta}^*\big)\big|^2 \leq C_2 M^{-2(p+1)} \tag{49}$$

for some $C_2 > 0$.

Combining (42)–(49), we claim that, given $\eta > 0$, there exist constants $C_3(\eta) > 0$, $C_4(\eta) > 0$, and constants $C_l > 0$, $l = 5, \ldots, 8$, not depending on η, such that uniformly on $\mathcal{A}_M(\alpha_n, \bar{\alpha}_n)$

$$\begin{aligned} &\frac{1}{n}\tilde{L}_\delta(\boldsymbol{\beta}) - \frac{1}{n}\tilde{L}_\delta(\boldsymbol{\beta}^*) \\ &\geq \Gamma_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*) - \sqrt{\Gamma_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*)}\Big(C_3(\eta)\sqrt{\frac{M \log n}{n}} + C_5 M^{-(p+1)} + C_6 \xi_n\Big) \\ &\quad - \xi_n\Big(C_4(\eta)\sqrt{\frac{M \log n}{n}} + C_7 M^{-(p+1)} + C_8 \xi_n\Big) \end{aligned} \tag{50}$$

with probability at least $1 - O(n^{-\eta})$.

From (50) and (41), and a careful choice of the constant C_0 in the definition (32) of α_n, and with M as in (31), we conclude that for any $\eta > 0$, there exists $C_9(\eta) > 0$ such that, uniformly in $\boldsymbol{\beta} \in \mathcal{A}_M(\alpha_n, \bar{\alpha}_n)$,

$$\frac{1}{n}\tilde{L}_\delta(\boldsymbol{\beta}) - \frac{1}{n}\tilde{L}_\delta(\boldsymbol{\beta}^*) \geq C_9(\eta)\frac{1}{M^2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2 \tag{51}$$

with probability at least $1 - O(n^{-\eta})$. From this, we can conclude that with probability at least $1 - O(n^{-\eta})$ there exists a local minimum $\hat{\boldsymbol{\beta}}$ of $\tilde{L}_\delta(\boldsymbol{\beta})$, which is also a global minimum within radius $\bar{\alpha}_n$ of $\boldsymbol{\beta}^*$ and which satisfies $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| = O(\alpha_n)$ with probability tending to 1.

5. Simulation study. In this section, we conduct simulation studies to examine the finite sample performance of the proposed nonlinear least squares (NLS) estimator, as well as to compare it with the two-stage estimator described in Section 3.5.

In the simulation, the true gradient function g is represented by 4 cubic B-spline functions with knots at $0.35, 0.60, 0.85, 1.10$ (without the boundary corrected splines) with respective coefficients $0.1, 1.2, 1.6, 0.4$ (shown by the blue curve in Figure 1). We set the initial value $x(0) = x_0 = 0.25$ in equation (2) to generate the true trajectory $x(\cdot)$. We consider two sampling density, namely, dense sampling where the number of measurements n is randomly sampled from $\{60, \ldots, 100\}$, and moderately dense sampling where n is randomly chosen from $\{20, \ldots, 60\}$. Then n observation times $\{t_1, \ldots, t_n\}$ are uniformly sampled from $[0, 1]$. Finally, the Y_j's are generated according to equation (3) with added noise $\varepsilon_i \sim \text{Normal}(0, \sigma_\varepsilon^2)$. We consider three noise levels, namely, $\sigma_\varepsilon = 0.01$ (low noise level), $\sigma_\varepsilon = 0.025$ (medium noise level), and $\sigma_\varepsilon = 0.05$ (high noise level). We simulate 500 independent data sets under each of the six combinations of sampling

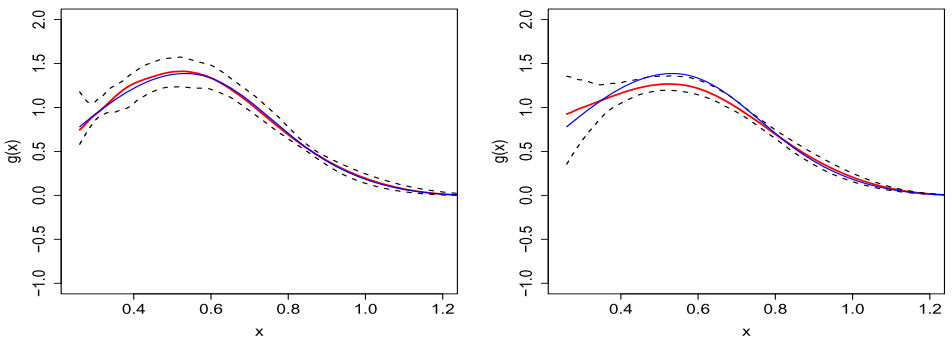

FIG. 1. *Simulation for dense sampling and* $\sigma_\varepsilon = 0.025$: *Pointwise median* (*red curve*) *and pointwise* 90% *percentile bands* (*broken black curves*) *of the estimated gradient functions overlayed on the true gradient function* (*blue curve*). *Left panel*: *proposed NLS estimator*; *Right panel*: *two-stage estimator.*

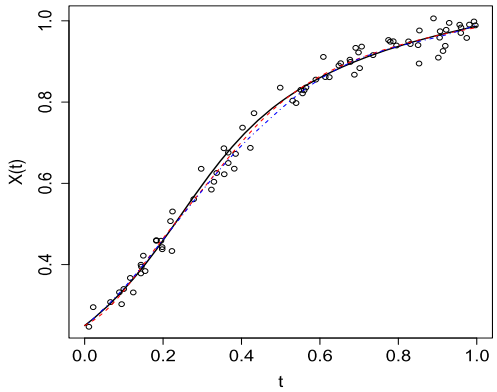

FIG. 2. *Simulation for dense sampling and* $\sigma_\varepsilon = 0.025$: *True trajectory* (*black solid curve*), *NLS fitted trajectory* (*red broken curve*), *two-stage fitted trajectory* (*blue dashed curve*) *and sample observations* (*black circles*) *for one replicate.*

density and noise level. We focus on the results for $n \sim \{60, \ldots, 100\}$ (dense sampling) and $\sigma_\varepsilon = 0.025$ (medium noise level) in the main text. More detailed results can be found in the SM [28]. The observed data from one replicate with $\sigma_\varepsilon = 0.025$ together with the true and estimated trajectories are shown in Figure 2.

