Classes of ODE solutions: smoothness, covering numbers, implications for noisy function fitting, and the curse of smoothness phenomenon*

Ying Zhu† and Mozhgan Mirzaei‡

First draft on ArXiv: November 2020. This draft: March 2021.

Abstract

Numerical methods for recovering ODE solutions from data largely rely on approximating the solutions using basis functions or kernel functions under a least square criterion. The accuracy of this approach hinges on the smoothness of the solutions. This paper provides a theoretical foundation for these methods by establishing novel results on the smoothness and covering numbers of ODE solution classes (as a measure of their "size"). Our results provide answers to "how do the degree of smoothness and the "size" of a class of ODEs affect the "size" of the associated class of solutions?" In particular, we show that: (1) for the first order ODEs, if the absolute values of all kth order derivatives are bounded by 1, then the solution can end up with derivatives whose magnitude grows factorially fast – "a curse of smoothness"; (2) our upper bounds for the covering numbers of the $(\beta+2)-$degree smooth solution classes are greater than those of the "standard" $(\beta+2)-$degree smooth class of univariate functions; (3) the mean squared error of least squares fitting in noisy settings has a convergence rate no larger than $\left(\frac{1}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$ if $n=\Omega\left(\left(\beta\sqrt{\log\left(\beta\vee 1\right)}\right)^{4\beta+10}\right)$, and under this condition, the rate $\left(\frac{1}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$ is minimax optimal in the case of $y^{'}(x)=f(x,\,y(x))$; (4) more generally, for the higher order Picard type ODEs, $y^{(m)}(x)=f\left(x,\,y(x),\,y^{'}(x),\,...,y^{(m-1)}(x)\right)$, the covering number of the underlying solution class is bounded from above by the product of the covering number of the class $\mathcal{F}$ that f ranges over and the covering number of the set where initial values lie.

1 Introduction

Ordinary differential equations (ODE) enjoy a long standing history in mathematics and have numerous applications in science and engineering. ODE also play a crucial role in social science and business, for example, the famous Solow growth model in economics, and the famous Bass product diffusion model in marketing. Since the COVID-19 pandemic, lots of attention has been given to the compartmental models in epidemiology (see, e.g., [24]). The compartmental models have also been used in forecasting election outcomes ([23]) as well as modeling the information diffusion (see, e.g., [3]).

As analytical solutions for ODEs are often unavailable, developing numerical methods for solving initial value or boundary value problems has been an active research area in ODEs. An important

*AMS subject classifications: Primary 68Q32; secondary: 34A45, 34F05, 62G08, 65L70. Keywords: Ordinary differential equations; covering numbers; local complexity; smooth functions; kernel methods

†Assistant Professor, Department of Economics, UC San Diego. Corresponding author. yiz012@ucsd.edu.

‡PhD, former student at Department of Mathematics, UC San Diego. momirzae@ucsd.edu

technique (belonging in the family of collocation methods) tries to approximate the solutions using basis functions or kernel functions (see [5, 12, 14]). For example, Mehrkanoon, et al. [14] consider approximate solutions in the form $\hat{y}(x) = w^T \varphi(x) + b$ where $\varphi : \mathbb{R} \mapsto \mathbb{R}^p$ is the feature map and $w \in \mathbb{R}^p$ are the unknown weights (solved from a least squares support vector machines formulation at collocation points).

Another active area in ODEs involves the recovery of solutions from their noisy measurements:

$$Y_i = y(x_i) + \varepsilon_i, \quad i = 1, ..., n. \tag{1}$$

In the above, y is a solution to an ODE, $\{Y_i\}_{i=1}^n$ are noisy measurements of $\{y(x_i)\}_{i=1}^n$ evaluated at a collection of fixed design points $\{x_i\}_{i=1}^n$ (both Y_i and x_i are observed), and $\{\varepsilon_i\}_{i=1}^n$ are unobserved noise terms. In the literature, function fitting with noisy measurements of ODE solutions is typically performed via the least squares procedure:

$$\hat{y} \in \arg\min_{\tilde{y} \in \Pi} \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - \tilde{y}(x_i)\right)^2 \tag{2}$$

where Π is a suitably chosen function class that contains y. For example, Volkening et al. [23] minimize the sum of squared deviations between the averaged polling data and the fitted solutions to a compartmental model. In this application, analytical solutions to the ODEs are available and a nonlinear least squares procedure can be applied to recover the solutions from noisy measurements (so Π would be the class of functions with the same form as the ODE solution, indexed by different parameters). In many problems, analytical solutions for ODEs are unavailable. While integration based nonlinear least squares overcome this issue, they often suffer many computational issues (see discussions in [22]). Therefore, researchers often resort to approximate the underlying solutions using polynomials and spline bases (see [11, 15, 16, 17, 18, 22]). For example, in clinical studies of AIDs, plasma viral load and CD4+ T cell counts are measured with additive noise. With these measurements, Liang and Wu [11] apply local polynomial regressions with the least squares criterion to recover the ODE solutions and their first derivatives (in their case, Π corresponds to the class of three times differentiable functions with bounded derivatives).

In either of the two areas described above, researchers rely on approximating the underlying solutions using basis functions or kernel functions. This strategy assumes some ambient space of smooth functions that contains the solutions and seeks a close enough estimator from the restricted class of functions. The choice of the function class and the accuracy of the approximations hinge on the smoothness of y. For example, in terms of the local polynomial estimators used in [11] for y and y', in order for such estimators to have well behaved biases, boundedness on the second and third derivatives of y is needed, respectively. As suggested by [11], higher order local polynomials can be applied and if so, boundedness on higher order derivatives of y would be needed. Because the smoothness of y ultimately depends on the smoothness of f, this prompts us to ask the following fundamental question:

- ***(Question 1)*** *in general, how does the degree of smoothness of f affect the smoothness of y in $y'(x) = f(y(x))$ and $y'(x) = f(x, y(x))$; for example, if f in $y'(x) = f(y(x))$ is $\beta + 1$ times differentiable with all derivatives bounded by 1 (where β is a non-negative integer), can one simply reason that all derivatives of y also stay "nicely" bounded?*

To our knowledge, it appears that no rigorous answers have been provided to the question above. Our results regarding ***Question 1*** can be summarized as follows:

- ***(Lemma 3.1)*** *For $y'(x) = f(y(x))$, for all $k = 0, ..., \beta+1$, if $\left|f^{(k)}(x)\right| \leq 1$, then $\left|y^{(k+1)}(x)\right| \leq k!$. These factorial bounds can be attained (and hence tight).*

- ***(Lemma 3.2)*** *For* $y^{'}(x) = f(x, y(x))$*, if the absolute values of all* k*th order partial derivatives are bounded by* 1*, then* $\left|y^{(k+1)}(x)\right| \leq 2^k k!$ *for* $k = 0, ..., \beta + 1$.[1]

Noisy recovery is generally more difficult than noiseless recovery. In terms of (2), the worst case bound for the average squared error of $\hat{y}$ depends on the maximum "correlation" between the noise vector $\{\varepsilon_i\}_{i=1}^n$ and the random error vector $\{\hat{y}(x_i) - y(x_i)\}_{i=1}^n$. To see this, note that since $y \in \Pi$ (y is feasible) and $\hat{y}$ is optimal, we have

$$\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \hat{y}(x_i)\right)^2 \leq \frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - y(x_i)\right)^2,$$

which yields

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}(x_i) - y(x_i)\right)^2 \leq \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\left(\hat{y}(x_i) - y(x_i)\right). \tag{3}$$

Controlling this right-hand-side (RHS) term in (3) can be reduced to controlling the "local complexity" (one of the most important notions in machine learning theory) associated with Π, which depends ultimately on the "size" of Π. In noisy recovery problems, researchers typically have prior knowledge on the smoothness structure of f and use such information to guide the choice of Π in (2); therefore, a natural choice for Π is

$$\mathcal{Y} = \left\{y : y^{(m)}(x) = f\left(x, y(x), y^{'}(x), ..., y^{(m-1)}(x)\right), f \in \mathcal{F}\right\} \tag{4}$$

with $y = y^{(0)}$. In many contexts, $\mathcal{Y}$ is a set consisting of infinitely many smooth functions. The notion "covering number", dated back to the seminal work of Kolmogorov, Tikhomirov, and others (see, e.g., [7]), provides a way to measure the size of a class with an infinite number of elements. In this paper, we derive *nonasymptotic* bounds on the covering numbers of $\mathcal{Y}$ to address the following novel question:

- ***(Question 2)*** *how does the "size" of a class of* f *affect the "size" of the associated solution class of* y*?*

Like ***Question 1***, we are not aware of any rigorous results regarding ***Question 2*** from the existing literature. Below provides an overview of our results regarding ***Question 2***.

- ***(Theorem 2.1)*** *We establish a general upper bound on the covering number of solution classes associated with (4). This result implies, the covering number of the underlying solution class* $\mathcal{Y}$ *is bounded from above by the product of the covering number of the class* $\mathcal{F}$ *that* f *ranges over and the covering number of the set where initial values lie, as long as* f *satisfies a Lipschitz condition with respect to the* $y, ..., y^{(m-1)}$ *coordinates in* l_2-*norm (i.e., the Picard Lipschitz condition). This general bound yields a sharp scaling in some problems (**Corollary 2.1**).*

- ***(Theorems 3.1-3.2)*** *For the first order ODEs, if the absolute values of all* k*th order (partial) derivatives of* f *are bounded by* 1*, then the general bound in **Theorem 2.1** may be improved by exploiting the factorial bounds in **Lemmas 3.1 and 3.2**. Our upper bounds for the covering numbers of the* $(\beta + 2)-$*degree smooth solution classes,* $\mathcal{Y}$*, are greater than those of the "standard"* $(\beta + 2)-$*degree smooth class,* $\mathcal{S}_{\beta+2}$*, of univariate functions (where all derivatives are bounded by* 1*). We also discuss a lower bound on the covering numbers of solution classes associated with* $y^{'}(x) = f(x, y(x))$ *in **Lemma 3.3**.*

[1] The bound "1" is assumed to avoid notation cluttering and can be relaxed easily.

The abovementioned bounds have important implications for analyzing techniques for solving ODEs as well as noisy recovery. In this paper, we focus on the noisy recovery problem and examine the implications on the convergence rate of the mean squared error $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}(x_i)-y(x_i)\right)^2\right]$. Let us first take a digression from discussing these implications. For noisy recovery of a function in the standard $(\gamma+1)-$degree smooth class of $m-$variate functions, the common wisdom is that large m brings a curse (resulting in a large convergence rate of the least squares) and large γ brings a bless (resulting in a small convergence rate). While there is no ambiguity from the "curse of dimension" here, this paper discovers that the "bless of smoothness" comes at the price of assuming a condition on the sample size n. In particular, for noisy recovery of a function in the standard $(\beta+2)-$degree smooth class, $\mathcal{S}_{\beta+2}$, the well-known minimax optimal rate $\left(\frac{1}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$ (which decreases in β) coincides with the convergence rate of the least squares when $n=\Omega\left((\beta\vee 1)^{2\beta+5}\right)$. We view this requirement on the sample size, $n=\Omega\left((\beta\vee 1)^{2\beta+5}\right)$, a "curse of smoothness" for least squares estimations of elements in $\mathcal{S}_{\beta+2}$.[2] A more detailed discussion is given in Section 3.1.

Coming back to the ODE problems, with the assistance of ***Theorems 3.1-3.2*** and ***Lemma 3.3***, we ask the following novel questions regarding $y^{'}(x)=f(y(x))$ and $y^{'}(x)=f(x,\,y(x))$:

- ***(Question 3)*** *If $f\in\mathcal{S}_{\beta+1}$, what is the condition on n for the least squares $\hat{y}$ to have $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}(x_i)-y(x_i)\right)^2\right]$ bounded above by $\left(\frac{1}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$, and what can we say about the optimality of $\left(\frac{1}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$?*

Here are the answers to ***Question 3***.

- ***(Theorem 3.3)*** *The mean squared error of $\hat{y}$ has a convergence rate no larger than $\left(\frac{1}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$ if $n=\Omega\left(\left(\beta\sqrt{\log(\beta\vee 1)}\right)^{4\beta+10}\right)$, and under this condition, the rate $\left(\frac{1}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$ is minimax optimal in the case of $y^{'}(x)=f(x,\,y(x))$. Relative to the standard smooth class $\mathcal{S}_{\beta+2}$ (where the least squares procedure needs $n=\Omega\left((\beta\vee 1)^{2\beta+5}\right)$ for the optimal rate $\left(\frac{1}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$), the requirement on the sample size in the ODE context is much greater.* ***Theorem 3.3*** *also presents upper bounds (with rates no smaller than $\left(\frac{1}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$) on the mean squared error of $\hat{y}$ under more general situations where the condition $n=\Omega\left(\left(\beta\sqrt{\log(\beta\vee 1)}\right)^{4\beta+10}\right)$ is not imposed.*

Having provided the theoretical implications of (2) in ***Theorem 3.3***, we further ask whether there is a practical implementation of (2) with comparable theoretical guarantees. We exploit the smoothness structures of solutions in ***Lemmas 3.1–3.2*** directly by considering some ambient reproducing kernel Hilbert spaces that contain $\mathcal{Y}$, and doing so allows us to develop a variant of the classical kernel ridge regression procedure. For all practical purposes, the performance guarantee for this procedure (provided in ***Theorem 3.4***) turns out quite comparable to the general bounds in ***Theorem 3.3***. The practical implementation associated with ***Theorem 3.4*** provides an example of how our theoretical results can be used to guide the use of kernel functions in noisy function fitting.

[2] We also discover that the "curse of smoothness" is exacerbated as the dimension m increases. This observation appears unmentioned in [7] and can have important finite sample implications.

Besides recovering the ODE solutions, one could also recover the first derivatives of ODE solutions with a least squares approach. This paper derives bounds for the covering number of the class consisting of the first derivative $y^{'}$ of $y \in \mathcal{Y}$ as well. Therefore, our arguments for Theorems 3.3 and 3.4 can be easily extended for analyzing the recovery of the first derivatives of ODE solutions.

2 A general upper bound on the covering number

General notation. The l_q-norm of a $K-$dimensional vector θ is denoted by $|\theta|_q$, $1 \leq q \leq \infty$ where $|\theta|_q := \left(\sum_{j=1}^{K} |\theta_j|^q\right)^{1/q}$ when $1 \leq q < \infty$ and $|\theta|_q := \max_{j=1,...,K} |\theta_j|$ when $q = \infty$. Define $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ that places a weight $\frac{1}{n}$ on each observation x_i for $i = 1, ..., n$, and the associated $\mathcal{L}^2(\mathbb{P}_n)-$norm of the vector $v := \{v(x_i)\}_{i=1}^{n}$, denoted by $|v|_n$, is given by $\left[\frac{1}{n}\sum_{i=1}^{n} (v(x_i))^2\right]^{\frac{1}{2}}$. For two functions f and g on $[a, b]^d \subseteq \mathbb{R}^d$, we denote the supremum metric by $|f - g|_\infty := \sup_{x \in [a, b]^d} |f(x) - g(x)|$. For functions $f(n)$ and $g(n)$, $f(n) \gtrsim g(n)$ means that $f(n) \geq cg(n)$ for a universal constant $c \in (0, \infty)$; similarly, $f(n) \precsim g(n)$ means that $f(n) \leq c^{'} g(n)$ for a universal constant $c^{'} \in (0, \infty)$; and $f(n) \asymp g(n)$ means that $f(n) \gtrsim g(n)$ and $f(n) \precsim g(n)$ hold simultaneously. As a general rule for this paper, the various c and C constants (all $\precsim 1$) denote positive universal constants that are independent of the sample size n and the smoothness parameter β, and may vary from place to place.

Definition (covering numbers). Given a set $\mathbb{T}$, a set $\left\{\mathrm{t}^1, \mathrm{t}^2, ..., \mathrm{t}^N\right\} \subset \mathbb{T}$ is called a $\delta-$cover of $\mathbb{T}$ with respect to a metric ρ if for each $\mathrm{t} \in \mathbb{T}$, there exists some $i \in \{1, ..., N\}$ such that $\rho(\mathrm{t}, \mathrm{t}^i) \leq \delta$. The cardinality of the smallest $\delta-$cover is denoted by $N_\rho(\delta; \mathbb{T})$, namely, the $\delta-$covering number of $\mathbb{T}$. For example, $N_\infty(\delta, \mathcal{F})$ denotes the $\delta-$covering number of a function class $\mathcal{F}$ with respect to the supremum metric $|\cdot|_\infty$.

Let us define

$$Y(x) := \begin{bmatrix} y(x) \\ y^{'}(x) \\ \vdots \\ y^{(m-1)}(x) \end{bmatrix}^T \quad \text{and} \quad Y_0 := \begin{bmatrix} y_{(0)} \\ y_{(1)} \\ \vdots \\ y_{(m-1)} \end{bmatrix}^T.$$

We consider the ODE

$$y^{(m)}(x) = f(x, Y(x)) \tag{5}$$

with the initial values

$$y_{(0)} = y(a_0), \, y_{(1)} = y^{'}(a_0), \, ..., \, y_{(m-1)} = y^{(m-1)}(a_0)$$

such that $|Y_0|_2 \leq C_0$ and

$$(x, Y(x)) \in \{(x, Y) : x \in [a_0, a_0 + a], |Y|_2 \leq b + C_0\} := \bar{\Gamma}$$

(where $a, b > 0$). Assume $f(x, Y)$ is continuous on $\bar{\Gamma}$ and satisfies the Picard Lipschitz condition

$$\left|f(x, Y) - f\left(x, \tilde{Y}\right)\right| \leq L\left|Y - \tilde{Y}\right|_2 \tag{6}$$

for all $(x,\, Y) := \left(x,\, y,\, ...,\, y^{(m-1)}\right)$ and $\left(x,\, \tilde{Y}\right) := \left(x,\, \tilde{y},\, ...,\, \tilde{y}^{(m-1)}\right)$ in $\bar{\Gamma}$. Then by the Picard local existence theorem (see, e.g., [4]), there is a solution to (5) in $\mathcal{C}\left[a_0,\, a_0+\alpha\right]$ where $\alpha \leq \min\left\{a,\, \frac{b}{M}\right\}$ and $M = \max_{(x,\, Y)\in\bar{\Gamma}} \left|f\left(x,\, Y\right)\right|$.

The Picard Lipschitz condition (6) is not necessary for the existence of a solution as suggested by the Cauchy-Peano existence theorem (see, e.g., [4]). However, (6) allows us to establish a general upper bound on the covering number of higher order ODEs. In what follows, we consider the ODE (5) with $\left|Y_0\right|_2 \leq C_0$ and $\left(x,\, Y\left(x\right)\right) \in \bar{\Gamma}$.

Theorem 2.1. *In terms of (5), suppose f is continuous on $\bar{\Gamma}$, and satisfies (6) for all $(x,\, Y)$, $\left(x,\, \tilde{Y}\right) \in \bar{\Gamma}$. If f ranges over a function class $\mathcal{F}$ with $N_{\infty}\left(\delta,\, \mathcal{F}\right)$ on $\bar{\Gamma}$, and $\max_{(x,\, Y)\in\bar{\Gamma}} \left|f\left(x,\, Y\right)\right| \leq M$ for all $f \in \mathcal{F}$, then we have*

$$\log N_{\infty}\left(\delta,\, \mathcal{Y}_k\right) \leq \log N_{\infty}\left(\frac{\delta}{L_{\max}},\, \mathcal{F}\right) + m\log\left(\frac{2C_0 L_{\max}}{\delta}+1\right) \tag{7}$$

where $L_{\max} = \sup_{x\in[a_0,\, a_0+\alpha]}\left\{\exp\left(x\sqrt{L^2+1}\right)\left[1+\int_0^x \exp\left(-s\sqrt{L^2+1}\right)ds\right]\right\}$ with $\alpha = \min\left\{a,\, \frac{b}{M}\right\}$, and $\mathcal{Y}_0 = \mathcal{Y}$ ($k=0$) is the class consisting of solutions (to (5) with f ranging over $\mathcal{F}$) on $\left[a_0,\, a_0+\alpha\right]$ and $\mathcal{Y}_k$ (the non-negative integer $k \leq m-1$) is the class consisting of the kth derivative $y^{(k)}$ of $y \in \mathcal{Y}_0$.

The proof for Theorem 2.1 is provided in Section A.1 of the supplementary materials. In the case where $m=1$, we can use the sharper constant $L_{\max} = \sup_{x\in[0,\, \alpha]}\left\{\exp\left(Lx\right)\left[1+\int_0^x \exp\left(-Ls\right)ds\right]\right\}$.

If $m\log\left(\frac{2C_0 L_{\max}}{\delta}+1\right) \precsim \log N_{\infty}\left(\frac{\delta}{L_{\max}},\, \mathcal{F}\right)$, Theorem 2.1 implies that the solution class $\mathcal{Y}$ is "essentially" no larger than the class f ranges over. Below we discuss the implications of Theorem 2.1 for ODEs parameterized by a finite $K-$dimensional vector of coefficients.

2.1 Implications on parametric ODEs

Corollary 2.1. *Consider the ODE*

$$y^{(m)}\left(x\right) = f\left(x,\, Y\left(x\right);\, \theta\right) \text{ with } \left(x,\, Y\left(x\right)\right) \in \bar{\Gamma} \text{ and initial values } Y_0, \tag{8}$$

where f is parameterized by a finite $K-$dimensional vector of coefficients

$$\theta \in \mathbb{B}_q\left(1\right) := \left\{\theta \in \mathbb{R}^K :\, \left|\theta\right|_q \leq 1\right\}$$

with $q \in [1,\, \infty]$. Suppose f is continuous on $\bar{\Gamma}$, $\max_{(x,\, Y)\in\bar{\Gamma}} \left|f\left(x,\, Y;\, \theta\right)\right| \leq M$, and

$$\left|f\left(x,\, Y;\, \theta\right) - f\left(x,\, \tilde{Y};\, \theta\right)\right| \leq L\left|Y-\tilde{Y}\right|_2 \tag{9}$$

for all $(x,\, Y)$, $\left(x,\, \tilde{Y}\right) \in \bar{\Gamma}$ and $\theta \in \mathbb{B}_q\left(1\right)$; moreover,

$$\left|f\left(x,\, Y;\, \theta\right) - f\left(x,\, Y;\, \theta^{'}\right)\right| \leq L_K\left|\theta-\theta^{'}\right|_q, \tag{10}$$

for all $(x,\, Y) \in \bar{\Gamma}$ and $\theta,\, \theta^{'} \in \mathbb{B}_q\left(1\right)$. Let $\mathcal{Y}$ be the class consisting of solutions to (8) with $\theta \in \mathbb{B}_q\left(1\right)$ on $\left[a_0,\, a_0+\min\left\{a,\, \frac{b}{M}\right\}\right]$. Then we have

$$\log N_{\infty}\left(\delta,\, \mathcal{Y}\right) \leq K\log\left(1+\frac{2L_{\max}L_K}{\delta}\right) + m\log\left(\frac{2C_0 L_{\max}}{\delta}+1\right). \tag{11}$$

The proof for Corollary 2.1 is provided in Section A.2 of the supplementary materials.

