Bernoulli **32**(3), 2026, 1880–1906
https://doi.org/10.3150/25-BEJ1936

Nonparametric estimation of ordinary differential equations: Snake and stubble

CHRISTOF SCHÖTZ[1,2,a]

[1] *Munich Climate Center and Earth System Modelling Group, Department of Aerospace and Geodesy, TUM School of Engineering and Design, Technical University of Munich, Munich, Germany,* [a] *christof.schoetz@tum.de*
[2] *Potsdam Institute for Climate Impact Research, Potsdam, Germany*

We study nonparametric estimation in dynamical systems described by ordinary differential equations (ODEs). Specifically, we focus on estimating the unknown function $f\colon \mathbb{R}^d \to \mathbb{R}^d$ that governs the system dynamics through the ODE $\dot{u}(t) = f(u(t))$, where observations $Y_{j,i} = u_j(t_{j,i}) + \varepsilon_{j,i}$ of solutions u_j of the ODE are made at times $t_{j,i}$ with independent noise $\varepsilon_{j,i}$. We introduce two novel models—the Stubble model and the Snake model—to mitigate the issue of observation location dependence on f, an inherent difficulty in nonparametric estimation of ODE systems. In the Stubble model, we observe many short solutions with initial conditions that adequately cover the domain of interest. Here, we study an estimator based on multivariate local polynomial regression and univariate polynomial interpolation. In the Snake model, we observe few long trajectories that traverse the domain of interest. Here, we study an estimator that combines univariate local polynomial estimation with multivariate polynomial interpolation. For both models, we establish error bounds of order $n^{-\frac{\beta}{2(\beta+1)+d}}$ for β-smooth functions f in an infinite-dimensional function class of Hölder-type and establish minimax optimality for the Stubble model in general and for the Snake model under some conditions via comparison to lower bounds from parallel work.

Keywords: Minimax optimal; nonparametric regression; ordinary differential equations; rate of convergence

1. Introduction

Many phenomena in physics and other sciences are naturally described by differential equations [23]. In practice, these systems are often observed through discrete and noisy measurements, which necessitates statistical methods to infer the underlying dynamics from the data. A fundamental example of such a model is given by the equations

$$Y_i = u(t_i) + \varepsilon_i\,, i = 1, \dots, n \qquad \dot{u}(t) = f(u(t))\,,$$

where $\dot{u}$ denotes the time derivative, i.e., $(\mathrm{d}u)/(\mathrm{d}t)$. Here, $u\colon \mathbb{R} \to \mathbb{R}^d$ is the solution to an ordinary differential equation (ODE) described by an unknown function $f\colon \mathbb{R}^d \to \mathbb{R}^d$. Observations Y_i are made at times $t_1 \le \cdots \le t_n$ and include independent measurement noise ε_i added to the true system state $u(t_i)$.

Such models support various objectives: reconstructing the solution u over the observed interval $[t_1, t_n]$ (*reanalysis*, borrowing terminology from climate modeling); predicting future states $u(t)$ for $t > t_n$ (*forecasting*); or estimating f itself (*learning the dynamics* of the system). This work focuses on the estimation of f, a fundamental task that also enables the other objectives, as an estimate $\hat{f}$ allows for the construction of an estimator $\hat{u}$ by solving $\dot{\hat{u}}(t) = \hat{f}(\hat{u}(t))$.

We aim to estimate f in a nonparametric setting, assuming f belongs to an infinite-dimensional class of smooth functions without imposing a specific functional form. This contrasts with parametric models, where f is assumed to belong to a finite-dimensional function class (e.g., polynomials of fixed degree). Parametric models have been studied extensively [3,7,10,17], and it is well-established that

the parametric $\sqrt{n}$-rate of convergence can be achieved. For a detailed review of statistical methods for dynamical systems, primarily in parametric contexts, see [6,15,18].

In the nonparametric setting, different algorithms for estimation have been proposed, e.g., [5,9,12]. In [14], theoretical results for learning dynamics nonparametrically are shown, but in a density estimation context that is rather different from the regression-type model studied here. To the best of the author's knowledge, only [13] takes a theoretical view on the ODE regression problem (additionally to proposing a new algorithm based on reproducing kernel Hilbert spaces). The authors show an error bound of order $n^{-1/4}$ (root mean squared error) for the reanalysis problem [13, Theorem 1]. This result cannot be optimal as standard nonparametric regression for u (ignoring the ODE constraints) yields an error of smaller order, namely $n^{-\beta/(2\beta+1)}$, where β describes the smoothness of u (in the case of [13, Theorem 1], we have $\beta = 3$).

Estimation in the nonparametric setting is significantly more challenging than in the parametric case. When f is a smooth function with bounded derivatives but otherwise unrestricted, each observation Y_i provides information about $f(x)$ only for x close to $u(t_i)$, i.e., the information is local. By contrast, in parametric models where f is known to be a polynomial of degree at most N, each observation informs all coefficients of the polynomial, making the information global. What further complicates nonparametric ODE estimation, as opposed to standard nonparametric regression (which also suffers from locality of information), is that the observation locations $u(t_i)$ themselves depend on the unknown f.

To address these challenges, we introduce two models: the *Snake model* and the *Stubble model*. For each model, we propose estimators and analyze their rates of convergence. In both cases, we derive an error bound (in root mean squared error) of order

$$n^{-\frac{\beta}{2(\beta+1)+d}} \tag{1}$$

for β-smooth functions f under optimal conditions. This error bound is minimax optimal as we have corresponding lower bounds in the parallel work [22].

1.1. Contributions

We now give more details on the contributions of this work.

A general model. In general, we assume the true model function $f^\star$ to be an element of Hölder-type smoothness class denoted as $\bar{\Sigma}^{d\to d}(\beta, L_{[\![0,\beta]\!]})$, which means that the k-th derivative of f for $k \in [\![0,\beta]\!] := \{0,\dots,\beta\}$ is bounded by $L_k \in \mathbb{R}_{>0}$, i.e., $\sup_{x\in\mathbb{R}^d} \|D^k f(x)\|_{\mathsf{op}} \le L_k$, where $\|\cdot\|_{\mathsf{op}}$ is the operator norm. Then, for given initial conditions $x_1,\dots,x_m$, we observe the solution $t \mapsto U(f^\star, x_j, t)$ of the ODE $\dot{u}(t) = f^\star(u(t))$ with $U(f^\star, x_j, 0) = x_j$ at time points $i\Delta t$ with time step $\Delta t \in \mathbb{R}_{>0}$, i.e.,

$$Y_{j,i} = U(f^\star, x_j, i\Delta t) + \varepsilon_{j,i}\,,\ j \in [\![1,m]\!], i \in [\![1,n_j]\!]\,.$$

In total, we have $n = \sum_{j=1}^m n_j$ observations. The noise variables $\varepsilon_{j,i}$ are assumed to be independent, centered, and to have a finite second moment. Because of the dependency of the location of the observation $u(t_{j,i})$ on $f^\star$, in general, consistent estimation in all of $[0,1]^d$ is impossible. Thus, we introduce the Stubble model and the Snake model, which restrict the general model. In the following, we use the notation $\asymp$ and $\preccurlyeq$ to mean *asymptotically equal* and *asymptotically lower than* (up to a positive constant), respectively (see Notation 2.3).

The Stubble model. In the Stubble model, we observe many ($m \asymp n$) short ($n_j \asymp 1$) solutions. We assume that their initial conditions x_j cover the domain of interest $[0,1]^d$ suitably. For this model, we

construct an estimator $\hat{f}$ based on a multivariate local polynomial estimator and a univariate polynomial interpolation, which is similar to the Adams–Bashforth method for numerical solutions of ODEs [4, Chapter 24]. We obtain

$$\mathbf{E}\left[\left\|\hat{f}(x) - f^\star(x)\right\|_2^2\right] \preccurlyeq \left(\Delta t^2 n\right)^{-\frac{2\beta}{2\beta+d}} + \Delta t^{2\beta} \tag{2}$$

for all $x \in [0,1]^d$, see Corollary 3.13. The rate is shown to be minimax optimal by comparing it to lower bounds in [22].

The Snake model. In the Snake model, we observe few ($m \asymp 1$) long ($n_j \asymp n$) solutions. We require the trajectories $U(f^\star, x_j, [0, n_j \Delta t])$ to cover the domain of interest $[0,1]^d$ suitably. For this model, we construct an estimator $\hat{f}$ based on a univariate local polynomial estimator and a multivariate polynomial interpolation. Then

$$\sup_{x\in[0,1]^d} \left\|\hat{f}(x) - f^\star(x)\right\|_2^2 = \mathbf{O_P}\left(\delta^{2\beta} + (\Delta t \log n)^{\frac{2\beta}{2(\beta+1)+1}}\right), \tag{3}$$

see Corollary 4.18, where we want to view $\delta \in \mathbb{R}_{>0}$ for now as the largest distance between a point $x \in [0,1]^d$ and its closest state $U(f^\star, x_j, t)$, $t \in [0, n_j \Delta t]$. This interpretation is true for the case $\beta = 1$, but if $\beta > 1$, the definition is more complex and the result more restrictive. By comparing it to lower bounds in [22], the rate is shown to be minimax optimal if δ and Δt are in a certain relation. For the rate to be optimal, we essentially require observations to be rather dense in time, and temporally distant parts of the trajectories $U(f^\star, x_j, [0, n_j \Delta t])$ to be distant enough in state space.

Connection and further results. Note the complementary nature of the two models and their estimators. In spite of this, in the optimal setting regarding Δt and δ, we obtain an error bound of the order given in (1) in both models (up to a log-factor when considering the sup-norm), see Corollaries 3.15 and 4.20. For both models, we begin our discussion with the case $\beta = 1$, which allows for simple estimators and a gentle introduction to the main ideas as well as slightly stronger and more specific results. See Theorem 3.2 and Corollary 3.4 for the Stubble model, and Theorem 4.4 and Corollary 4.6 for the Snake model. For the case of general smoothness $\beta \in \mathbb{N}$, we first show *black-box* results for estimation strategies that can be used with an arbitrary regression estimator, Theorems 3.11 and 4.17, respectively. If the chosen regression estimator achieves the minimax rate for a standard nonparametric regression problem, the estimation strategies achieve the rates (2) and (3), respectively. In Corollaries 3.13 and 4.18, we apply the general results with the (minimax optimal) local polynomial estimator as regression estimator.

Technical novelties. To achieve our results, we introduce several key technical contributions: We identify the previously mentioned locality problem in nonparametric ODE estimation, provide two solutions via the Snake and Stubble models, and construct new estimation algorithms with proven error bounds for these models. We establish general-purpose *black-box* theorems (Theorem 3.11 and 4.17) that transfer results of nonparametric regression to our ODE framework. To be able to apply multivariate local polynomial regression results in the ODE context, we derive explicit bounds on the time derivatives of ODE solutions with Hölder-smooth model functions using tree derivative operators, and on derivatives with respect to initial conditions using partition derivative operators. In the Snake model, we use multivariate polynomial interpolation over general design points to globally extend local regression results.

1.2. Overview

The remaining article is structured as follows. In Section 2, we introduce basic concepts for the study of ordinary differential equations (Section 2.1), give a full formal description of the general model (Section 2.2), argue why this model is not useful without further restriction (Section 2.3), and introduce some notation for standard regression problems that allows us to formulate estimation strategies referring to arbitrary regression estimators (Section 2.4). Sections 3 and 4 formally introduce the Stubble and Snake model, respectively, and describe the estimators $\hat{f}$ of $f^\star$, upper bounds on the error $\|\hat{f}(x) - f^\star(x)\|_2$, and their proofs. Both sections are separated into two parts: The first one (Sections 3.1 and 4.1) concerns the Lipschitz case ($\beta = 1$). The second one (Sections 3.2 and 4.2) concerns the general case ($\beta \in \mathbb{N}$). A discussion of the main results can be found in Section 5. The supplement [21] provides extended discussion of the models and results, reviews notation for multivariate calculus and smoothness classes, and contains proofs omitted from the main text.

2. Preliminaries

In this section, we recall some basic concepts related to ODEs, we introduce a general statistical model for observing solutions of ODEs, we explain why this model requires further restriction in order to be a useful model, and we introduce some terminology for a standard regression problem that will later allow us to construct ODE estimators in a generic (black-box) fashion.

2.1. Ordinary differential equations

We introduce some basic terminology and fundamental properties concerning ODEs. See also [2,11].

Notation 2.1.

(i) Let $\mathbb{Z}$ be the set of integers and $\mathbb{R}$ the set of reals. For $\mathbb{K} \in \{\mathbb{Z}, \mathbb{R}\}$ and $a \in \mathbb{R}$, denote $\mathbb{K}_{>a} = \{x \in \mathbb{K} \mid x > a\}$. Define $\mathbb{K}_{\geq a}$, $\mathbb{K}_{<a}$, $\mathbb{K}_{\leq a}$ accordingly.
(ii) Set $\mathbb{N} := \mathbb{Z}_{\geq 1}, \mathbb{N}_0 := \mathbb{Z}_{\geq 0}$. Let $a, b \in \mathbb{Z}$ with $a \leq b$. Set $[\![a, b]\!] := \mathbb{Z}_{\geq a} \cap \mathbb{Z}_{\leq b}$. Set $[\![a]\!] := [\![1, a]\!]$.

Let $d \in \mathbb{N}$. Let $f\colon \mathbb{R}^d \to \mathbb{R}^d$. For a differentiable function $u\colon \mathbb{R} \to \mathbb{R}^d$, denote its derivative as $\dot{u}\colon \mathbb{R} \to \mathbb{R}^d$. Then

$$\dot{u}(t) = f(u(t)) \qquad \text{for } t \in \mathbb{R} \tag{4}$$

is an *autonomous*, *first-order*, *ordinary differential equation*. It is of first-order, as (4) only depends on the first derivative of u. It is autonomous, as the right-hand side term $f(u(t))$ only depends on t via u. In contrast, a non-autonomous, first-order ODE has the form $\dot{u}(t) = g(t, u(t))$ for a function $g\colon \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$. Any differentiable $u\colon \mathbb{R} \to \mathbb{R}^d$ that fulfills (4) is a *solution* to the ODE. The domain of u is called *time*. The codomain of u is called *state space*. A single element of the state space is a *state*. The image of u is called *trajectory*. We call f the *model function* of the ODE. Let $x \in \mathbb{R}^d$. Consider the requirement

$$u(0) = x\,. \tag{5}$$

We call x the *initial conditions*. We call (4) together with (5) *initial value problem* (IVP). If u fulfills (4) and (5), it is a solution to the IVP.