We fit the proposed NLS estimator $\hat{g}(\cdot)$ with M cubic B-spline basis functions with equally spaced knots on $[0.1, 1.1]$. We consider $M = 3, 4, 5$ and choose M by Mallow's C_p as described in Section 2.4. Under dense sampling, when $\sigma_\varepsilon = 0.025$, out of the 500 replicates, 222 times the model with $M = 3$ is chosen, 220 times the model with $M = 4$ (the true model) is chosen and 58 times the model with $M = 5$ is chosen. The corresponding numbers for $\sigma_\varepsilon = 0.01$ are 30 (for $M = 3$), 443 (for $M = 4$) and 27 (for $M = 5$); and when $\sigma_\varepsilon = 0.05$: 339 (for $M = 3$), 102 (for $M = 4$) and 59 (for $M = 5$). C_p has a tendency to select simpler models when noise level is high due to bias-variance trade-off. However, as can be seen by Figure S.5, the estimated gradient functions are reasonably unbiased even under the high noise setting ($\sigma_\varepsilon = 0.05$).

We also consider the two-stage estimator, where in the first stage, the sample trajectory $x(\cdot)$ and its derivative $x'(\cdot)$ are estimated by applying local linear and local quadratic smoothing with Gaussian kernel, respectively, to the observed data $\{(t_j, Y_j)\}_{j=1}^n$. The bandwidths are chosen by cross-validation. In the second stage, the true model is used to estimate g through a least-squares regression of $\hat{x}'(\cdot)$ versus $\hat{x}(\cdot)$.

Figure 1 shows the true gradient function g along with the pointwise median and 90% percentile bands of the estimated gradient functions (computed based on the 500 replicates under dense sampling and $\sigma_\varepsilon = 0.025$) for the proposed NLS estimator and the two-stage estimator, respectively. It is noticeable that the pointwise median of $\hat{g}$ is closer to the true g for the NLS estimator than the two-stage estimator, which indicates a lesser degree of bias for the NLS estimator. Moreover, NLS

estimator shows less variability near the left end of the domain where observations are scarce. These are also reflected in the median relative integrated squared errors (M-RISE) where the proposed NLS estimator has M-RISE = 5.756% and the two-stage estimator has M-RISE = 7.416%. Here, RISE is calculated as $\frac{\int(\hat{g}-g)^2}{\int g^2}$. These phenomena persist across all simulation settings we considered (see Section S5 of the SM [28]). Indeed, under the high noise setting ($\sigma_\varepsilon = 0.05$), the improvement of the NLS estimator over the two-stage estimator is even more: There is a 35% improvement in terms of M-RISE (10.75% of NLS vs. 16.48% of two-stage). We also conduct a simulation to mimic the Berkeley growth data (see Section 6). The proposed estimator works very well under that setting and the results are reported in the SM [28].

6. Application: Berkeley growth data. We apply the proposed methodology to the Berkeley growth data [34]. Although in the literature, there are many studies of growth curves [16, 23], most of them try to model either the growth trajectories [i.e., $x(\cdot)$] or the rate of growth [i.e., $x'(\cdot)$]. On the contrary, our goal is to estimate the gradient function, that is, the functional relationship between $x'(\cdot)$ and $x(\cdot)$ which provides insights of the growth dynamics, such as at which range of height the growth rate tends to be the highest.

Specifically, we fit the proposed model to each of the 54 female subjects in this data set. For each girl, her heights were measured at 31 time points from 1 year old to 18 years old. We use M B-spline basis functions with equally spaced knots. We consider $M = 4, 5, 6, 7$. By Mallow's C_p criterion, in 24 out of 54 subjects, the model with $M = 6$ is chosen, and for the rest 30 subjects, the model with $M = 7$ is chosen. Figure 3 shows the fitted gradient functions for these 54 subjects. From this figure, we can see that, most girls experienced two growth spurts, one

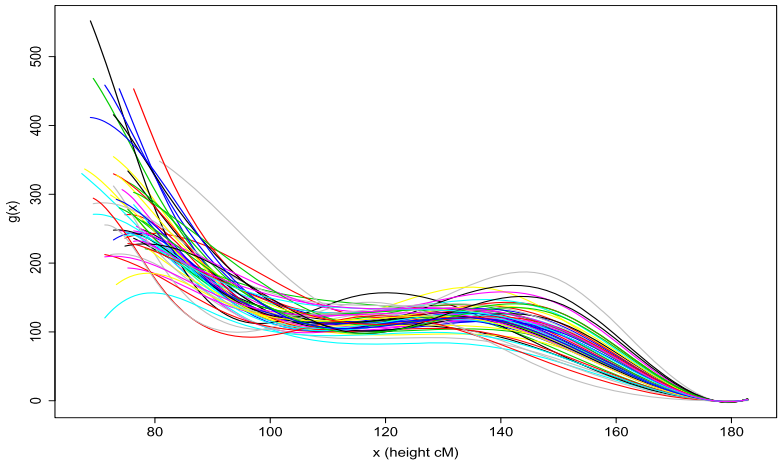

FIG. 3. *Berkeley Growth Data*: *fitted gradient functions for* 54 *female subjects.*

at the birth (when their heights are shortest) and another when they were around either 130 cm tall or 150 cm tall. Moreover, Figure S.1 in the SM [28] shows the fitted gradient functions with the two-standard-error bands [by equation (18)] for 25 girls. Figure S.2 shows the observed (red dots) and fitted (black curve) growth trajectories for these 25 girls. It can be seen that, the fitted trajectories fit the observed data very well.

7. Discussion. In this paper, we propose a nonparametric estimator for the gradient function of a first order autonomous differential equation when the trajectory is strictly monotone. We fit the model through a nonlinear least squares approach. We consider the case when the measurements are on a discrete set of time points with sub-Gaussian observational noise. We show that, if the gradient function g is $p(>3)$ times differentiable, then the optimal rate of the proposed estimator $\hat{g}$ is $O(n^{-2p/(2p+3)})$, which is the same as the optimal rate for the estimation of the derivative of a trajectory, assuming that the latter is $p+1$ times differentiable. Indeed, as discussed in Section 3.3, if the estimators of the gradient function g are restricted to a class of uniformly Lipschitz function (which includes the proposed estimator), then the minimax rate for estimation of g is of the order $n^{-2p/(2p+3)}$, and thus the proposed estimator is optimal within this class of estimators. We conjecture that the Lipschitz requirement on the estimators of g is not necessary and the minimax rate for estimation of g is indeed of the order $n^{-2p/(2p+3)}$. In addition, we carry out simulation studies to show that the proposed nonlinear least squares estimator of the gradient function performs well and is superior than a two-stage estimator.