In (11), the part "$K\log\left(1+\frac{2L_{\max}L_K}{\delta}\right)$" is related to the "size" of the ball that θ ranges over, and the part "$m\log\left(\frac{C_0L_{\max}}{\delta}+1\right)$" is related to the "size" of the ball that the initial value Y_0 ranges over. The part "$m\log\left(\frac{2C_0L_{\max}}{\delta}+1\right)$" reveals an interesting feature of differential equations: even if the equation is fixed and known (for example, $y^{'}(x)=y(x)$ with solutions $y(x)=c^{*}e^{x}$), the solution class is still not a singleton (and has infinitely many solutions) unless a fixed initial value is given. As a consequence, in the simple example $y^{'}=y$, the noisy recovery of a solution to this ODE still requires the estimation of c^{*}.[3]

Bound (11) suggests that if the class f ranges over is a "parametric" class, then the associated solution class $\mathcal{Y}$ also behaves like a "parametric" class. Suppose the class of ODEs (8) has a fixed initial value. Then the term "$m\log\left(\frac{C_0L_{\max}}{\delta}+1\right)$" can be dropped while the scaling of "$K\log\left(1+\frac{2L_{\max}L_K}{\delta}\right)$" can be attained as the following example suggests. Consider the simple ODE $y^{'}(x)=-\theta y(x)$ with $\theta\in[0,\,1]$, $x\in[0,\,1]$ and initial value $y(0)=1$, which has solutions in the form $y(x)=e^{-\theta x}$. It can be easily verified that $\log N_{\infty}(\delta,\,\mathcal{Y})\asymp\log\left(\frac{c}{\delta}+c^{'}\right)$ for some positive universal constants c and $c^{'}$.

While (7) provides a sharp scaling in the example above, this general bound may be improved in other contexts as we will see in Section 3.2.

2.2 Noisy fitting with plain vanilla methods, issues, and remedies

Let us consider (1) with y a solution to the following ODE

$$y^{'}(x;\,\theta^{*})=f(x,\,y(x;\,\theta^{*});\,\theta^{*}),\qquad y(0;\,\theta^{*})=y_0 \tag{13}$$

where $x\in[0,\,1]$, $|y_0|\leq C_0$, and f is parameterized by a finite $K-$dimensional vector of coefficients $\theta\in\mathbb{B}_q(1)$, $q\in[1,\,\infty]$. Suppose we have noisy measurements (Y) of y sampled at n points $\{x_i\}_{i=1}^{n}$. If one can solve for $y(x;\,\theta^{*},\,y_0^{*})$ in an explicit form from (13), then the estimator $\hat{y}_{NLS}=y\left(x;\,\hat{\theta},\,\hat{y}_0\right)$ can be formed with the plain vanilla nonlinear least squares (NLS)

$$\left(\hat{\theta},\,\hat{y}_0\right)\in\arg\min_{\tilde{\theta}\in\mathbb{B}_q(1),\,|\tilde{y}_0|\leq C_0}\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i-y\left(x_i;\,\tilde{\theta},\,\tilde{y}_0\right)\right)^2. \tag{14}$$

In many ODEs, the solutions are in implicit forms so it is not possible to write down (14). In the situation where a quality estimator $\hat{y}_0$ of y_0 is available, it is natural to exploit the *Picard iteration* as below:

$$y_{r+1}(x;\,\theta)=y_0+\int_0^x f(s,\,y_r(s;\,\theta);\,\theta)\,ds,\quad \text{integer } r\geq 0,\,y(0;\,\theta)=y_0. \tag{15}$$

[3] In terms of the homogeneous higher order linear ODEs

$$a_0(x)\,y+a_1(x)\,y^{(1)}+a_2(x)\,y^{(2)}+\cdots+a_m(x)\,y^{(m)}=0, \tag{12}$$

it is well understood that the solutions of (12) (where $a_j(\cdot)$s are fixed continuous functions) form a vector space of dimension m. A classical result on the VC dimension of a vector space of functions ([6, 20]) implies that the solution class associated with (12) has VC dimension at most m, when $a_j(\cdot)$s $(j=0,...,m)$ are fixed continuous functions.

An estimator based on (15) performs the following steps: first, we compute

$$\begin{aligned} \hat{y}_1(x;\,\theta) &= \hat{y}_0 + \int_0^x f(s,\,\hat{y}_0;\theta)\,ds, \\ \hat{y}_2(x;\,\theta) &= \hat{y}_0 + \int_0^x f(s,\,\hat{y}_1(s;\,\theta)\,;\,\theta)\,ds, \\ &\vdots \\ \hat{y}_{R+1}(x;\,\theta) &= \hat{y}_0 + \int_0^x f(s,\,\hat{y}_R(s;\,\theta)\,;\,\theta)\,ds; \end{aligned} \tag{16}$$

second, we solve the following program

$$\hat{\theta} \in \arg\min_{\theta\in\mathbb{B}_q(1)} \frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \hat{y}_{R+1}(x_i;\theta)\right)^2; \tag{17}$$

third, we form the estimator $\hat{y}_{Picard} = \hat{y}_{R+1}\left(x;\,\hat{\theta},\,\hat{y}_0\right)$.

For completeness, **Proposition C.1** in Section C.1 and **Proposition C.2** in Section C.2 of the supplementary materials establish upper bounds on the average squared errors of $\hat{y}_{NLS}$ and $\hat{y}_{Picard}$ with high probability guarantees, respectively. We choose not to highlight these propositions here as the abovementioned procedures are not widely used in practice for reasons discussed in the following.

Issues with the plain vanilla methods and remedies. The methods discussed above have serious practical issues. Clearly it is not possible to apply (14) if $y(x;\,\theta^*,\,y_0)$ cannot be solved in an explicit form from (13), while the implementation of (17) requires a quality estimator $\hat{y}_0$ of y_0. Even if these methods can be applied, they suffer numerous computational issues. For both (14) and (17), the objective functions can be highly nonconvex. The integrals can be hard to compute analytically in (16); one could use numerical integrations to approximate the integrals in (16) at a given θ, but doing so would mean that we have to perform $R+1$ numerical integrations for every single candidate θ. Such a method is not efficient computationally.

Because of these issues, researchers have resorted to function fitting techniques such as splines and alternative smoothing methods to obtain estimates for y and $y^{'}$. As we have discussed in Section 1, this strategy typically assumes some smooth class that contains the solutions and seeks a close enough estimator from the restricted class of smooth functions. For studying the theoretical behavior of algorithms used for recovering ODE solutions, ***Question 1*** raised in Section 1 is fundamental; moreover, answers to ***Question 1*** can be useful for studying ***Question 2*** (for which Theorem 2.1 has provided a partial answer). We will show how to exploit our answers to ***Question 1*** to improve the general bound (7) when $\mathcal{F}$ corresponds to a smooth function class. The next section examines these questions in depth.

3 First order ODEs with smooth f

Given that f is often a smooth function in many ODE applications and prior knowledge on the smoothness structure of f is available, researchers typically use such information to guide the choice of Π in (2). This fact motivates us to study in Section 3.2 the properties of the class of solutions y when f ranges over the class of functions with various degrees of smoothness. Our results reveal several sharp contrasts between the standard $(\beta+2)-$degree smooth class of univariate functions and the $(\beta+2)-$degree smooth solution classes (of also univariate functions). Before Section 3.2, we provide a summary of results for the standard smooth class of univariate functions and their implications in Section 3.1.

3.1 A summary of results on standard smooth functions

Let $p = (p_1, ..., p_d)$ and $[p] = \sum_{k=1}^{d} p_k$ where p_ks are non-negative integers. We write

$$D^p h(z_1, ..., z_d) := \partial^{[p]} h / \partial z_1^{p_1} ... \partial z_d^{p_d}.$$

Definition (a typical smooth class). Given a non-negative integer γ, we let $\mathcal{S}_{\gamma+1,d}\left(\rho, [\underline{a}, \overline{a}]^d\right)$ denote the class of functions such that any function $h \in \mathcal{S}_{\gamma+1,d}\left(\rho, [\underline{a}, \overline{a}]^d\right)$ satisfies: (1) h is continuous on $[\underline{a}, \overline{a}]^d$, and all partial derivatives of h exist for all p with $[p] \leq \gamma$; (2) $|D^p h(X)| \leq \rho$ for all $X \in [\underline{a}, \overline{a}]^d$ and all p with $[p] \leq \gamma$, where $D^0 h(X) = h(X)$ and ρ is a constant independent of the smoothness parameter γ; (3) $\left|D^p h(X) - D^p(X')\right| \leq \rho \left|X - X'\right|_\infty$ for all $X, X' \in [\underline{a}, \overline{a}]^d$ and all p with $[p] = \gamma$. When $\rho = 1$, $d = 1$, $\underline{a} = 0$ and $\overline{a} = 1$, we use the shortform $\mathcal{S}_{\gamma+1} := \mathcal{S}_{\gamma+1,1}(1, [0, 1])$, the standard smooth class of univariate functions.

Kolmogorov and Tikhomirov (1959). For $\mathcal{S}_{\gamma+1}$, the standard smooth class of degree $\gamma + 1$ on $[0, 1]$, the best upper and lower bounds (to our knowledge) for $\log N_\infty(\delta, \mathcal{S}_{\gamma+1})$ due to Kolmogorov and Tikhomirov [7] take the forms

$$\log N_\infty(\delta, \mathcal{S}_{\gamma+1}) \quad \precsim \quad \delta^{\frac{-1}{\gamma+1}} + (\gamma+1)\log\frac{1}{\delta}, \tag{18}$$

$$\log N_\infty(\delta, \mathcal{S}_{\gamma+1}) \quad \succsim \quad \delta^{\frac{-1}{\gamma+1}}. \tag{19}$$

Kolmogorov and Tikhomirov [7] derive the lower bound on the packing number $M_\infty(\delta, \mathcal{S}_{\gamma+1})$ of $\mathcal{S}_{\gamma+1}$ and (19) comes from the fact that $M_\infty(2\delta, \mathcal{S}_{\gamma+1}) \leq N_\infty(\delta, \mathcal{S}_{\gamma+1})$. Upper bound (18) is useful for analyzing the "local complexity" associated with $\mathcal{S}_{\gamma+1}$, as we discuss below.

Definition (local complexity). Given a radius $\tilde{r}_n > 0$ and a function class $\bar{\mathcal{F}}$, define the *local complexity*

$$\mathcal{G}_n(\tilde{r}_n; \bar{\mathcal{F}}) := \mathbb{E}_\varepsilon\left[\sup_{h\in\Lambda(\tilde{r}_n; \bar{\mathcal{F}})}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i h(x_i)\right|\right], \tag{20}$$

where $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, $\varepsilon = \{\varepsilon_i\}_{i=1}^n$, $\Lambda(\tilde{r}_n; \bar{\mathcal{F}}) = \left\{h \in \bar{\mathcal{F}} : |h|_n \leq \tilde{r}_n\right\}$, and $|h|_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(h(x_i))^2}$. Bounds on local complexity are often used to analyze the distance between an unknown function from a class and an estimator of this function constrained in the same class. Therefore, the underlying $\bar{\mathcal{F}}$ in (20) takes the following form

$$\bar{\mathcal{F}} := \{g = g_1 - g_2 : g_1, g_2 \in \mathcal{F}\}. \tag{21}$$

Various papers have studied localized forms of the Rademacher complexity (where ε_i in (20) is a Rademacher random variable) and Gaussian complexity (where ε_i in (20) is a normal random variable); see, e.g., [1, 8, 9, 21].

Implications on noisy fitting. In what follows, let us consider an application of (20). Suppose that one only observes noisy measurements Y_i of $g^*(x_i)$ in the following form:

$$Y_i = g^*(x_i) + \varepsilon_i, \quad i = 1, ..., n, \tag{22}$$

where $\{x_i\}_{i=1}^n$ is a collection of fixed design points and the unobserved noise terms $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. draws from $\mathcal{N}(0,\,1)$. The assumption var $(\varepsilon_i) = 1$ is simply to facilitate the exposition (so that we can focus on the impact of smoothness) and can be relaxed. The normality assumption can also be relaxed, for instance, to allow ε_i only subject to a sub-Gaussian tail condition in both (20) and (22). Examples of sub-Gaussian random variables include bounded variables, mixture of normal and bounded variables, etc.

Suppose that $g^*(\cdot) \in \mathcal{F}$. To find a function $\hat{g}(\cdot)$ that fits $(Y_i,\,x_i)$, a classical approach is based on the least squares

$$\hat{g} \in \arg\min_{\tilde{g}\in\mathcal{F}} \frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \tilde{g}\left(x_i\right)\right)^2. \tag{23}$$

Because $g^* \in \mathcal{F}$ (g^* is feasible) and $\hat{g}$ is optimal, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}(x_i) - g^*(x_i)\right)^2 \leq \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\left(\hat{g}(x_i) - g^*(x_i)\right). \tag{24}$$

Bounding the mean squared error $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}(x_i) - g^*(x_i)\right)^2\right]$ from above can be reduced to: (1) seeking a sharp enough bound $\mathcal{U}_n\left(\tilde{r}_n;\,\bar{\mathcal{F}}\right)$ on $\mathcal{G}_n(\tilde{r}_n;\,\bar{\mathcal{F}})$, and (2) seeking $\tilde{r}_n > 0$ that satisfies

$$\tilde{r}_n^2 \asymp \mathcal{U}_n\left(\tilde{r}_n;\,\bar{\mathcal{F}}\right). \tag{25}$$

We refer the readers to the textbook by [26] (Chapter 13, 2019) for a detailed explanation on how this argument works.

It is well known that the local complexity $\mathcal{G}_n\left(\tilde{r}_n;\,\bar{\mathcal{F}}\right)$ in (20) can be bounded by exploiting the Dudley's entropy integral (see, e.g., [10, 21, 26]). In the case where $\mathcal{F} = \mathcal{S}_{\gamma+1}$ in (21), the entropy integral approach gives that

$$\mathcal{G}_n(\tilde{r}_n;\,\bar{\mathcal{F}}) \leq \frac{16}{\sqrt{n}}\int_{\frac{\tilde{r}_n^2}{4}}^{\tilde{r}_n}\sqrt{\log N_\infty\left(\delta,\,\bar{\mathcal{F}}\right)}d\delta + \frac{\tilde{r}_n^2}{4} \leq \underbrace{c\left(\tilde{r}_n\sqrt{\frac{\gamma+1}{n}} + \frac{1}{\sqrt{n}}\tilde{r}_n^{\frac{2\gamma+1}{2\gamma+2}}\right) + \frac{\tilde{r}_n^2}{4}}_{\mathcal{U}_n\left(\tilde{r}_n;\,\bar{\mathcal{F}}\right)}$$

where the second inequality follows from (18). Setting $\mathcal{U}_n\left(\tilde{r}_n;\,\bar{\mathcal{F}}\right) = \tilde{r}_n^2$ yields

$$\tilde{r}_n^2 \asymp \max\left\{\frac{\gamma\vee 1}{n},\,\left(\frac{1}{n}\right)^{\frac{2(\gamma+1)}{2(\gamma+1)+1}}\right\} = \begin{cases}\frac{\gamma\vee 1}{n}, & n \precsim (\gamma\vee 1)^{2\gamma+3}\\ \left(\frac{1}{n}\right)^{\frac{2(\gamma+1)}{2(\gamma+1)+1}}, & n \succsim (\gamma\vee 1)^{2\gamma+3}.\end{cases} \tag{26}$$

Therefore, an upper bound on $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}(x_i) - g^*(x_i)\right)^2\right]$ scales as $\frac{\gamma\vee 1}{n}$ when $n \precsim (\gamma\vee 1)^{2\gamma+3}$ and $\left(\frac{1}{n}\right)^{\frac{2(\gamma+1)}{2(\gamma+1)+1}}$ when $n \succsim (\gamma\vee 1)^{2\gamma+3}$.

We can also derive (26) using the approach due to Mendelson [13], which bounds the local complexity through eigenvalues of the empirical kernel matrix associated with a reproducing kernel Hilbert space (RKHS). In particular, we consider the following RKHS (that contains $\mathcal{S}_{\gamma+1}$)

$$\mathcal{H}_{\gamma+1} = \left\{f:\,[0,\,1]\to\mathbb{R}|f^{(\gamma)}\text{ is absolutely continuous and }\int_0^1\left[f^{(\gamma+1)}\left(t\right)\right]^2dt \leq 1\right\} \tag{27}$$

and then apply the Mendelson approach to arrive at (26). Suppose $n \geq \gamma + 1$. The scaling $\frac{\gamma \vee 1}{n}$ is associated with the space of polynomials of degree γ, and the bound $\left(\frac{1}{n}\right)^{\frac{2(\gamma+1)}{2(\gamma+1)+1}}$ is associated with the space

$$\mathcal{H}_{\gamma+1,0} = \{f : [0, 1] \to \mathbb{R} | \text{ for all non-negative integer } k \leq \gamma,\ f^{(k)}(0) = 0,$$
$$f^{(\gamma)} \text{ is abs. cont. and } \int_0^1 \left[f^{(\gamma+1)}(t)\right]^2 dt \leq 1\}.$$

This can be seen from the fact that $f \in \mathcal{H}_{\gamma+1}$ has the expansion $f(x) = \sum_{k=0}^{\gamma} f^{(k)}(0)\frac{x^k}{k!} + \int_0^1 f^{(\gamma+1)}(t)\frac{(x-t)_+^{\gamma}}{\gamma!}dt$, where $(w)_+ = w \vee 0$. The scheme above is important as it allows one to implement (23) via spline bases and kernel functions, very popular techniques for the problems of fitting and interpolating functions (see, e.g., [2, 19, 25]). For example, if $\mathcal{F} = \mathcal{H}_2$ in (23), we have the cubic spline estimator $\hat{g}(x) = \hat{\alpha}_0 + \hat{\alpha}_1 x + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{\pi}_i \mathcal{K}_1(x, x_i)$, where $\mathcal{K}_1(x_i, x_j) = \int_0^1 (x_i - t)_+ (x_j - t)_+ dt$ and

$$(\hat{\alpha}, \hat{\pi}) = \arg \min_{(\alpha,\pi)\in\mathbb{R}^2\times\mathbb{R}^n} \left\{\frac{1}{2n}\left|Y - Z\alpha - \sqrt{n}\mathbb{K}_1\pi\right|_2^2\right\},$$
$$\text{s.t.} \pi^T \mathbb{K}_1 \pi \leq 1, \quad \forall k = 0, ..., \beta + 1, \tag{28}$$

with $\alpha = (\alpha_0, \alpha_1)^T$, Z an $n \times 2$ matrix whose ith row is $(1, x_i)$, and $\mathbb{K}_1$ the $n \times n$ kernel matrix consisting of entries $\mathcal{K}_1(x_i, x_j)$.

It is worth pointing out that in the broad statistics, machine learning, and econometrics community, researchers generally take $O\left(n^{\frac{-2(\gamma+1)}{2(\gamma+1)+1}}\right)$ as the upper bound for $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{g}(x_i) - g^*(x_i))^2\right]$ in terms of (23). This practice assumes $\log N_\infty(\delta, \mathcal{S}_{\gamma+1}) \asymp \delta^{\frac{-1}{\gamma+1}}$ based on (18) and (19). In deriving (18), Kolmogorov and Tikhomirov [7] partition $[0, 1]$ into $s \asymp \delta^{\frac{1}{\gamma+1}}$ intervals and consider a grid of points $(x_0, x_1, ..., x_s)$. The term $(\gamma + 1)\log\frac{1}{\delta}$ in (18) comes from counting the number of distinct values of $\left(\left[\frac{f^{(k)}(x_0)}{\delta_k}\right], k = 0, ..., \gamma, f \in \mathcal{S}_{\gamma+1}\right)$ given the prespecified accuracy δ_k for each kth derivative ($\delta_0 = \delta$). Assuming $(\gamma + 1)\log\frac{1}{\delta}$ is negligible has a similar effect on $\log N_\infty(\delta, \mathcal{S}_{\gamma+1})$ as assuming $f^{(k)}(0) = 0$ for all $k \leq \gamma$ and $f \in \mathcal{S}_{\gamma+1}$.[4]

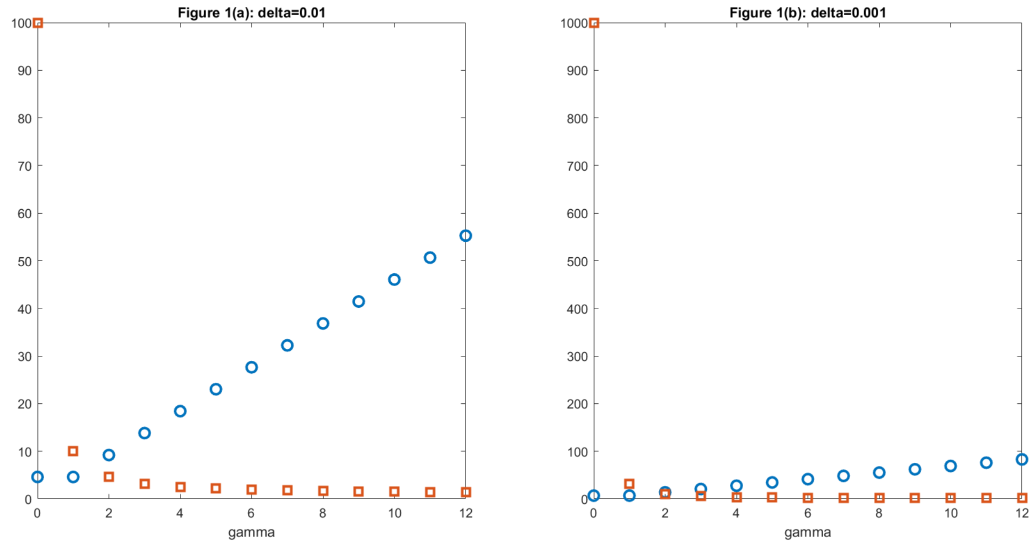
For small γ, (18) and (19) would be comparable; nonasymptotically, for large enough (still finite) γ, $(\gamma + 1)\log\frac{1}{\delta}$ can be greater than $\delta^{\frac{-1}{\gamma+1}}$ even with small δ. Figure 1 exhibits the growth of $(\gamma \vee 1)\log\frac{1}{\delta}$ and $\delta^{\frac{-1}{\gamma+1}}$ as γ increases for $\delta = 0.01$ and $\delta = 0.001$. Taking $n^{\frac{-2(\gamma+1)}{2(\gamma+1)+1}}$ as the upper bound for the mean squared error of (23) essentially assumes that the sample size $n \gtrsim (\gamma \vee 1)^{2\gamma+3}$ as in (26).