Assume that f is (globally) Lipschitz continuous. Then the IVP (4), (5) has a unique solution (Picard–Lindelöf theorem, also known as the Cauchy–Lipschitz theorem). Denote this solution as $U(f,x,\cdot)\colon \mathbb{R} \to \mathbb{R}^d, t \mapsto U(f,x,t)$. The function $U(f,\cdot,\cdot)\colon \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ is called *flow* of the ODE (4). Note the *semigroup property* of the flow,

$$U(f,x,0) = x \qquad \text{and} \qquad U(f,x,s+t) = U(f,U(f,x,s),t)$$

for all $x \in \mathbb{R}^d$, $s,t \in \mathbb{R}$. For a given time step $\Delta t \in \mathbb{R}_{>0}$, the mapping $x \mapsto U(f,x,\Delta t)$ is called *propagator*. If u is a solution to the ODE (4), then $u(t+\Delta t) = U(f,u(t),\Delta t)$ for all $t \in \mathbb{R}$. Furthermore, we call

$$\Upsilon(f,\Delta t,x) := U(f,x,\Delta t) - x = U(f,x,\Delta t) - U(f,x,0)$$

the *increment*. If u is a solution to the ODE (4), then $u(t+\Delta t) = u(t) + \Upsilon(f,\Delta t,u(t))$ for all $t \in \mathbb{R}$.

The setting of autonomous, first-order ODEs is not a strong restriction: Consider the d-dimensional, non-autonomous ODE of order ℓ,

$$v^{(\ell)}(t) = g(t, v(t), v^{(1)}(t), \dots, v^{(\ell-1)}(t)), \tag{6}$$

where $v^{(k)}$ for $k \in [\![0,\ell]\!]$ denotes the k-th derivative of the ℓ-times differentiable function $v\colon \mathbb{R} \to \mathbb{R}^d$ and $g\colon \mathbb{R} \times (\mathbb{R}^d)^\ell \to \mathbb{R}^d$ is a function. We represent the derivatives and the time variable with new variables: $u_k = v^{(k-1)}$ for $k \in [\![\ell]\!]$ and $u_{\ell+1}(t) = t$. For $u(t) = (u_1(t), \dots, u_{\ell+1}(t))^\top$, we obtain the ODE

$$\begin{pmatrix} \dot u_1(t) \\ \vdots \\ \dot u_{\ell-1}(t) \\ \dot u_\ell(t) \\ \dot u_{\ell+1}(t) \end{pmatrix} = \begin{pmatrix} u_2(t) \\ \vdots \\ u_\ell(t) \\ g(u_{\ell+1}(t), u_1(t), \dots, u_\ell(t)) \\ 1 \end{pmatrix}. \tag{7}$$

It is of the form $\dot u(t) = f(u(t))$ for a suitably chosen $f\colon \mathbb{R}^{\tilde d} \to \mathbb{R}^{\tilde d}$, where $\tilde d := d\ell + 1$. Hence, (7) is a $\tilde d$-dimensional, autonomous, first-order ODE. If u is a solution to the ODE (7), then u_1 is a solution of the ODE (6).

2.2. Formal description of the general ODE model

We introduce a general model for observations from solutions of an ODE in two forms—standard and generic. While the standard form is similar to the model presented in [13], the generic form offers greater generality.

Notation 2.2.

(i) Let $a,b \in \mathbb{Z}$ with $a \le b$. For $i \in [\![a,b]\!]$, let x_i be some object. Set $x_{[\![a,b]\!]} := (x_i)_{i \in [\![a,b]\!]}$.

(ii) Let $d \in \mathbb{N}$. For $x \in \mathbb{R}^d$, let $\|x\|_2$ denote the Euclidean norm. Let $p,k \in \mathbb{N}$. For a k-multilinear operator $A\colon (\mathbb{R}^d)^k \to \mathbb{R}^p$, denote the operator norm as

$$\|A\|_{\mathsf{op}} := \sup_{v_1,\dots,v_k \in \mathbb{R}^d, \|v_1\|_2 = \dots = \|v_1\|_2 = 1} \|A(v_1,\dots,v_k)\|_2 .$$

Let $d_{\mathsf x}, d_{\mathsf y} \in \mathbb{N}$. Let $f\colon \mathbb{R}^{d_{\mathsf x}} \to \mathbb{R}^{d_{\mathsf y}}$. Denote the sup norm of f as $|f|_\infty := \sup_{x \in \mathbb{R}^{d_{\mathsf x}}} \|f(x)\|_2$. For k-multilinear operators, like the k-th derivative $D^k f(x)$, we abuse the notation slightly

and set $|D^k f|_\infty := \sup_{x\in\mathbb{R}^{d_x}} \left\|D^k f(x)\right\|_{\mathsf{op}}$. For more details on derivatives and their norms, see [21, section 2].

(iii) Let $d_x, d_y, \beta \in \mathbb{N}$. Denote the set of β-times continuously differentiable functions $f\colon \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ as $\mathcal{D}^\beta(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$. Let $L_{[\![0,\beta]\!]} \subseteq \mathbb{R}_{>0} \cup \{\infty\}$. Denote by $\bar{\Sigma}^{d_x\to d_y}(\beta, L_{[\![0,\beta]\!]}) \subseteq \mathcal{D}^\beta(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ the set of β-times continuously differentiable functions $f\colon \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ with $\left|D^k f\right|_\infty \le L_k$ for $k \in [\![0,\beta]\!]$. Let $L \in \mathbb{R}_{>0}$. Denote by $\Sigma^{d_x\to d_y}(\beta, L) \subseteq \mathcal{D}^{\beta-1}(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ the set of $(\beta-1)$-times continuously differentiable functions $f\colon \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ with

$$\left\|D^{\beta-1} f(x) - D^{\beta-1} f(\tilde{x})\right\|_{\mathsf{op}} \le L\|x - \tilde{x}\|_2$$

for all $x, \tilde{x} \in \mathbb{R}^{d_x}$. More details on derivatives and smoothness classes are described in [21, section 2].

(iv) For a measurable space $\mathcal{Y}$ denote the set of probability distributions on $\mathcal{Y}$ as $\mathfrak{P}(\mathcal{Y})$.

Let $d \in \mathbb{N}$ be the dimension of the state space $\mathbb{R}^d$. Let $\beta \in \mathbb{N}$ be the smoothness parameter. Let $L_{[\![0,\beta]\!]} \subseteq \mathbb{R}_{>0}$ be the Lipschitz parameters. Let $\mathcal{F}_{d,\beta} := \bar{\Sigma}^{d\to d}(\beta, L_{[\![0,\beta]\!]})$ be the smoothness class. Let the true model function be $f^\star \in \mathcal{F}_{d,\beta}$. Let $m \in \mathbb{N}$ be the number of observed solutions and let $x_1, \ldots, x_m \in \mathbb{R}^d$ be their initial conditions. For $j \in [\![m]\!]$, let $n_j \in \mathbb{N}$ be the number of observations for the j-th solution. Denote the total number of observations as $n := \sum_{j=1}^m n_j$. For $j \in [\![m]\!]$, let $T_j \in \mathbb{R}_{>0}$ be the observation duration of the j-th solution. For $j \in [\![m]\!], i \in [\![0, n_j]\!]$, let $t_{j,i} \in \mathbb{R}_{\ge 0}$ be the observation times such that $0 = t_{j,0} \le t_{j,1} \le \cdots \le t_{j,n_j} = T_j$.

In the *generic general ODE model*, the observations are given as follows: Let $\mathcal{Y}$ be a measurable space, the space of observations. Let $G_n\colon (\mathbb{R}^d)^n \to \mathfrak{P}(\mathcal{Y}^n)$ be a Markov kernel. We call G_n the *data generating process*. It maps n states to the distribution of their *observations*:

$$(Y_{j,i})_{j\in[\![m]\!], i\in[\![n_j]\!]} \sim G_n\Big(U\big(f^\star, x_j, t_{j,i}\big)_{j\in[\![m]\!], i\in[\![n_j]\!]}\Big)\,.$$

Let $\mathfrak{G}_n$ be a set of such data generating processes that make up our statistical model.

We define the *standard general ODE model* as an instance of the generic one: Let $\sigma \in \mathbb{R}_{\ge 0}$ be the variance bound of the noise. Let the noise variables $(\epsilon_{j,i})_{j\in[\![m]\!], i\in[\![n_j]\!]}$ be independent $\mathbb{R}^d$-valued random variables such that $\mathbf{E}[\epsilon_{j,i}] = 0$ and $\mathbf{E}[\|\epsilon_{j,i}\|_2^2] \le \sigma^2$. For $j \in [\![m]\!], i \in [\![n_j]\!]$, let the observations be

$$Y_{j,i} := U(f^\star, x_j, t_{j,i}) + \epsilon_{j,i}\,. \tag{8}$$

Equation (8) defines a data generating process G_n. The restrictions on $\epsilon_{j,i}$ define the set $\mathfrak{G}_n$.

In any form of general ODE model, we observe $Y_{j,i}$ and know x_j and $t_{j,i}$, but $f^\star$ is unknown and to be estimated. We assume d, β, $L_{[\![0,\beta]\!]}$, and σ to be fixed and we are interested in upper bounds when $n \to \infty$ for the mean squared error for estimators $\hat{f}$ of $f^\star$ on the domain of interest $[0,1]^d$.

2.3. Need for restriction of the general ODE model

We argue that the general ODE model introduced in Section 2.2 is not suitable for estimation of the model function in a fixed domain of interest and give first informal descriptions of two possible remedies of the problem: the Snake model and the Stubble model.

Notation 2.3. Let $(a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}_{>0}$. Write $a_n \preccurlyeq b_n$ for $\limsup_{n\to\infty} a_n/b_n < \infty$, $a_n \prec b_n$ for $\limsup_{n\to\infty} a_n/b_n = 0$, and $a_n \asymp b_n$ for $a_n \preccurlyeq b_n$ and $b_n \preccurlyeq a_n$. Define $\succ$ and $\succcurlyeq$ accordingly.

For the general ODE model, consistent estimation is not possible when considering the domain of interest $[0,1]^d$. We can easily construct settings, where we will never make any observation, say, in the ball $\mathrm{B}^d(0,1/3) = \{x \in \mathbb{R}^d \mid \|x\|_2 \leq 1/3\}$, i.e., $u(t_{j,i}) \notin \mathrm{B}^d(0,1/3)$ for all $j \in [\![m]\!], i \in [\![0, n_j]\!]$, no matter how large n is (e.g., a model function that results in periodic circular trajectories centered at the origin with radius $2/3$ when started from initial conditions $x_j \in \mathbb{R}^2 \times \{0\}^{d-2}$ with $\|x_j\|_2 = 2/3$). As we are considering a nonparametric class of model functions $f^\star \in \mathcal{F}_{d,\beta}$, we have, for the maximal risk of any estimator $\hat{f}$,

$$\sup_{f^\star \in \mathcal{F}_{d,\beta}} \mathbf{E}_{f^\star}\left[\left\|f^\star(0) - \hat{f}(0)\right\|_2^2\right] \succcurlyeq 1\,.$$

This behavior is different from parametric ODE models (finite-dimensional class $\mathcal{F}_{d,\beta}$): In the nonparametric setting, observations contain only local information (information about $f^\star(x)$ for x close to $u(t_{j,i})$), whereas in a parametric setting, observations typically contain global information so that consistent estimation is possible under mild restrictions [3,7,10,17].

In this work, we introduce two restrictions to the general nonparametric ODE model that make consistent estimation possible:

In the *Stubble model* (Section 3), we require the initial conditions $(x_j)_{j \in [\![m]\!]}$ to suitably cover the domain of interest $[0,1]^d$. By doing so, we ensure that we obtain information from every part of the domain of interest. In this model, we assume to observe many short trajectories, i.e., $m \succ 1$ and $\max_{j \in [\![m]\!]} n_j \preccurlyeq 1$, which looks like *stubble*.

In the *Snake model* (Section 4), we directly require the trajectories $U(f^\star, x_j, [0,T_j])$ to cover the domain of interest $[0,1]^d$. In this model, we assume to observe few long trajectories, i.e., $\min_{j \in [\![m]\!]} n_j \succ 1$. In the extreme case, we have $m = 1$. Consistent estimation on all of $[0,1]^d$ with one observed solution is only possible, if its trajectory covers $[0,1]^d$ suitably. Thus, this trajectory must not intersect itself, as otherwise it would mean that the solution is periodic. This behavior of self-avoidance in a bounded domain resembles the video game *Snake* [8, chapter 22].

2.4. Generic and standard regression model

In this article, we present estimation strategies based on black-box regression estimators. To formalize what we mean by *black-box regression estimators* and to be able to talk about their properties, we introduce some terminology in this section and give some examples at the end of the section.

Notation 2.4.

(i) Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of real-valued random variables. Let $(a_n)_{n \in \mathbb{N}} \subseteq \mathbb{R}_{>0}$. We use the *big O in probability* notation: By $X_n = \mathbf{O_P}(a_n)$, we mean that for all $\epsilon \in \mathbb{R}_{>0}$, there are $B \in \mathbb{R}_{>0}$ and $n_0 \in \mathbb{N}$ such that

$$\forall n \in \mathbb{N}_{\geq n_0} \colon \mathbf{P}\left(\left|\frac{X_n}{a_n}\right| > B\right) < \epsilon\,.$$

(ii) Let $d \in \mathbb{N}$, $k \in [\![d]\!]$, $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$. Denote the projection to the k-th dimension as Π_k, i.e., $\Pi_k v = v_k$.