In this paper, we consider an L_2 loss function. If the response is categorial or has specific characteristics, such as being counts or binary, then it may be more appropriate to model the underlying continuous process through a generalized linear model framework. Whether we can extend the theoretical results under such a framework may depend on the specific features of the error distribution and the link function. This is a topic of future research.

APPENDIX

In this section, we provide technical details for the proofs of the main results. Specifically, in Appendix A.1, we present results on perturbation analysis of differential equations that are central to controlling the bias in the estimates. In Appendix A.2, we verify that condition (vi) of A2 is satisfied by a B-spline basis of sufficiently high order. In Appendix A.3, we prove Lemma 4.2. Details of the estimation procedure are provided in Appendix A.4. Further technical details are given in the supplementary material [28].

A.1. Properties of sample trajectories and their derivatives. Throughout this subsection, with slight abuse of notation, we use $x(\cdot)$ to mean $x(\cdot;\boldsymbol{\beta})$, unless

stated otherwise. Since $x(\cdot)$ satisfies the ODE

$$x(t) = x_0 + \int_0^t \sum_{k=1}^{M} \beta_k \phi_k\big(x(s)\big)\, ds, \qquad t \in [0,1], \tag{51}$$

differentiating with respect to $\boldsymbol{\beta}$ we obtain the linear differential equations

$$\frac{d}{dt} x^{\beta_\ell}(t) = x^{\beta_\ell}(t) \sum_{k=1}^{M} \beta_k \phi_k'\big(x(t)\big) + \phi_\ell\big(x(t)\big), \qquad x^{\beta_\ell}(0) = 0, \tag{52}$$

for $\ell = 1, \ldots, M$, where $x^{\beta_\ell}(t) := \frac{\partial x(t)}{\partial \beta_\ell}$. The Hessian of $x(\cdot)$ with respect to $\boldsymbol{\beta}$ is given by the matrix $(x^{\beta_\ell,\beta_{\ell'}})_{\ell,\ell'=1}^{M}$, where $x^{\beta_\ell,\beta_{\ell'}}(t) := \frac{\partial^2}{\partial\beta_\ell \partial\beta_{\ell'}} x(t)$, which satisfies the system of ODEs, for $\ell, \ell' = 1, \ldots, M$:

$$\begin{aligned} &\frac{d}{dt} x^{\beta_\ell,\beta_{\ell'}}(t) \\ &\quad = \Bigg[x^{\beta_\ell,\beta_{\ell'}}(t) \sum_{k=1}^{M} \beta_k \phi_k'\big(x(t)\big) + x^{\beta_\ell}(t) \phi_{\ell'}'\big(x(t)\big) \\ &\qquad + x^{\beta_{\ell'}}(t) \phi_\ell'\big(x(t)\big) + x^{\beta_\ell}(t) x^{\beta_{\ell'}}(t) \sum_{k=1}^{M} \beta_k \phi_k''\big(x(t)\big) \Bigg], \\ &\qquad\qquad x^{\beta_\ell,\beta_{\ell'}}(0) = 0. \end{aligned} \tag{53}$$

With $a := x(\delta)$ and $x^a(t)$ denoting $\frac{\partial}{\partial a} x(t)$, we also have

$$\frac{d}{dt} x^a(t) = g_{\boldsymbol{\beta}}'\big(x(t)\big) x^a(t), \qquad x^a(\delta) = 1. \tag{54}$$

Note that (52), (53) and (54) are linear differential equations. If the function $g_{\boldsymbol{\beta}} := \sum_{k=1}^{M} \beta_k \phi_k$ is positive on the domain then the gradients of the trajectories can be solved explicitly as follows:

$$x^{\beta_\ell}(t) = g_{\boldsymbol{\beta}}\big(x(t)\big) \int_{x_0}^{x(t)} \frac{\phi_\ell(u)}{(g_{\boldsymbol{\beta}}(u))^2}\, du, \tag{55}$$

$$\begin{aligned} x^{\beta_\ell,\beta_{\ell'}}(t) &= g_{\boldsymbol{\beta}}\big(x(t)\big) \int_0^t \frac{1}{g_{\boldsymbol{\beta}}(x(s))} x^{\beta_\ell}(s) x^{\beta_{\ell'}}(s) g_{\boldsymbol{\beta}}''\big(x(s)\big)\, ds \\ &\quad + g_{\boldsymbol{\beta}}\big(x(t)\big) \int_0^t \frac{1}{g_{\boldsymbol{\beta}}(x(s))} \big[x^{\beta_\ell}(s) \phi_{\ell'}'\big(x(s)\big) + \phi_\ell'\big(x(s)\big) x^{\beta_{\ell'}}(s) \big]\, ds \end{aligned} \tag{56}$$

and

$$x^a(t) = \frac{g_{\boldsymbol{\beta}}(x(t))}{g_{\boldsymbol{\beta}}(a)}, \qquad t \in [\delta, 1-\delta]. \tag{57}$$

The following result on the perturbation of the solution path due to a perturbation in the gradient function is derived from [9].

PROPOSITION A.1. *Consider the initial value problem*:

$$r' = f(t, r), \qquad r(t_0) = r_0, \tag{58}$$

where $r \in \mathbb{R}^d$. *On the augmented phase space* Ω, *say, let the mappings* f *and* δf *be continuous and continuously differentiable with respect to the state variable. Assume that for* $(t_0, r_0) \in \Omega$, *the initial value problem* (58), *and the perturbed problem*

$$r' = f(t, r) + \delta f(t, r), \qquad r(t_0) = r_0,$$

have the solutions r *and* $\overline{r} = r + \delta r$, *respectively. If* f *is such that* $\|f_r(t, \cdot)\|_\infty \le \chi(t)$ *for a function* $\chi(\cdot)$ *bounded on* $[t_0, t_1]$, *and* $\|\delta f(t, \cdot)\|_\infty \le \tau(t)$ *for some nonnegative function* $\tau(\cdot)$ *on* $[t_0, t_1]$, *then*

$$\|\delta r(t)\| \le \int_{t_0}^{t} \exp\left(\int_s^t \chi(u)\, du\right) \tau(s)\, ds \qquad \text{for all } t \in [t_0, t_1].$$

We use the above result to compute bounds for the trajectories and their derivatives corresponding to the different values of the parameter $\boldsymbol{\beta}$ in a neighborhood of the point $\boldsymbol{\beta}^*$. In order to keep the exposition simple, we assume that $g_{\boldsymbol{\beta}}(x) = g_{\boldsymbol{\beta}}(x_{1,M})$ for $x > x_{1,M}$ and $g_{\boldsymbol{\beta}}(x) = g_{\boldsymbol{\beta}}(x_{0,M})$ for $x < x_{0,M}$ with a differentiability requirement at the points $x_{0,M}$ and $x_{1,M}$.