Information theoretic lower bounds and minimax optimality. Under the assumption $(\gamma + 1)\log\frac{1}{\delta} \precsim \delta^{\frac{-1}{\gamma+1}}$, one could use a Yang-Barron version of Fano's method (see [27]) to show that

$$\inf_{\tilde{g}} \sup_{g\in\mathcal{S}_{\gamma+1}} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{g}(x_i) - g(x_i)\right]^2\right] \gtrsim \min\left\{1, \left(\frac{1}{n}\right)^{\frac{2(\gamma+1)}{2(\gamma+1)+1}}\right\};$$

[4] On a related note, we conjecture that the lower bound (19) can be sharpened based on the construction in [7].

Figure 1: "○": $(\gamma \vee 1)\log\frac{1}{\delta}$; "□": $\delta^{\frac{-1}{\gamma+1}}$. (a): $\delta = 0.01$; (b): $\delta = 0.001$



see [26] (Chapter 15, 2019) for more details on this derivation. The bound above implies the rate $\left(\frac{1}{n}\right)^{\frac{2(\gamma+1)}{2(\gamma+1)+1}}$ is minimax optimal. Again, we emphasize that this derivation assumes $(\gamma+1)\log\frac{1}{\delta} \leq \delta^{\frac{-1}{\gamma+1}}$ in (18).

Considering $\gamma = \beta + 1$, in sum, under the assumption $n \gtrsim (\beta \vee 1)^{2\beta+5}$, the minimax optimal rate $\left(\frac{1}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$ (which decreases in β) coincides with the convergence rate of the least squares. We can view this requirement on the sample size, $n \gtrsim (\beta \vee 1)^{2\beta+5}$, a "curse of smoothness" for least squares estimations of elements in $\mathcal{S}_{\beta+2}$. In our subsequent results for ODE solution classes, we will show that the "curse of smoothness" is exacerbated in the nonautonomous systems.

3.2 Results for ODE solution classes

In this subsection, we consider the ODEs in the forms

$$\begin{aligned} \textit{autonomous systems: } y'(x) &= f(y(x)), \qquad y(x_0) = y_0, &(29)\\ \textit{nonautonomous systems: } y'(x) &= f(x, y(x)), \qquad y(x_0) = y_0. &(30)\end{aligned}$$

Extensions of the analyses in this section to higher order ODEs are certainly possible but are very laborious.

3.2.1 Nonasymptotic bounds on derivatives and covering numbers

We have established a general bound on the covering numbers of ODE solution classes in Theorem 2.1, which states that the covering number of the underlying solution class is bounded from above by the product of the covering number of the class $\mathcal{F}$ that f ranges over and the covering number of the set where initial values lie. For the nonautonomous system (30), it appears there is still room to improve upon (7). After all, y is a univariate function even though f in (30) is a bivariate function (on the other hand, improvement upon (7) may be more limited for the autonomous system (29)). This observation motivates the following lemmas which build upon or extend the classical Picard smoothness structure (6) to higher degree smoothness structures. These results are not only useful for potentially tightening bounds on the covering numbers but also provide guidance on the design of function fitting algorithms (as we will see in Section 3.2.2).

Lemma 3.1 (autonomous systems). *Let β be a non-negative integer. In terms of* (29)*, assume that f is is continuous on $[y_0 - b,\, y_0 + b]$ and $\beta-$times differentiable, and for all $y,\, \tilde{y} \in [y_0 - b,\, y_0 + b]$,*

$$\left|f^{(k)}(y)\right| \;\leq\; 1, \quad \forall\, 0 \leq k \leq \beta, \tag{31}$$

$$\left|f^{\beta}(y) - f^{\beta}(\tilde{y})\right| \;\leq\; |y - \tilde{y}|, \tag{32}$$

where $f^{(0)} = f$. Suppose $y\,(x) \in [y_0 - b,\, y_0 + b]$ for $x \in [x_0 - a,\, x_0 + a]$.

(i) For all $x,\, x^{'} \in [x_0 - \alpha,\, x_0 + \alpha]$ with $\alpha = a \wedge b$, we have

$$\left|y^{(k)}(x)\right| \leq (k-1)!$$

for all $0 < k \leq \beta + 1$ and

$$\left|y^{(\beta+1)}(x) - y^{(\beta+1)}(x^{'})\right| \leq (\beta+1)!|x - x^{'}|.$$

(ii) There exists an ODE with a solution such that the absolute value of the kth derivative of this solution equals $(k-1)!$ for all $k = 1, ..., \beta + 2$. For example, $y^{'} = e^{-y-\frac{1}{2}}$ is such an ODE.[5]

Lemma 3.2 (nonautonomous systems). *In terms of (30), assume that f is continuous on $\Upsilon = [x_0 - a,\, x_0 + a] \times [y_0 - b,\, y_0 + b]$ and all partial derivatives D^p of f exist for all p with $[p] = p_1 + p_2 \leq \beta$; $|D^p f\,(x,\, y)| \leq 1$ for all p with $[p] \leq \beta$ and $(x,\, y) \in \Upsilon$, where $D^0 f\,(x,\, y) = f\,(x,\, y)$; and*

$$\left|D^p f\,(x,\, y) - D^p f\left(x^{'},\, \tilde{y}\right)\right| \leq \max\left\{\left|x - x^{'}\right|,\, |y - \tilde{y}|\right\} \qquad \forall\,(x,\, y)\,,\, \left(x^{'},\, \tilde{y}\right) \in \Upsilon, \tag{33}$$

for all p with $[p] = \beta$. Then for all $x,\, x^{'} \in [x_0 - \alpha,\, x_0 + \alpha]$ with $\alpha = a \wedge b$, we have

$$\left|y^{(k)}(x)\right| \leq 2^{k-1}(k-1)!$$

for all $0 \leq k \leq \beta + 1$ and

$$\left|y^{(\beta+1)}(x) - y^{(\beta+1)}(x^{'})\right| \leq 2^{\beta+1}\,(\beta+1)!|x - x^{'}|.$$

The proofs for Lemmas 3.1 and 3.2 are provided in Section A.3 of the supplementary materials. Recalling *Question 1*, now we see from Lemma 3.1(ii) that the derivatives of y may not stay "nicely" bounded even if the absolute values of the derivatives of f are all bounded by 1. Based on Lemmas 3.1–3.2 and Theorem 2.1, the following two theorems reveal that our upper bounds for the covering

[5]Note that $f(y) = e^{-y-\frac{1}{2}}$ is continuous on $\left[-\frac{1}{2},\, \frac{1}{2}\right]$, $\beta-$times differentiable, and satisfies (31)–(32) for all $y \in \left[-\frac{1}{2},\, \frac{1}{2}\right]$.

numbers of the $(\beta+2)-$degree smooth solution classes are greater than those of the standard $(\beta+2)-$degree smooth class of univariate functions.

Let us first introduce several definitions used throughout the rest of this section. To lighten notations, from now on, let us assume the initial condition $y(0)=y_0$ in (29)–(30), with $|y_0|\le C_0$. Given $C_0>0$, $b>0$, and $\alpha=\min\{1,\,b\}$, let $\overline{C}:=C_0+b$, $\Xi:=[0,\,1]\times\left[-\overline{C},\,\overline{C}\right]$, and $L_{\max}:=\sup_{x\in[0,\,\alpha]}\left\{\exp(x)\left[1+\int_0^x\exp(-s)\,ds\right]\right\}$.

Theorem 3.1 (autonomous systems). *In terms of (29), assume f ranges over $\mathcal{S}_{\beta+1,\,1}\left(1,\,\left[-\overline{C},\,\overline{C}\right]\right)$. For a given $\frac{\delta}{5}\in(0,\,\alpha)$, we have*

$$\begin{aligned}\log N_{\infty}(\delta,\,\mathcal{Y}) &\le \min\left\{\min_{\gamma\in\{0,\dots,\beta\}} Z_1(\delta,\gamma),\ \min_{\gamma\in\{0,\dots,\beta\}} Z_2\left(\frac{\delta}{L_{\max}},\gamma\right)\right\}, &(34)\\ \log N_{\infty}(\delta,\,\mathcal{Y}_1) &\le \min_{\gamma\in\{0,\dots,\beta\}} Z_3(\delta,\gamma), &(35)\end{aligned}$$

where $\mathcal{Y}$ is the class of solutions (to (29) with $f\in\mathcal{S}_{\beta+1,\,1}\left(1,\,\left[-\overline{C},\,\overline{C}\right]\right)$) on $[0,\,\alpha]$, $\mathcal{Y}_1$ is the class consisting of the first derivative y' of $y\in\mathcal{Y}$, and

$$\begin{aligned}Z_1(\delta,\gamma) &= \log\left(\prod_{i=0}^{\gamma} i!\right)+\frac{\gamma+3}{2}\log\frac{5}{\delta}+\alpha\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+2}}\log 2+\log\left(4\overline{C}\right),\\ Z_2(\delta,\gamma) &= \frac{\gamma+2}{2}\log\frac{5}{\delta}+2\overline{C}\,(\log 2)\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+1}}+\log 4+\log\left(\frac{C_0}{\delta}+1\right),\\ Z_3(\delta,\gamma) &= \log\left(\prod_{i=0}^{\gamma} i!\right)+\frac{\gamma+2}{2}\log\frac{5}{\delta}+\alpha\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+1}}\log 2+\log 4.\end{aligned}$$

Theorem 3.2 (nonautonomous systems). *In terms of (30), assume f ranges over $\mathcal{S}_{\beta+1,\,2}(1,\,\Xi)$. For a given $\frac{\delta}{5}\in(0,\,\alpha)$, we have*

$$\begin{aligned}\log N_{\infty}(\delta,\,\mathcal{Y}) &\le \min\left\{\min_{\gamma\in\{0,\dots,\beta\}} W_1(\delta,\gamma),\ \min_{\gamma\in\{0,\dots,\beta\}} W_2\left(\frac{\delta}{L_{\max}},\gamma\right)\right\}, &(36)\\ \log N_{\infty}(\delta,\,\mathcal{Y}_1) &\le \min_{\gamma\in\{0,\dots,\beta\}} W_3(\delta,\gamma) &(37)\end{aligned}$$

where $\mathcal{Y}$ is the class of solutions (to (30) with $f\in\mathcal{S}_{\beta+1,\,2}(1,\,\Xi)$) on $[0,\,\alpha]$, $\mathcal{Y}_1$ is the class consisting of the first derivative y' of $y\in\mathcal{Y}$, and

$$\begin{aligned}W_1(\delta,\gamma) &= \log\left(\prod_{i=0}^{\gamma} i!\right)+\frac{\gamma^2+\gamma}{2}\log 2+\frac{\gamma+3}{2}\log\frac{5}{\delta}+\alpha\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+2}}\log 4+\log\left(4\overline{C}\right),\\ W_2(\delta,\gamma) &= \frac{2^{\gamma+2}}{\gamma+1}\log\frac{5}{\delta}+20\left(\overline{C}\vee 1\right)(\log 2)\left(\frac{\delta}{5}\right)^{\frac{-2}{\gamma+1}}+4\log 2+\log\left(\frac{C_0}{\delta}+1\right),\\ W_3(\delta,\gamma) &= \log\left(\prod_{i=0}^{\gamma} i!\right)+\frac{\gamma^2+\gamma}{2}\log 2+\frac{\gamma+2}{2}\log\frac{5}{\delta}+\alpha\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+1}}\log 4+\log 4.\end{aligned}$$

The proofs for Theorems 3.1 and 3.2 are provided in Section A.4 of the supplementary materials. The bounds related to Z_2 and W_2 are based on Theorem 2.1. The bounds related to Z_1, Z_3, W_1 and W_3 are based on Lemmas 3.1–3.2, which imply that $\mathcal{Y}\subseteq\mathcal{AS}^{\dagger}_{\gamma+2,\overline{C}}$ for the autonomous systems and $\mathcal{Y} \subseteq \mathcal{S}^{\dagger}_{\gamma+2,\overline{C}}$ for the nonautonomous systems, with $\mathcal{AS}^{\dagger}_{\gamma+2,\overline{C}}$ and $\mathcal{S}^{\dagger}_{\gamma+2,\overline{C}}$ defined as follows.

Definition (nonstandard smooth classes)

- Given $\overline{C}$ and a non-negative integer γ, we let $\mathcal{AS}^{\dagger}_{\gamma+2,\overline{C}}$ denote the class of functions such that any function $h \in \mathcal{AS}^{\dagger}_{\gamma+2,\overline{C}}$ satisfies the following properties: (1) h is continuous on $[0,\, \alpha]$ and differentiable $\gamma+1$ times; (2) $|h(x)| \leq \overline{C}$, and $\left|h^{(k)}(x)\right| \leq (k-1)!$ for all $k = 1, ..., \gamma+1$ and $x \in [0,\, \alpha]$; (3) $\left|h^{(\gamma+1)}(x) - h^{(\gamma+1)}(x^{'})\right| \leq (\gamma+1)! \left|x - x^{'}\right|$ for all $x,\, x^{'} \in [0,\, \alpha]$.

- We let $\mathcal{S}^{\dagger}_{\gamma+2,\overline{C}}$ denote the class of functions such that any function $h \in \mathcal{S}^{\dagger}_{\gamma+2,\overline{C}}$ satisfies the following properties: (1) h is continuous on $[0,\, \alpha]$ and differentiable $\gamma+1$ times; (2) $|h(x)| \leq \overline{C}$, and $\left|h^{(k)}(x)\right| \leq 2^{k-1}\,(k-1)!$ for all $k = 1, ..., \gamma+1$ and $x \in [0,\, \alpha]$; (3) $\left|h^{(\gamma+1)}(x) - h^{(\gamma+1)}(x^{'})\right| \leq 2^{\gamma+1}\,(\gamma+1)! \left|x - x^{'}\right|$ for all $x,\, x^{'} \in [0,\, \alpha]$.

Comparing the autonomous system with the nonautonomous system

The function $Z_2(\delta, \gamma)$ in Theorem 3.1 associated with the autonomous system comes from a straightforward application of (18), Theorem 2.1 (specialized to $m = 1$)[6], as well as the fact $\mathcal{S}_{\beta+1} \subseteq \mathcal{S}_{\beta} \subseteq \mathcal{S}_{\beta-1} \cdots$.

In comparison, first notice that the function $W_2(\delta, \gamma)$ in Theorem 3.2 associated with the nonautonomous system involves $\frac{2^{\gamma+2}}{\gamma+1} \log \frac{5}{\delta}$, substantially larger in order than $\frac{\gamma+2}{2} \log \frac{5}{\delta}$ in $Z_2(\delta, \gamma)$. The term $\frac{2^{\gamma+2}}{\gamma+1} \log \frac{5}{\delta}$ has to do with the fact that a function of two variables has 2^{γ} γth partial derivatives. While Kolmogorov and Tikhomirov [7] explicitly derived $(\gamma+1) \log \frac{1}{\delta}$ for the univariate function class, moving from univariate functions to multivariate functions, this type of log terms appears unmentioned in [7] possibly because γ is assumed to be very small. However, it does not take a large γ for $\left(\frac{\delta}{5}\right)^{\frac{-2}{\gamma+1}}$ to be dominated by $\frac{2^{\gamma+2}}{\gamma+1} \log \frac{5}{\delta}$ even with a reasonably small accuracy parameter δ; for example, with $\frac{\delta}{5} = 0.001$, $\left(\frac{\delta}{5}\right)^{\frac{-2}{\gamma+1}} \leq \frac{2^{\gamma+2}}{\gamma+1} \log \frac{5}{\delta}$ for any $\gamma \geq 3$. Consequently, the log term $\frac{2^{\gamma+2}}{\gamma+1} \log \frac{5}{\delta}$ plays a role in the rates of convergence of least squares (2) and sample size requirement, just like the term $(\gamma+1) \log \frac{1}{\delta}$ in (18) discussed in Section 3.1.

Second, the derivations of $Z_2(\delta, \gamma)$ and $W_2(\delta, \gamma)$ are based on Theorem 2.1; therefore, $W_2(\delta, \gamma)$ involves $\left(\frac{\delta}{5}\right)^{\frac{-2}{\gamma+1}}$ (since f in (30) is a function of two variables) while $Z_2(\delta, \gamma)$ involves $\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+1}}$ (since f in (29) is a univariate function). On the other hand, the derivations of $Z_1(\delta, \gamma)$ and $W_1(\delta, \gamma)$ are based on Lemmas 3.1 and 3.2, and consequently, $\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+2}}$ shows up in both $W_1(\delta, \gamma)$ and $Z_1(\delta, \gamma)$; after all, even though f in (30) is a function of two variables, the solution y is a univariate function, and $\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+2}}$ in $W_1(\delta, \gamma)$ reflects this characteristic. The term $\log\left(\prod_{i=0}^{\gamma} i!\right)$ in $Z_1(\delta, \gamma)$ of Theorem 3.1 and $W_1(\delta, \gamma)$ of Theorem 3.2 comes from the factorial part in Lemmas 3.1 and 3.2. The term $\frac{\gamma^2+\gamma}{2} \log 2$ in $W_1(\delta, \gamma)$ of Theorem 3.2 comes from the exponential part in Lemma 3.2. Note that these terms do not appear in (18) for the standard smooth classes, and reflect the higher complexity of $\mathcal{AS}^{\dagger}_{\beta+2,\overline{C}}$ and $\mathcal{S}^{\dagger}_{\beta+2,\overline{C}}$.

Based on the above comparison, for large enough β, the nonautonomous system (30) benefits more from $W_1(\delta, \gamma)$ than the autonomous system (29) from $Z_1(\delta, \gamma)$. Recall that by Lemma 3.1(ii),

[6] In this case, we can set $L_{\max} = \sup_{x \in [0,\, \alpha]} \left\{\exp(x) \left[1 + \int_0^x \exp(-s)\, ds\right]\right\}$.

the bound $(k-1)!$ on the absolute value of the kth derivative of $y(x)$ (for all $k=1,...,\beta+1$) and the bound $(\beta+1)!\left|x-x^{'}\right|$ on $\left|y^{(\beta+1)}(x)-y^{(\beta+1)}(x^{'})\right|$ are tight in (29). While a tight bound is hard to obtain for (30), it turns out that the bounds $2^{k-1}(k-1)!$ from Lemma 3.2 only make $W_1(\delta,\gamma)$ in Theorem 3.2 differ from $Z_1(\delta,\gamma)$ in Theorem 3.1 by an extra term $\frac{\gamma^2+\gamma}{2}\log 2$. For any $\gamma\geq 5$, $\log\left(\prod_{i=0}^{\gamma} i!\right)\geq\frac{\gamma^2+\gamma}{2}\log 2$ and therefore the factorial part $\left(\prod_{i=0}^{\gamma} i!\right)$ dominates the exponential part $\left(2^{\frac{\gamma^2+\gamma}{2}}\right)$.

For large enough $\gamma\in\{0,...,\beta\}$ and a range of δ, terms like $\log\left(\prod_{i=0}^{\gamma} i!\right)$ can dominate $\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+2}}$ in $Z_1(\delta,\gamma)$ and $W_1(\delta,\gamma)$, and terms like $\frac{2^{\gamma+2}}{\gamma+1}\log\frac{5}{\delta}$ and even $\frac{\gamma+3}{2}\log\frac{5}{\delta}$ can dominate $\left(\frac{\delta}{5}\right)^{\frac{-2}{\gamma+1}}$ and $\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+1}}$ in $W_2(\delta,\gamma)$ and $Z_2(\delta,\gamma)$. Because of this, we choose γs that minimize the Z functions in Theorem 3.1, and the W functions in Theorem 3.2, using the fact that $\mathcal{AS}^{\dagger}_{\beta+2,\overline{C}}\subseteq\mathcal{AS}^{\dagger}_{\beta+1,\overline{C}}\subseteq\mathcal{AS}^{\dagger}_{\beta,\overline{C}}\cdots$, $\mathcal{S}^{\dagger}_{\beta+2,\overline{C}}\subseteq\mathcal{S}^{\dagger}_{\beta+1,\overline{C}}\subseteq\mathcal{S}^{\dagger}_{\beta,\overline{C}}\cdots$, etc. These previously mentioned log terms would clearly have implications on the rates of convergence of least squares (2) and sample size requirement (as we will see in Section 3.2.2).

The best lower bound?

We can easily obtain a lower bound on the covering numbers of solution classes associated with the nonautonomous systems by restricting (30) to a class of separable ODEs, and then apply the construction in [7] which gives (19). This result is stated formally below.