Let $d_{\mathrm{x}}, d_{\mathrm{y}} \in \mathbb{N}$ be the dimensions of the predictor and the target, respectively. Let $\mathcal{F}$ be a set of measurable functions of the form $\mathbb{R}^{d_{\mathrm{x}}} \to \mathbb{R}^{d_{\mathrm{y}}}$, the *smoothness class*. Let $f^\star \in \mathcal{F}$ be the true *regression function*. Let $n \in \mathbb{N}$ be the number of observations. Let $\mathcal{X} \subseteq \mathbb{R}^{d_{\mathrm{x}}}$ be a Borel-measurable set, the *domain of interest*. For $i \in [\![n]\!]$, let $x_i \in \mathcal{X}$ be the observation locations.

In the *generic* regression model, the observations are given as follows: Let $\mathcal{Y}$ be a measurable space, the space of observations. Let $G_n\colon (\mathbb{R}^{d_y})^n \to \mathfrak{P}(\mathcal{Y}^n)$ be a Markov kernel. We call G_n the *data generating process*. It maps the true values of the regression function to the distribution of the *observations*:

$$(Y_1, \dots, Y_n) \sim G_n(f^\star(x_1), \dots, f^\star(x_n))\,.$$

Let $\mathfrak{G}_n$ be a set of such data generating processes that make up our statistical model. We denote this statistical model as $\mathfrak{P}^{d_x \to d_y}(\mathcal{F}, \mathcal{X}, \mathcal{Y}, (x_i)_{i\in[\![n]\!]}, \mathfrak{G}_n)$.

The *standard regression model* is an instance of the generic regression model with the following specifications: Let $T \in \mathbb{R}_{>0}$ be the extent of the *domain of interest* $\mathcal{X} := [0,T]^{d_x}$. Let $\sigma \in \mathbb{R}_{>0}$ be the *variance bound* of the noise. For $i \in [\![n]\!]$, let ε_i be independent $\mathcal{Y} := \mathbb{R}^{d_y}$-valued random variables with $\mathbf{E}[\epsilon_i] = 0$ and $\mathbf{E}[\|\epsilon_i\|_2^2] \le \sigma^2$. We call ε_i the *noise*. For $i \in [\![n]\!]$, set the observations as

$$Y_i = f^\star(x_i) + \varepsilon_i\,. \tag{9}$$

Equation (9) defines a data generating process G_n. The restrictions on ϵ_i define the set $\mathfrak{G}_n$. We denote this statistical model as $\mathfrak{P}^{d_x \to d_y}(\mathcal{F}, T, (x_i)_{i\in[\![n]\!]}, \sigma)$.

The standard task in such models is to estimate $f^\star$ with an estimator $\hat{f}$ given the *data* $\mathcal{D} := (x_i, Y_i)_{i\in[\![n]\!]}$. An *estimator for the regression function* $f^\star$ in $\mathfrak{P}^{d_x \to d_y}(\mathcal{F}, \mathcal{X}, \mathcal{Y}, (x_i)_{i\in[\![n]\!]}, \mathfrak{G}_n)$ is any measurable map

$$\mathcal{E}^{d_x \to d_y}(\cdot,\cdot)\colon (\mathcal{X}\times\mathcal{Y})^n \times \mathcal{X} \to \mathbb{R}^{d_y}\,.$$

Given the data $\mathcal{D}$, it estimates the regression function $f^\star$ as $x \mapsto \hat{f}(x) := \mathcal{E}^{d_x \to d_y}(\mathcal{D}, x)$. We define the term *regression estimator for the derivative* $Df^\star$ of the regression function accordingly.

In the case $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$, we say that $\mathcal{E}^{d_x \to d_y}$ is the *componentwise* estimator $\mathcal{E}^{d_x \to 1}$, if we apply $\mathcal{E}^{d_x \to 1}$ in each target dimension:

$$\mathcal{E}^{d_x \to d_y}(\mathcal{D}, \cdot)\colon \mathcal{X} \to \mathbb{R}^{d_y},\ x \mapsto \left(\mathcal{E}^{d_x \to 1}((x_i, \Pi_1 Y_i)_{i\in[\![n]\!]}, x), \dots, \mathcal{E}^{d_x \to 1}((x_i, \Pi_{d_y} Y_i)_{i\in[\![n]\!]}, x)\right)\,.$$

The *maximal risk* for the model $\mathfrak{P}_n := \mathfrak{P}^{d_x \to d_y}(\mathcal{F}, \mathcal{X}, \mathcal{Y}, (x_i)_{i\in[\![n]\!]}, \mathfrak{G}_n)$ at a point $x_0 \in \mathcal{X}$ of an estimator $\hat{f} = \mathcal{E}^{d_x \to d_y}(\mathcal{D}, \cdot)$ is

$$r(\hat{f}, \mathfrak{P}_n, x_0) := \sup_{G_n \in \mathfrak{G}_n, f^\star \in \mathcal{F}} \mathbf{E}_{G_n}\left[\left\|\hat{f}(x_0) - f^\star(x_0)\right\|_2^2\right]\,, \tag{10}$$

where $\mathbf{E}_{G_n}$ is the expectation with respect to $(Y_1, \dots, Y_n)$ generated by G_n. We say that $r_0^{\mathsf{pr},\mathsf{sup}}(\hat{f}, \mathfrak{P}_n) \in \mathbb{R}_{>0}$ is a sup-norm error bound in probability for an estimator $\hat{f} = \mathcal{E}^{d_x \to d_y}(\mathcal{D}, \cdot)$ if, for all $G_n \in \mathfrak{G}_n$ and $f^\star \in \mathcal{F}$, for $n \to \infty$, we have

$$\sup_{x\in[0,T]^{d_x}} \left\|\hat{f}(x) - f^\star(x)\right\|_2 = \mathbf{O}_{\mathbf{P}}\left(r_0^{\mathsf{pr},\mathsf{sup}}(\hat{f}, \mathfrak{P}_n)\right)\,, \tag{11}$$

where $\hat{f}$ depends on $(Y_1, \dots, Y_n)$ generated by G_n. We define the sup-norm error bound of the first derivative accordingly and denote it by $r_1^{\mathsf{pr},\mathsf{sup}}(\widehat{Df}, \mathfrak{P}_n)$.

An example of a regression estimator $\hat{f}$ suitable for the Hölder-type smoothness class $\mathcal{F} = \Sigma^{d_x \to d_y}(\beta, L)$ with $\beta, L \in \mathbb{R}_{>0}$ (see [21, Definition 2.2]) in the standard regression model $\mathfrak{P}_n = \mathfrak{P}^{d_x \to d_y}(\mathcal{F}, T, (x_i)_{i\in[\![n]\!]}, \sigma)$, is the componentwise local polynomial estimator described in [21, section 3]. *Suitable* here means that it achieves the lowest values possible for $r(\hat{f}, \mathfrak{P}_n, x_0)$ for large enough $n \in \mathbb{N}$, up to a constant.

Further examples of (nonparametric) regression estimators that may be used in the general estimation strategies constructed in this article include orthogonal series estimators [24, chapter 1.7], wavelets [1], penalized splines [25], and neural networks [20].

3. The stubble model and estimation of the increment map

In the Stubble model, we observe many solutions to an ODE with different initial conditions. For each solution, we have only a few observations. The initial conditions are known and located so that they cover the domain of interest.

In this section, we first present an explicit estimation procedure for a Lipschitz-continuous class of model functions with optimal rate of convergence of the mean squared error at a point. The estimation procedure is based on a local constant estimator for the increment map and a linear interpolation for the model function. In the second part of this section, we generalize these results: For a general Hölder-smoothness class, we present a black-box estimation strategy based on an arbitrary multivariate regression estimator for different increment maps and a univariate polynomial interpolation for the model function. If the nonparametric estimator enjoys certain optimality criteria with respect to the standard nonparametric regression problem, it induces an optimal procedure for nonparametric ODE estimation.

3.1. Lipschitz case

In this specific instance of the general ODE estimation model, we consider Lipschitz-continuous functions f, initial conditions x_j on a uniform grid, and an estimation procedure for $f^\star$ that is based on the local constant estimator for the increments $\Upsilon(f^\star, \Delta t, \cdot)$.

3.1.1. Model

The following is a restriction of the standard general model of Section 2.2.

Let $d \in \mathbb{N}$. Set $\beta = 1$. Let $L_0, L_1 \in \mathbb{R}_{>0}$. Set $\mathcal{F} := \bar{\Sigma}^{d\to d}(1, L_0, L_1)$. Let $f^\star \in \mathcal{F}$ and $n_0 \in \mathbb{N}$. Set $n := n_0^d$. Let $x_i \in [0,1]^d$ for $i \in [\![n]\!]$ form a uniform grid in $[0,1]^d$, i.e.,

$$\{x_1, \dots, x_n\} = \left\{ \left(\frac{k_1}{n_0}, \dots, \frac{k_d}{n_0} \right)^\top \middle| \, k_1, \dots, k_d \in [\![n_0]\!] \right\} .$$

Let $\sigma \in \mathbb{R}_{\geq 0}$. Let $\epsilon_j := \epsilon_{j,1}$, $j \in [\![n]\!]$ be independent $\mathbb{R}^d$-valued random variables such that $\mathbf{E}[\epsilon_j] = 0$ and $\mathbf{E}[\|\epsilon_j\|_2^2] \leq \sigma^2$. Let $\Delta t \in \mathbb{R}_{>0}$. Set

$$Y_j := Y_{j,1} := U(f^\star, x_j, \Delta t) + \epsilon_j .$$

We observe Y_j and know x_j and Δt, but $f^\star$ is unknown and to be estimated. We assume d, L_0, L_1, and σ to be fixed. We are interested in upper bounds for the mean squared error at a point $x_0 \in [0,1]^d$ depending on the asymptotics of n and Δt.

3.1.2. Estimator

For $f \in \mathcal{F}$, $\Delta t \in \mathbb{R}_{\geq 0}$, $x \in \mathbb{R}^d$, recall the definition of the increment map $\Upsilon(f, \Delta t, x) = U(f, x, \Delta t) - x$ and set $\iota^\star(x) := \Upsilon(f^\star, \Delta t, x)$. Let $\hat{\iota}$ be the componentwise (see Section 2.4) local constant estimator (Nadaraya–Watson) of $\iota^\star$ with kernel K using the data $(x_j, Y_j)_{j \in [\![n]\!]}$ and a bandwidth of optimal order.

See [21, section 3] with $\beta = 1, \ell = 0, s = 0$ for details on the local constant estimator. For $x \in \mathbb{R}^d$, define the scaled increment estimator of $f^\star$ as

$$\hat{f}(x) := \frac{\hat{\iota}(x)}{\Delta t}. \tag{12}$$

3.1.3. Result

Assumption 3.1. StrictKernel: The kernel $K\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ is Lipschitz-continuous, nonnegative, nontrivial ($K \neq 0$), and there is $C_{\mathsf{ker}} \in \mathbb{R}_{>0}$ such that $K(x) \leq C_{\mathsf{ker}} \mathbb{1}_{[0,1]}(x)$ for all $x \in \mathbb{R}_{\geq 0}$.

Theorem 3.2. *Use the model of Section 3.1.1 and the estimator of Section 3.1.2. Assume* StrictKernel. *Assume* $\Delta t \preccurlyeq 1$. *Then*

$$\mathbf{E}\left[\left\|\hat{f}(x_0) - f^\star(x_0)\right\|_2^2\right] \preccurlyeq \left(\Delta t^2 n\right)^{-\frac{2}{2+d}} + \Delta t^2.$$

Remark 3.3. By [22, Corollary 3.11], the error rate in Theorem 3.2 is minimax optimal.

The following corollary minimizes the error bound with respect to Δt, i.e., it shows the asymptotic behavior of Δt that allows the best estimates of f for the same amount of data.

Corollary 3.4. *Use the model of Section 3.2.1 and the estimator of Section 3.1.2. Assume* StrictKernel. *Assume* $\Delta t \asymp n^{-\frac{1}{4+d}}$. *Then*

$$\mathbf{E}\left[\left\|\hat{f}(x_0) - f^\star(x_0)\right\|_2^2\right] \preccurlyeq n^{-\frac{2}{4+d}}.$$

3.1.4. Proof

Lemma 3.5. *Assume* $f \in \mathcal{F}$. *Let* $\iota := \Upsilon(f, \Delta t, \cdot)$. *Then* $\iota \in \bar{\Sigma}^{d\to d}(1, \tilde{L}_0, \tilde{L}_1)$ *with* $\tilde{L}_0 = \Delta t L_0$ *and* $\tilde{L}_1 = \exp(L_1 \Delta t) - 1$.