We first deduce that the range of the trajectories $x(t; \boldsymbol{\beta}, x_{0,\delta})$ is contained in the set $D_0 = [x_{0,M}, x_{1,M}]$, for all $t \in [\delta, 1-\delta]$ and for all $\boldsymbol{\beta} \in \mathcal{B}(\alpha_n) := \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \le \alpha_n\}$. Let $\gamma_n = \max\{\sup_{y \in D} |g_{\boldsymbol{\beta}^*}(y) - g(y)|, \sup_{y \in D_0} |g_{\boldsymbol{\beta}^*}(y) - g_{\boldsymbol{\beta}}(y)|\}$. Then $\gamma_n = O(M^{-p}) + O(\alpha_n M^{1/2})$. Also, let $\xi_n = \max_{j=0,1} |\hat{x}_{j,\delta} - x_{j,\delta}|$. As in the proof of Proposition 3.1, we can easily show that $[x_{0,\delta}, x_{1,\delta}] \subset [x_{0,M}, x_{1,M}]$ for sufficiently large M almost surely. On the other hand, by using the perturbation bound given by Proposition A.1 progressively over small subintervals of the interval $[\delta, 1-\delta]$, it can be shown that

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}(\alpha_n)} \sup_{t \in [\delta, 1-\delta]} \left|x(t; \boldsymbol{\beta}, x_{0,\delta}) - x_g(t; x_{0,\delta})\right| \le C_1 \gamma_n + C_2 \xi_n,$$

for appropriate $C_1, C_2 > 0$ depending on g and g' but not on α_n. Now, using Lemma 3.1, the condition on α_n as given in Theorem 4.1, and the definitions of $\hat{x}_{j,\delta}$, $x_{j,\delta}$ and $x_{j,M}$, for $j = 0, 1$, we conclude that for large enough M, the range of $x(t; \boldsymbol{\beta}, x_{0,\delta})$ is contained in D_0 for all $t \in [\delta, 1-\delta]$ and for all $\boldsymbol{\beta} \in \mathcal{B}(\alpha_n)$. The scenario is depicted in Figure 4, where the dashed curves indicate the envelop of the trajectories $x(t; \boldsymbol{\beta}, x_{0,\delta})$, while the solid curve indicates the trajectory $x_g(t; x_{0,\delta})$.

Next, we provide bounds for trajectories and their derivatives. In the following, $\|\cdot\|_\infty$ is used to denote the sup-norm over $D_0 = [x_{0,M}, x_{1,M}]$. First, by A2 we have the following:

$$\|g_{\boldsymbol{\beta}}^{(j)} - g_{\boldsymbol{\beta}^*}^{(j)}\|_\infty = O(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| M^{j+1/2}), \qquad j = 0, 1, 2, \tag{59}$$

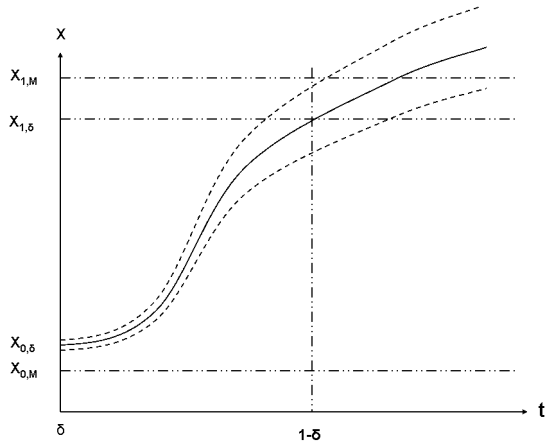

FIG. 4. *Schematic diagram of the trajectory* $x_g(t; x_{0,\delta})$ (*solid curve*) *and the envelop of trajectories* $x(t; \boldsymbol{\beta}, x_{0,\delta})$ (*boundaries indicated by dashed curves*).

where $g^{(j)}$ and $g^{(j)}_{\boldsymbol{\beta}^*}$ denote the jth derivative of g and $g_{\boldsymbol{\beta}^*}$, respectively. Next, again from A2, for M large enough, solutions $\{x(t; \boldsymbol{\beta}) : t \in [\delta, 1-\delta]\}$ exist for all $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \le \alpha_n$. This also implies that the solutions $x^{\beta_\ell}(\cdot; \boldsymbol{\beta})$ and $x^{\beta_\ell, \beta_{\ell'}}(\cdot; \boldsymbol{\beta})$ exist on $[\delta, 1-\delta]$ for all $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \le \alpha_n$, since they follow linear differential equations where the coefficient functions depend on $X(t; \boldsymbol{\beta})$. Moreover, by *Gronwall's lemma* [9], (59) and the fact that $\|g^{(j)}_{\boldsymbol{\beta}^*}\|_\infty = O(1)$ for $j = 0, 1, 2$ (again by A2).

Hence, if $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| M^{3/2} = o(1)$, then using Proposition A.1, the fact that $\|g^{(j)}_{\boldsymbol{\beta}^*}\|_\infty = O(1)$ for $j = 0, 1, 2$, and the expressions for the ODEs for the partial derivatives, we obtain

$$\big\|x(\cdot; \boldsymbol{\beta}^*) - x_g(\cdot)\big\|_\infty = O\big(M^{-p}\big). \tag{60}$$

Note that we improve this bound in Appendix A.2 to $O(M^{-(p+1)})$ by a more refined calculation. Similar derivations as in (60) can be used to prove the following whenever $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| M^{3/2} = o(1)$:

$$\big\|x(\cdot; \boldsymbol{\beta}) - x(\cdot; \boldsymbol{\beta}^*)\big\|_\infty = O\big(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|\big); \tag{61}$$

$$\max_{1 \le \ell \le M} \big\|x^{\beta_\ell}(\cdot; \boldsymbol{\beta})\big\|_\infty = O\big(M^{-1/2}\big); \tag{62}$$

$$\max_{1 \le \ell \le M} \big\|x^{\beta_\ell}(\cdot; \boldsymbol{\beta}) - x^{\beta_\ell}\big(\cdot; \boldsymbol{\beta}^*\big)\big\|_\infty = O\big(M^{1/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|\big); \tag{63}$$

$$\max_{1 \le \ell, \ell' \le M} \big\|x^{\beta_\ell, \beta_{\ell'}}\big(\cdot; \boldsymbol{\beta}^*\big)\big\|_\infty = O(1); \tag{64}$$

$$\max_{1 \le \ell, \ell' \le M} \big\|x^{\beta_\ell, \beta_{\ell'}}(\cdot; \boldsymbol{\beta}) - x^{\beta_\ell, \beta_{\ell'}}\big(\cdot; \boldsymbol{\beta}^*\big)\big\|_\infty = O\big(M\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| + M^{-1}\big). \tag{65}$$

The M^{-1} term in the bound in (65) can be dropped if the basis functions $\{\phi_{k,M}\}$ are in C^3.