Lemma 3.3 (lower bound based on separable ODEs). *Let us consider a special case of (30):*

$$\text{separable systems: } y^{'}(x)=f(x), \qquad y(0)=y_0. \tag{38}$$

In terms of (38), suppose $f\in\mathcal{S}_{\beta+1}$. Let $\mathcal{Y}^{sep}$ be the class of solutions (to (38) with $f\in\mathcal{S}_{\beta+1}$) on $[0,\,1]$ and $\mathcal{Y}_1^{sep}$ be the class consisting of the first derivative $y^{'}$ of $y\in\mathcal{Y}_{sep}$. We simply have

$$\log N_{\infty}\left(\delta,\,\mathcal{Y}^{sep}\right) \succsim \delta^{\frac{-1}{\beta+2}}, \tag{39}$$

$$\log N_{\infty}\left(\delta,\,\mathcal{Y}_1^{sep}\right) \succsim \delta^{\frac{-1}{\beta+1}}. \tag{40}$$

In the case of (38), $\mathcal{Y}_1^{sep}=\mathcal{S}_{\beta+1}$, and $\mathcal{Y}^{sep}$ and $\mathcal{S}_{\beta+2}$ have similar size (with differences only in the constants). Let $\mathcal{Y}$ be the class of solutions (to (30) with $f\in\mathcal{S}_{\beta+1,\,2}(1,\,\Xi)$). Since $\mathcal{Y}^{sep}\subseteq\mathcal{Y}$, we must have

$$\log N_{\infty}\left(\delta,\,\mathcal{Y}\right)\geq\log N_{\infty}\left(\delta,\,\mathcal{Y}^{sep}\right)\succsim\delta^{\frac{-1}{\beta+2}}.$$

Similarly, letting $\mathcal{Y}_1$ be the class consisting of the first derivative $y^{'}$ of $y\in\mathcal{Y}$, we must have

$$\log N_{\infty}\left(\delta,\,\mathcal{Y}_1\right)\geq\log N_{\infty}\left(\delta,\,\mathcal{Y}_1^{sep}\right)\succsim\delta^{\frac{-1}{\beta+1}}.$$

The questions remain, whether, one can find a solution class whose size attains (36). Before answering this question, a more fundamental question is whether one can find a smooth class whose size attains (18). We conjecture that the lower bound (19) can be sharpened based on the construction in [7]. The problems only get more challenging in the ODE context and we hope to seek insights from the community.

3.2.2 Implications of the previous results

In what follows, we apply the results in Section 3.2.1 to study the theoretical behavior of least squares in noisy function fitting problems.

Problem setup. Let us consider (1) where in **problem (i):** $y(\cdot)$ is a solution to (29) and $f \in \mathcal{S}_{\beta+1,\,1}\left(1,\,\left[-\overline{C},\,\overline{C}\right]\right)$; **in problem (ii):** $y(\cdot)$ is a solution to (30) and $f \in \mathcal{S}_{\beta+1,\,2}(1,\,\Xi)$. We seek rates of convergence for the mean squared errors of the least squares

$$\hat{y} \in \arg\min_{\tilde{y}\in\mathcal{Y}} \frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \tilde{y}(x_i)\right)^2 \tag{41}$$

where in problem (i), $\mathcal{Y}$ is the class of solutions (to (29) with $f \in \mathcal{S}_{\beta+1,\,1}\left(1,\,\left[-\overline{C},\,\overline{C}\right]\right)$) on $[0,\,\alpha]$; in problem (ii), $\mathcal{Y}$ is the class of solutions (to (30) with $f \in \mathcal{S}_{\beta+1,\,2}(1,\,\Xi)$) on $[0,\,\alpha]$.

Definition. Let us define the following functions

$$\begin{aligned}
\mathcal{A}_1(\tilde{r}_n) &= c_0 \min_{\gamma\in\{0,\ldots,\beta\}}\left(\frac{\tilde{r}_n}{\sqrt{n}}\sqrt{1\vee\log\left(\prod_{i=0}^{\gamma} i!\right)} + \frac{1}{\sqrt{n}}\tilde{r}_n^{\frac{2\gamma+3}{2\gamma+4}}\right),\\
\mathcal{A}_2(\tilde{r}_n) &= c_1 \min_{\gamma\in\{0,\ldots,\beta\}}\left(\tilde{r}_n\sqrt{\frac{\gamma\vee 1}{n}} + \frac{1}{\sqrt{n}}\tilde{r}_n^{\frac{2\gamma+1}{2\gamma+2}}\right),\\
\mathcal{B}_1(\tilde{r}_n) &= c_0 \min_{\gamma\in\{0,\ldots,\beta\}}\left(\tilde{r}_n\sqrt{\frac{1}{n}\log\left(\prod_{i=0}^{\gamma} i!\right)} + \tilde{r}_n\sqrt{\frac{\gamma^2\vee 1}{n}} + \frac{1}{\sqrt{n}}\tilde{r}_n^{\frac{2\gamma+3}{2\gamma+4}}\right),\\
\mathcal{B}_2(\tilde{r}_n) &= c_1 \min_{\gamma\in\{1,\ldots,\beta\}}\left(\tilde{r}_n\sqrt{\frac{2^{\gamma}}{n(\gamma\vee 1)}} + \frac{1}{\sqrt{n}}\tilde{r}_n^{\frac{\gamma}{\gamma+1}}\right), \qquad (\beta>0),
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{M}_1^a(\gamma) &= \max\left\{\frac{\sigma^2}{n}\log\left(\prod_{i=0}^{\gamma} i!\right),\,\frac{\sigma^2}{n},\,\left(\frac{\sigma^2}{n}\right)^{\frac{2(\gamma+2)}{2(\gamma+2)+1}}\right\},\\
\mathcal{M}_2^a(\gamma) &= \max\left\{\frac{\sigma^2(\gamma\vee 1)}{n},\,\left(\frac{\sigma^2}{n}\right)^{\frac{2(\gamma+1)}{2(\gamma+1)+1}}\right\},\\
\mathcal{M}_1(\gamma) &= \max\left\{\frac{\sigma^2}{n}\log\left(\prod_{i=0}^{\gamma} i!\right),\,\frac{\sigma^2(\gamma^2\vee 1)}{n},\,\left(\frac{\sigma^2}{n}\right)^{\frac{2(\gamma+2)}{2(\gamma+2)+1}}\right\},\\
\mathcal{M}_2(\gamma) &= \max\left\{\frac{\sigma^2 2^{\gamma}}{n(\gamma\vee 1)},\,\left(\frac{\sigma^2}{n}\right)^{\frac{\gamma+1}{\gamma+2}}\right\},
\end{aligned}$$

and

$$\tilde{r}_{n,a}^2 = \min\left\{\mathcal{M}_1^a(\gamma_{1a}^*),\,\mathcal{M}_2^a(\gamma_{2a}^*)\right\}, \quad \text{(autonomous systems)} \tag{42}$$

$$\tilde{r}_{n}^2 = \min\left\{\mathcal{M}_1(\gamma_{1}^*),\,\mathcal{M}_2(\gamma_{2}^*)\right\}, \quad \text{(nonautonomous systems)} \tag{43}$$

where γ_{1a}^*, γ_{2a}^*, γ_1^*, and γ_2^* are the minimizers that give $\mathcal{A}_1(\tilde{r}_n)$, $\mathcal{A}_2(\tilde{r}_n)$, $\mathcal{B}_1(\tilde{r}_n)$ and $\mathcal{B}_2(\tilde{r}_n)$, respectively.

Theorem 3.3. *Under the* **problem setup**,
(i) if $\overline{C} \precsim 1$ *and* $\max\{\alpha, C_0\} \succsim \tilde{r}_{n,a}$, *then*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}(x_i)-y(x_i)\right)^2\right] \precsim \tilde{r}_{n,a}^2 + \exp\left\{-cn\sigma^{-2}\tilde{r}_{n,a}^2\right\}; \quad \textit{(autonomous systems)}$$

(ii) if $\overline{C} \precsim 1$, $\frac{\sigma^2}{n} \precsim 1$,[7] *and* $\max\{\alpha, C_0\} \succsim \tilde{r}_n$, *then*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}(x_i)-y(x_i)\right)^2\right] \precsim \tilde{r}_n^2 + \exp\left\{-cn\sigma^{-2}\tilde{r}_n^2\right\}. \quad \textit{(nonautonomous systems)}$$

(iii) If $\frac{n}{\sigma^2} \succsim \left(\gamma\sqrt{\log(\gamma \vee 1)}\right)^{4(\gamma+2)+2}$ *with* $\gamma \in \{0, ..., \beta\}$, *for both autonomous systems and nonautonomous systems, we have*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}(x_i)-y(x_i)\right)^2\right] \precsim \left(\frac{\sigma^2}{n}\right)^{\frac{2(\gamma+2)}{2(\gamma+2)+1}} + \exp\left\{-c\left(\frac{n}{\sigma^2}\right)^{\frac{1}{2(\gamma+2)+1}}\right\}. \tag{44}$$

(iv) For the nonautonomous systems, if $\frac{n}{\sigma^2} \succsim \left(\beta\sqrt{\log(\beta \vee 1)}\right)^{4(\beta+2)+2}$, *(44) holds with* $\gamma = \beta$ *and the rate* $\left(\frac{\sigma^2}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$ *is (minimax) optimal.*

The proof for Theorem 3.3 is provided in Section A.5 of the supplementary materials.

The bounds in Theorem 3.3 exploit the covering number results in Theorems 3.1 and 3.2. Note that neither $\mathcal{M}_1^a$ nor $\mathcal{M}_2^a$ for the autonomous systems outperform each other in all situations across different sample sizes (n) and degrees of smoothness (β) for f in the ODEs. These bounds for Theorem 3.3(i) may seem cumbersome but hard to simplify (even asymptotically). On the other hand, $\mathcal{M}_2$ dominates $\mathcal{M}_1$ quickly as γ increases; as a consequence, for the nonautonomous systems, we may consider setting $\tilde{r}_n^2 = \mathcal{M}_1(\gamma_1^*)$ in Theorem 3.3(ii).

Remark. The assumptions $\max\{\alpha, C_0\} \succsim \tilde{r}_{n,a}$ and $\max\{\alpha, C_0\} \succsim \tilde{r}_n$ in Theorem 3.3 simply exclude the trivial case where b and C_0 are "too small". Without these assumptions, as long as f is bounded from above, we would simply bound the convergence rate of $\sqrt{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}(x_i)-y(x_i)\right)^2\right]}$ with $\max\{\alpha, C_0\}$.

Recalling *Question 3*, we now see from Theorem 3.3(iii) that, the convergence rate for the mean squared error of $\hat{y}$ is roughly bounded from above by $\left(\frac{\sigma^2}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$ if $n \succsim \sigma^2\left(\beta\sqrt{\log(\beta \vee 1)}\right)^{4(\beta+2)+2}$, and under this condition, the rate $\left(\frac{\sigma^2}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$ is (minimax) optimal for the nonautonomous systems.

[7]In fact, $\left(\frac{\sigma^2}{n}\right)^{\frac{\gamma+1}{\gamma+2}}$ in $\mathcal{M}_2(\gamma)$ is derived for $\gamma > 0$ in Section A.5 for technical reasons. However, since we are only interested in (43), under the condition $\frac{\sigma^2}{n} \precsim 1$, having $\gamma = 0$ in $\mathcal{M}_2(\gamma)$ will not affect the result in part (ii).

Practical fitting via kernel functions

In Theorem 3.3, we have applied the covering number results from Theorems 3.1 and 3.2 (which are based on Theorem 2.1 and Lemmas 3.1–3.2) to analyze the theoretical behavior of the least squares. Alternatively, we can also exploit the factorial bounds in Lemmas 3.1–3.2 directly by considering the following spaces (that contain $\mathcal{Y}$):

$$\begin{aligned}\mathcal{AH}^{\dagger}_{\beta+2} = \{h:\,[0,\,1]\to\mathbb{R}|h^{(k-1)}\text{ is abs. cont. and}\\ \int_0^1\left[h^{(k)}\left(t\right)\right]^2dt\le[(k-1)!]^2\;\;\forall k\in\{1,...,\beta+2\}\},\end{aligned}\tag{45}$$

$$\begin{aligned}\mathcal{H}^{\dagger}_{\beta+2} = \{h:\,[0,\,1]\to\mathbb{R}|h^{(k-1)}\text{ is abs. cont. and}\\ \int_0^1\left[h^{(k)}\left(t\right)\right]^2dt\le\left[2^{k-1}\left(k-1\right)!\right]^2\;\;\forall k\in\{1,...,\beta+2\}\}.\end{aligned}\tag{46}$$

Importantly, working with the spaces (45) and (46) allows us to implement (41) via kernel functions. In particular, let us consider

$$\hat{y}\in\arg\min_{\tilde{y}\in\mathcal{E}}\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i-\tilde{y}\left(x_i\right)\right)^2\tag{47}$$

where in problem (i), we let $\mathcal{E}=\mathcal{AH}^{\dagger}_{\beta+2}$; in problem (ii), we let $\mathcal{E}=\mathcal{H}^{\dagger}_{\beta+2}$.

For a given $k\in\{0,...,\beta+1\}$, let $\mathbb{K}_k\in\mathbb{R}^{n\times n}$ consist of entries in the following form

$$\begin{aligned}\mathcal{K}_k\left(x_i,\,x_j\right) &= x_i\wedge x_j,\quad k=0\\ \mathcal{K}_k\left(x_i,\,x_j\right) &= \int_0^1\frac{(x_i-t)^k_+}{k!}\frac{(x_j-t)^k_+}{k!}dt,\quad k>0\end{aligned}$$

with $(x)_+=x\vee 0$. For problem (i), solving program (47) is equivalent to solving the following problem

$$\begin{aligned}(\hat{\alpha},\,\hat{\pi}) = \arg\min_{(\alpha,\,\pi)\in\mathbb{R}\times\mathbb{R}^n}\frac{1}{2n}\left|Y-\alpha 1_n-\sqrt{n}\mathbb{K}_{\beta+1}\pi\right|_2^2\\ \text{s.t.}\pi^T\mathbb{K}_k\pi\le C\left(k!\right)^2,\quad\forall k=0,...,\beta+1\end{aligned}\tag{48}$$

(where 1_n is a column vector of 1s) and form $\hat{y}\left(\cdot\right)=\hat{\alpha}+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\pi}_i\mathcal{K}_{\beta+1}\left(\cdot,\,x_i\right)$. For problem (ii), we simply replace (48) by

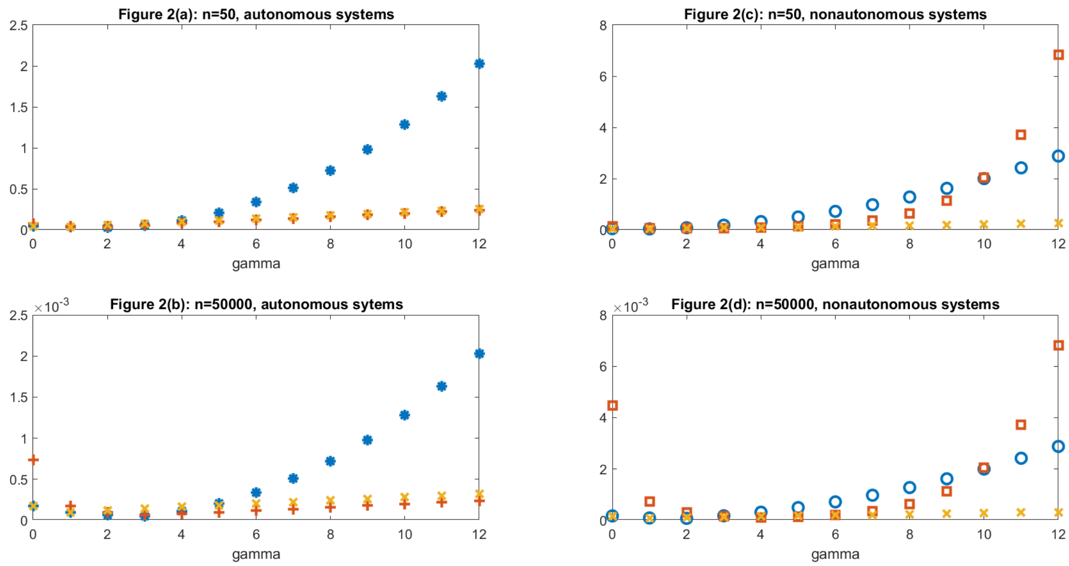
$$\pi^T\mathbb{K}_k\pi\le C\left(2^k k!\right)^2\quad\forall k=0,...,\beta+1.\tag{49}$$

In this section, we may take $C=1$; more generally in practice, if we have no prior knowledge on C, we can use cross validation to select C.

The typical kernel fitting algorithm such as (28) in Sobolev space taking the form of (27) has only one constraint related to the squared RKHS norm. In comparison, (48) and (49) exploit the unique structures of (45) and (46) as a result of Lemmas 3.1–3.2.

Theorem 3.4. *Assume* $\overline{C}\precsim 1$ *and* $\max\{\alpha,\,C_0\}\succsim\bar{r}_n$. *Under the* **problem setup**, *in terms of (47) and for both autonomous systems and nonautonomous systems, we have*

Figure 2: "$*$": $\mathcal{M}_1^a(\gamma)$; "$+$": $\mathcal{M}_2^a(\gamma)$; "$\bigcirc$": $\mathcal{M}_1(\gamma)$; "$\square$": $\mathcal{M}_2(\gamma)$; "$\times$": $\mathcal{M}_3(\gamma)$



$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}(x_i)-y(x_i)\right)^2\right]\precsim\bar{r}_n^2+\exp\left\{-cn\sigma^{-2}\bar{r}_n^2\right\} \tag{50}$$

where

$$\bar{r}_n^2=\min_{\gamma\in\{0,\dots,\beta\}}\max\left\{\frac{\sigma^2\left((\gamma\wedge n)\vee 1\right)}{n},\ \left[2^{\gamma+1}(\gamma+1)!\right]^{\frac{2}{2(\gamma+2)+1}}\left(\frac{\sigma^2}{n}\right)^{\frac{2(\gamma+2)}{2(\gamma+2)+1}}\right\}$$
$$\precsim\min_{\gamma\in\{0,\dots,\beta\}}\max\left\{\frac{\sigma^2\left((\gamma\wedge n)\vee 1\right)}{n},\ \left(\frac{(\gamma\vee 1)\,\sigma^2}{n}\right)^{\frac{2(\gamma+2)}{2(\gamma+2)+1}}\right\}.$$

The proof for Theorem 3.4 is provided in Section A.6 of the supplementary materials.

In contrast with (44) in Theorem 3.3(iii), one theoretical drawback of bound (50) is that it cannot achieve the rate $\left(\frac{\sigma^2}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$ even if $n\succsim\sigma^2\left(\beta\sqrt{\log(\beta\vee 1)}\right)^{4(\beta+2)+2}$. However, when n is this large, the factor $\beta\vee 1$ is much less substantial than $\frac{\sigma^2}{n}$. For all practical purposes, the guarantees for (47) via kernel fitting are quite comparable as those in Theorem 3.3(i)-(ii). Figure 2 exhibits the growth of $\mathcal{M}_1^a(\gamma)$, $\mathcal{M}_2^a(\gamma)$, $\mathcal{M}_1(\gamma)$, $\mathcal{M}_2(\gamma)$, and $\mathcal{M}_3(\gamma):=\left(\frac{(\gamma\vee 1)\sigma^2}{n}\right)^{\frac{2(\gamma+2)}{2(\gamma+2)+1}}$ as γ increases for $n=50$ and $n=50000$. These plots suggest that it is hard for the minimizers γ_{1a}^*, γ_{2a}^* in (42), γ_1^*, and γ_2^* in (43), and the minimizer yielding $\bar{r}_n^2$, to take large γ values unless n is huge. For small γs, the factor $\gamma\vee 1$ does not contribute much in $\left(\frac{(\gamma\vee 1)\sigma^2}{n}\right)^{\frac{2(\gamma+2)}{2(\gamma+2)+1}}$.

4 Conclusion

Numerical methods for recovering ODE solutions from data largely rely on approximating the solutions with basis functions or kernel functions under a least squares criterion. This strategy assumes some ambient space of smooth functions that contains the solutions and seeks a close enough estimator from the restricted class of functions. The accuracy of these methods hinges on the smoothness of the solutions. In this paper, we provide a theoretical foundation for these methods by establishing novel results on the smoothness and covering numbers of ODE solution classes (as a measure of their "size") as well as the rates of convergence for least squares fitting in noisy settings.

We choose to focus on ODEs in this paper as this is a natural and necessary first step before one delves into more complicated dynamic systems. A potential but perhaps challenging extension would be to explore the "size" of solution classes associated with partial differential equations. As for other extensions, because of the deep connections between ODEs and contraction mapping, it may be worthwhile to extend the analyses in this paper to study the "size" of fixed point classes (which play an important role in reinforcement learning).

References

[1] P. Bartlett and S. Mendelson (2002). "Gaussian and Rademacher complexities: risk bounds and structural results". *Journal of Machine Learning Research*, 3, 463–482.

[2] A. Berlinet and C. Thomas-Agnan (2004). *Reproducing kernel Hilbert spaces in probability and statistics.* Kluwer Academic, Norwell, MA.

[3] M. Cinelli, W. Quattrociocchi, A. Galeazzi, et al. (2020). "The COVID-19 social media infodemic". *Sci Rep*, 10, 16598.

[4] E. A. Coddington and N. Levinson (1955). *Theory of ordinary differential equations.* McGraw-Hill, New York, NY.

[5] P. Deuflhard and F. Bornemann (2000). *Scientific computing with ordinary differential equations.* Springer, New York.

[6] Dudley, R. M. (1978). "Central limit theorems for empirical measures". *Annals of Probability*, 6, 899–929.

[7] A. Kolmogorov and B. Tikhomirov (1959). "ϵ-entropy and ϵ-capacity of sets in functional spaces". *Uspekhi Mat. Nauk.*, 86, 3-86. Appeared in English as American Mathematical Society Translations, 17, 277–364, 1961.

[8] V. Koltchinskii (2001). "Rademacher penalties and structural risk minimization". *IEEE Transactions on Information Theory*, 47(5), 1902–1914.

[9] V. Koltchinskii (2006). "Local Rademacher complexities and oracle inequalities in risk minimization". *Annals of Statistics*, 34, 2593–2656.

[10] M. Ledoux and M. Talagrand (1991). *Probability in Banach spaces: isoperimetry and processes.* Springer, New York, NY.

[11] H. Liang and H. Wu (2008). "Parameter estimation for differential equation models using a framework of measurement error in regression models". *Journal of the American Statistical Association*, 103:484, 1570–1583.