Proof. By the definition of U and the fundamental theorem of calculus,

$$U(f, x, t) = x + \int_0^t f(U(f, x, s))\mathrm{d}s.$$

Plugging this into the definition of ι yields

$$\iota(x) = U(f, x, \Delta t) - U(f, x, 0) = \int_0^{\Delta t} f(U(f, x, s))\mathrm{d}s. \tag{13}$$

Thus, as f is uniformly bounded by L_0,

$$|\iota(x)| \leq \int_0^{\Delta t} |f(U(f, x, s))|\,\mathrm{d}s \leq \int_0^{\Delta t} L_0 \mathrm{d}s = \Delta t L_0.$$

This is the value of $\tilde{L}_0$. [21, Lemma 5.9] yields the value of $\tilde{L}_1$ by $|D\iota|_\infty \leq \exp(L_1 \Delta t) - 1$. □

Lemma 3.6. *Assume* $\Delta t \preccurlyeq 1$. *Then, for all* $x_0 \in [0,1]^d$,

$$\mathbf{E}\left[\left\|\hat{\iota}(x_0) - \iota^\star(x_0)\right\|_2^2\right] \preccurlyeq \left(\Delta t^d n^{-1}\right)^{\frac{2}{2+d}}.$$

Proof. According to Lemma 3.5, $\iota^\star$ is Lipschitz-continuous with constant $\tilde{L}_1 = \exp(L_1\Delta t) - 1$. As we consider L_1 to be constant and $\Delta t \preccurlyeq 1$, we obtain $\tilde{L}_1 \preccurlyeq \Delta t$. By [21, Corollary 3.4] and [21, Proposition 3.9], the local constant estimator $\hat{\iota}$ with optimal bandwidth then fulfills

$$\mathbf{E}\left[\left\|\hat{\iota}(x_0) - \iota^\star(x_0)\right\|_2^2\right] \preccurlyeq \Delta t^{\frac{2d}{2+d}} n^{-\frac{2}{2+d}} . \qquad \square$$

Lemma 3.7. *Let $\Delta t \in \mathbb{R}_{>0}$. Assume $f \in \mathcal{F}$. Let $\iota := \Upsilon(f, \Delta t, \cdot)$. Then*

$$\sup_{x\in[0,1]^d} \left\| \frac{\iota(x)}{\Delta t} - f(x) \right\|_2 \le \frac{1}{2} L_0 L_1 \Delta t .$$

Proof. Let $x \in [0,1]^d$. With (13), as f is L_1-Lipschitz,

$$\begin{aligned} \left\| \frac{\iota(x)}{\Delta t} - f(x) \right\|_2 &= \left\| \frac{1}{\Delta t} \int_0^{\Delta t} f(U(f,x,s)) - f(x) \mathrm{d}s \right\|_2 \\ &\le \frac{L_1}{\Delta t} \int_0^{\Delta t} \|U(f,x,s) - x\|_2 \, \mathrm{d}s . \end{aligned}$$

Using $U(f,x,s) = x + \int_0^s f(U(f,x,r))\mathrm{d}r$, we obtain

$$\begin{aligned} \int_0^{\Delta t} \|U(f,x,s) - x\|_2 \, \mathrm{d}s &\le \int_0^{\Delta t} \int_0^s \|f(U(f,x,r))\|_2 \, \mathrm{d}r\mathrm{d}s \\ &\le \int_0^{\Delta t} \int_0^s L_0 \mathrm{d}r\mathrm{d}s \;=\; \frac{1}{2} L_0 \Delta t^2 . \end{aligned}$$

Combining the above inequalities yields

$$\left\| \frac{\iota(x)}{\Delta t} - f(x) \right\|_2 \le \frac{1}{2} L_0 L_1 \Delta t .$$

As this inequality holds for all $x \in [0,1]^d$, we have finished the proof. $\square$

Proof of Theorem 3.2. Recall the definition of $\hat{f}$ in (12) and combine the bounds in Lemma 3.6 and Lemma 3.7 to obtain

$$\begin{aligned} \mathbf{E}\left[\left\|\hat{f}(x_0) - f^\star(x_0)\right\|_2^2\right] &= \mathbf{E}\left[\left\| \frac{\hat{\iota}(x_0) - \iota^\star(x_0)}{\Delta t} + \frac{\iota^\star(x_0)}{\Delta t} - f^\star(x_0) \right\|_2^2\right] \\ &\preccurlyeq \frac{1}{\Delta t^2} \mathbf{E}\left[\left\|\hat{\iota}(x_0) - \iota^\star(x_0)\right\|_2^2\right] + \left\| \frac{\iota^\star(x_0)}{\Delta t} - f^\star(x_0) \right\|_2^2 \\ &\preccurlyeq \Delta t^{-\frac{4}{2+d}} n^{-\frac{2}{2+d}} + \Delta t^2 . \end{aligned}$$

$\square$

3.2. General case

We generalize the Lipschitz setting of Section 3.1 in the following way: We consider an arbitrary smoothness parameter $\beta \in \mathbb{N}$ for the model function $f^\star$. The initial conditions are not restricted to a uniform grid, but must still be *uniform enough* in some sense. We present an estimation strategy that can be used with any regression estimator. The convergence rate results are given in a black-box fashion, i.e., they depend on the convergence rates of the chosen regression estimator. If that estimator achieves the optimal rate of convergence for a standard nonparametric regression problem, the resulting ODE estimator is also optimal in a minimax sense.

Notation 3.8.

(i) Let $d \in \mathbb{N}$. Let $\alpha \in \mathbb{N}_0^d$. Denote $|\alpha| := \sum_{i=1}^d \alpha_i$, and $\alpha! := \prod_{i=1}^d \alpha_i!$, For $x \in \mathbb{R}^d$, denote $x^\alpha := \prod_{i=1}^d x_i^{\alpha_i}$.

(ii) For $\ell, d \in \mathbb{N}$, denote $\mathcal{P}_{d,\ell}$ be the set of all polynomials of degree at most ℓ defined over $\mathbb{R}^d$,

$$\mathcal{P}_{d,\ell} := \left\{ p \colon \mathbb{R}^d \to \mathbb{R} \,\middle|\, p(x) = \sum_{\alpha \in \mathbb{N}_0^d, |\alpha| \le \ell} \beta_\alpha x^\alpha \,,\, \beta_\alpha \in \mathbb{R} \right\}.$$

(iii) Let A be a set. Denote the indicator function of A as $\mathbb{1}_A(\cdot)$.

3.2.1. Model

The following is a restriction of the general ODE model of Section 2.2. Let $d \in \mathbb{N}$, $\beta \in \mathbb{N}$, and $L_{[\![0,\beta]\!]} \subseteq \mathbb{R}_{>0}$. Let the smoothness class be $\mathcal{F}_{d,\beta} \subseteq \bar{\Sigma}^{d\to d}(\beta, L_{[\![0,\beta]\!]})$, see [21, Definition 2.2]. Let $f^\star \in \mathcal{F}_{d,\beta}$. Let $m \in \mathbb{N}$ and $x_1, \ldots, x_m \in \mathbb{R}^d$. Let $\Delta t \in \mathbb{R}_{>0}$. Set $n_j = \beta$ for all $j \in [\![m]\!]$. Set $t_{j,i} = t_i = i\Delta t$ for $i \in [\![0,\beta]\!]$. In the *generic general Stubble model*, the observations are given as follows: Let $\mathfrak{G}_n$ be a set of data generating processes that give rise to the observations

$$(Y_{j,i})_{j\in[\![m]\!],i\in[\![\beta]\!]} \sim G_n\Big(U\big(f^\star, x_j, i\Delta t\big)_{j\in[\![m]\!],i\in[\![\beta]\!]}\Big)\,. \tag{14}$$

We define the *standard general Stubble model* as an instance of the generic one: Let $\sigma \in \mathbb{R}_{\ge 0}$. Let $(\epsilon_{j,i})_{j\in[\![m]\!],i\in[\![\beta]\!]}$ be independent $\mathbb{R}^d$-valued random variables such that $\mathbf{E}[\epsilon_{j,i}] = 0$ and $\mathbf{E}[\|\epsilon_{j,i}\|_2^2] \le \sigma^2$. For $j \in [\![m]\!]$, $i \in [\![\beta]\!]$, let

$$Y_{j,i} = U(f^\star, x_j, i\Delta t) + \epsilon_{j,i}\,.$$

In both versions of the general Stubble model, we observe $Y_{j,i}$, and know x_j and Δt, but $f^\star$ is unknown and to be estimated. We assume d, $L_{[\![0,\beta]\!]}$, and σ to be fixed and we are interested in asymptotic upper bounds for the mean squared error at a point $x_0 \in [0,1]^d$ depending on n and Δt.

Remark 3.9. The results below are the same (up to a multiplicative constant) if we allow $n_j \ge \beta$ as long as $m \asymp n$.

3.2.2. Estimation

For $i \in [\![\beta]\!]$, let $\iota_i^\star = \Upsilon(f^\star, t_i, \cdot)$. Let $\mathfrak{G}_m^i$ be the set of data generating processes obtained by projecting the observations $(Y_{j,i})_{j\in[\![m]\!],i\in[\![\beta]\!]}$ created according to (14) with $G_n \in \mathfrak{G}_n$ to $(Y_{i,1}, \ldots, Y_{i,m})$. Let

$L \in \mathbb{R}_{>0}$. Let $\mathcal{E}_i^{d\to d}$ be an arbitrary regression estimator for the generic regression model

$$\mathfrak{P}^i := \mathfrak{P}^{d\to d}(\Sigma^{d\to d}(\beta, L), [0,1]^d, \mathcal{Y}, (x_j)_{j\in[\![m]\!]}, \mathfrak{G}_m^i),$$

see Section 2.4. Estimate the increment maps $\iota_i^\star$ by $\hat{\iota}_i := \mathcal{E}^{d\to d}((x_j, Y_{j,i} - x_j)_{j\in[\![m]\!]}, \cdot)$. For convenience, set $\hat{\iota}_0(x) := \iota_0^\star(x) := 0$ for all $x \in \mathbb{R}^d$. Denote the maximal risk of the regression estimator, see (10), as

$$r(d, \beta, L, \mathfrak{G}_n, (x_j)_{j\in[\![m]\!]}) := \max_{i\in[\![\beta]\!]} \sup_{x\in[0,1]^d} r(\mathcal{E}_i^{d\to d}, \mathfrak{P}^i, x).$$

For $x_0 \in \mathbb{R}^d$, let $\hat{p}(x_0, \cdot) := (\hat{p}_1(x_0, \cdot), \dots, \hat{p}_d(x_0, \cdot))$, where $\hat{p}_k(x_0, \cdot) \in \mathcal{P}_{1,\beta}$ is the (univariate) polynomial of at most degree β that interpolates $(t_i, \Pi_k \hat{\iota}_i(x_0))_{i\in[\![0,\beta]\!]}$. To estimate $f^\star$, we locally approximate $U(f^\star, x_0, t)$ by $x_0 + \hat{p}(x_0, t)$, i.e., we consider $\dot{\hat{p}}(x_0, t) \approx f^\star(x_0 + \hat{p}(x_0, t))$. As $f^\star(x_0) = f^\star(x_0 + \hat{p}(x_0, 0))$, this approach yields the estimator

$$\hat{f}(x_0) := \dot{\hat{p}}(x_0, 0). \tag{15}$$

Remark 3.10.

(i) The value of L is to be specified later and depends on Δt, β, and $L_{[\![0,\beta]\!]}$. Typically the estimator $\mathcal{E}_i^{d\to d}$ does not depend on L, but the risk $r(\mathcal{E}_i^{d\to d}, \mathfrak{P}^i, x)$ does.

(ii) If we use the standard general Stubble model, set $\beta = 1$, and use the componentwise local constant estimator for $\mathcal{E}^{d\to d}$, we obtain the procedure of Section 3.1.

(iii) The polynomial interpolation in the second step can be viewed as an Adams–Bashforth method: a linear multistep method for solving ordinary differential equations numerically, see [4, Chapter 24].

3.2.3. Result

Theorem 3.11. *Use the generic general Stubble model of Section 3.2.1 and the estimator of Section 3.2.2. Assume $\Delta t \preccurlyeq 1$. Then*

$$\mathbf{E}\left[\left\|\hat{f}(x_0) - f^\star(x_0)\right\|_2^2\right] \preccurlyeq \Delta t^{-2} r(d, \beta, C\Delta t, \mathfrak{G}_n, (x_j)_{j\in[\![m]\!]}) + \Delta t^{2\beta}.$$

for some constant $C \in \mathbb{R}_{>0}$ depending only on β and $L_{[\![0,\beta]\!]}$.

The minimax rate for the squared error in the standard regression problem with smoothness class $\Sigma^{d\to d}(\beta, L)$ is $L^{\frac{2d}{2\beta+d}} n^{-\frac{2\beta}{2\beta+d}}$ up to some constants. It is achieved, for example, by the local polynomial estimator with optimal bandwidth under some conditions on x_i, which are fulfilled by a uniform grid. See [21, Corollary 3.4]. Using the standard general Stubble model and considering $m \asymp n$, we then have

$$r(d, \beta, C\Delta t, \mathfrak{G}_n, (x_j)_{j\in[\![m]\!]}) \preccurlyeq \Delta t^{\frac{2d}{2\beta+d}} n^{-\frac{2\beta}{2\beta+d}}.$$

The conditions for the optimal rate for the local polynomial estimator with kernel K of [21, section 3] are as follows.

Assumption 3.12.

- EIGENVALUE: There is $C_{\mathsf{egv}} \in \mathbb{R}_{>0}$ such that $\lambda_{\min}(B(x))^{-1} \le C_{\mathsf{egv}}$ for all $x \in [0,T]^d$, where $B(x)$ is defined in [21, Eq. (5)].

- COVER: There is $C_{\mathsf{cvr}} \in \mathbb{R}_{>0}$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\mathrm{B}^d(x,r)}(x_i) \leq \max\left(\frac{1}{n}, C_{\mathsf{cvr}}\left(\frac{r}{T}\right)^d\right),$$

  for all $x \in [0,T]^d$, $r \in \mathbb{R}_{>0}$.
- KERNEL: The support of the kernel $K\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ fulfills $\operatorname{supp}(K) \subseteq [0,1]$. Furthermore, there is $C_{\mathsf{ker}} \in \mathbb{R}_{>0}$ such that $K(z) \leq C_{\mathsf{ker}}$ for all $z \in \mathbb{R}_{\geq 0}$.

The following corollary shows the result of Theorem 3.11 using minimax optimal estimators $\mathcal{E}_i^{d\to d}$ in the standard general Stubble model.

Corollary 3.13. *Use the standard general Stubble model of Section 3.2.1. Use the estimator of Section 3.2.2 with the componentwise local polynomial estimator of degree $\ell := \beta - 1$ with kernel K and optimal bandwidth for all estimators $\mathcal{E}_i^{d\to d}$. Assume $\Delta t \preccurlyeq 1$. Assume* COVER, EIGENVALUE, *and* KERNEL. *Then, for all $x_0 \in [0,1]^d$,*

$$\mathbf{E}\left[\left\|\hat{f}(x_0) - f^\star(x_0)\right\|_2^2\right] \preccurlyeq \left(\Delta t^2 n\right)^{-\frac{2\beta}{2\beta+d}} + \Delta t^{2\beta}.$$

Remark 3.14.

(i) Of course, the local polynomial estimator can be replaced by any estimator that achieves the same (optimal) rate of convergence.
(ii) By [22, Corollary 3.11], the error rate in Corollary 3.13 is minimax optimal for $\mathcal{F}_{d,\beta} = \bar{\Sigma}^{d\to d}(\beta, L_{[\![0,\beta]\!]})$.