To illustrate the key arguments, we prove (61), (62) and (63). First, from (59) and Proposition A.1, we have $\|x(\cdot;\boldsymbol{\beta}) - x(\cdot;\boldsymbol{\beta}^*)\|_\infty = O(M^{1/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|)$. Using this and (59) in (55), recalling that $g_{\boldsymbol{\beta}^*}$ is bounded away from zero and is bounded above on the domain, and the fact that $\|\phi_\ell^{(j)}\|_\infty = O(M^{1/2} + j)$, for $j = 0, 1, 2$, and is supported on an interval of length $O(M^{-1})$, we obtain (62). Now, with an application of the mean value theorem to the difference $x(t;\boldsymbol{\beta}) - x(t;\boldsymbol{\beta}^*)$, we obtain (61).

Next, from (55), $|x^{\beta_\ell}(t;\boldsymbol{\beta}) - x^{\beta_\ell}(t;\boldsymbol{\beta}^*)|$ can be bounded by

$$
\begin{aligned}
&\bigl|g_{\boldsymbol{\beta}}\bigl(x(t;\boldsymbol{\beta})\bigr) - g_{\boldsymbol{\beta}^*}\bigl(x(t;\boldsymbol{\beta})\bigr)\bigr|\left|\int_{x_0}^{x(t;\boldsymbol{\beta})} \frac{\phi_\ell(u)}{(g_{\boldsymbol{\beta}}(u))^2}\,du\right| \\
&\quad + \bigl|g_{\boldsymbol{\beta}^*}\bigl(x(t;\boldsymbol{\beta})\bigr)\bigr| \int_{x_0}^{x(t;\boldsymbol{\beta})} \left|\frac{\phi_\ell(u)}{(g_{\boldsymbol{\beta}}(u))^2} - \frac{\phi_\ell(u)}{(g_{\boldsymbol{\beta}^*}(u))^2}\right| du \\
&\quad + \bigl|g_{\boldsymbol{\beta}^*}\bigl(x(t;\boldsymbol{\beta})\bigr)\bigr|\left|\int_{x(t;\boldsymbol{\beta})}^{x(t;\boldsymbol{\beta}^*)} \frac{\phi_\ell(u)}{(g_{\boldsymbol{\beta}^*}(u))^2}\,du\right| \\
&\quad + \bigl|g_{\boldsymbol{\beta}^*}\bigl(x(t;\boldsymbol{\beta})\bigr) - g_{\boldsymbol{\beta}^*}\bigl(x(t;\boldsymbol{\beta}^*)\bigr)\bigr|\left|\int_{x_0}^{x(t;\boldsymbol{\beta}^*)} \frac{\phi_\ell(u)}{(g_{\boldsymbol{\beta}^*}(u))^2}\,du\right|.
\end{aligned}
\tag{66}
$$

Now, it is easily seen using (59), (61) and the properties of ϕ_ℓ and ϕ_ℓ', that among the four terms in (66), the first two terms are $O(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|)$, the third term is $O(M^{1/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|)$, and the last term is $O(M^{-1/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|)$, thus yielding (63). Proof of (64) and (65) follows a similar pattern involving the representation (56).

A.2. Verification of A2(vi) for B-spline basis. In this subsection, we verify that the condition A2(vi) is satisfied if $\{\phi_{k,M}\}_{k=1}^M$ is a normalized B-spline basis with equally spaced knots on $[x_{0,M}, x_{1,M}]$ and of order $d \geq \max\{3, p\}$. In particular, we show that the rate of approximation of $x_g(t)$ by $x(t;\boldsymbol{\beta}^*)$ with a carefully chosen $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ satisfies the requirement that $\sup_{t\in[\delta,1-\delta]} |x_g(t) - x(t;\boldsymbol{\beta}^*)| = O(M^{-(p+1)})$ and the conditions $\sup_{y\in[x_{0,M},x_{1,M}]} |g^{(j)}(y) - g_{\boldsymbol{\beta}^*}^{(j)}(y)| = O(M^{-p+j})$ for $j = 0, 1, 2$. The result is proved through the following lemmas proved in the supplementary material [28].

LEMMA A.1. *Suppose that* $\{\phi_{k,M}\}_{k=1}^M$ *has combined support* $[x_{0,\delta}, x_{1,\delta}] = [x(\delta), x(1-\delta)]$ *and satisfies* (ii)–(v) *of* A2 *and* $\boldsymbol{\beta}^*$ *furthermore has the property that*

$$
\sup_{y\in[x_{0,\delta},x_{1,\delta}]} \left|\int_{x_{0,\delta}}^{y} \frac{g(u) - g_{\boldsymbol{\beta}^*}(u)}{g(u)}\,du\right| = a_M \tag{67}
$$

such that $c_0 M^{-(p+1)} \le a_M \ll M^{-p-\epsilon}$, *uniformly in* M, *for some* $\epsilon \in (0, 1]$ *and some* $c_0 > 0$. *Then, if* $x(\delta; \boldsymbol{\beta}^*) = x(\delta)$, *there exists* $C > 0$ *such that*

$$\sup_{t \in [\delta, 1-\delta]} \left| x_g(t) - x\big(t; \boldsymbol{\beta}^*\big) \right| \le C a_M. \tag{68}$$

LEMMA A.2. *Suppose that* A1 *holds with* $p \ge 2$. *Let* $\{\phi_{k,M}\}_{k=1}^M$ *denote the normalized B-spline basis of order* $\ge \max\{3, p\}$ *with equally spaced knots on the interval* $[x_{0,M}, x_{1,M}]$. *Then there exists a* $\boldsymbol{\beta}^* \in \mathbb{R}^M$ *such that* $g_{\boldsymbol{\beta}^*} = \sum_{k=1}^M \beta_k^* \phi_{k,M}$ *satisfies*

$$\sup_{y \in [x_{0,\delta}, x_{1,\delta}]} \left| \int_{x_{0,\delta}}^{y} \frac{g(u) - g_{\boldsymbol{\beta}^*}(u)}{g(u)}\, du \right| = O\big(M^{-(p+1)}\big). \tag{69}$$