[12] A. J. Meade and A. A. Fernadez (1994). "The numerical solution of linear ordinary differential equations by feedforward neural networks". *Mathematical and Computer Modelling*,19(12), 1–25.

[13] S. Mendelson (2002). "Geometric parameters of kernel machines". *Proceedings of COLT*, 29–43.

[14] S. Mehrkanoon, T. Falck and J. A. K. Suykens (2012). "Approximate solutions to ordinary differential equations using least squares support vector machines". *IEEE Transactions on Neural Networks and Learning Systems*, 23(9), 1356–1367.

[15] A. A. Poyton, M. S. Varziri, K. B. McAuley, P. J. McLellan, J. O. Ramsay (2006) "Parameter estimation in continuous-time dynamic models using principal differential analysis". *Computers & Chemical Engineering*, 30(4), 698–708.

[16] J. O. Ramsay and B. W. Silverman (2005). *Fitting differential equations to functional data: Principal differential analysis. In: Functional Data Analysis*. Springer, New York, NY.

[17] J. O. Ramsay (1996). "Principal differential analysis: Data reduction by differential operators". *Journal of the Royal Statistical Society, Series B*, 58, 495–508.

[18] J. Ramsay, G. Hooker, D. Campbell and J. Cao (2007). "Parameter estimation for differential equations: A generalized smoothing approach". *Journal of the Royal Statistical Society*, 69, 741–796.

[19] B. Schölkopf and A. Smola (2002) *Learning with kernels*. MIT Press, Cambridge, MA.

[20] J. M. Steele (1978). "Empirical discrepancies and sub-additive processes". *Annals of Probability*, 6, 118–127.

[21] S. van de Geer (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

[22] J. M. Varah (1982). "A spline least squares method for numerical parameter estimation in differential equations". *SIAM Journal on Scientific and Statistical Computing*, 3(1), 28–46.

[23] A. Volkening, D. F. Linder, M. A. Porter, and G. A. Rempala (2020). "Forecasting elections using compartmental models of infection". *SIAM Review*, 62(4), 837–865.

[24] E. Vynnycky and R. White (2010). *An introduction to infectious diseases modelling*. Oxford University Press.

[25] G. Wahba (1990). *Spline models for observational data. SIAM, CBMS-NSF Regional Conference Series in Applied Mathematics*. Philadelphia, PA.

[26] M. J. Wainwright (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.

[27] Y. Yang and A. Barron (1999). "Information-theoretic determination of minimax rates of convergence". *Annals of Statistics*, 27(5), 1564–1599.

Note: The supplement of this paper has its own list of references at the end.

Supplementary materials for "Classes of ODE solutions: smoothness, covering numbers, implications for noisy function fitting, and the curse of smoothness phenomenon"

by Ying Zhu[8] and Mozhgan Mirzaei[9]

A Proofs for the main results

A.1 Theorem 2.1

Proof. Let $L_{\max}=\sup_{x\in[a_0,\,a_0+\alpha]}\left\{\exp\left(x\sqrt{L^2+1}\right)\left[1+\int_{a_0}^{x}\exp\left(-s\sqrt{L^2+1}\right)ds\right]\right\}$. For a given $\delta>0$, we consider the smallest $\frac{\delta}{L_{\max}}-$covering of $\mathcal{F}$ with respect to the sup-norm. We also consider the smallest $\frac{\delta}{L_{\max}}-$covering of $\mathbb{B}_2\left(C_0\right):=\left\{\theta\in\mathbb{R}^m:\left|\theta\right|_2\leq C_0\right\}$ (where the initial values lie) with respect to the l_2-norm. Note that the standard volumetric argument yields

$$\log N_2\left(\frac{\delta}{L_{\max}},\,\mathbb{B}_2\left(C_0\right)\right)\leq m\log\left(\frac{2C_0L_{\max}}{\delta}+1\right).$$

By Theorem B.2 in Section B, for any $y\in\mathcal{Y}$ with $Y_0\in\mathbb{B}_2\left(C_0\right)$, we can find an element (indexed by i) from the smallest $\frac{\delta}{L_{\max}}-$covering of $\mathcal{F}$ and an element (indexed by $i^{'}$) from the smallest $\frac{\delta}{L_{\max}}-$covering of $\mathbb{B}_2\left(C_0\right)$ such that

$$\left|y^{(k)}\left(x\right)-y_{(i,i')}^{(k)}\left(x\right)\right|\leq\frac{\delta}{L_{\max}}\exp\left(x\sqrt{L^2+1}\right)\left[1+\int_0^x\exp\left(-s\sqrt{L^2+1}\right)ds\right]\leq\delta\qquad\forall x\in\left[a_0,\,a_0+\alpha\right]$$

for all $k=0,...,m-1$, where $y_{(i,i')}$ is a solution to the ODE associated with f_i and the initial value $Y_{0,\,i'}$ from the covering sets, and $y_{(i,i')}^{(k)}$ is the kth derivative of $y_{(i,i')}$ ($k\leq m-1$). Consequently, we obtain a $\delta-$cover of $\mathcal{Y}_k$. We conclude that

$$\log N_{\infty}\left(\delta,\,\mathcal{Y}_k\right)\leq\log N_{\infty}\left(\frac{\delta}{L_{\max}},\,\mathcal{F}\right)+m\log\left(\frac{2C_0L_{\max}}{\delta}+1\right).$$

□

A.2 Corollary 2.1

Proof. Let $N_q\left(\delta,\,\mathbb{B}_q\left(1\right)\right)$ denote the covering number of $\mathbb{B}_q\left(1\right)$ with respect to the l_q-norm and $N_2\left(\frac{\delta}{L_{\max}},\,\mathbb{B}_2\left(C_0\right)\right)$ denote the covering number of $\mathbb{B}_2\left(C_0\right)$ with respect to the l_2-norm. For a given $\delta>0$, let us consider the smallest $\frac{\delta}{L_{\max}L_K}-$covering $\left\{\theta^1,...,\theta^N\right\}$ with respect to the l_q-norm. By (10), note that $\left\{f_{\theta^1},\,f_{\theta^2},...,\,f_{\theta^N}\right\}$ forms a $\frac{\delta}{L_{\max}}-$cover of $\mathcal{F}$ with respect to the sup-norm. We also consider the smallest $\frac{\delta}{L_{\max}}-$covering $\left\{Y_{0,1},...,Y_{0,N'}\right\}$ for the interval $\mathbb{B}_2\left(C_0\right)$ where the initial values lie. By Theorem B.2 in Section B, for a solution y to the ODE with f parameterized by any

[8] Assistant Professor, Department of Economics, UC San Diego. Corresponding author. yiz012@ucsd.edu.
[9] PhD, former student at Department of Mathematics, UC San Diego. momirzae@ucsd.edu

$\theta \in \mathbb{B}_q(1)$ (and subject to (10)) and $Y_0 \in \mathbb{B}_2(C_0)$, we can find $i \in \{1, ..., N\}$ and $i' \in \left\{1, ..., N'\right\}$ such that

$$\left|y_{(i,i')}(x) - y(x)\right| \leq \frac{\delta}{C_{\max}} \exp(x) \left[1 + \int_0^x \exp(-s)\, ds\right] \leq \delta \qquad \forall x \in [0,\, \alpha]$$

where $y_{(i,i')}$ is a solution to the ODE with f parameterized by θ^i and the initial value being $y_{0,i'}$. Consequently, we obtain a $\delta-$cover of $\mathcal{Y}$. By the standard volumetric argument which yields

$$\log N_2(\delta,\, \mathbb{B}_q(1)) \leq K \log\left(1 + \frac{2L_{\max}L_K}{\delta}\right),$$

and

$$\log N_2\left(\frac{\delta}{L_{\max}},\, \mathbb{B}_2(C_0)\right) \leq \log\left(\frac{2C_0L_{\max}}{\delta} + 1\right),$$

we conclude that

$$\log N_\infty(\delta,\, \mathcal{Y}) \leq K \log\left(1 + \frac{2L_{\max}L_K}{\delta}\right) + \log\left(\frac{2C_0L_{\max}}{\delta} + 1\right).$$

□

A.3 Lemmas 3.1 and 3.2

In part (i) of Lemma 3.1, by the mean value theorem, (31) with $k = 0$ (i.e., $\left|f^{(0)}(y(x))\right| = |f(y(x))| = \left|y'(x)\right| \leq 1$) and (32) imply that

$$\left|f^{\beta}(y(x)) - f^{\beta}(y(x'))\right| \leq |y(x) - y(x')| \leq |x - x'|; \tag{51}$$

moreover, (31) with $1 \leq k \leq \beta$ implies that

$$\left|f^{k-1}(y(x)) - f^{k-1}(y(x'))\right| \leq |y(x) - y(x')| \leq |x - x'|, \quad \forall\, 1 \leq k \leq \beta. \tag{52}$$

In Lemma 3.2, by the mean value theorem, the assumption that $|f(x,\, y(x))| = \left|y'(x)\right| \leq 1$ and (33) imply that

$$\left|D^p f(x,\, y(x)) - D^p f\left(x',\, y\left(x'\right)\right)\right| \leq \max\left\{\left|x - x'\right|,\, \left|y(x) - y\left(x'\right)\right|\right\} \leq \left|x - x'\right| \tag{53}$$

for all p with $[p] = \beta$; moreover, the assumption that $|D^p f(x,\, y)| \leq 1$ for all p with $1 \leq [p] \leq \beta$ implies that

$$\left|D^{p-1} f(x,\, y(x)) - D^{p-1} f\left(x',\, y\left(x'\right)\right)\right| \leq \left|x - x'\right| + \left|y(x) - y\left(x'\right)\right| \leq 2\left|x - x'\right|, \tag{54}$$

for all p with $1 \leq [p] \leq \beta$.

Remark. The following proofs can easily handle situations where the bound “1” on the absolute values of the derivatives and the Lipschitz conditions for f in Lemmas 3.1 and 3.2 is replaced with general constants.

We first prove parts (i)-(ii) and then part (iii). Let us begin with some intuitions for the autonomous system. When $\beta = 0$, $\left|y^{'}(x)\right| = |f(y(x))| \leq 1$ for all x and $y(x)$ on $[x_0 - a,\, x_0 + a] \times [y_0 - b,\, y_0 + b]$; moreover, we have

$$\left|y^{'}(x) - y^{'}\left(x^{'}\right)\right| = \left|f(y(x)) - f\left(y\left(x^{'}\right)\right)\right| \leq \left|x - x^{'}\right|.$$

When $\beta = 1$, note that

$$\left|y^{(2)}(x)\right| = \left|\frac{\partial y^{(1)}(x)}{\partial x}\right| = \left|f^{(1)}(y(x))y^{(1)}(x)\right| \leq 1$$

and

$$\begin{aligned}\left|y^{(2)}(x) - y^{(2)}\left(x^{'}\right)\right| =& \left|f^{(1)}(y(x))y^{(1)}(x) - f^{(1)}\left(y\left(x^{'}\right)\right)y^{(1)}\left(x^{'}\right)\right| \\ \leq& \left|f^{(1)}(y(x))y^{(1)}(x) - f^{(1)}\left(y\left(x^{'}\right)\right)y^{(1)}(x)\right| \\ &+ \left|f^{(1)}\left(y\left(x^{'}\right)\right)y^{(1)}(x) - f^{(1)}\left(y\left(x^{'}\right)\right)y^{(1)}\left(x^{'}\right)\right| \\ \leq& \left|y(x) - y\left(x^{'}\right)\right| + \left|y^{(1)}(x) - y^{(1)}\left(x^{'}\right)\right| \\ \leq& 2\left|x - x^{'}\right|.\end{aligned}$$

When $\beta = 2$, we have

$$\begin{aligned}\left|y^{(3)}(x)\right| =& \left|f^{(2)}(y(x))\left(y^{(1)}(x)\right)^2 + f^{(1)}(y(x))y^{(2)}(x)\right| \\ \leq& \left|f^{(2)}(y(x))\right| + \left|f^{(1)}(y(x))\right| \leq 2\end{aligned}$$

and

$$\left|y^{(3)}(x) - y^{(3)}\left(x^{'}\right)\right| \leq 6\left|x - x^{'}\right|.$$

After trying $\beta = 1, ..., 3$, we observe the following pattern:

$$\begin{aligned}&y^{(1)}(x) = f(y(x)) \\ &- - - - - - - - - > (0) \\ &y^{(2)}(x) = f^{(1)}(y(x)).y^{(1)}(x) \\ &\qquad\quad = f^{(1)}(y(x)).f(y(x)) \\ &- - - - - - - - - > (1,0) \\ &y^{(3)}(x) = f^{(2)}(y(x)).\left(f(y(x))\right)^2 + \left(f^{(1)}(y(x))\right)^2.f(y(x)) \\ &- - - - - - - - - - > (2,0,0), (1,1,0) \\ &y^{(4)}(x) = f^{(3)}(y(x)).\left(f(y(x))\right)^3 + 2f^{(2)}(y(x)).f^{(1)}(y(x)).\left(f(y(x))\right)^2 + \\ &\qquad\quad 2f^{(2)}(y(x)).f^{(1)}(y(x)).\left(f(y(x))\right)^2 + \left(f^{(1)}(y(x))\right)^3.f(y(x)) \\ &- - - - - - - - - - - > (3,0,0,0), 2(2,1,0,0), 2(2,1,0,0), (1,1,1,0)\end{aligned} \tag{55}$$

In what follows, we derive b_ks (for all $k = 1, ..., \beta + 1$) such that

$$\left|y^{(k)}(x)\right| \leq b_k,$$

and $L_{\beta+1}$ such that

$$\left|y^{(\beta+1)}(x) - y^{(\beta+1)}\left(x^{'}\right)\right| \leq L_{\beta+1}\left|x - x^{'}\right|.$$

Proof of Lemma 3.1(i). For a $\beta-$times differentiable function f, define

$$f^{(a_1,\cdots,a_k)}(x) := f^{(a_1)}(y(x)).f^{(a_2)}(y(x)).\cdots.f^{(a_k)}(y(x)),$$

where $a_1 + a_2 + \cdots + a_k = k-1$ for all $k = 1, ..., \beta+1$, and $a_i \geq 0$ are all integers. We first show that

$$\frac{d}{dx}f^{(a_1,\cdots,a_k)}(x) = \sum_{j=1}^{k} f^{(a_1,\ldots,a_{j-1},a_j+1,0,a_{j+1},\ldots,a_k)}(x). \tag{56}$$

The equality (56) follows from the derivations below:

$$\begin{aligned}
&\frac{d}{dx}(f^{(a_1)}(y(x)).f^{(a_2)}(y(x)).\cdots.f^{(a_k)}(y(x))) \\
&= \sum_{j=1}^{k} f^{(a_1)}(y(x))\ldots.f^{(a_{j-1})}(y(x)).\frac{d}{dx}(f^{(a_j)}(y(x))).f^{(a_{j+1})}(y(x))\ldots f^{a_k}(y(x)) \\
&\hspace{20em}\text{by product rule} \\
&= \sum_{j=1}^{k} f^{(a_1)}(y(x))\ldots.\left[f^{(a_j+1)}(y(x)).y^{'}(x)\right]\ldots f^{a_k}(y(x)) \\
&\hspace{20em}\text{by chain rule} \\
&= \sum_{j=1}^{k} f^{(a_1)}(y(x))\ldots.\left[f^{(a_j+1)}(y(x)).f(y(x))\right]\ldots f^{a_k}(y(x)).
\end{aligned}$$

Now by induction on k, we show that if f is $\beta-$times differentiable, then for each $1 \leq k \leq \beta+1$, we have

$$y^{(k)}(x) = \sum_{i=1}^{(k-1)!} f^{(a_1^i,\cdots,a_k^i)}(x), \qquad \text{and for all } i \text{ in } (a_1^i,\cdots,a_k^i),\ a_1^i+\ldots+a_k^i = k-1. \tag{57}$$

For the base case, $k = 1 \Rightarrow y^{'}(x) = f(y(x)) = f^{(a_1)}(x) = \sum_{i=1}^{0!} f^{(a_1)}(x)$ where $a_1 = 0$. Now assume $k \leq \beta+1$ and the induction hypothesis holds for $k-1$. Then

$$\begin{aligned}
y^{(k)}(x) &= \frac{d}{dx}\left(y^{(k-1)}(x)\right) \\
&= \frac{d}{dx}\left(\sum_{i=1}^{(k-2)!} f^{(a_1^i,\ldots,a_{k-1}^i)}(x)\right) \text{ where } \forall i,\ a_1^i+\ldots+a_{k-1}^i = k-2, \\
&= \sum_{i=1}^{(k-2)!}\sum_{j=1}^{k-1} f^{\left(a_1^i,\ldots,a_{j-1}^i,a_j+1,0,a_{j+1}^i,\ldots,a_{k-1}^i\right)}(x) \qquad \text{by (56)}.
\end{aligned}$$

Notice that $\left(a_1^i, \ldots, a_{j-1}^i, a_j^i + 1, 0, a_{j+1}^i, \ldots, a_{k-1}^i\right)$ has exactly k terms, and adds up to $a_1^i + \ldots + a_{k-1}^i + 1 + 0 = k - 2 + 1 + 0 = k - 1$, and in total there are $(k-2)!(k-1) = (k-1)!$ terms. This completes the induction. By (31), we have $\left|y^{(k)}(t)\right| \leq (k-1)!$.

To show the second claim $\left|y^{(\beta+1)}(x) - y^{(\beta+1)}(x^{'})\right| \leq (\beta+1)!|x - x^{'}|$, we use (57). Note that

$$\begin{aligned} \left|y^{(\beta+1)}(x) - y^{(\beta+1)}(x^{'})\right| &= \left|\sum_{i=1}^{\beta!} f^{\left(a_1^i,\ldots,a_{\beta+1}^i\right)}(x) - \sum_{i=1}^{\beta!} f^{\left(a_1^i,\ldots,a_{\beta+1}^i\right)}(x^{'})\right| \\ &\leq \sum_{i=1}^{\beta!} \left|f^{\left(a_1^i,\ldots,a_{\beta+1}^i\right)}(x) - f^{\left(a_1^i,\ldots,a_{\beta+1}^i\right)}(x^{'})\right|, \end{aligned}$$

where

$$\begin{aligned} & \left|f^{\left(a_1^i,\ldots,a_{\beta+1}^i\right)}(x) - f^{\left(a_1^i,\ldots,a_{\beta+1}^i\right)}(x^{'})\right| \\ = & \left|\sum_{j=1}^{\beta+1} f^{(a_1)}(y(x)).\ldots.f^{(a_{j-1})}(y(x)).\left(f^{(a_j)}(y(x)) - f^{(a_j)}(y(x^{'}))\right).f^{(a_{j+1})}(y(x^{'}))\ldots f^{(a_{\beta+1})}(y(x^{'}))\right| \\ \leq & \sum_{j=1}^{\beta+1} \left|f^{(a_1)}(y(x))\ldots f^{(a_{j-1})}(y(x))\right| \left|\left(f^{(a_j)}(y(x)) - f^{(a_j)}(y(x^{'}))\right)\right| \left|f^{(a_{j+1})}(y(x^{'}))\ldots f^{(a_{\beta+1})}(y(x^{'}))\right| \\ \leq & \sum_{j=1}^{\beta+1} |x - x^{'}| \end{aligned}$$

and the third line in the above follows from (51) and (52). Hence,

$$\begin{aligned} \left|y^{(\beta+1)}(x) - y^{(\beta+1)}(x^{'})\right| &= \sum_{i=1}^{\beta!} \left|f^{\left(a_1^i,\ldots,a_{\beta+1}^i\right)}(x) - f^{\left(a_1^i,\ldots,a_{\beta+1}^i\right)}(x^{'})\right| \\ &\leq \beta!\,(\beta+1)\,|x - x^{'}| = (\beta+1)!|x - x^{'}|. \end{aligned}$$

□

Proof of Lemma 3.1(ii). In view of (57), if we can find an infinitely differentiable function f on $\mathbb{R}$ such that $\left|f^{(k)}(y(x))\right| = 1$ for some $y(x)$ and all k, then $y^{(k)}(x) = (k-1)!$ for all $k = 1, ..., \beta + 2$. Note that $y^{'} = e^{-y-\frac{1}{2}}$ is such an example. For $y \in \left[-\frac{1}{2}, \frac{1}{2}\right]$, $\left|\frac{d^k\left(e^{-y-\frac{1}{2}}\right)}{dy^k}\right| \leq 1$ for all k, satisfying (31) and (32). To argue formally, let us consider f with $f^{2j+1}(y(x)) = -1$ and $f^{2j}(y(x)) = 1$ for all $j = 0, 1, ...$ and some $y(x)$. We can verify in (55) that $y^{(1)}(x) = 0!$, $y^{(2)}(x) = -1!$, $y^{(3)}(x) = 2!$, and $y^{(4)}(x) = -3!$. Now performing induction on k,we show that if $y^{(2j+1)}(x) = (2j)!$ ($y^{(2j+2)}(x) = -(2j+1)!$), then $y^{(2j+2)}(x) = -(2j+1)!$ (respectively, $y^{(2j+3)}(x) = (2j+2)!$) for all $j = 0, 1, ...$ Note that whenever $y^{(2j+1)}(x) = (2j)!$, each positive summand "1" in $y^{(2j+1)}(x) = \sum_{i=1}^{(2j)!} f^{(a_1^i,\cdots,a_{2j+1}^i)}(x)$ must be due to either any number of even derivatives of f or an even number of odd derivatives of f. Taking the derivative of any even derivative of f (the derivative of any odd derivative of f), evaluated at $y(x)$, gives "-1" (respectively, "1"), by the construction of f. This operation clearly makes each summand in $\sum_{i=1}^{(2j+1)!} f^{(a_1^i,\cdots,a_{2j+2}^i)}(x) = y^{(2j+2)}(x)$ equal to "-1" and therefore $y^{(2j+2)}(x) = -(2j+1)!$. The argument for $y^{(2j+3)}(x) = (2j+2)!$ from