The following corollary minimizes the error bound with respect to Δt, i.e., it shows the asymptotic behavior of Δt that allows the best estimates of f for the same amount of data.

Corollary 3.15. *Use the setting and assumptions of Corollary 3.13. Assume*

$$\Delta t \asymp n^{-\frac{1}{2(\beta+1)+d}}.$$

Then

$$\mathbf{E}\left[\left\|\hat{f}(x_0) - f^\star(x_0)\right\|_2^2\right] \preccurlyeq n^{-\frac{2\beta}{2(\beta+1)+d}}.$$

3.2.4. Proof

For $x_0 \in \mathbb{R}^d$, let $p^\star(x_0,\cdot) \in \mathcal{P}_{1,\beta}$ be the (univariate) polynomial of at most degree β that interpolates $(t_i, \iota_i^\star(x_0))_{i\in[\![0,\beta]\!]}$. By the definition of $\hat{f}$ in (15), we can write

$$\hat{f}(x_0) - f^\star(x_0) = \dot{\hat{p}}(x_0,0) - \dot{p}^\star(x_0,0) + \dot{p}^\star(x_0,0) - f^\star(x_0).$$

Using this, we split the error into two parts,

$$\mathbf{E}\left[\left\|\hat{f}(x_0) - f^\star(x_0)\right\|_2^2\right] \leq 2\mathbf{E}\left[\left\|\dot{\hat{p}}(x_0,0) - \dot{p}^\star(x_0,0)\right\|_2^2\right] + 2\left\|\dot{p}^\star(x_0,0) - f^\star(x_0)\right\|_2^2. \tag{16}$$

We start with the second part: The polynomial $t \mapsto p^\star(x_0,t)$ interpolates the translated solution $\bar{u}^\star(x_0,t) := U(f^\star,x_0,t) - x_0$. By [21, Corollary 5.7], we have $\left|D^{\beta+1}\bar{u}^\star(x_0,\cdot)\right|_\infty \le C_L$, where C_L depends only on β and $L_{[\![0,\beta]\!]}$. By the definition of $\bar{u}^\star(x_0,t)$, it fulfills $\dot{\bar{u}}^\star(x_0,0) = f^\star(x_0)$. Thus, [21, Lemma 4.2] provides a constant C_β depending only on β such that

$$\begin{aligned}\left\|\dot{p}^\star(x_0,0) - f^\star(x_0)\right\|_2 &= \left\|\dot{p}^\star(x_0,0) - \dot{\bar{u}}^\star(x_0,0)\right\|_2 \\ &\le C_\beta C_L \Delta t^\beta .\end{aligned}$$

In other words,

$$\left\|\dot{p}^\star(x_0,0) - f^\star(x_0)\right\|_2^2 \preccurlyeq \Delta t^{2\beta} . \tag{17}$$

Let us turn our attention to the first part of the right-hand side of (16): By [21, Corollary 5.12], $\left|D^\beta \iota_i^\star\right|_\infty \le C\Delta t$ for some constant C depending only on $d, \beta, L_{[\![0,\beta]\!]}$ and the maximal Δt as $f^\star \in \bar{\Sigma}^{d\to d}(\beta, L_{[\![0,\beta]\!]})$ and $\Delta t \preccurlyeq 1$. As all $\mathcal{E}_i^{d\to d}$ achieve the rate of convergence $r(d,\beta,L,\mathfrak{G}_n,(x_j)_{j\in[\![m]\!]})$ on $\Sigma^{d\to d}(\beta,L)$, we obtain

$$\mathbf{E}\left[\max_{i\in[\![0,\beta]\!]}\left\|\hat{\iota}_i(x_0) - \iota_i^\star(x_0)\right\|_2^2\right] \le (\beta+1)\max_{i\in[\![\beta]\!]}\mathbf{E}\left[\left\|\hat{\iota}_i(x_0) - \iota_i^\star(x_0)\right\|_2^2\right] \preccurlyeq r(d,\beta,C\Delta t,\mathfrak{G}_n,(x_j)_{j\in[\![m]\!]}) . \tag{18}$$

Applying [21, Lemma 4.3] with the bound (18) yields

$$\mathbf{E}\left[\left\|\hat{\dot{p}}(x_0,0) - \dot{p}^\star(x_0,0)\right\|_2^2\right] \preccurlyeq \Delta t^{-2} r(d,\beta,C\Delta t,\mathfrak{G}_n,(x_j)_{j\in[\![m]\!]}) . \tag{19}$$

We combine (17) and (19) and get

$$\mathbf{E}\left[\left\|\hat{f}(x_0) - f^\star(x_0)\right\|_2^2\right] \preccurlyeq \Delta t^{-2} r(d,\beta,C\Delta t,\mathfrak{G}_n,(x_j)_{j\in[\![m]\!]}) + \Delta t^{2\beta} .$$

4. The snake model and estimation of the observed solutions

In the Snake model, we observe one (or a few) solution(s) to an ODE. For each solution, we have many observations. The associated estimation problem depends strongly on how the trajectories of the solutions are located in the state space.

In this section, we first present an explicit estimation procedure for a Lipschitz-continuous class of model functions and an upper bound on the maximal risk in sup norm. The estimation procedure is based on a local linear estimator for the solutions and a nearest neighbor interpolation for the model function. It is rate-optimal in some settings. In the second part of this section, we generalize these results: For a general Hölder-smoothness class, we present a black-box estimation strategy based on a generic nonparametric estimator for the solutions and a multivariate polynomial interpolation for the model function. If the nonparametric estimator enjoys certain optimality criteria with respect to the standard nonparametric regression problem, it induces a procedure for nonparametric ODE estimation that is optimal in some settings.

4.1. Lipschitz case

In this specific instance of the general ODE estimation model, we consider Lipschitz-continuous model functions $f^\star$ and an estimation procedure $\hat{f}$ that is based on local linear estimators $\hat{u}$ and $\hat{\dot{u}}$ for the

observed solution $U(f^\star, x_1, \cdot)$ and its derivative, and a nearest neighbor interpolation of $\hat{u}(t) \mapsto \hat{\hat{u}}(t)$ for the model function. We obtain an upper bound on the expected sup norm error. For this, we either require the trajectories $U(f^\star, x_1, [0, T_1])$ to cover the fixed domain of interest $[0,1]^d$ in a suitable way or we adapt our domain of interest to these trajectories, i.e., the upper bound only holds close to the trajectory.

Notation 4.1. Let $d \in \mathbb{N}$, $r \geq 0$, and $z \in \mathbb{R}^d$. Denote the closed ball with radius r around z as

$$\mathrm{B}^d(z, r) := \left\{ x \in \mathbb{R}^d \,\middle|\, \|x - z\|_2 \leq r \right\} .$$

Let $\mathcal{Z} \subseteq \mathbb{R}^d$. Denote the closed ball with radius r around $\mathcal{Z}$ as

$$\mathrm{B}^d(\mathcal{Z}, r) := \bigcup_{z \in \mathcal{Z}} \mathrm{B}^d(z, r) .$$

4.1.1. Model

The following is a restriction of the general model of Section 2.2. Let $d \in \mathbb{N}_{\geq 2}$ and $L_0, L_1 \in \mathbb{R}_{>0}$. Set $\mathcal{F} := \bar{\Sigma}^{d \to d}(1, L_0, L_1)$. Let $f^\star \in \mathcal{F}$. Set $m = 1$. Let $x_1 \in \mathbb{R}^d$ and $n = n_1 \in \mathbb{N}$. Let $\Delta t \in \mathbb{R}_{>0}$ possibly changing with n. Set $t_{1,i} := t_i := i\Delta t$ for $i = [\![0, n]\!]$ and $T := T_1 := n\Delta t$. Let $\sigma \in \mathbb{R}_{\geq 0}$. Let $\epsilon_i := \epsilon_{1,i}$, $i \in [\![n]\!]$ be independent $\mathbb{R}^d$-valued random variables such that $\mathbf{E}[\epsilon_i] = 0$ and $\mathbf{E}[\|\epsilon_i\|_2^2] \leq \sigma^2$. Set

$$Y_i := Y_{1,i} := U(f^\star, x_1, t_i) + \epsilon_i \qquad \text{for } i \in [\![n]\!] .$$

We observe Y_i and know x_1 and Δt, but $f^\star$ is unknown and to be estimated. We assume d, L_0, L_1, and σ to be fixed. For an estimator $\hat{f}$, we are interested in asymptotic upper bounds for the mean squared sup-norm of $f^\star - \hat{f}$ in the domain of interest $[0,1]^d$ depending on the asymptotics of n and Δt.

4.1.2. Estimator

Let $\hat{u} \colon [0, T] \to \mathbb{R}$ be the componentwise local linear estimator (as described in [21, Eq. (8)] with $\ell = 1$, $s = 0$, $d = 1$) of $u^\star := U(f^\star, x_1, \cdot)$ using the data $(t_i, Y_i)_{i \in [\![n]\!]}$. Similarly, denote by $\hat{\hat{u}}$ the local linear estimator of the first derivative of $U(f^\star, x_1, \cdot)$ ($\ell = 1$, $s = 1$, $d = 1$ in terms of [21, section 3]). We assume that the bandwidth h is chosen of optimal order for the error bound in sup-norm, see [21, Corollary 3.7]. Let $\hat{f}$ be the nearest neighbor interpolation of $(\hat{u}, \hat{\hat{u}})$: For $x \in \mathbb{R}^d$ and a continuous function $u \colon [0, T] \to \mathbb{R}^d$, let

$$t_{\mathsf{NN}}(u, T, x) \in \operatorname*{arg\,min}_{t \in [0,T]} \|u(t) - x\|_2$$

with an arbitrary choice if the minimizer is not unique. Then, for $x \in \mathbb{R}^d$, we define the estimator of $f^\star$,

$$\hat{f}(x) := \hat{\hat{u}}(t_{\mathsf{NN}}(\hat{u}, T, x)) . \tag{20}$$

4.1.3. Result

For our results in sup-norm, we require the noise to be sub-Gaussian, which is a standard assumption for regression results in sup-norm, see [24, section 1.6.2].

Definition 4.2 (Sub-Gaussian). A real-valued random variable Z is called *sub-Gaussian* if $\mathbf{E}[|Z|] < \infty$ and there is $v \in \mathbb{R}_{>0}$ such that

$$\mathbf{P}(|Z - \mathbf{E}[Z]| \geq t) \leq 2\exp\left(-\frac{t^2}{2v}\right)$$

for all $t \in \mathbb{R}_{>0}$. In this case, v is called *variance parameter*.

Assumption 4.3. SUBGAUSSIAN: In each component, the noise ε_i is sub-Gaussian with variance parameter σ^2.

For $\mathcal{X} \subseteq \mathbb{R}^d$ and a function $u\colon [0,T] \to \mathbb{R}^d$, define $\delta_{\mathsf{NN}}(u,T,\mathcal{X})$ as smallest radius so that the ball around $u([0,T])$ includes $\mathcal{X}$:

$$\delta_{\mathsf{NN}}(u,T,\mathcal{X}) := \sup_{x\in\mathcal{X}} \inf_{t\in[0,T]} \|u(t) - x\|_2 .$$

The results come in two versions with respect to the domain of interest: We either fix the domain of interest, e.g., $\mathcal{X} = [0,1]^d$ and make the error bound depend on $\delta_{\mathsf{NN}}(u^\star,T,\mathcal{X})$, or we fix δ (or a sequence $\delta = \delta_n$) and adapt our domain of interest to be $\mathcal{X} = \mathrm{B}^d(u^\star([0,T]),\delta)$.

Theorem 4.4. *Use the model of Section 4.1.1 and the estimator of Section 4.1.2. Assume* STRICTKERNEL, SUBGAUSSIAN. *Assume*

$$\left(\frac{n}{\log(n)}\right)^{-\frac{1}{4}} \preccurlyeq T \preccurlyeq n\log(n)^{\frac{1}{4}} . \tag{21}$$

(i) *Let $\delta \in \mathbb{R}_{\geq 0}$, potentially changing with n. Let $\mathcal{X} := \mathrm{B}^d(u^\star([0,T]),\delta)$. Then*

$$\mathbf{E}\left[\sup_{x\in\mathcal{X}} \left\|\hat{f}(x) - f^\star(x)\right\|_2^2\right] \preccurlyeq \delta^2 + \left(\frac{T\log n}{n}\right)^{\frac{2}{5}} .$$

(ii) *Let $\delta = \delta_{\mathsf{NN}}(u^\star,T,[0,1]^d)$. Then*

$$\mathbf{E}\left[\sup_{x\in[0,1]^d} \left\|\hat{f}(x) - f^\star(x)\right\|_2^2\right] \preccurlyeq \delta^2 + \left(\frac{T\log n}{n}\right)^{\frac{2}{5}} .$$

Remark 4.5.

(i) Compare the upper error bound in Theorem 4.4 (ii) with the lower error bound [22, Corollary 4.11] (for simplicity, we here ignore all factors in the error bounds that are polynomial in $\log(n)$):

$$\delta + \delta^{\frac{d-1}{4+d}}\left(\frac{T}{n}\right)^{\frac{1}{4+d}} \preccurlyeq \mathbf{E}\left[\sup_{x\in[0,1]^d} \left\|\hat{f}(x) - f^\star(x)\right\|_2^2\right]^{\frac{1}{2}} \preccurlyeq \delta + \left(\frac{T}{n}\right)^{\frac{1}{5}} .$$

Direct calculations yield

$$\left(\frac{T}{n}\right)^{\frac{1}{5}} \leq \delta \quad\Leftrightarrow\quad \left(\frac{T}{n}\right)^{\frac{1}{5}} \leq \delta^{\frac{d-1}{4+d}}\left(\frac{T}{n}\right)^{\frac{1}{4+d}} \quad\Leftrightarrow\quad \delta^{\frac{d-1}{4+d}}\left(\frac{T}{n}\right)^{\frac{1}{4+d}} \leq \delta .$$

Thus, the error rate in Theorem 4.4 (ii) is minimax optimal if

$$\left(\frac{T}{n}\right)^{\frac{1}{5}} \preccurlyeq \delta\,, \tag{22}$$

ignoring log-factors.