A.3. Proof of Lemma 4.2. By a Taylor expansion, we have for $j = 1, \ldots, n$,

$$\begin{aligned} x(T_j; \boldsymbol{\beta}) - x\big(T_j; \boldsymbol{\beta}^*\big) &= x^{\boldsymbol{\beta}}\big(T_j; \boldsymbol{\beta}^*\big)^T \big(\boldsymbol{\beta} - \boldsymbol{\beta}^*\big) \\ &\quad + \big(x^{\boldsymbol{\beta}}\big(T_j; \tilde{\boldsymbol{\beta}}(T_j)\big) - x^{\boldsymbol{\beta}}\big(T_j; \boldsymbol{\beta}^*\big)\big)^T \big(\boldsymbol{\beta} - \boldsymbol{\beta}^*\big), \end{aligned}$$

where $\|\tilde{\boldsymbol{\beta}}(T_j) - \boldsymbol{\beta}^*\| \le \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|$ for all j. From this, it follows that, for all $\boldsymbol{\beta} \in \mathcal{A}_M(\alpha_n, \bar{\alpha}_n)$,

$$\begin{aligned} &\Gamma_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \\ &\ge \frac{3}{4}\big(\boldsymbol{\beta} - \boldsymbol{\beta}^*\big)^T \left[\frac{1}{n} \sum_{j=1}^n x^{\boldsymbol{\beta}}\big(T_j; \boldsymbol{\beta}^*\big) x^{\boldsymbol{\beta}}\big(T_j; \boldsymbol{\beta}^*\big)^T \mathbf{1}_{[\delta, 1-\delta]}(T_j) \right] \big(\boldsymbol{\beta} - \boldsymbol{\beta}^*\big) \\ &\quad - 3\big\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\big\|^2 \frac{1}{n} \sum_{j=1}^n \big\|x^{\boldsymbol{\beta}}\big(T_j; \tilde{\boldsymbol{\beta}}(T_j)\big) - x^{\boldsymbol{\beta}}\big(T_j; \boldsymbol{\beta}^*\big)\big\|^2 \mathbf{1}_{[\delta, 1-\delta]}(T_j), \end{aligned} \tag{70}$$

where we have used $|2ab| \le a^2/4 + 4b^2$. Using Proposition 3.1 and Lemma A.3 (stated below), given $\eta > 0$, there exists $C_{10}(\eta) > 0$ such that

$$\begin{aligned} &\big(\boldsymbol{\beta} - \boldsymbol{\beta}^*\big)^T \left[\frac{1}{n} \sum_{j=1}^n x^{\boldsymbol{\beta}}\big(T_j; \boldsymbol{\beta}^*\big) x^{\boldsymbol{\beta}}\big(T_j; \boldsymbol{\beta}^*\big)^T \mathbf{1}_{[\delta, 1-\delta]}(T_j) \right] \big(\boldsymbol{\beta} - \boldsymbol{\beta}^*\big) \\ &\quad \ge C_{10}(\eta) \frac{1}{M^2} \big\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\big\|^2 \end{aligned}$$

for all $\boldsymbol{\beta} \in \mathcal{A}_M(\alpha_n, \bar{\alpha}_n)$, with probability at least $1 - n^{-\eta}$. Now, another application of the mean value theorem yields that for $T_j \in [\delta, 1-\delta]$,

$$\begin{aligned} &\big\|x^{\boldsymbol{\beta}}\big(T_j; \tilde{\boldsymbol{\beta}}(T_j)\big) - x^{\boldsymbol{\beta}}\big(T_j; \boldsymbol{\beta}^*\big)\big\|^2 \\ &\quad \le \big\|\tilde{\boldsymbol{\beta}}(T_j) - \boldsymbol{\beta}^*\big\|^2 \big\|x^{\boldsymbol{\beta}\boldsymbol{\beta}^T}\big(T_j; \boldsymbol{\beta}^*\big)\big\|_F^2 \\ &\qquad + \big\|\tilde{\boldsymbol{\beta}}(T_j) - \boldsymbol{\beta}^*\big\|^2 \sum_{1 \le k, k' \le M} \big|X^{\beta_k, \beta_{k'}}\big(T_j; \bar{\boldsymbol{\beta}}^k(T_j)\big) - X^{\beta_k, \beta_{k'}}\big(T_j; \boldsymbol{\beta}^*\big)\big|^2, \end{aligned}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $\|\bar{\boldsymbol{\beta}}^k(T_j) - \boldsymbol{\beta}^*\| \le \|\tilde{\boldsymbol{\beta}}(T_j) - \boldsymbol{\beta}^*\|$ for all $1 \le k \le M$ and $1 \le j \le n$. Now, using (64) and (65), and combining the last three displays, we get (40).

LEMMA A.3. *Suppose that* A1–A4 *hold. Let*

$$\bar{G}_{*n} := \frac{1}{F_T(1-\delta) - F_T(\delta)} \frac{1}{n} \sum_{j=1}^{n} x^{\boldsymbol{\beta}}(T_j; \boldsymbol{\beta}^*)\big(x^{\boldsymbol{\beta}}(T_j; \boldsymbol{\beta}^*)\big)^T \mathbf{1}_{[\delta,1-\delta]}(T_j).$$

Then, given $\eta > 0$, there exists constants $c_1'(\eta), c_2'(\eta) > 0$ such that, with probability $1 - n^{-\eta}$, uniformly in $\boldsymbol{\gamma} \in \mathbb{S}^{M-1}$,

$$\boldsymbol{\gamma}^T \bar{G}_{*n} \boldsymbol{\gamma} \ge \boldsymbol{\gamma}^T G_* \boldsymbol{\gamma} - c_1'(\eta) \sqrt{\boldsymbol{\gamma}^T G_* \boldsymbol{\gamma}} \sqrt{\frac{M \log n}{n}} \ge c_2'(\eta) M^{-2}. \tag{71}$$