$y^{(2j+2)}(x) = -(2j+1)!$ is very similar. $\square$

Now we show Lemma 3.2. As before, let us gain some intuitions first. For $\beta = 0$, $\left|y^{'}(x)\right| = |f(x,\, y(x))| \leq 1$ for all x and $y(x)$ on $[x_0 - a,\, x_0 + a] \times [y_0 - b,\, y_0 + b]$; moreover, we have

$$\begin{aligned}\left|y^{'}(x) - y^{'}\left(x^{'}\right)\right| &= \left|f(x, y(x)) - f\left(x^{'}, y\left(x^{'}\right)\right)\right| \\ &\leq \left|f(x, y(x)) - f\left(x^{'}, y(x)\right)\right| + \left|f\left(x^{'}, y(x)\right) - f\left(x^{'}, y\left(x^{'}\right)\right)\right| \\ &\leq 2\left|x - x^{'}\right|.\end{aligned}$$

For $\beta = 1$, note that

$$\begin{aligned}\left|y^{(2)}(x)\right| = \left|\frac{\partial y^{(1)}(x)}{\partial x}\right| &= \left|f^{(1)}_y(x, y(x))y^{(1)}(x) + f^{(1)}_x(x, y(x))\right| \\ &\leq \left|f^{(1)}_y(x, y(x))y^{(1)}(x)\right| + \left|f^{(1)}_x(x, y(x))\right| \leq 2\end{aligned}$$

and

$$\begin{aligned}\left|y^{(2)}(x) - y^{(2)}\left(x^{'}\right)\right| \leq& \left|f^{(1)}_y(x, y(x))y^{(1)}(x) - f^{(1)}_y\left(x^{'}, y\left(x^{'}\right)\right)y^{(1)}\left(x^{'}\right)\right| \\ &+ \left|f^{(1)}_x(x, y(x)) - f^{(1)}_x\left(x^{'}, y\left(x^{'}\right)\right)\right| \\ \leq& \left|f^{(1)}_y(x, y(x))y^{(1)}(x) - f^{(1)}_y\left(x^{'}, y\left(x^{'}\right)\right)y^{(1)}(x)\right| \\ &+ \left|f^{(1)}_y\left(x^{'}, y\left(x^{'}\right)\right)y^{(1)}(x) - f^{(1)}_y\left(x^{'}, y\left(x^{'}\right)\right)y^{(1)}\left(x^{'}\right)\right| \\ &+ \left|f^{(1)}_x(x, y(x)) - f^{(1)}_x\left(x^{'}, y\left(x^{'}\right)\right)\right| \\ \leq& 6\left|x - x^{'}\right|.\end{aligned}$$

In what follows, we derive b_ks (for all $k = 1, ..., \beta + 1$) such that

$$\left|y^{(k)}(x)\right| \leq b_k$$

and $L_{\beta+1}$ such that

$$\left|y^{(\beta+1)}(x) - y^{(\beta+1)}\left(x^{'}\right)\right| \leq L_{\beta+1}\left|x - x^{'}\right|.$$

Proof of Lemma 3.2. Writing $\vartheta(x) = (x, y(x))$ and $\eta_j(t_1, t_2) = \frac{\partial^{a_j+b_j}f}{(\partial t_1)^{a_j}(\partial t_2)^{b_j}}(t_1, t_2)$, we have

$$\begin{aligned}&\frac{d}{dx}\vartheta(x) = (1, y^{'}(x)) = (1, f(x, y(x))), \\ &\frac{\partial \eta_j}{\partial t_1}(t_1, t_2) = \frac{\partial^{a_j+b_j+1}f}{(\partial t_1)^{a_j+1}(\partial t_2)^{b_j}}(t_1, t_2), \\ &\frac{\partial \eta_j}{\partial t_2}(t_1, t_2) = \frac{\partial^{a_j+b_j+1}f}{(\partial t_1)^{a_j}(\partial t_2)^{b_j+1}}(t_1, t_2).\end{aligned}$$

Hence,

$$\begin{aligned}
&\frac{d}{dx}\left(\frac{\partial^{a_j+b_j} f}{(\partial t_1)^{a_j}(\partial t_2)^{b_j}}(x,y(x))\right)\\
&=\frac{d}{dx}\left(\eta_j\circ\vartheta(x)\right)\\
&=\left(\frac{\partial\eta_j}{\partial t_1}(x,y(x)),\frac{\partial\eta_j}{\partial t_2}(x,y(x))\right)\cdot(1,f(x,y(x)))\\
&=\frac{\partial^{a_j+b_j+1} f}{(\partial t_1)^{a_j+1}(\partial t_2)^{b_j}}(x,y(x))+\frac{\partial^{a_j+b_j+1} f}{(\partial t_1)^{a_j}(\partial t_2)^{b_j+1}}(x,y(x))f(x,y(x)).
\end{aligned} \tag{58}$$

For a $\beta-$times differentiable function f, define

$$f^{(a_1,b_1\cdots,a_k,b_k)}(x):=\prod_{i=1}^{k}\frac{\partial^{a_i+b_i} f}{(\partial t_1)^{a_i}(\partial t_2)^{b_i}}(x,y(x))$$

where $\sum_{i=1}^{k}a_i+\sum_{i=1}^{k}b_i=k-1$ for all $k=1,...,\beta+1$, and a_i, $b_i\geq 0$ are all integers. We have

$$\begin{aligned}
&\frac{d}{dx}f^{(a_1,b_1\cdots,a_k,b_k)}(x)\\
&=\sum_{j=1}^{k}\frac{\partial^{a_1+b_1} f}{(\partial t_1)^{a_1}(\partial t_2)^{b_1}}(x,y(x))\cdot\ldots\cdot\frac{d}{dx}\left(\frac{\partial^{a_j+b_j} f}{(\partial t_1)^{a_j}(\partial t_2)^{b_j}}(x,y(x))\right)\cdot\ldots\cdot\frac{\partial^{a_k+b_k} f}{(\partial t_1)^{a_k}(\partial t_2)^{b_k}}(x,y(x))\\
&=\sum_{j=1}^{k}\frac{\partial^{a_1+b_1} f}{(\partial t_1)^{a_1}(\partial t_2)^{b_1}}(x,y(x))\cdot\ldots\cdot\\
&\quad\left[\frac{\partial^{a_j+b_j+1} f}{(\partial t_1)^{a_j+1}(\partial t_2)^{b_j}}(x,y(x))+\frac{\partial^{a_j+b_j+1} f}{(\partial t_1)^{a_j}(\partial t_2)^{b_j+1}}(x,y(x))\cdot f(x,y(x))\right]\cdot\ldots\cdot\\
&\quad\frac{\partial^{a_k+b_k} f}{(\partial t_1)^{a_k}(\partial t_2)^{b_k}}(x,y(x))
\end{aligned}$$

where the second equality comes from (58); that is,

$$\begin{aligned}
&\frac{d}{dx}f^{a_1,b_1,\ldots,a_k,b_k}(x)=\\
&\sum_{j=1}^{k}f^{(a_1,b_1,\ldots,a_{j-1},b_{j-1},a_j+1,b_j,a_{j+1},b_{j+1},\ldots,a_k,b_k)}+\sum_{j=1}^{k}f^{(a_1,b_1,\ldots,a_{j-1},b_{j-1},a_j,b_j+1,0,0,a_{j+1},b_{j+1},\ldots,a_k,b_k)}.
\end{aligned} \tag{59}$$

Given (59), now we show that

$$y^{(k)}(x)=\sum_{i=1}^{M_{k-1}}f^{(a_1^i,b_1^i,\ldots,a_k^i,a_k^i)}(x) \tag{60}$$

satisfies the following properties: (1) the integer $M_{k-1}\leq 2^{k-1}(k-1)!$ for each $1\leq k\leq\beta+1$, and (2) $\forall i$, $a_1^i+b_1^i+\ldots+a_k^i+b_k^i=k-1$.

The base case is obvious. Notice that by (59), differentiating a term of the form $f^{(a_1^i,b_1^i,\ldots,a_k^i,b_k^i)}(x)$ gives us $2k$ terms of the form $f^{(c_1^i,d_1^i,\ldots,c_m^i,d_m^i)}(x)$ where $m=k$ or $k+1$, and $c_1^i+d_1^i+\ldots+c_m^i+d_m^i=$

$1+(a_1^i+b_1^i+\ldots+a_k^i+b_k^i)=1+(k-1)=k$. So the total number of terms in $y^{(k+1)} \leq 2k$(number of terms in y^k) $\leq 2k(k-1)!2^{k-1}$. This proves the induction hypothesis. By the assumption that $|f(x,\, y(x))| \leq 1$, we have $\left|y^{(k)}(x)\right| \leq 2^{k-1}(k-1)!$.

To show the second claim $\left|y^{(\beta+1)}(x)-y^{(\beta+1)}(x^{'})\right| \leq 2^{\beta+1}\,(\beta+1)!|x-x^{'}|$, we use (60). Note that

$$\begin{aligned}|y^{(\beta+1)}(x)-y^{(\beta+1)}(x^{'})| &= \left|\sum_{i=1}^{M_\beta} f^{(a_1^i,b_1^i,\ldots,a_{\beta+1}^i,b_{\beta+1}^i)}(x) - \sum_{i=1}^{M_\beta} f^{(a_1^i,b_1^i,\ldots,a_{\beta+1}^i,b_{\beta+1}^i)}(x^{'})\right| \\ &\leq \sum_{i=1}^{M_\beta}\left|f^{(a_1^i,b_1^i,\ldots,a_{\beta+1}^i,b_{\beta+1}^i)}(x) - f^{(a_1^i,b_1^i,\ldots,a_{\beta+1}^i,b_{\beta+1}^i)}(x^{'})\right|\end{aligned}$$

where the integer $M_\beta \leq 2^\beta\beta!$ and

$$\begin{aligned}&\left|f^{(a_1^i,b_1^i,\ldots,a_{\beta+1}^i,b_{\beta+1}^i)}(x) - f^{(a_1^i,b_1^i,\ldots,a_{\beta+1}^i,b_{\beta+1}^i)}(x^{'})\right| \\ =&\Big|\sum_{j=1}^{\beta+1}\frac{\partial^{a_1^i+b_1^i}f}{(\partial t_1)^{a_1^i}(\partial t_2)^{b_1^i}}(x,y(x))\cdot\ldots \\ &\quad\Big|\frac{\partial^{a_{j-1}^i+b_{j-1}^i}f}{(\partial t_1)^{a_{j-1}^i}(\partial t_2)^{b_{j-1}^i}}(x,y(x))\Big|\cdot \\ &\quad\left(\frac{\partial^{a_j^i+b_j^i}f}{(\partial t_1)^{a_j^i}(\partial t_2)^{b_j^i}}(x,y(x)) - \frac{\partial^{a_j^i+b_j^i}f}{(\partial t_1)^{a_j^i}(\partial t_2)^{b_j^i}}(x^{'},y(x^{'}))\right)\cdot \\ &\quad\Big|\frac{\partial^{a_{j+1}^i+b_{j+1}^i}f}{(\partial t_1)^{a_{j+1}^i}(\partial t_2)^{b_{j+1}^i}}(x^{'},y(x^{'}))\Big|\cdot\ldots \\ &\quad\frac{\partial^{a_{\beta+1}^i+b_{\beta+1}^i}f}{(\partial t_1)^{a_{\beta+1}^i}(\partial t_2)^{b_{\beta+1}^i}}(x^{'},y(x^{'}))\Big| \\ \leq&\sum_{j=1}^{\beta+1}\Big|\frac{\partial^{a_1^i+b_1^i}f}{(\partial t_1)^{a_1^i}(\partial t_2)^{b_1^i}}(x,y(x))\Big|\cdot\ldots \\ &\quad\Big|\frac{\partial^{a_{j-1}^i+b_{j-1}^i}f}{(\partial t_1)^{a_{j-1}^i}(\partial t_2)^{b_{j-1}^i}}(x,y(x))\Big|\cdot \\ &\quad\underbrace{\left|\left(\frac{\partial^{a_j^i+b_j^i}f}{(\partial t_1)^{a_j^i}(\partial t_2)^{b_j^i}}(x,y(x)) - \frac{\partial^{a_j^i+b_j^i}f}{(\partial t_1)^{a_j^i}(\partial t_2)^{b_j^i}}(x^{'},y(x^{'}))\right)\right|}_{\leq 2\left|x-x^{'}\right| \text{ by (53) and (54)}}\cdot \\ &\quad\Big|\frac{\partial^{a_{j+1}^i+b_{j+1}^i}f}{(\partial t_1)^{a_{j+1}^i}(\partial t_2)^{b_{j+1}^i}}(x^{'},y(x^{'}))\Big|\cdot\ldots \\ &\quad\Big|\frac{\partial^{a_{\beta+1}^i+b_{\beta+1}^i}f}{(\partial t_1)^{a_{\beta+1}^i}(\partial t_2)^{b_{\beta+1}^i}}(x^{'},y(x^{'}))\Big| \leq 2\,(\beta+1)\left|x-x^{'}\right|.\end{aligned}$$

Hence,

$$\left|y^{(\beta+1)}(x)-y^{(\beta+1)}(x^{'})\right| \leq \left(2^\beta\beta!\right)\left(2\,(\beta+1)\left|x-x^{'}\right|\right) \leq 2^{\beta+1}\,(\beta+1)!|x-x^{'}|.$$

□

A.4 Theorems 3.1 and 3.2

A.4.1 Main argument

In what follows, we show Theorem 3.2 and point out the minor differences in the proof for Theorem 3.1 at the end.

Theorem 3.2

Proof. **Term related to** $W_1(\delta,\gamma)$: In Lemma A.1, we establish the following bound:

$$N_\infty\left(\delta,\, \mathcal{S}^{\dagger}_{\beta+2,\overline{C}}\right) \leq \exp\left[\log\left(\overline{C}\prod_{i=0}^{\gamma} i!\right) + \frac{\gamma^2+\gamma}{2}\log 2 + \frac{\gamma+3}{2}\log\frac{5}{\delta} + \alpha\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+2}}\log 4 + \log 4\right] \tag{61}$$

which is valid for all $\gamma \in \{0,...,\beta\}$. By Lemma 3.2, $\mathcal{Y} \subseteq \mathcal{S}^{\dagger}_{\beta+2,\overline{C}}$, and therefore, choosing γ that minimizes the RHS of (61) yields

$$\log N_\infty\left(\delta,\, \mathcal{Y}\right) \leq \min_{\gamma\in\{0,\ldots,\beta\}} W_1\left(\delta,\gamma\right). \tag{62}$$

Term related to $W_2(\delta,\gamma)$: In Lemma A.2, we establish the following bound:

$$N_\infty\left(\delta,\, \mathcal{S}_{\beta+1,\,2}\left(1,\,\Xi\right)\right) \leq \exp\left[\frac{2^{\gamma+2}}{\gamma+1}\log\frac{5}{\delta} + 20\left(\overline{C}\vee 1\right)\left(\frac{\delta}{5}\right)^{\frac{-2}{\gamma+1}}\log 2 + 4\log 2\right]$$

which is valid for all $\gamma \in \{0,...,\beta\}$. Let $L_{\max} = \sup_{x\in[0,\,\alpha]}\left\{\exp\left(x\right)\left[1+\int_0^x \exp\left(-s\right)ds\right]\right\}$. For a given $\delta > 0$, let us consider the smallest $\frac{\delta}{L_{\max}}-$covering $\{f_1,...,f_N\}$ of $\mathcal{S}_{\beta+1,\,2}\left(1,\,\Xi\right)$ with respect to the sup-norm such that

$$\log N_\infty\left(\frac{\delta}{L_{\max}},\, \mathcal{S}_{\beta+1,\,2}\left(1,\,\Xi\right)\right) \leq \frac{2^{\gamma+2}}{\gamma+1}\log\frac{5L_{\max}}{\delta} + 20\left(\overline{C}\vee 1\right)\left(\frac{\delta}{5L_{\max}}\right)^{\frac{-2}{\gamma+1}}\log 2 + 4\log 2 \tag{63}$$

valid for all $\gamma \in \{0,...,\beta\}$. We also consider the smallest $\frac{\delta}{L_{\max}}-$covering $\left\{y_{0,1},...,y_{0,N'}\right\}$ for the interval $[-C_0,\,C_0]$ where the initial value lies. Note that

$$\log N_\infty\left(\frac{\delta}{L_{\max}},\,[-C_0,\,C_0]\right) \leq \log\left(\frac{C_0L_{\max}}{\delta}+1\right).$$

By Theorem B.1 in Section B, for a solution y to the ODE associated with any $f \in \mathcal{S}_{\beta+1,\,2}\left(1,\,\Xi\right)$ and $y_0 \in [-C_0,\,C_0]$, we can find $i \in \{1,...,N\}$ and $i' \in \left\{1,...,N'\right\}$ such that

$$\left|y\left(x\right) - y_{(i,i')}\left(x\right)\right| \leq \frac{\delta}{L_{\max}}\exp\left(x\right)\left[1+\int_0^x \exp\left(-s\right)ds\right] \leq \delta \qquad \forall x\in[0,\,\alpha]$$

where $y_{(i,i')}$ is a solution to the ODE associated with f_i and the initial value $y_{0,i'}$ from the covering sets. Consequently, we obtain a $\delta-$cover of $\mathcal{Y}$. Choosing γ that minimizes the RHS of (63), we conclude that

$$\log N_\infty\left(\delta,\,\mathcal{Y}\right) \leq \min_{\gamma\in\{0,\ldots,\beta\}} W_2\left(\frac{\delta}{L_{\max}},\gamma\right). \tag{64}$$

Combining (62) and (64) yields the bound on $\log N_\infty\left(\delta,\,\mathcal{Y}\right)$ in Theorem 3.2. We can apply argument almost identical to the above (related to $W_1(\delta,\gamma)$) to derive the bound on $\log N_\infty\left(\delta,\,\mathcal{Y}_1\right)$. In particular, we have

$$\log N_\infty\left(\delta,\,\mathcal{Y}_1\right) \leq \min_{\gamma\in\{0,\ldots,\beta\}} W_3\left(\delta,\gamma\right).$$

□

Theorem 3.1

Proof. With slight modifications, the arguments for Theorem 3.1 are almost identical to those for Theorem 3.2. Because most of these modifications are straightforward, we only point out the main differences.

1. By Lemma 3.1, $\mathcal{Y} \subseteq \mathcal{AS}^{\dagger}_{\beta+2,\overline{C}}$. Following the argument in the proof for Lemma A.1, we establish the following bound:

$$N_{\infty}\left(\delta,\, \mathcal{AS}^{\dagger}_{\beta+2,\overline{C}}\right) \leq \exp\left[\log\left(\overline{C}\prod_{i=0}^{\gamma} i!\right) + \frac{\gamma+3}{2}\log\frac{5}{\delta} + \alpha\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+2}}\log 2 + \log 4\right] \tag{65}$$

which is valid for all $\gamma \in \{0, ..., \beta\}$.

2. In applying Theorem 2.1, we exploit the following bound:

$$N_{\infty}\left(\delta,\, \mathcal{S}_{\beta+1,1}\left(1,\, \left[-\overline{C},\, \overline{C}\right]\right)\right) \leq \exp\left[\frac{\gamma+2}{2}\log\frac{5}{\delta} + 2\overline{C}\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+1}}\log 2 + \log 4\right]$$

which is valid for all $\gamma \in \{0, ..., \beta\}$. □

A.4.2 Lemma A.1

Lemma A.1. *In terms of* $\mathcal{S}^{\dagger}_{\beta+2,\overline{C}}$*, we have (61).*

Proof. By Lemma 3.2, we have shown that $\mathcal{Y} \subseteq \mathcal{S}^{\dagger}_{\beta+2,\overline{C}}$. Note that $\mathcal{S}^{\dagger}_{\beta+2,\overline{C}} \subseteq \mathcal{S}^{\dagger}_{\beta+1,\overline{C}} \subseteq \cdots \subseteq \mathcal{S}^{\dagger}_{2,\overline{C}}$. In what follows, we provide an upper bound on the covering number of $\mathcal{S}^{\dagger}_{\gamma+2,\overline{C}}$ for any given $\gamma \in \{0, ..., \beta\}$. The argument modifies the original proof for smooth classes in [5]. For every function $h \in \mathcal{S}^{\dagger}_{\gamma+2,\overline{C}}$ and $x,\, x+\Delta \in (0, \alpha)$, we have

$$h(x+\Delta) = h(x) + \Delta h'(x) + \frac{\Delta^2}{2!}h''(x) + \cdots + \frac{\Delta^{\gamma}}{\gamma!}h^{(\gamma)}(x) + \frac{\Delta^{\gamma+1}}{(\gamma+1)!}h^{(\gamma+1)}(z).$$

Let us define

$$R_h(x,\Delta) := h(x+\Delta) - h(x) - \Delta h'(x) - \frac{\Delta^2}{2!}h''(x) - \cdots - \frac{\Delta^{\gamma}}{\gamma!}h^{(\gamma)}(x) - \frac{\Delta^{\gamma+1}}{(\gamma+1)!}h^{(\gamma+1)}(x)$$

and we have

$$|R_h(x,\Delta)| = \frac{\Delta^{\gamma+1}}{(\gamma+1)!}\left|h^{(\gamma+1)}(x) - h^{(\gamma+1)}(z)\right| \leq (2|\Delta|)^{\gamma+2}. \tag{66}$$

As a consequence, we obtain

$$h(x+\Delta) = \sum_{k=0}^{\gamma+1}\frac{\Delta^k}{k!}h^{(k)}(x) + R_h(x,\Delta) \quad \text{where } |R_h(x,\Delta)| \leq (2|\Delta|)^{\gamma+2}.$$

Let us consider $h^{(i)} \in \mathcal{S}^{\dagger}_{\gamma+2-i,1}$ for $1 \leq i \leq \gamma+1$ and the above implies that

$$h^{(i)}(x+\Delta) = \sum_{k=0}^{\gamma+1-i}\frac{\Delta^k}{k!}h^{(i+k)}(x) + R_{h^{(i)}}(x,\Delta) \;\; \text{where } |R_{h^{(i)}}(x,\Delta)| \leq (2|\Delta|)^{\gamma+2-i}. \tag{67}$$

To bound $N_\infty\left(\delta, \mathcal{S}^\dagger_{\gamma+2,\overline{C}}\right)$ from above, we fix $\delta > 0$ and $x \in (0,\alpha)$. Suppose that for some $\delta_0,\ldots,\delta_{\gamma+1} > 0$, $h,\, g \in \mathcal{S}^\dagger_{\gamma+2,\overline{C}}$ satisfy

$$\left|h^{(k)}(x) - g^{(k)}(x)\right| \le \delta_k \quad \text{for all } k = 0,\ldots,\gamma+1.$$