(ii) Note that [22, Corollary 4.11] uses $m \succ 1$ many solutions. But in [22, Appendix E], it is argued that $m = 1$ suffices.

(iii) One can think of (22) as the requirement of large distances between temporally different parts of trajectory. If this is not the case, then the estimator seems suboptimal: For the estimation of $\hat{u}$ and $\hat{\dot{u}}$, it ignores information from observations that are distant in time but close in state space. An estimator that is optimal for regimes where different trajectory parts are not well-separated, must not ignore this information.

Corollary 4.6. *Use the model of Section 4.1.1 and the estimator of Section 4.1.2. Assume* StrictKernel, SubGaussian.

(i) *Let $\delta \in \mathbb{R}_{\geq 0}$, potentially changing with n. Let $\mathcal{X} := \mathrm{B}^d(u^\star([0,T]), \delta)$. Assume $T \preccurlyeq \left(\frac{n}{\log n}\right)^{\frac{d-1}{4+d}}$ and $\delta \preccurlyeq \left(\frac{n}{\log n}\right)^{-\frac{1}{4+d}}$. Then*

$$\mathbf{E}\left[\sup_{x\in\mathcal{X}} \left\|\hat{f}(x) - f^\star(x)\right\|_2^2\right] \preccurlyeq \left(\frac{n}{\log n}\right)^{-\frac{2}{4+d}}.$$

(ii) *Let $\delta = \delta_{\mathsf{NN}}(u^\star, T, [0,1]^d)$. Assume $T \preccurlyeq \left(\frac{n}{\log n}\right)^{\frac{d-1}{4+d}}$ and $\delta \preccurlyeq \left(\frac{n}{\log n}\right)^{-\frac{1}{4+d}}$. Then*

$$\mathbf{E}\left[\sup_{x\in[0,1]^d} \left\|\hat{f}(x) - f^\star(x)\right\|_2^2\right] \preccurlyeq \left(\frac{n}{\log n}\right)^{-\frac{2}{4+d}}.$$

Remark 4.7. The conditions on T and δ ensure an optimal trade-off between the ability to reconstruct $u^\star$ and coverage of $[0,1]^d$. They can be fulfilled for certain f, but it is rather restrictive. Furthermore, to cover the hypercube $[0,1]^d$ with $\mathrm{B}^d(u([0,T]), \delta)$, where u has a speed bounded by L, we require $T \succcurlyeq \delta^{-(d-1)}$. Hence, the conditions on δ and T in Theorem 4.4 are tight in the sense that one variable determines the other up to a constant.

4.1.4. Proof

Lemma 4.8. *Let $\delta > 0$. Let $x_0 \in \mathrm{B}^d(u^\star([0,T]), \delta)$. Then*

$$\left\|\hat{f}(x_0) - f^\star(x_0)\right\|_2 \leq L_1\left(\delta + 2\sup_{t\in[0,T]}\left\|\hat{u}(t) - u^\star(t)\right\|_2\right) + \sup_{t\in[0,T]}\left\|\hat{\dot{u}}(t) - \dot{u}^\star(t)\right\|_2.$$

Proof. Recall $u^\star = U(f^\star, x_1, \cdot)$. Denote $\tau^\star := t_{\mathsf{NN}}(u^\star, T, x_0)$ and $\hat{\tau} := t_{\mathsf{NN}}(\hat{u}, T, x_0)$. Using the triangle inequality and the definition of $\hat{f}$ (20) and $\dot{u}^\star$, we obtain

$$\begin{aligned}\left\|\hat{f}(x_0) - f^\star(x_0)\right\|_2 &\leq \left\|\hat{f}(x_0) - f^\star(u^\star(\hat{\tau}))\right\|_2 + \left\|f^\star(u^\star(\hat{\tau})) - f^\star(x_0)\right\|_2 \\ &= \left\|\hat{\dot{u}}(\hat{\tau}) - \dot{u}^\star(\hat{\tau})\right\|_2 + \left\|f^\star(u^\star(\hat{\tau})) - f^\star(x_0)\right\|_2.\end{aligned}$$

Let us consider the second term of the last line. As $f^\star$ is L_1-Lipschitz, we have

$$\left\|f^\star(u^\star(\hat\tau)) - f^\star(x_0)\right\|_2 \le L_1 \left\|u^\star(\hat\tau) - x_0\right\|_2 .$$

We use the triangle inequality in $\left\|u^\star(\hat\tau) - x_0\right\|_2 \le \left\|u^\star(\hat\tau) - \hat u(\hat\tau)\right\|_2 + \|\hat u(\hat\tau) - x_0\|_2$ and bound

$$\begin{aligned}\|\hat u(\hat\tau) - x_0\|_2 &\le \left\|\hat u(\tau^\star) - x_0\right\|_2 \\ &\le \left\|\hat u(\tau^\star) - u^\star(\tau^\star)\right\|_2 + \left\|u^\star(\tau^\star) - x_0\right\|_2 ,\end{aligned}$$

where we used the minimizing property of $\hat\tau$ and again the triangle inequality. Putting all previous bounds together, we obtain

$$\left\|\hat f(x_0) - f^\star(x_0)\right\|_2 \le \left\|\hat u(\hat\tau) - \dot u^\star(\hat\tau)\right\|_2 + L_1\left(\left\|u^\star(\hat\tau) - \hat u(\hat\tau)\right\|_2 + \left\|\hat u(\tau^\star) - u^\star(\tau^\star)\right\|_2 + \left\|u^\star(\tau^\star) - x_0\right\|_2\right).$$

The claim follows by taking supremum over all possible values of $\tau^\star, \hat\tau \in [0,T]$ and noting the definition of x_0 and $\tau^\star$. □

Proof of Theorem 4.4. As $f^\star \in \bar\Sigma(1, L_0, L_1)$, we have $u^\star \in \bar\Sigma(2, \infty, L_0, \tilde L_1)$, where $\tilde L_1$ depends on L_0 and L_1, by [21, Corollary 5.7]. We want to apply [21, Corollary 3.7] to obtain bounds on the sup-norm of the error for $\hat u$ and $\hat{\dot u}$. StrictKernel together with the uniform grid $t_i = i\Delta t$ implies Kernel, Eigenvalue, and Cover by [21, Proposition 3.9]. SubGaussian is assumed and (21) implies [21, Eq. (9)]. Thus, [21, Corollary 3.7] yields

$$\mathbf{E}\left[\sup_{t\in[0,T]} \left\|\hat u(t) - u^\star(t)\right\|_2^2\right] \preccurlyeq \left(\frac{T\log(n)}{n}\right)^{\frac45} \qquad \text{and} \qquad \mathbf{E}\left[\sup_{t\in[0,T]} \left\|\hat{\dot u}(t) - \dot u^\star(t)\right\|_2^2\right] \preccurlyeq \left(\frac{T\log(n)}{n}\right)^{\frac25}.$$

Together with Lemma 4.8, we obtain

$$\begin{aligned}\mathbf{E}\left[\sup_{x\in[0,1]^d} \left\|\hat f(x) - f^\star(x)\right\|_2^2\right] &\preccurlyeq \delta^2 + \mathbf{E}\left[\sup_{t\in[0,T]} \left\|\hat u(t) - u^\star(t)\right\|_2^2\right] + \mathbf{E}\left[\sup_{t\in[0,T]} \left\|\hat{\dot u}(t) - \dot u^\star(t)\right\|_2^2\right] \\ &\preccurlyeq \delta^2 + \left(\frac{T\log(n)}{n}\right)^{\frac45} + \left(\frac{T\log(n)}{n}\right)^{\frac25} \preccurlyeq \delta^2 + \left(\frac{T\log(n)}{n}\right)^{\frac25}.\end{aligned}$$

□

4.2. General case

We now want to consider estimation under higher-order smoothness. We generalize the estimation procedure of Section 4.1.2 as follows: We replace the local linear estimation of the solution $u^\star$ and its derivative $\dot u^\star$ by arbitrary regression estimators $\mathcal{E}^{1\to d}$ and $\widetilde{\mathcal{E}}^{1\to d}$, respectively. E.g., this could be local polynomial estimators of the appropriate degree. Furthermore, the nearest neighbor interpolation is replaced by multivariate polynomial interpolation of the appropriate degree. The convergence rate results are given in a black-box fashion, i.e., they depend on the convergence rates of the chosen regression estimators $\mathcal{E}^{1\to d}$ and $\widetilde{\mathcal{E}}^{1\to d}$.

Notation 4.9.

(i) Let $A \subseteq \mathbb{R}^d$. Denote the diameter of A as $\mathsf{diam}(A) := \sup_{a,a'\in A} \|a - a'\|_2$.

(ii) Let $A \subseteq \mathbb{R}^d$. Denote the convex hull of A as

$$\mathsf{ch}(A) := \left\{ \sum_{k=1}^{K} w_i a_i \,\middle|\, K \in \mathbb{N},\ a_1, \dots, a_K \in A,\ w_1, \dots, w_K \in [0,1], \sum_{k=1}^{K} w_i = 1 \right\}.$$

(iii) Let $A \subseteq \mathbb{R}^d$, $\mu \in \mathbb{R}_{\geq 0}$. Denote the relative μ-interior of the convex hull of A as

$$\mathsf{ch}_\mu(A) := \left\{ x \in \mathbb{R}^d \,\middle|\, \mathrm{B}^d(x, \mu\, \mathsf{diam}(A)) \subseteq \mathsf{ch}(A) \right\}.$$

4.2.1. Model

The following is a restriction of the general ODE model of Section 2.2. Let $d \in \mathbb{N}_{\geq 2}$, $\beta \in \mathbb{N}_{\geq 2}$, and $L_{[\![0,\beta]\!]} \subseteq \mathbb{R}_{>0}$. Let $\mathcal{F}_{d,\beta} \subseteq \bar{\Sigma}^{d\to d}(\beta, L_{[\![0,\beta]\!]})$, see [21, Definition 2.2]. Let $f^\star \in \mathcal{F}_{d,\beta}$. Set $m = 1$. Let $x_1 \in \mathbb{R}^d$ and $n = n_1 \in \mathbb{N}$. Let $T := T_1 \in \mathbb{R}_{>0}$ possibly changing with n. Set $t_{1,i} := t_i \in [0,T]$ for $i \in [\![0,n]\!]$ with $0 = t_0 \leq \cdots \leq t_n = T$. The values of t_i may change with n. In the *generic general Snake model*, the observations are given as follows: Let $\mathfrak{G}_n$ be a set of data generating processes that give rise to the observations

$$(Y_i)_{i\in[\![n]\!]} := (Y_{1,i})_{i\in[\![n]\!]} \sim G_n\Big(U\big(f^\star, x_1, t_i\big)_{i\in[\![n]\!]}\Big). \tag{23}$$

We define the *standard general Snake model* as an instance of the generic one: Let $\sigma \in \mathbb{R}_{\geq 0}$. Let $\epsilon_i := \epsilon_{1,i}$, $i \in [\![n]\!]$ be independent $\mathbb{R}^d$-valued random variables such that $\mathbf{E}[\epsilon_i] = 0$ and $\mathbf{E}[\|\epsilon_i\|_2^2] \leq \sigma^2$. Set

$$Y_i := Y_{1,i} := U(f^\star, x_1, t_i) + \epsilon_i \qquad \text{for } i \in [\![n]\!].$$

In both versions of the general Snake model, we observe Y_i and know x_1 and t_i, but $f^\star$ is unknown and to be estimated. We assume d, $L_{[\![0,\beta]\!]}$, and σ to be fixed. For an estimator $\hat{f}$, we are interested in asymptotic upper bounds for the sup-norm of $f^\star - \hat{f}$ in the domain of interest $[0,1]^d$ in probability depending on the asymptotics of n and T.

Remark 4.10.

(i) It does not matter whether x_1 is known or not: Below, we derive upper bounds on the estimation error for the potentially more difficult problem of unknown x_1. The lower bounds from [22, Corollary 4.11] are formulated with known x_1, which are then also lower bounds for the case of unknown x_1.

(ii) This model and following estimation procedure can easily be generalized to larger m as long as $\min_{j\in[\![m]\!]} n_j \succ 1$ fast enough and the $t_{j,i}$ behave in an appropriate way.

4.2.2. Multivariate polynomial interpolation

The estimation strategy for the model described in Section 4.2.1 uses multivariate polynomial interpolation. In this section, we introduce some basics of polynomial interpolation and some further objects that are required to describe the ODE estimator. Recall Notation 3.8 on polynomials.

Definition 4.11 (Polynomial interpolation). Let $\ell, d_{\mathsf{x}}, d_{\mathsf{y}} \in \mathbb{N}$.

(i) Set

$$N := N_{d_{\mathsf{x}},\ell} := \dim(\mathcal{P}_{d_{\mathsf{x}},\ell}) = \binom{\ell + d_{\mathsf{x}}}{d_{\mathsf{x}}}.$$

(ii) For $x \in \mathbb{R}^{d_\mathsf{x}}$, denote the vector of monomials of degree at most ℓ in d_x dimensions as

$$\psi(x) := \psi_\ell(x) := (x^\alpha)_{|\alpha| \le \ell} \in \mathbb{R}^N .$$

For $M \in \mathbb{N}$, $\mathbf{x} = (x_1, \dots, x_M) \in (\mathbb{R}^{d_\mathsf{x}})^M$, denote $\Psi(\mathbf{x}) := \Psi_\ell(\mathbf{x}) := (\psi(x_1), \dots, \psi(x_M))^\top \in \mathbb{R}^{M \times N}$.