Proof of Lemma A.3. Let $\mathbf{v}_j = x^{\boldsymbol{\beta}}(T_j; \boldsymbol{\beta}^*)$. Define $D(\boldsymbol{\gamma}) = \boldsymbol{\gamma}^T(\bar{G}_{*n} - G_*)\boldsymbol{\gamma}$. Notice that

$$\frac{1}{(F_T(1-\delta) - F_T(\delta))} \mathbb{E}_T\big[\mathbf{v}_j \mathbf{v}_j^T \mathbf{1}_{[\delta,1-\delta]}(T_j)\big] = \mathbb{E}_{\tilde{T}}\big[\mathbf{v}_j \mathbf{v}_j^T\big] = G_*,$$

where the first expectation is with respect to the distribution of T_1 and the second with respect to that of $\tilde{T}_1$. Hence, we can write $D(\boldsymbol{\gamma}) = n^{-1} \sum_{j=1}^n u_j(\boldsymbol{\gamma})$ where

$$u_j(\boldsymbol{\gamma}) = \boldsymbol{\gamma}^T \left(\mathbf{v}_j \mathbf{v}_j^T \frac{\mathbf{1}_{[\delta,1-\delta]}(T_j)}{F_T(1-\delta) - F_T(\delta)} - \mathbb{E}_{\tilde{T}}\big[\mathbf{v}_j \mathbf{v}_j^T\big] \right) \boldsymbol{\gamma}.$$

Note that, the random variables $u_j(\boldsymbol{\gamma})$ have zero conditional mean, are uniformly bounded, and are independent. Moreover, the functions $u_j(\boldsymbol{\gamma})$ are differentiable functions of $\boldsymbol{\gamma}$. Then, since by (62), $u_j(\boldsymbol{\gamma})$'s are uniformly bounded by some $K_1 > 0$,

$$\operatorname{Var}\left(\sum_{j=1}^n u_j(\boldsymbol{\gamma}) \right) = \sum_{j=1}^n \mathbb{E}\big[\big(u_j(\boldsymbol{\gamma})\big)^2\big] \le K_1 \sum_{j=1}^n \mathbb{E}\big|u_j(\boldsymbol{\gamma})\big| \le 2K_1 n \boldsymbol{\gamma}^T G_* \boldsymbol{\gamma}.$$

Thus, by Bernstein's inequality, for every $v > 0$ and $\boldsymbol{\gamma} \in \mathbb{S}^{M-1}$,

$$\mathbb{P}\left(\left| \sum_{j=1}^n u_j(\boldsymbol{\gamma}) \right| > v \right) \le 2 \exp\left(-\frac{v^2/2}{2K_1 n \boldsymbol{\gamma}^T G_* \boldsymbol{\gamma} + K_1 v/3} \right).$$

On the other hand, by (12), $\boldsymbol{\gamma}^T G_* \boldsymbol{\gamma} \ge cM^{-2}$ for some $c > 0$. By this, and the condition that $M^3 = o(n/\log n)$, it is easy to see that $\sqrt{\boldsymbol{\gamma}^T G_* \boldsymbol{\gamma}} \gg \sqrt{M \log n / n}$. Thus, using an entropy argument as in the proof of (47), we conclude that given $\eta > 0$ there exists $c_1'(\eta) > 0$ such that

$$\mathbb{P}\left(\sup_{\boldsymbol{\gamma} \in \mathbb{S}^{M-1}} \frac{|n^{-1} \sum_{j=1}^n u_j(\boldsymbol{\gamma})|}{\sqrt{\boldsymbol{\gamma}^T G_* \boldsymbol{\gamma}}} \le c_1'(\eta) \sqrt{\frac{M \log n}{n}} \right) > 1 - n^{-\eta}. \tag{72}$$

Recalling the definition of $D(\boldsymbol{\gamma})$, and again using the fact that $\boldsymbol{\gamma}^T G_* \boldsymbol{\gamma} \geq cM^{-2}$ and $M^3 = o(n/\log n)$, (71) follows from (72).

A.4. Estimation procedure. The Levenberg–Marquardt procedure proceeds by first specifying an initial estimate of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}^0$ (e.g., the two-stage estimator), and then successively solving the linearized regression problem (at step k):

$$\min_{\boldsymbol{\beta}} \sum_{j=1}^{n} \bigl(Y_j - x\bigl(t_j; \boldsymbol{\beta}^{(k-1)}\bigr) - \bigl(x^{\boldsymbol{\beta}}\bigl(t_j, \boldsymbol{\beta}^{(k-1)}\bigr)\bigr)^T \bigl(\boldsymbol{\beta} - \boldsymbol{\beta}^{(k-1)}\bigr)\bigr)^2 \chi_j,$$

where we use χ_j to denote $\mathbf{1}_{[\delta,1-\delta]}(t_j)$, and $x^{\boldsymbol{\beta}}$ means partial derivative of $x(\cdot;\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. This yields the solution $\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(k-1)} + \mathbf{A}_{k-1}^{-1}\mathbf{b}_{k-1}$, where $\mathbf{A}_{k-1} = \sum_{j=1}^{n} x^{\boldsymbol{\beta}}(t_j, \boldsymbol{\beta}^{(k-1)})(x^{\boldsymbol{\beta}}(t_j, \boldsymbol{\beta}^{(k-1)}))^T \chi_j$ and $\mathbf{b}_{k-1} = \sum_{j=1}^{n}(Y_j - x(t_j; \boldsymbol{\beta}^{(k-1)}))x^{\boldsymbol{\beta}}(t_j, \boldsymbol{\beta}^{(k-1)})\chi_j$. Iterations are continued until convergence of $\boldsymbol{\beta}^{(k)}$. Note that, at each step, $x(t;\boldsymbol{\beta})$ and $x^{\boldsymbol{\beta}}(t;\boldsymbol{\beta})$ are obtained by numerically solving the differential equations (51) and (52) using the fourth-order Runge–Kutta method.

Acknowledgment. The authors would like to thank the Associate Editor and the referee for their careful reading of the manuscript and for insightful suggestions that led to a significant improvement of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material for "Nonparametric estimation of dynamics of monotone trajectories" (DOI: 10.1214/15-AOS1409SUPP; .pdf). This document contains some proof details and additional simulation results and data analysis.

REFERENCES

[1] BRUNK, H. D. (1970). Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference* (*Proc. Sympos., Indiana Univ., Bloomington, Ind.*, 1969) (M. L. Puri, ed.) 177–197. Cambridge Univ. Press, London. MR0277070

[2] CAO, J., FUSSMANN, G. F. and RAMSAY, J. O. (2008). Estimating a predator-prey dynamical model with the parameter cascades method. *Biometrics* **64** 959–967. MR2526648

[3] CAO, J. and ZHAO, H. (2008). Estimating dynamic models for gene regulation networks. *Bioinformatics* **24** 1619–1624.