We can bound $|h(x+\Delta) - g(x+\Delta)|$ from above for some Δ such that $x+\Delta \in (0,\alpha)$ as follows:

$$\begin{aligned}
|h(x+\Delta) - g(x+\Delta)| &= \left|\sum_{k=0}^{\gamma+1} \frac{\Delta^k}{k!}\left(h^{(k)}(x) - g^{(k)}(x)\right) + R_h(x,\Delta) - R_g(x,\Delta)\right| \\
&\le \sum_{k=0}^{\gamma+1} \frac{|\Delta|^k \delta_k}{k!} + 2(2|\Delta|)^{\gamma+2}.
\end{aligned}$$

If $|\Delta| \le \frac{\left(\frac{\delta}{5}\right)^{\frac{1}{\gamma+2}}}{2}$ and $\delta_k = \left(\frac{\delta}{5}\right)^{1-\frac{k}{\gamma+2}}$, we then have

$$|h(x+\Delta) - g(x+\Delta)| \le \frac{\delta}{5}\left(\sum_{k=0}^{\gamma+1}\frac{1}{k!} + 2\right) \le \delta. \tag{68}$$

In other words, by considering a grid of points $\frac{\left(\frac{\delta}{5}\right)^{\frac{1}{\gamma+2}}}{2}-$apart in $(0,\alpha)$ and covering the kth derivative of functions in $\mathcal{S}^\dagger_{\gamma+2,\overline{C}}$ within $\delta_k = \left(\frac{\delta}{5}\right)^{1-\frac{k}{\gamma+2}}$ at each point, we can then obtain a δ cover in the sup-norm for $\mathcal{S}^\dagger_{\gamma+2,\overline{C}}$. Let $x_1 < \cdots < x_s$ be a $\frac{\left(\frac{\delta}{5}\right)^{\frac{1}{\gamma+2}}}{2}-$grid of points in $(0,1)$ with $s \le 2\alpha\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+2}}+2$. For each $h_0 \in \mathcal{S}^\dagger_{\gamma+2,\overline{C}}$, let us define

$$\mathbb{H}(h_0) := \left\{h \in \mathcal{S}^\dagger_{\gamma+2,\overline{C}} : \left\lfloor \frac{h^{(k)}(x_i)}{\delta_k}\right\rfloor = \left\lfloor \frac{h_0^{(k)}(x_i)}{\delta_k}\right\rfloor, 1 \le i \le s,\ 0 \le k \le \gamma+1\right\}$$

where $\lfloor x \rfloor$ means the largest integer smaller than or equal to x. Our earlier argument implies that the number of distinct sets $\mathbb{H}(h_0)$ with h_0 ranging over $\mathcal{S}^\dagger_{\gamma+2,\overline{C}}$ bounds the $\delta-$covering number of $\mathcal{S}^\dagger_{\gamma+2,\overline{C}}$ from above. Note that for $i = 1,\ldots,s$ and $k = 0,\ldots,\gamma+1$, $\mathbb{H}(h_0)$ depends on $\left\lfloor h_0^{(k)}(x_i)/\delta_k\right\rfloor$ only and the number of distinct sets $\mathbb{H}(h_0)$ is bounded above by the cardinality of

$$I := \left\{\left(\left\lfloor \frac{h^{(k)}(x_i)}{\delta_k}\right\rfloor, 1 \le i \le s \text{ and } 0 \le k \le \gamma+1\right) : h \in \mathcal{S}^\dagger_{\gamma+2,\overline{C}}\right\}. \tag{69}$$

Starting from x_1, let us count the number of possible values of the vector

$$\left(\left\lfloor \frac{h^{(k)}(x_1)}{\delta_k}\right\rfloor, 0 \le k \le \gamma+1\right)$$

with h ranging over $\mathcal{S}^\dagger_{\gamma+2,\overline{C}}$. Since $|h(x_1)| \le \overline{C}, \left|h^{(1)}(x_1)\right| \le 1, \left|h^{(k)}(x_1)\right| \le 2^{k-1}(k-1)!$ for $2 \le k \le \gamma+1$, this number is at most

$$\frac{\overline{C}}{\delta_0}\frac{1}{\delta_1}\frac{2}{\delta_2}\frac{2^2 2!}{\delta_3}\cdots\frac{2^{\gamma-1}(\gamma-1)!}{\delta_\gamma}\frac{2^\gamma \gamma!}{\delta_{\gamma+1}} \le \left(\frac{\delta}{5}\right)^{-\frac{\gamma+3}{2}}\overline{C}2^{\frac{(\gamma+1)\gamma}{2}}\prod_{i=0}^{\gamma} i! \tag{70}$$

Now we move to x_2. Given the values of $\left(\left\lfloor h^{(k)}\left(x_1\right)/\delta_k\right\rfloor, 0 \le k \le \gamma+1\right)$, we count the number of possible values of the vector

$$\left(\left\lfloor \frac{h^{(k)}\left(x_2\right)}{\delta_k}\right\rfloor, 0 \le k \le \gamma+1\right).$$

For each $0 \le k \le \gamma+1$, we define

$$A_k := \left\lfloor \frac{h^{(k)}\left(x_1\right)}{\delta_k}\right\rfloor \quad \text{such that } A_k\delta_k \le h^{(k)}\left(x_1\right) < \left(A_k+1\right)\delta_k.$$

Let us fix $0 \le i \le \gamma+1$. Applying (67) with $x = x_1$ and $\Delta = x_2 - x_1$ yields

$$\left|h^{(i)}\left(x_2\right) - \sum_{k=0}^{\gamma+1-i}\frac{\Delta^k}{k!}h^{(i+k)}\left(x_1\right)\right| \le |2\Delta|^{\gamma+2-i}.$$

Consequently we have

$$\begin{aligned}
&\left|h^{(i)}\left(x_2\right) - \sum_{k=0}^{\gamma+1-i}\frac{\Delta^k}{k!}A_{i+k}\right| \\
&\le \left|h^{(i)}\left(x_2\right) - \sum_{k=0}^{\gamma+1-i}\frac{\Delta^k}{k!}h^{(i+k)}\left(x_1\right)\right| + \left|\sum_{k=0}^{\gamma+1-i}\frac{\Delta^k}{k!}\left(h^{(i+k)}\left(x_1\right) - A_{i+k}\right)\right| \\
&\le |2\Delta|^{\gamma+2-i} + \sum_{k=0}^{\gamma+1-i}\frac{|\Delta|^k}{k!}\delta_{i+k} \\
&\le \left(\frac{\delta}{5}\right)^{1-\frac{i}{\gamma+2}} + \left(\frac{\delta}{5}\right)^{1-\frac{i}{\gamma+2}} = 2\delta_i
\end{aligned}$$

(recalling $|\Delta| = |x_2 - x_1| = \frac{\left(\frac{\delta}{5}\right)^{\frac{1}{\gamma+2}}}{2}$). Therefore, given the values of $\left(\left\lfloor h^{(k)}\left(x_1\right)/\delta_k\right\rfloor, 0 \le k \le \gamma+1\right)$, $h^{(i)}\left(x_2\right)$ takes values in an interval whose length is no greater than $2\delta_i$. Therefore, the number of possible values of $\left(\left\lfloor \frac{h^{(k)}(x_2)}{\delta_k}\right\rfloor, 0 \le k \le \gamma+1\right)$ is at most 2. The same argument goes through when x_1 and x_2 are replaced with x_j and x_{j+1} for any $j = 1, ..., s-1$. This result along with (70) gives

$$\begin{aligned}
|I| &\le 2^s\left(\frac{\delta}{5}\right)^{-\frac{\gamma+3}{2}}\overline{C}2^{\frac{(\gamma+1)\gamma}{2}}\prod_{i=0}^{\gamma} i! \\
&\le 4^{\alpha\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+2}}+1}\left(\frac{\delta}{5}\right)^{-\frac{\gamma+3}{2}}\overline{C}2^{\frac{(\gamma+1)\gamma}{2}}\prod_{i=0}^{\gamma} i! \\
&\le \exp\left[\log\left(\overline{C}\prod_{i=0}^{\gamma} i!\right) + \frac{\gamma^2+\gamma}{2}\log 2 + \frac{\gamma+3}{2}\log\frac{5}{\delta} + \alpha\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+2}}\log 4 + \log 4\right]
\end{aligned}$$

where I is defined in (69). $\square$

A.4.3 Lemma A.2

Lemma A.2. *In terms of* $\mathcal{S}_{\beta+1,2}\left(1, \Xi\right)$*, we have (63).*

Proof. Note that $\mathcal{S}_{\beta+1,2}(1,\,\Xi)\subseteq\mathcal{S}_{\beta,2}(1,\,\Xi)\subseteq\cdots\subseteq\mathcal{S}_{1,2}(1,\,\Xi)$. In what follows, we provide an upper bound on the covering number of $\mathcal{S}_{\gamma+1,2}(1,\,\Xi)$ for any given $\gamma\in\{0,...,\beta\}$. As we discussed in the main paper, while Kolmogorov and Tikhomirov [5] explicitly derived $(\gamma+1)\log\frac{1}{\delta}$ (from counting the number of distinct values in the first interval on [0, 1]) for the univariate function class, moving from univariate functions to multivariate functions, this type of log terms appears unmentioned in [5] possibly because γ is assumed to be very small. However, these log terms would clearly have nonasymptotic implications (on the sample size, as discussed in Section 3.2.2). Therefore, in what follows, we carefully extend the arguments from univariate functions to bivariate functions and derive the log term explicitly.

Let $p=(p_1,p_2)$ and $[p]=p_1+p_2$ where p_1 and p_2 are non-negative integers. We write $D^pf(w_1,w_2)=\partial^{[p]}f/\partial w_1^{p_1}\partial w_2^{p_2}$ with $w=(w_1,\,w_2)$ and $\Delta^p=\Delta_1^{p_1}\Delta_2^{p_2}$ with $\Delta=(\Delta_1,\,\Delta_2)$. For every function $f\in\mathcal{S}_{\gamma+1,2}(1,\,\Xi)$, $w_1,\,w_1+\Delta_1\in(0,1)$, and $w_2,\,w_2+\Delta_2\in\left(-\overline{C},\,\overline{C}\right)$, we have

$$f(w+\Delta)=\sum_{k=0}^{\gamma}\sum_{p:[p]=k}\frac{\Delta^pD^pf(w)}{k!}+\underbrace{\sum_{p:[p]=\gamma}\frac{\Delta^pD^pf(z)}{\gamma!}-\sum_{p:[p]=\gamma}\frac{\Delta^pD^pf(w)}{\gamma!}}_{R_{0,f}(w,\Delta)}$$

for some $z=(z_1,\,z_2)\in(0,1)\times\left(-\overline{C},\,\overline{C}\right)$. Because a function of two variables has 2^{γ} γth partial derivatives, we have

$$|R_{0,f}(w,\Delta)|\leq\frac{|2\Delta|_{\infty}^{\gamma+1}}{\gamma!}. \tag{71}$$

Similarly, letting $w+\Delta:=(w_1+\Delta_1,\,w_2+\Delta_2)$ and $D^{\tilde{p}}f(w+\Delta)\in\mathcal{S}_{\gamma+1-[\tilde{p}],1}(1,\,\Xi)$ for $1\leq[\tilde{p}]\leq\gamma$, we have

$$D^{\tilde{p}}f(w+\Delta)=\sum_{k=0}^{\gamma-[\tilde{p}]}\sum_{p:[p]=k}\frac{\Delta^pD^{p+\tilde{p}}f(w)}{k!}+\underbrace{\sum_{p:[p]=\gamma-[\tilde{p}]}\frac{\Delta^pD^{p+\tilde{p}}f(\tilde{z})}{(\gamma-[\tilde{p}])!}-\sum_{p:[p]=\gamma-[\tilde{p}]}\frac{\Delta^pD^{p+\tilde{p}}f(w)}{(\gamma-[\tilde{p}])!}}_{R_{[\tilde{p}],f}(w,\Delta)} \tag{72}$$

for some $\tilde{z}=(\tilde{z}_1,\,\tilde{z}_2)\in(0,1)\times\left(-\overline{C},\,\overline{C}\right)$, where

$$\left|R_{[\tilde{p}],f}(w,\Delta)\right|\leq\frac{|2\Delta|_{\infty}^{\gamma+1-[\tilde{p}]}}{(\gamma-[\tilde{p}])!}.$$

Suppose that for some $\delta_0,\ldots,\delta_\gamma>0$, $f,\,g\in\mathcal{S}_{\gamma+1,2}(1,\,\Xi)$ satisfy

$$|D^pf(w)-D^pg(w)|\leq\delta_k\quad\text{for all } p \text{ with } [p]=k\in\{0,\ldots,\gamma\}.$$

We can bound $|f(w+\Delta)-g(w+\Delta)|$ from above for some Δ such that $w+\Delta\in(0,1)\times\left(-\overline{C},\,\overline{C}\right)$ as follows:

$$\begin{aligned}|f(w+\Delta)-g(w+\Delta)| &= \left|\sum_{k=0}^{\gamma}\sum_{p:[p]=k}\frac{\Delta^p}{k!}\left(D^pf(w)-D^pg(w)\right)+R_{0,f}(w,\Delta)-R_{0,g}(w,\Delta)\right|\\ &\leq \sum_{k=0}^{\gamma}\frac{|2\Delta|_{\infty}^{k}\,\delta_k}{k!}+2\frac{|2\Delta|_{\infty}^{\gamma+1}}{\gamma!}\end{aligned}$$

where the second line follows from (71) and the fact that a function of two variables has 2^k kth partial derivatives. If $|\Delta|_\infty \leq \frac{\left(\frac{\delta}{5}\right)^{\frac{1}{\gamma+1}}}{2}$ and $\delta_k = \left(\frac{\delta}{5}\right)^{1-\frac{k}{\gamma+1}}$, we then have

$$|h(x+\Delta) - g(x+\Delta)| \leq \frac{\delta}{5}\left(\sum_{k=0}^{\gamma} \frac{1}{k!} + 2\right) \leq \delta. \tag{73}$$

Let $w_{1,1} < \cdots < w_{1,s_1}$ be a $\frac{\left(\frac{\delta}{5}\right)^{\frac{1}{\gamma+1}}}{2}$−grid of points in $(0,1)$ with $s_1 \leq 2\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+1}} + 2$ and $w_{2,1} < \cdots < w_{2,s_2}$ be a $\frac{\left(\frac{\delta}{5}\right)^{\frac{1}{\gamma+1}}}{2}$−grid of points in $\left(-\overline{C},\, \overline{C}\right)$ with $s_2 \leq 4\overline{C}\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+1}} + 2$. For each $f_0 \in \mathcal{S}_{\gamma+1,2}(1,\, \Xi)$, let us define

$$\begin{aligned}\mathbb{H}(f_0) := \{f \in \mathcal{S}_{\gamma+1,2}(1,\, \Xi) : &\left\lfloor \frac{D^p f(w_{1,i}, w_{2,j})}{\delta_k} \right\rfloor = \left\lfloor \frac{D^p f_0(w_{1,i}, w_{2,j})}{\delta_k} \right\rfloor, \\ &1 \leq i \leq s_1,\, 1 \leq j \leq s_2,\, p \text{ with } [p] = k \in \{0, \ldots, \gamma\}\}\end{aligned}$$

where $\lfloor x \rfloor$ means the largest integer smaller than or equal to x. Our earlier argument implies that the number of distinct sets $\mathbb{H}(f_0)$ with f_0 ranging over $\mathcal{S}_{\gamma+1,2}(1,\, \Xi)$ bounds the δ−covering number of $\mathcal{S}_{\gamma+1,2}(1,\, \Xi)$ from above. Note that for $i = 1, \ldots, s_1$, $j = 1, \ldots, s_2$, and p with $[p] = k \in \{0, \ldots, \gamma\}$, $\mathbb{H}(f_0)$ depends on $\left\lfloor \frac{D^p f_0(w_{1,i}, w_{2,j})}{\delta_k} \right\rfloor$ only and the number of distinct sets $\mathbb{H}(f_0)$ is bounded above by the cardinality of

$$\begin{aligned} I &= \{(\left\lfloor \frac{D^p f(w_{1,i}, w_{2,j})}{\delta_k} \right\rfloor, 1 \leq i \leq s_1,\, 1 \leq j \leq s_2,\, p \text{ with } [p] = k \in \{0, \ldots, \gamma\}) \\ &\quad : f \in \mathcal{S}_{\gamma+1,2}(1,\, \Xi)\}. \end{aligned} \tag{74}$$

Starting from $(w_{1,1}, w_{2,1})$, let us count the number of possible values of the vector

$$\left(\left\lfloor \frac{D^p f(w_{1,1}, w_{2,1})}{\delta_k} \right\rfloor, p \text{ with } [p] = k \in \{0, \ldots, \gamma\}\right) \tag{75}$$

with f ranging over $\mathcal{S}_{\gamma+1,2}(1,\, \Xi)$. Given a function of two variables has 2^k kth partial derivatives, this number is at most

$$\begin{aligned} \left(\frac{1}{\delta_0}\right)^{2^0}\left(\frac{1}{\delta_1}\right)^{2^1} \cdots \left(\frac{1}{\delta_\gamma}\right)^{2^\gamma} &\leq \left(\frac{5}{\delta}\right)^{\sum_{j=0}^{\gamma} 2^j - \sum_{j=0}^{\gamma} \frac{j}{\gamma+1} 2^j} \\ &= \left(\frac{5}{\delta}\right)^{2^{\gamma+1} - 1 - \frac{(\gamma-1)2^{\gamma+1}+2}{\gamma+1}} \\ &\leq \left(\frac{5}{\delta}\right)^{\frac{2^{\gamma+2}}{\gamma+1}}. \end{aligned} \tag{76}$$

The diagram below shows how we move from $(w_{1,1}, w_{2,1})$ to (w_{1,s_1}, w_{2,s_2}) to count the the number of possible values of $\left\lfloor \frac{D^p f(w_{1,i}, w_{2,j})}{\delta_k} \right\rfloor$ given that of its previous adjacent pair of points:

$$\begin{array}{ccc} (w_{1,1}, w_{2,1}) \rightarrow & (w_{1,2}, w_{2,1}) \rightarrow \cdots \rightarrow & (w_{1,s_1}, w_{2,1}) \\ & & \downarrow \\ (w_{1,1}, w_{2,2}) \leftarrow \cdots \leftarrow & (w_{1,s_1-1}, w_{2,2}) \leftarrow & (w_{1,s_1}, w_{2,2}) \\ \vdots & \vdots & \vdots \end{array}$$

Note that in each row j, we fix $w_{2,j}$ and cycle through all $w_{1,i}$s; moreover, every time we move from one pair of points to the next, we only change one coordinate and keep the other fixed. Because of this construction, we can argue in a similar way as in the proof for Lemma A.1 that,

$$\begin{aligned}
|I| &\leq 2^{s_1 s_2} \left(\frac{\delta}{5}\right)^{\frac{2^{\gamma+2}}{\gamma+1}} \\
&\leq 2^{\left[2\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+1}}+2\right]\left[4\overline{C}\left(\frac{\delta}{5}\right)^{\frac{-1}{\gamma+1}}+2\right]} \left(\frac{5}{\delta}\right)^{\frac{2^{\gamma+2}}{\gamma+1}} \\
&\leq \exp\left[\frac{2^{\gamma+2}}{\gamma+1}\log\frac{5}{\delta} + 20\left(\overline{C}\vee 1\right)\left(\frac{\delta}{5}\right)^{\frac{-2}{\gamma+1}}\log 2 + 4\log 2\right]
\end{aligned}$$

where I is defined in (74). □

A.5 Theorem 3.3

Preliminary

In terms of our least squares (41), the *basic inequality*

$$\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \hat{y}\left(x_i\right)\right)^2 \leq \frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - y\left(x_i\right)\right)^2$$

yields

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}(x_i) - y(x_i)\right)^2 \leq \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\left(\hat{y}(x_i) - y(x_i)\right). \tag{77}$$

To bound the right-hand-side of (77), it suffices to bound $\mathcal{G}_n(\tilde{r}_n;\, \bar{\mathcal{F}})$ defined in (20) with $\mathcal{F} = \mathcal{Y}$ in (21).

One way to bound $\mathcal{G}_n(\tilde{r}_n;\, \bar{\mathcal{F}})$ is to seek a sharp enough $\tilde{r}_n > 0$ that satisfies the *critical inequality*

$$\mathcal{G}_n\left(\tilde{r}_n;\, \bar{\mathcal{F}}\right) \precsim \frac{\tilde{r}_n^2}{\sigma}. \tag{78}$$

It is known that the complexity $\mathcal{G}_n\left(\tilde{r}_n;\, \bar{\mathcal{F}}\right)$ can be bounded above by the Dudley's entropy integral (see, e.g., [7, 9, 11]). Let $N_n(\delta,\, \Lambda(\tilde{r}_n;\, \bar{\mathcal{F}}))$ denote the $\delta-$covering number of the set $\Lambda(\tilde{r}_n;\, \bar{\mathcal{F}})$ in the $|\cdot|_n$ norm. Then by Corollary 13.7 in [11], the critical radius condition (78) holds for any $\tilde{r}_n \in (0,\, \sigma]$ such that

$$\frac{c}{\sqrt{n}}\int_{\frac{\tilde{r}_n^2}{4\sigma}}^{\tilde{r}_n}\sqrt{\log N_n(\delta,\, \Lambda(\tilde{r}_n;\, \bar{\mathcal{F}}))}d\delta \leq \frac{\tilde{r}_n^2}{\sigma}. \tag{79}$$

Proof. In what follows, we show Theorem 3.3(ii) as the argument for Theorem 3.3(i) is nearly identical; at the end, we show Theorems 3.3(iii)-(iv).