(iii) For $x \in \mathbb{R}^{d_\mathsf{x}}$, $\mathbf{x} \in (\mathbb{R}^{d_\mathsf{x}})^N$ with $\Psi(\mathbf{x}) \in \mathbb{R}^{N \times N}$, $\mathbf{y} \in (\mathbb{R}^{d_\mathsf{y}})^N = \mathbb{R}^{N \times d_\mathsf{y}}$, denote the *componentwise polynomial interpolation* as

$$I(\mathbf{x}, \mathbf{y}, \cdot) := I_\ell(\mathbf{x}, \mathbf{y}, \cdot) \colon \mathbb{R}^{d_\mathsf{x}} \to \mathbb{R}^{d_\mathsf{y}},\ x \mapsto \begin{cases} \psi(x)^\top \Psi(\mathbf{x})^{-1} \mathbf{y} & \text{if } \Psi(\mathbf{x}) \text{ is invertible,} \\ 0 & \text{otherwise.} \end{cases}$$

Remark 4.12.

(i) Let $\mathbf{x} = (x_1, \dots, x_N) \in (\mathbb{R}^{d_\mathsf{x}})^N$ and $\mathbf{y} = (y_1, \dots, y_N) \in (\mathbb{R}^{d_\mathsf{y}})^N$. The function $I(\mathbf{x}, \mathbf{y}, \cdot)$ is an *interpolation* as $I(\mathbf{x}, \mathbf{y}, x_k) = y_k$. It is *polynomial* as $x \mapsto \Pi_k I(\mathbf{x}, \mathbf{y}, x) \in \mathcal{P}_{d_\mathsf{x}, \ell}$ for each $k \in [\![d_\mathsf{y}]\!]$. It is *componentwise* as

$$I(\mathbf{x}, \mathbf{y}, x) = (I(\mathbf{x}, (\Pi_k y_1, \dots \Pi_k y_N), x))_{k \in [\![d_\mathsf{y}]\!]} .$$

(ii) The definition of $I(\mathbf{x}, \mathbf{y}, \cdot) = 0$ if $\Psi(\mathbf{x})$ is not invertible is for convenience only. The value 0 is not of importance.

(iii) If $d_\mathsf{x} = 1$, the matrix $\Psi(\mathbf{x}) \in \mathbb{R}^{N \times N}$ is also called *Vandermonde matrix*. It is invertible if and only if all $x_k \in \mathbb{R}$ are distinct. In the multivariate case, conditions for invertibility of $\Psi(\mathbf{x})$ (and hence, unique existence of the polynomial interpolation) are more complex, see [16, Theorem 4.1].

The geometric configuration of the base points $\mathbf{x}$ for the multivariate polynomial interpolation influences the approximation quality. We next define a normalization that removes location and scale from $\mathbf{x}$ to isolate the geometric configuration.

Definition 4.13 (Normalization). Let $d, N \in \mathbb{N}$. For $\mathbf{x} = (x_1, \dots, x_N) \in (\mathbb{R}^d)^N$ with $\mathsf{diam}(\mathbf{x}) > 0$, let

$$\eta_{\mathbf{x}} \colon \mathbb{R}^d \to \mathbb{R}^d, x \mapsto \frac{x - \frac{1}{N} \sum_{k=1}^N x_k}{\mathsf{diam}(\mathbf{x})} .$$

Furthermore, denote $\eta_{\mathbf{x}}(\tilde{\mathbf{x}}) = (\eta_{\mathbf{x}}(\tilde{x}_1), \dots, \eta_{\mathbf{x}}(\tilde{x}_M))$ for $\tilde{\mathbf{x}} = (\tilde{x}_1, \dots, \tilde{x}_M) \in (\mathbb{R}^d)^M$, $M \in \mathbb{N}$.

Next, we define the set $\mathcal{H}$ of sets of base points $\mathbf{x}$ that can be used in a polynomial interpolation to suitably approximate the function value at x, the set $\mathcal{G}$ of points x with $\mathcal{H} \neq \varnothing$, and an (almost) optimal choice χ of $\mathbf{x} \in \mathcal{H}$.

Definition 4.14. Let $d, \ell \in \mathbb{N}$, $N = \dim(\mathcal{P}_{d,\ell})$. Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $s, \delta \in \mathbb{R}_{>0}$.

(i) Let $\mu \in \mathbb{R}_{\ge 0}$. Define $\mathcal{H}_{\ell,s,\mu}(\mathcal{X}, x)$ as the set of all sets $\mathbf{x} \in \mathcal{X}^N$ such that $x \in \mathsf{ch}_\mu(\mathbf{x})$ and $\left\| \Psi(\eta_{\mathbf{x}}(\mathbf{x}))^{-1} \right\|_{\mathsf{op}} \le s$.

(ii) Fix a constant $D \in (1, \infty)$. Fix a constant $\mu \in (0, 1/N)$. Define $\mathcal{G}_{\ell,\delta,s}(\mathcal{X})$ as the set of all points $x \in \mathbb{R}^d$ with following property: There is a set of points $\mathbf{x} \in \mathcal{H}_{\ell,s,\mu}(\mathcal{X}, x)$ with $D^{-1}\delta \le \mathsf{diam}(\mathbf{x}) \le D\delta$.

(iii) Let $\chi_{\ell,s}(\mathcal{X},x)$ be an element $\mathbf{x}_0 \in \mathcal{H}_{\ell,2s,0}(\mathcal{X},x)$ such that

$$\left\|\Psi(\eta_{\mathbf{x}_0}(\mathbf{x}_0))^{-1}\right\|_{\mathsf{op}} \mathsf{diam}(\mathbf{x}_0)^{\ell+1} \le 2 \min_{\mathbf{x}\in\mathcal{H}_{\ell,2s,0}(\mathcal{X},x)} \left\|\Psi(\eta_{\mathbf{x}}(\mathbf{x}))^{-1}\right\|_{\mathsf{op}} \mathsf{diam}(\mathbf{x})^{\ell+1}.$$

Remark 4.15. The parameters μ, s, D are fixed and do not change with n. They ensure the stability of the multivariate polynomial interpolation and are used in some technical arguments of the main proof.

(i) For the polynomial interpolation, we have to solve a system of linear equations given by the matrix $\Psi(\mathbf{x})$. The requirements $\left\|\Psi(\eta_{\mathbf{x}}(\mathbf{x}))^{-1}\right\|_{\mathsf{op}} \le s$ and $\mathsf{diam}(\mathbf{x}) \le D\delta$ ensure stability and precision of the obtained approximation.

(ii) The lower bound on the diameter of the interpolation base points, $D^{-1}\delta \le \mathsf{diam}(\mathbf{x})$, ensures the stability of the normalization $\eta_{\mathbf{x}}$, where we divide by $\mathsf{diam}(\mathbf{x})$.

(iii) Multivariate polynomial approximation results are known for points $x_0 \in \mathbb{R}^d$ in the convex hull of the base points $\mathbf{x} \in (\mathbb{R}^d)^N$, see [21, Lemma 4.6]. To ensure that x_0 belongs also to the convex hull of sufficiently accurate estimates of the base points $\hat{\mathbf{x}}$, we employ the concept of the μ-interior of the convex hull: Specifically, $x_0 \in \mathsf{ch}_\mu(\mathbf{x})$ with $\mu > 0$ implies $x_0 \in \mathsf{ch}(\hat{\mathbf{x}})$ if the estimation error is small enough relative to $\mathsf{diam}(\mathbf{x})$.

4.2.3. Estimator

Let $L \in \mathbb{R}_{>0}$. Let $\mathcal{E}^{1\to d}$ and $\widetilde{\mathcal{E}}^{1\to d}$ be arbitrary regression estimators for the regression function and its derivative in the generic regression model $\mathfrak{P} := \mathfrak{P}^{1\to d}(\Sigma^{1\to d}(\beta+1,L),[0,T],(t_i)_{i\in[\![n]\!]},\mathfrak{G}_n)$, see Section 2.4. With these estimators, we estimate the observed ODE-solution and its derivative,

$$\hat{u}(t) := \mathcal{E}^{1\to d}\Big((t_i,Y_i)_{i\in[\![n]\!]},t\Big) \qquad \text{and} \qquad \hat{\dot{u}}(t) := \widetilde{\mathcal{E}}^{1\to d}\Big((t_i,Y_i)_{i\in[\![n]\!]},t\Big).$$

Denote the maximal sup-norm risk in probability of the regression estimators, see (11), as

$$\Gamma := \Gamma(n,\beta+1,L,T) := r_0^{\mathsf{pr},\sup}(\mathcal{E}^{1\to d},\mathfrak{P}) \qquad \text{and} \qquad \Lambda := \Lambda(n,\beta+1,L,T) \quad := r_1^{\mathsf{pr},\sup}(\widetilde{\mathcal{E}}^{1\to d},\mathfrak{P}).$$

Let $\ell := \beta - 1$, $s \in \mathbb{R}_{>0}$, and $x \in \mathbb{R}^d$. Assume $\chi_{\ell,s}(\hat{u}([0,T]),x)$ exists. Let $\hat{\mathbf{x}}(x) := (\hat{u}(\hat{\tau}_1),\dots,\hat{u}(\hat{\tau}_N))$, where $\hat{\tau}_i$ is chosen so that $\hat{\mathbf{x}}(x) = \chi_{\ell,s}(\hat{u}([0,T]),x)$. Let $\hat{\mathbf{y}}(x) := (\hat{\dot{u}}(\hat{\tau}_1),\dots,\hat{\dot{u}}(\hat{\tau}_N))$. Define the estimator

$$\hat{f}(x) := I_\ell(\hat{\mathbf{x}}(x),\hat{\mathbf{y}}(x),x). \tag{24}$$

Remark 4.16. The estimator is only defined for $x \in \mathcal{G}_{\ell,\infty,s}(\hat{u}([0,T]))$, which is a subset of the convex hull of $\hat{u}([0,T])$. If it is not defined, a fallback such as nearest neighbor (see Section 4.1) could be applied in practice.

4.2.4. Results

For $\mathcal{X} \subseteq \mathbb{R}^d$ and a function $u\colon [0,T] \to \mathbb{R}^d$, define

$$\delta_{\mathsf{NN}}^{\ell,s}(u,T,\mathcal{X}) := \inf\left\{\delta' \ge 0\colon \mathcal{X} \subseteq \mathcal{G}_{\ell,\delta',s}(u([0,T]))\right\}.$$

Theorem 4.17. *Use the generic general Snake model of Section 4.2.1 and the estimator of Section 4.2.3. Let $\delta \in \mathbb{R}_{\ge 0}$, potentially changing with n. Assume $T \prec n$, $\Gamma \prec 1$, $\Gamma \prec \delta$, $\Gamma \preccurlyeq \Lambda \preccurlyeq 1$. Set $L := \beta! \sup_{k\in[\![\beta]\!]} L_k{}^{\beta+1}$.*

(i) *Set* $\mathcal{X} := \mathcal{G}_{\ell,\delta,s}(u^\star([0,T]))$. *Then*

$$\sup_{x\in\mathcal{X}} \left\|\hat{f}(x) - f^\star(x)\right\|_2^2 = \mathbf{O_P}\left(\delta^{2\beta} + \Lambda(n,\beta+1,L,T)^2\right).$$

(ii) *Set* $\delta := \delta_{\mathsf{NN}}^{\ell,s}(u^\star,T,[0,1]^d)$. *Then*

$$\sup_{x\in[0,1]^d} \left\|\hat{f}(x) - f^\star(x)\right\|_2^2 = \mathbf{O_P}\left(\delta^{2\beta} + \Lambda(n,\beta+1,L,T)^2\right).$$

The proof of this result can be found in [21, section 6]. In the standard general Snake model, we want to use the componentwise local polynomial estimators of [21, section 3] as the estimators $\mathcal{E}^{1\to d}$ and $\widetilde{\mathcal{E}}^{1\to d}$ ([21, section 3]: $d = 1$, $\ell = \beta$, and $s = 0$ and $s = 1$, respectively). We assume KERNEL, EIGENVALUE, and COVER for the observation times t_i and SUBGAUSSIAN for the noise ε_i. Furthermore, we require [21, Eq. (9)]. Then, assuming $L \asymp 1$, [21, Corollary 3.7] yields

$$\Gamma(n,\beta+1,L,T) \preccurlyeq \left(\frac{T\log n}{n}\right)^{\frac{\beta+1}{2(\beta+1)+1}} \quad \text{and} \quad \Lambda(n,\beta+1,L,T) \preccurlyeq \left(\frac{T\log n}{n}\right)^{\frac{\beta}{2(\beta+1)+1}}.$$

Corollary 4.18. *Use the standard general Snake model of Section 4.2.1 and the estimator of Section 4.2.3 with the componentwise local polynomial estimator of degree* β *(see [21, section 3]) as* $\mathcal{E}^{1\to d}$ *and* $\widetilde{\mathcal{E}}^{1\to d}$. *Let* $\delta \in \mathbb{R}_{\geq 0}$, *potentially changing with* n, *and assume*

$$\left(\frac{T\log n}{n}\right)^{\frac{\beta+1}{2(\beta+1)+1}} \prec \delta.$$

Assume KERNEL, EIGENVALUE, COVER, *where the symbols* d *and* x_i *in the assumptions are* $d = 1$ *and* $x_{[\![1,n]\!]} = t_{[\![1,n]\!]}$. *Assume* SUBGAUSSIAN *and*

$$\left(\frac{\log(n)}{n}\right)^{\frac{1}{2(\beta+1)}} \preccurlyeq T \prec n.$$

(i) *Set* $\mathcal{X} := \mathcal{G}_{\ell,\delta,s}(u^\star([0,T]))$. *Then*

$$\sup_{x\in\mathcal{X}} \left\|\hat{f}(x) - f^\star(x)\right\|_2^2 = \mathbf{O_P}\left(\delta^{2\beta} + \left(\frac{T\log n}{n}\right)^{\frac{2\beta}{2(\beta+1)+1}}\right).$$

(ii) *Set* $\delta := \delta_{\mathsf{NN}}^{\ell,s}(u^\star,T,[0,1]^d)$. *Then*

$$\sup_{x\in[0,1]^d} \left\|\hat{f}(x) - f^\star(x)\right\|_2^2 = \mathbf{O_P}\left(\delta^{2\beta} + \left(\frac{T\log n}{n}\right)^{\frac{2\beta}{2(\beta+1)+1}}\right). \tag{25}$$

Remark 4.19.