[4] CAVALIER, L. (2008). Nonparametric statistical inverse problems. *Inverse Probl.* **24** 034004, 19. MR2421941

[5] CAVALIER, L., GOLUBEV, Y., LEPSKI, O. and TSYBAKOV, A. (2004). Block thresholding and sharp adaptive estimation in severely ill-posed inverse problems. *Theory Probab. Appl.* **48** 426–446. MR2141349

[6] CHEN, J. and WU, H. (2008). Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to HIV-1 dynamics. *J. Amer. Statist. Assoc.* **103** 369–384. MR2420240

[7] CHEN, J. and WU, H. (2008). Estimation of time-varying parameters in deterministic dynamic models. *Statist. Sinica* **18** 987–1006. MR2440076

[8] DE BOOR, C. (1978). *A Practical Guide to Splines. Applied Mathematical Sciences* **27**. Springer, New York. MR0507062

[9] DEUFLHARD, P. and BORNEMANN, F. (2002). *Scientific Computing with Ordinary Differential Equations. Texts in Applied Mathematics* **42**. Springer, New York. MR1912409

[10] DONOHO, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.* **2** 101–126. MR1325535

[11] ERICKSON, R. O. (1976). Modelling of plant growth. *Annual Review of Plant Physiology* **27** 407–434.

[12] FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications. Monographs on Statistics and Applied Probability* **66**. Chapman & Hall, London. MR1383587

[13] GARDNER, T. S., DI BERNARDO, D., LORENZ, D. and COLLINS, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301** 102–105.

[14] GASSER, T. and MÜLLER, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Stat.* **11** 171–185. MR0767241

[15] HALL, P. and MA, Y. (2014). Quick and easy one-step parameter estimation in differential equations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 735–748. MR3248674

[16] HAUSPIE, R. C., WACHHOLDER, A., BARON, G., CANTRAINE, F., SUSANNE, C. and GRAFFAR, M. (1980). A comparative study of the fit of four different functions to longitudinal data of growth in height of Belgian girls. *Annals of Human Biology* **7** 347–358.

[17] JANK, W. and SHMUELI, G. (2006). Functional data analysis in electronic commerce research. *Statist. Sci.* **21** 155–166. MR2324075

[18] JOHNSTONE, I. M., KERKYACHARIAN, G., PICARD, D. and RAIMONDO, M. (2004). Wavelet deconvolution in a periodic setting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 547–573. MR2088290

[19] KELLY, C. and RICE, J. (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics* **46** 1071–1085.

[20] LIU, B. and MÜLLER, H.-G. (2009). Estimating derivatives for samples of sparsely observed functions, with application to online auction dynamics. *J. Amer. Statist. Assoc.* **104** 704–717. MR2541589

[21] LYCHE, T. and SCHUMAKER, L. L. (1975). Local spline approximation methods. *J. Approx. Theory* **15** 294–325. MR0397249

[22] MAMMEN, E. (1991). Estimating a smooth monotone regression function. *Ann. Statist.* **19** 724–740. MR1105841

[23] MILANI, S. (2000). Kinetic models for normal and impaired growth. *Annals of Human Biology* **27** 1–18.

[24] MITRINOVIĆ, D. S., PEČARIĆ, J. E. and FINK, A. M. (1991). *Inequalities Involving Functions and Their Integrals and Derivatives. Mathematics and Its Applications* (*East European Series*) **53**. Kluwer Academic, Dordrecht. MR1190927

[25] MÜLLER, H.-G., STADTMÜLLER, U. and SCHMITT, T. (1987). Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika* **74** 743–749. MR0919842

[26] NICOL, F. (2013). Functional principal component analysis of aircraft trajectories. In *ISIATM* 2013, 2*nd International Conference on Interdisciplinary Science for Innovative Air Traffic Management*, Toulouse, France.

[27] NOCEDAL, J. and WRIGHT, S. J. (2006). *Numerical Optimization*, 2nd ed. Springer, New York. MR2244940

[28] PAUL, D., PENG, J. and BURMAN, P. (2016). Supplement to "Nonparametric estimation of dynamics of monotone trajectories." DOI:10.1214/15-AOS1409SUPP.

[29] RAMSAY, J. O. (1988). Monotone regression splines in action (with discussions). *Statist. Sci.* **3** 425–461.

[30] RAMSAY, J. O. (1998). Estimating smooth monotone functions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 365–375. MR1616049

[31] RAMSAY, J. O., HOOKER, G., CAMPBELL, D. and CAO, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 741–796. MR2368570

[32] SILK, W. K. and ERICKSON, R. O. (1979). Kinametics of plant growth. *J. Theoret. Biol.* **76** 481–501.

[33] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642

[34] TUDDENHAM, R. D. and SNYDER, M. M. (1954). Physical growth of California boys and girls from birth to eighteen years. *University of California Publications in Child Development* **1** 183–364.

[35] TURCHIN, P. (2003). *Complex Population Dynamics*: *A Theoretical/Empirical Synthesis*. *Monographs in Population Biology* **35**. Princeton Univ. Press, Princeton, NJ. MR2080584

[36] VARAH, J. M. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Statist. Comput.* **3** 28–46. MR0651865

[37] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* (Y. C. Eldar and G. Kutyniok, eds.) 210–268. Cambridge Univ. Press, Cambridge. MR2963170

[38] WANG, S., JANK, W., SHMUELI, G. and SMITH, P. (2008). Modeling price dynamics in eBay auctions using differential equations. *J. Amer. Statist. Assoc.* **103** 1101–1118. MR2528829

[39] WRIGHT, I. W. and WEGMAN, E. J. (1980). Isotonic, convex and related splines. *Ann. Statist.* **8** 1023–1035. MR0585701

[40] WU, H. and DING, A. (1999). Population HIV-1 dynamics in vivo: Applicable models and inferential tools for virological data from AIDS clinical trials. *Biometrics* **55** 410–418.

[41] WU, H., DING, A. and DEGRUTTOLA, V. (1998). Estimation of HIV dynamic parameters. *Stat. Med.* **17** 2463–2485.

[42] WU, H., LU, T., XUE, H. and LIANG, H. (2014). Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *J. Amer. Statist. Assoc.* **109** 700–716. MR3223744

[43] XIA, X. (2003). Estimation of HIV/AIDS parameters. *Automatica J. IFAC* **39** 1983–1988. MR2142834

[44] XUE, H., MIAO, H. and WU, H. (2010). Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. *Ann. Statist.* **38** 2351–2387. MR2676892

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, DAVIS
DAVIS, CALIFORNIA 95656
USA
E-MAIL: debpaul@ucdavis.edu
jiepeng@ucdavis.edu
pburman@ucdavis.edu