(i) From [22, Corollary 4.11], we obtain

$$\mathbf{E}\left[\sup_{x\in[0,1]^d} \left\|\hat{f}(x) - f^\star(x)\right\|_2^2\right] \succcurlyeq \delta^{2\beta} + \left(\delta^{d-1}\frac{T\log(n)}{n}\right)^{\frac{2\beta}{2(\beta+1)+d}}.$$

if $\log(n) \asymp \log(\delta^{-(d-1)} n T^{-1})$ and

$$\max\left(\left(\frac{T}{n}\right)^{2\beta+d+1}, T^{-1}\right) \preccurlyeq \delta^{d-1} \preccurlyeq \min\left(1, \frac{n}{T}\right).$$

Thus, for $\mathcal{F}_{d,\beta} = \bar{\Sigma}^{d\to d}(\beta, L_{[\![0,\beta]\!]})$, the rate given in (25) is minimax optimal if $\delta^{2\beta}$ is the dominating term the error bound, i.e., if

$$\delta \succcurlyeq \left(\frac{T \log n}{n}\right)^{\frac{1}{2(\beta+1)+1}}.$$

This condition is fulfilled in the setting of Corollary 4.20 below. See Remark 4.5 for a discussion of the condition for $\beta = 1$.

(ii) In Section 4.1, we can extrapolate beyond the convex hull of $\hat{u}([0,T])$ in the case of $\beta = 1$ using nearest neighbors. In contrast, our general results are only available for interpolation, i.e., for $x \in \mathsf{ch}(\hat{u}([0,T]))$ — to be precise, only for the even more restrictive assumption $x \in \mathcal{G}_{\ell,\delta,s}(u^\star([0,T]))$. Extrapolation seems also possible using polynomials of degree > 0, but the technicalities seem more difficult.

(iii) In practice, finding $\chi_{\ell,s}$ may be computationally demanding. Furthermore, using only results of interpolation may be inconveniently restrictive. Thus, one may want to replace the polynomial interpolation step by a polynomial regression step on the $k \geq N$ nearest neighbors of $\hat{u}(\tau_j)$ for some chosen time points $\tau_1, \ldots, \tau_J \in [0,T]$. Another alternative to polynomials would be Gaussian process interpolation (or regression) [19], potentially also restricted to k nearest neighbors for performance reasons.

The lowest upper bound on the error is obtained if the two terms in (25) are balanced. Additionally, we require $T\delta^{d-1} \succcurlyeq 1$ to be able to cover the domain of interest with $\mathrm{B}^d(U(f^\star, x_1, [0,T]), \delta)$, which is necessary for $[0,1]^d \subseteq \mathcal{G}_{\ell,\delta,s}(u^\star([0,T]))$.

Corollary 4.20. *Use the setting and assumptions of Corollary 4.18. Set* $\delta := \delta^{\ell,s}_{\mathsf{NN}}(u^\star, T, [0,1]^d)$. *Assume* $T \preccurlyeq \left(\frac{n}{\log n}\right)^{\frac{d-1}{2(\beta+1)+d}}$ *and* $\delta \preccurlyeq \left(\frac{n}{\log n}\right)^{-\frac{1}{2(\beta+1)+d}}$. *Then*

$$\sup_{x\in[0,1]^d} \left\|\hat{f}(x) - f^\star(x)\right\|_2^2 = \mathbf{O_P}\left(\left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2(\beta+1)+d}}\right).$$

5. Discussion

In this section, we revisit our main results obtained for the Snake and Stubble models, presenting a direct comparison between them. For a broader discussion—including how these models relate to other well-known frameworks, possible extensions of the results introduced here, and open problems—see Supplement Section 1.

So far, we have derived upper bounds on the error for estimating the model function of an autonomous, first-order ODE in two distinct settings:

1. In the Stubble model, characterized by $m \asymp n$ short ($n_j \asymp 1$) trajectories with equidistant measurement times, we established a bound on the pointwise mean squared error (Corollary 3.13),

$$\mathbf{E}\left[\left\|\hat{f}(x_0) - f^\star(x_0)\right\|_2^2\right] \preccurlyeq \left(\Delta t^2 n\right)^{-\frac{2\beta}{2\beta+d}} + \Delta t^{2\beta},$$

which is minimax optimal. For optimal asymptotics of Δt (Corollary 3.15), we have

$$\mathbf{E}\left[\left\|\hat{f}(x_0) - f^\star(x_0)\right\|_2^2\right] \preccurlyeq n^{-\frac{2\beta}{2(\beta+1)+d}}.$$

2. In the Snake model, involving a single trajectory ($m = 1$) with $n_1 \asymp n$ measurements and a flexible time-step design under sub-Gaussian noise, we derived a bound on the sup-norm error in probability (Corollary 4.18),

$$\sup_{x\in[0,1]^d} \left\|\hat{f}(x) - f^\star(x)\right\|_2^2 = \mathbf{O_P}\left(\delta^{2\beta} + \left(\frac{T\log n}{n}\right)^{\frac{2\beta}{2(\beta+1)+1}}\right),$$

which is minimax optimal if $\delta \succcurlyeq \left(\frac{T\log n}{n}\right)^{\frac{1}{2(\beta+1)+1}}$. For optimal asymptotics of T and δ (Corollary 4.20), we have

$$\sup_{x\in[0,1]^d} \left\|\hat{f}(x) - f^\star(x)\right\|_2^2 = \mathbf{O_P}\left(\left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2(\beta+1)+d}}\right).$$

Thus, although the two models are complementary, they yield the same error rate (up to a log-factor) in the case of optimal asymptotics.

The error bounds in the Stubble model can be extended to the sup-norm loss, yielding a bound of the form

$$\mathbf{E}\left[\sup_{x\in[0,1]^d} \left\|\hat{f}(x) - f^\star(x)\right\|_2^2\right] \preccurlyeq \left(\Delta t^2 \frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+d}} + \Delta t^{2\beta}.$$

This extension is obtained by adapting the proof of Theorem 3.11, replacing pointwise error bounds with sup-norm error bounds and substituting the pointwise maximal risk r with its sup-norm counterpart. A suitable minimax optimal regression estimator for the increment maps $\iota^\star$ is provided in [21, Theorem 3.6]. It should be noted that achieving the optimal rate also necessitates assuming SubGaussian, akin to the Snake model.

In the Snake model, the sup-norm error bound also applies to the pointwise error. However, one might anticipate the possibility of eliminating the $\log n$ factor from the error when considering pointwise error, as observed in standard nonparametric regression [21, Section 3]. Yet, in the ODE context, this appears more challenging because the proof requires bounding the maximal error in the first regression step independently of the target error measure.

The estimation strategies, like the models themselves, are complementary. In the Stubble model, the procedure begins with a nonparametric regression to estimate the increment map $\mathbb{R}^d \to \mathbb{R}^d$, followed by polynomial interpolation $\mathbb{R} \to \mathbb{R}^d$, from which we obtain an estimate of the model function. In contrast, in the Snake model, we first estimate the ODE solution and its derivative via nonparametric regression $\mathbb{R} \to \mathbb{R}^d$, and then apply polynomial interpolation $\mathbb{R}^d \to \mathbb{R}^d$ to estimate f.

Each model comes with its own advantages and limitations, both in terms of practical applicability and theoretical assumptions.

The primary advantage of the Stubble model is that it yields a complete and coherent estimation procedure within the framework presented in this article. For each smoothness order, the estimator is practical and achieves minimax optimality under relatively mild conditions. We require that the initial conditions adequately cover the domain of interest. As discussed in Section 2.3, such a covering assumption is unavoidable in general. On the downside, the Stubble model may be seen as less practical in certain settings, as it relies on having access to many trajectories with precisely known initial conditions.

In contrast, in the Snake model, all state vectors may be observed noisily. However, the theoretical results for the Snake model are currently incomplete: they do not fully cover regimes characterized by dense spatial coverage (small δ). When the domain of interest is fixed, ensuring that a trajectory adequately covers the entire region can also be restrictive, though such coverage assumptions are generally unavoidable. Additionally, the Snake estimator becomes impractical in settings with higher-order smoothness, as the required multivariate polynomial interpolation introduces considerable computational complexity. In contrast, the estimator in the Lipschitz case is feasible.

In summary, the Stubble model offers a theoretically clean and computationally practical approach under controlled data collection conditions. The Snake model, by contrast, is—in some aspects—more flexible in terms of the observational setup, but its error bounds are not optimal in all regimes, and the proposed estimator for higher-order smoothness is impractical.

Supplementary Material

Supplement to "Nonparametric estimation of ordinary differential equations: Snake and stubble" (DOI: 10.3150/25-BEJ1936SUPP; .pdf). The supplement [21] provides extended discussion of the models and results, reviews notation for multivariate calculus and smoothness classes, and contains proofs omitted from the main text.

References

[1] Antoniadis, A. (2007). Wavelet methods in statistics: Some recent developments and their applications. *Stat. Surv.* **1** 16–55. MR2520413 https://doi.org/10.1214/07-SS014

[2] Arnold, V.I. (2006). *Ordinary Differential Equations*. *Universitext*. Berlin: Springer-Verlag. Translated from the Russian by Roger Cooke, Second printing of the 1992 edition. MR2242407

[3] Brunel, N.J.-B. (2008). Parameter estimation of ODE's via nonparametric estimators. *Electron. J. Stat.* **2** 1242–1267. MR2471285 https://doi.org/10.1214/07-EJS132

[4] Butcher, J.C. (2016). *Numerical Methods for Ordinary Differential Equations*, 3rd ed. Chichester: John Wiley & Sons, Ltd. With a foreword by J. M. Sanz-Serna. MR3559553 https://doi.org/10.1002/9781119121534

[5] Chen, R.T.Q., Rubanova, Y., Bettencourt, J. and Duvenaud, D. (2018). Neural ordinary differential equations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems. NIPS'18* 6572–6583. Red Hook, NY, USA: Curran Associates Inc.

[6] Dattner, I. (2021). Differential equations in data analysis. *Wiley Interdiscip. Rev.: Comput. Stat.* **13** Paper No. e1534, 20. MR4336079 https://doi.org/10.1002/wics.1534

[7] Dattner, I. and Klaassen, C.A.J. (2015). Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. *Electron. J. Stat.* **9** 1939–1973. MR3391125 https://doi.org/10.1214/15-EJS1053

[8] Donovan, T. (2010). *Replay: The History of Video Games*. Yellow Ant.

[9] Gottwald, G.A. and Reich, S. (2021). Supervised learning from noisy observations: Combining machine-learning techniques with data assimilation. *Phys. D* **423** Paper No. 132911, 15. MR4249157 https://doi.org/10.1016/j.physd.2021.132911
[10] Gugushvili, S. and Klaassen, C.A.J. (2012). $\sqrt{n}$-Consistent parameter estimation for systems of ordinary differential equations: Bypassing numerical integration via smoothing. *Bernoulli* **18** 1061–1098. MR2948913 https://doi.org/10.3150/11-BEJ362
[11] Hartman, P. (2002). *Ordinary Differential Equations*. *Classics in Applied Mathematics* **38**. Philadelphia: Society for Industrial and Applied Mathematics (SIAM). Corrected reprint of the second (1982) edition [Birkhäuser, Boston, MA; MR0658490 (83e:34002)], With a foreword by Peter Bates. MR1929104 https://doi.org/10.1137/1.9780898719222
[12] Heinonen, M., Yildiz, C., Mannerström, H., Intosalmi, J. and Lähdesmäki, H. (2018). Learning unknown ODE models with Gaussian processes. In *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 1959–1968. PMLR.
[13] Lahouel, K., Wells, M., Rielly, V., Lew, E., Lovitz, D. and Jedynak, B.M. (2024). Learning nonparametric ordinary differential equations from noisy data. *J. Comput. Phys.* **507** Paper No. 112971, 22. MR4729758 https://doi.org/10.1016/j.jcp.2024.112971
[14] Marzouk, Y., Ren, Z., Wang, S. and Zech, J. (2024). Distribution learning via neural differential equations: A nonparametric statistical perspective. *J. Mach. Learn. Res.* **25** Paper No. [232], 61. https://doi.org/4796493
[15] McGoff, K., Mukherjee, S. and Pillai, N. (2015). Statistical inference for dynamical systems: A review. *Stat. Surv.* **9** 209–252. MR3422438 https://doi.org/10.1214/15-SS111
[16] Olver, P.J. (2006). On multivariate interpolation. *Stud. Appl. Math.* **116** 201–240. MR2200737 https://doi.org/10.1111/j.1467-9590.2006.00335.x
[17] Qi, X. and Zhao, H. (2010). Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *Ann. Statist.* **38** 435–481. MR2589327 https://doi.org/10.1214/09-AOS724
[18] Ramsay, J. and Hooker, G. (2017). *Dynamic Data Analysis: Modeling Data with Differential Equations*. *Springer Series in Statistics*. New York: Springer. MR3645102 https://doi.org/10.1007/978-1-4939-7190-9
[19] Rasmussen, C.E. and Williams, C.K.I. (2005). *Gaussian Processes for Machine Learning*. The MIT Press. https://doi.org/10.7551/mitpress/3206.001.0001
[20] Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist.* **48** 1875–1897. MR4134774 https://doi.org/10.1214/19-AOS1875
[21] Schötz, C. (2026). Supplement to "Nonparametric estimation of ordinary differential equations: Snake and stubble." https://doi.org/10.3150/25-BEJ1936SUPP
[22] Schötz, C. and Siebel, M. (2024). Lower bounds for nonparametric estimation of ordinary differential equations. https://doi.org/10.48550/arXiv.2407.14993
[23] Strogatz, S. (2024). *Nonlinear Dynamics and Chaos*, 3rd ed. New York: Chapman and Hall/CRC. https://doi.org/10.1201/9780429398490
[24] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. New York: Springer. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR2724359 https://doi.org/10.1007/b13794
[25] Xiao, L. (2019). Asymptotic theory of penalized splines. *Electron. J. Stat.* **13** 747–794. MR3925516 https://doi.org/10.1214/19-ejs1541

Received January 2025 and revised September 2025