Claim (ii). Let us begin with the part $\mathcal{M}_1$. Since $\mathcal{Y} \subseteq \mathcal{S}^{\dagger}_{\beta+2,\overline{C}}$, we let $\mathcal{F} = \mathcal{S}^{\dagger}_{\beta+2,\overline{C}}$ in (21). Because we are working with $\bar{\mathcal{F}}$ in (21) in terms of $\mathcal{F} = \mathcal{S}^{\dagger}_{\beta+2,\overline{C}}$, all the coefficients $2^{k-1}\,(k-1)!$ associated with the derivatives for $k = 1, ..., \beta+1$, the coefficient $2^{\beta+1}\,(\beta+1)!$ associated with the Lipschitz condition, as well as the function value itself are multiplied by 2. Slight modifications of the proof for Lemma A.1 give

$$\begin{aligned}
|I| &\leq 3^s \left(\frac{\delta}{7}\right)^{-\frac{\gamma+3}{2}} \overline{C} 2^{\frac{(\gamma+1)\gamma}{2}+\gamma+2} \prod_{i=0}^{\gamma} i! \\
&\leq 9^{\left(\frac{\delta}{7}\right)^{\frac{-1}{\gamma+2}}+1} \left(\frac{\delta}{7}\right)^{-\frac{\gamma+3}{2}} \overline{C} 2^{\frac{(\gamma+1)\gamma}{2}+\gamma+2} \prod_{i=0}^{\gamma} i! \\
&\leq \exp\left[\log\left(\overline{C}\prod_{i=0}^{\gamma} i!\right) + \frac{\gamma^2+3\gamma}{2}\log 2 + \frac{\gamma+3}{2}\log\frac{7}{\delta} + \left(\frac{\delta}{7}\right)^{\frac{-1}{\gamma+2}}\log 9 + \log 36\right]
\end{aligned}$$

valid for all $\gamma = 0, ..., \beta$. Note that

$$\begin{aligned}
& \frac{1}{\sqrt{n}}\int_{\frac{\tilde{r}_n^2}{4\sigma}}^{\tilde{r}_n} \sqrt{\log N_n(\delta,\, \Lambda(\tilde{r}_n;\, \bar{\mathcal{F}}))}d\delta \\
\leq\ & \frac{1}{\sqrt{n}}\int_0^{\tilde{r}_n} \sqrt{\log N_\infty\left(\delta,\, \bar{\mathcal{F}}\right)}d\delta \\
\leq\ & \frac{\tilde{r}_n}{\sqrt{n}}\left[\log\left(\overline{C}\prod_{i=0}^{\gamma} i!\right) + \frac{\gamma^2+3\gamma}{2}\log 2\right]^{\frac{1}{2}} + \left(\frac{\gamma+3}{2n}\right)^{\frac{1}{2}}\int_0^{\tilde{r}_n}\sqrt{\log\frac{7}{\delta}}d\delta \\
& + \left(\frac{\log 9}{n}\right)^{\frac{1}{2}}\int_0^{\tilde{r}_n}\sqrt{\left(\frac{\delta}{7}\right)^{\frac{-1}{\gamma+2}}}d\delta + \tilde{r}_n\sqrt{\frac{\log 36}{n}} \\
\leq\ & \underbrace{c_0\left\{\tilde{r}_n\sqrt{\frac{1}{n}\log\left(\prod_{i=0}^{\gamma} i!\right)} + \tilde{r}_n\sqrt{\frac{\gamma^2}{n}} + \tilde{r}_n\sqrt{\frac{1}{n}} + \frac{1}{\sqrt{n}}\tilde{r}_n^{\frac{2\gamma+3}{2\gamma+4}}\right\}}_{\mathcal{T}_1(\gamma,\tilde{r}_n)}
\end{aligned}$$

valid for all $\gamma = 0, ..., \beta$. Therefore, we can take $\min_{\gamma\in\{0,...,\beta\}} \mathcal{T}_1(\gamma, \tilde{r}_n) = \mathcal{B}_1(\tilde{r}_n)$ and let γ_1^* be the minimizer. Setting $\mathcal{B}_1(\tilde{r}_n) \asymp \frac{\tilde{r}_n^2}{\sigma}$ yields $\mathcal{M}_1(\gamma_1^*)$.

We now show the part $\mathcal{M}_2$. For a given $\delta > 0$, let us consider the smallest $\frac{\delta}{2L_{\max}}$−covering $\left\{f^1, ..., f^N\right\}$ (w.r.t. the sup-norm) of $\mathcal{S}_{\beta+1,\,2}(1,\, \Xi)$ and the smallest $\frac{\delta}{2L_{\max}}$−covering $\left\{y_{0,1}, ..., y_{0,N'}\right\}$ for the interval $[-C_0,\, C_0]$ where the initial value lies. By Theorem B.1 in Section B and arguments similar to those in the proof for Corollary 2.1, for any $f,\, \tilde{f} \in \mathcal{S}_{\beta+1,\,2}(1,\, \Xi)$ and initial values $y_0,\, \tilde{y}_0 \in [-C_0,\, C_0]$, we can find some $f^i,\, f^j \in \left\{f^1, ..., f^N\right\}$ and $y_{0,i'},\, y_{0,j'} \in \left\{y_{0,1}, ..., y_{0,N'}\right\}$ such that

$$\begin{aligned}
& \left|y(x) - \tilde{y}(x) - \left(y_{(i,i')}(x) - y_{(j,j')}(x)\right)\right| \\
\leq\ & \left|y(x) - y_{(i,i')}(x)\right| + \left|\tilde{y}(x) - y_{(j,j')}(x)\right| \\
\leq\ & \delta
\end{aligned}$$

where y, $\tilde{y}$, $y_{(i,i')}$, and $y_{(j,j')}$ are solutions to the ODE associated with $\{f,\, y_0\}$, $\left\{\tilde{f},\, \tilde{y}_0\right\}$, $\left\{f^i,\, y_{0,i'}\right\}$ and $\left\{f^j,\, y_{0,j'}\right\}$, respectively. Thus, we obtain a δ−cover of $\bar{\mathcal{F}}$ in terms of $\mathcal{F} = \mathcal{Y}$ in (21). The rest of arguments are very similar to those for $\mathcal{M}_1$. For $\beta > 0$, we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\int_0^{\tilde{r}_n}\sqrt{\log N_n(\delta,\,\Lambda(\tilde{r}_n;\,\bar{\mathcal{F}}))}d\delta &\leq \frac{1}{\sqrt{n}}\int_0^{\tilde{r}_n}\sqrt{\log N_\infty\left(\delta,\,\bar{\mathcal{F}}\right)}d\delta \\ &\leq \underbrace{c_1\tilde{r}_n\sqrt{\frac{2^{\gamma+3}}{n\left(\gamma+2\right)}}+\frac{1}{\sqrt{n}}\tilde{r}_n^{\frac{\gamma}{\gamma+1}}}_{\mathcal{T}_2(\gamma,\tilde{r}_n)}\end{aligned}$$

valid for all $\gamma = 0, ..., \beta$. In the second inequality, we have used (64). Therefore, we can take $\min_{\gamma\in\{0,...,\beta\}}\mathcal{T}_2\left(\gamma,\tilde{r}_n\right) = \mathcal{B}_2\left(\tilde{r}_n\right)$ and let γ_2^* be the minimizer. Setting $\mathcal{B}_2\left(\tilde{r}_n\right) \asymp \frac{\tilde{r}_n^2}{\sigma}$ yields $\mathcal{M}_2\left(\gamma_2^*\right)$.

Now, we can take $\tilde{r}_n^2 = \min\left\{\mathcal{M}_1\left(\gamma_1^*\right), \mathcal{M}_2\left(\gamma_2^*\right)\right\}$. By Theorem 13.5 in [11], in terms of (41), we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}(x_i) - y(x_i)\right)^2 \precsim \tilde{r}_n^2$$

with probability at least $1 - \exp\left(\frac{-n\tilde{r}_n^2}{2\sigma^2}\right)$. Integrating the tail bound yields

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}(x_i) - y(x_i)\right)^2\right] \precsim \tilde{r}_n^2 + c_1^{'}\exp\left\{-c_2^{'}n\sigma^{-2}\tilde{r}_n^2\right\}.$$

Claim (i). Following almost identical arguments as above, we have, for the autonomous systems,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}(x_i) - y(x_i)\right)^2 \precsim \tilde{r}_{n,a}^2$$

with probability at least $1 - \exp\left(\frac{-n\tilde{r}_{n,a}^2}{2\sigma^2}\right)$. Integrating the tail bound yields

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}(x_i) - y(x_i)\right)^2\right] \precsim \tilde{r}_{n,a}^2 + c_1^{'}\exp\left\{-c_2^{'}n\sigma^{-2}\tilde{r}_{n,a}^2\right\}.$$

Claim (iii). Note that if $\frac{n}{\sigma^2} \succsim \left(\gamma\sqrt{\log\left(\gamma\vee 1\right)}\right)^{4(\gamma+2)+2}$ where $\gamma \in \{0, ..., \beta\}$, we have

$$\begin{aligned}\mathcal{M}_1^a\left(\gamma\right) &\asymp \left(\frac{\sigma^2}{n}\right)^{\frac{2(\gamma+2)}{2(\gamma+2)+1}}, \quad \text{and} \quad \tilde{r}_{n,a}^2 = \mathcal{M}_1^a\left(\gamma\right), \\ \mathcal{M}_1\left(\gamma\right) &\asymp \left(\frac{\sigma^2}{n}\right)^{\frac{2(\gamma+2)}{2(\gamma+2)+1}}, \quad \text{and} \quad \tilde{r}_n^2 = \mathcal{M}_1\left(\gamma\right).\end{aligned} \tag{80}$$

Claim (iv). Note that every member in $\mathcal{S}_{\beta+2}$ (the standard smooth class of degree $\beta+2$) can be expressed as a solution to the ODE (38) and clearly, $\mathcal{S}_{\beta+2} \subseteq \mathcal{Y}$ where $\mathcal{Y}$ is the class of solutions to (30) with $f \in \mathcal{S}_{\beta+1,\,2}\left(1,\,\Xi\right)$. As we have discussed in Section 3.1, by the existing minimax results on $\mathcal{S}_{\beta+2}$ (see, e.g., [11], Chapter 15), the minimax lower bound has a scaling $\left(\frac{\sigma^2}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$. When $\frac{n}{\sigma^2} \succsim \left(\beta\sqrt{\log\left(\beta\vee 1\right)}\right)^{4(\beta+2)+2}$, we have (80) with $\gamma = \beta$ and the rate $\left(\frac{\sigma^2}{n}\right)^{\frac{2(\beta+2)}{2(\beta+2)+1}}$ is clearly minimax optimal for the nonautonomous systems. □

A.6 Theorem 3.4

Preliminary

An alternative approach for bounding $\mathcal{G}_n(\tilde{r}_n;\, \bar{\mathcal{F}})$ is developed by [8]. This approach can be formulated as follows.

Given a radius $\tilde{r}_n > 0$ and a function class $\bar{\mathcal{F}}$, define

$$\mathcal{G}_n(\tilde{r}_n;\, \bar{\mathcal{F}}) := \mathbb{E}_\varepsilon \left[\sup_{h \in \Lambda(\tilde{r}_n;\, \bar{\mathcal{F}})} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(x_i) \right| \right], \tag{81}$$

where $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}\,(0,\, 1)$, $\varepsilon = \{\varepsilon_i\}_{i=1}^n$, and

$$\Lambda(\tilde{r}_n;\, \bar{\mathcal{F}}) = \left\{ h \in \bar{\mathcal{F}} :\, |h|_n \leq \tilde{r}_n,\, |h|_{\mathcal{H}} \leq 1 \right\}$$

where $\bar{\mathcal{F}}$ is defined in (21) and $|\cdot|_{\mathcal{H}}$ is a norm associated with some underlying RKHS. Let $\mathcal{K}$ be the kernel function of this RKHS and $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq 0$ be the eigenvalues of the matrix $\mathbb{K}$ consisting of entries $K_{ij} = \frac{1}{n}\mathcal{K}\,(x_i,\, x_j)$. Introduce the critical radius condition

$$\mathcal{G}_n\left(\tilde{r}_n;\, \bar{\mathcal{F}}\right) \precsim \mathcal{R}\frac{\tilde{r}_n^2}{\sigma}, \tag{82}$$

where $\mathcal{R}$ is the radius of the RKHS of interest. For any $\tilde{r}_n > 0$, one has

$$\mathcal{G}_n(\tilde{r}_n;\, \bar{\mathcal{F}}) \precsim \sqrt{\frac{1}{n}} \sqrt{\sum_{i=1}^{n} \left(\tilde{r}_n^2 \wedge \mu_i\right)} \tag{83}$$

(see [8]). As a consequence, the critical radius condition (83) is satisfied for any $\tilde{r}_n > 0$ such that

$$\sqrt{\frac{1}{n}} \sqrt{\sum_{i=1}^{n} \left(\tilde{r}_n^2 \wedge \mu_i\right)} \precsim \frac{\mathcal{R}\tilde{r}_n^2}{\sigma}.$$

Therefore, in terms of

$$\hat{g} \in \arg \min_{\tilde{g} \in \mathcal{F},\, |\tilde{g}|_{\mathcal{H}} \leq \mathcal{R}} \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - \tilde{g}\,(x_i)\right)^2,$$

one has

$$\frac{1}{n} \sum_{i=1}^{n} \left(\hat{g}(x_i) - g^*(x_i)\right)^2 \precsim \mathcal{R}^2 \tilde{r}_n^2$$

with probability at least $1 - c^{'} \exp\left(\frac{-c^{''} n \mathcal{R}^2 \tilde{r}_n^2}{\sigma^2}\right)$.

Proof. In what follows, we argue for the nonautonomous systems as the argument for the autonomous systems is nearly identical.

Given that $\mathcal{Y} \subseteq \mathcal{S}^{\dagger}_{\beta+2,\overline{C}}$, $\mathcal{Y}$ is contained in $\mathcal{H}^{\dagger}_{\beta+2}$, defined in (46). Clearly, we have

$$\begin{aligned} \mathcal{H}^{\dagger}_{\beta+2} \subseteq \mathcal{H}_{\gamma+2} := \{ h :\, [0,\, 1] \to \mathbb{R} | h^{(\gamma+1)} \text{ is abs. cont. with} \\ \int_0^1 \left[h^{(\gamma+2)}\,(t)\right]^2 dt \leq \left[2^{\gamma+1}\,(\gamma+1)!\right]^2 \} \end{aligned}$$

for all $\gamma \in \{0, ..., \beta\}$, and

$$\mathcal{H}_{\gamma+2} = \mathcal{H}_{\gamma+2,1} + \mathcal{H}_{\gamma+2,2} \tag{84}$$

where $\mathcal{H}_{\gamma+2,1}$ is an RKHS of polynomials of degree $\gamma + 1$ and

$$\mathcal{H}_{\gamma+2,2} := \{h : [0,\, 1] \to \mathbb{R} | h^{(k)}(0) = 0 \text{ for all } k \leq \gamma,$$
$$h^{(\gamma+1)} \text{ is abs. cont. with } \int_0^1 \left[h^{(\gamma+2)}(t)\right]^2 dt \leq \left[2^{\gamma+1}(\gamma+1)!\right]^2\}.$$

As a result, we can equip $\mathcal{H}_{\gamma+2}$ with the following norm

$$|h|_{\mathcal{H}} = \left[\sum_{k=0}^{\gamma+1}\left(h^{(k)}(0)\right)^2 + \int_0^1 \left[h^{(\gamma+2)}(t)\right]^2 dt\right]^{\frac{1}{2}}.$$

Note that any $h \in \mathcal{Y}$ has

$$|h|_{\mathcal{H}} \leq c\left[2^{\gamma+1}(\gamma+1)!\right]^2$$

for some positive universal constant c so we may let $\mathcal{R}_\gamma := c\left[2^{\gamma+1}(\gamma+1)!\right]^2$ be the radius associated with $\mathcal{H}_{\gamma+2}$. Because of (84), we can generate $\mathcal{H}_{\gamma+2}$ with the kernel

$$\mathcal{K}\left(x,\, x'\right) = \sum_{k=0}^{\gamma+1} \frac{x^k}{k!}\frac{x'^k}{k!} + \int_0^1 \frac{(x-t)_+^{\gamma+1}}{(\gamma+1)!}\frac{\left(x'-t\right)_+^{\gamma+1}}{(\gamma+1)!}dt$$

where $(w)_+ = w \vee 0$. As a consequence, we have

$$\mathcal{G}_n(\tilde{r}_n;\, \bar{\mathcal{F}}) \precsim \underbrace{\max\left\{\tilde{r}_n\sqrt{\frac{(\gamma \wedge n) \vee 1}{n}},\, \frac{1}{\sqrt{n}}\tilde{r}_n^{\frac{2\gamma+3}{2\gamma+4}}\right\}}_{\mathcal{T}_3(\gamma, \tilde{r}_n)}$$

valid for all $\gamma \in \{0, ..., \beta\}$. Setting $\mathcal{T}_3(\gamma, \tilde{r}_n) \asymp \left[2^{\gamma+1}(\gamma+1)!\right]\tilde{r}_n^2\sigma^{-1}$ yields

$$\mathcal{R}_\gamma^2\tilde{r}_n^2 \asymp \max\left\{\frac{\sigma^2\left((\gamma \wedge n) \vee 1\right)}{n},\, \left[2^{\gamma+1}(\gamma+1)!\right]^{\frac{2}{2(\gamma+2)+1}}\left(\frac{\sigma^2}{n}\right)^{\frac{2(\gamma+2)}{2(\gamma+2)+1}}\right\}$$

valid for all $\gamma \in \{0, ..., \beta\}$. Taking the minimum of the RHS over all $\gamma \in \{0, ..., \beta\}$ yields $\bar{r}_n^2$ in Theorem 3.4. Hence,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}(x_i) - y^*(x_i)\right)^2 \precsim \bar{r}_n^2$$

with probability at least $1 - c'\exp\left(\frac{-c'' n\bar{r}_n^2}{\sigma^2}\right)$. Integrating the tail bound yields the claim in Theorem 3.4. □

B Supporting lemmas and proofs

Theorem B.1 (Gronwall inequality for first order ODEs). *Consider the following pair of ODEs:*

$$y^{'}(x) = f(x, y(x)), \quad y(0) = y_0,$$

and

$$z^{'}(x) = g(x, z(x)), \quad z(0) = z_0,$$

with $|y_0|, |z_0| \leq C_0$, *and* $(x, y(x)), (x, z(x)) \in \bar{\Gamma} := [0, 1] \times [-C_0 - b, C_0 + b]$. *Suppose* f *and* g *are continuous on* $\bar{\Gamma}$*; for all* $(x, y), (x, \tilde{y}) \in \bar{\Gamma}$,

$$|f(x, y) - f(x, \tilde{y})| \leq L |y - \tilde{y}|. \tag{85}$$

Assume there is a continuous function $\varphi : [0, a] \mapsto [0, \infty)$ *such that*

$$|f(x, y(x)) - g(x, y(x))| \leq \varphi(x). \tag{86}$$

Then we have

$$|y(x) - z(x)| \leq \exp(Lx) \int_0^x \exp(-Ls) \varphi(s) \, ds + \exp(Lx) |y_0 - z_0|$$

for $x \in \left[0, \min\left\{a, \frac{b}{M}\right\}\right]$ *where* $M = \max\left\{\max_{(x, y) \in \bar{\Gamma}} |f(x, y)|, \max_{(x, z) \in \bar{\Gamma}} |g(x, z)|\right\}$.

Remark. Theorem B.1 is a slight modification of Theorem 2.1 in [4], which gives a variant of the Gronwall inequality [3].

In the following result, we extend Theorem B.1 to higher order ODEs. Let

$$Y(x) = \begin{bmatrix} y(x) \\ y^{'}(x) \\ \vdots \\ y^{(m-1)}(x) \end{bmatrix} \quad \text{and} \quad Z(x) = \begin{bmatrix} z(x) \\ z^{'}(x) \\ \vdots \\ z^{(m-1)}(x) \end{bmatrix}$$

with

$$Y_0 := \begin{bmatrix} y_{(0)} \\ y_{(1)} \\ \vdots \\ y_{(m-1)} \end{bmatrix} \quad \text{and} \quad Z_0 := \begin{bmatrix} z_{(0)} \\ z_{(1)} \\ \vdots \\ z_{(m-1)} \end{bmatrix}.$$

In addition, let

$$\bar{\Gamma} := \{(x, Y) : x \in [a_0, a_0 + a], |Y|_2 \leq b + C_0\}.$$

Theorem B.2 (Gronwall inequality for higher order ODEs). *Consider the following pair of ODEs:*

$$y^{(m)}(x) = f\left(x, y(x), y^{'}(x), ..., y^{(m-1)}(x)\right),$$
$$y(a_0) = y_{(0)}, y^{'}(a_0) = y_{(1)}, ..., y^{(m-1)}(a_0) = y_{(m-1)}, \tag{87}$$

and

$$z^{(m)}(x) = g\left(x,\, z(x),\, z^{'}(x),\, ...,\, z^{(m-1)}(x)\right),$$
$$z(a_0) = z_{(0)},\, z^{'}(a_0) = z_{(1)},\, ...,\, z^{(m-1)}(a_0) = z_{(m-1)}, \tag{88}$$

with $|Y_0|_2\,,\ |Z_0|_2 \le C_0$ *and* $(x,\, Y(x))\,,\ (x,\, Z(x)) \in \bar{\Gamma}$. *Suppose* f *and* g *are continuous on* $\bar{\Gamma}$*; and*

$$\left|f(x,\, Y) - f\left(x,\, \tilde{Y}\right)\right| \le L\left|Y - \tilde{Y}\right|_2 \tag{89}$$

for all $(x,\, Y) := \left(x,\, y,\, ...,\, y^{(m-1)}\right)$ *and* $\left(x,\, \tilde{Y}\right) := \left(x,\, \tilde{y},\, ...,\, \tilde{y}^{(m-1)}\right)$ *in* $\bar{\Gamma}$. *Assume there is a continuous function* $\varphi : [a_0,\, a_0 + a] \mapsto [0,\, \infty)$ *such that*

$$|f(x,\, Y(x)) - g(x,\, Y(x))| \le \varphi(x)\,. \tag{90}$$

Then we have

$$\left|y^{(k)}(x) - z^{(k)}(x)\right| \le \exp\left(x\sqrt{L^2+1}\right)\int_{a_0}^{x} \exp\left(-s\sqrt{L^2+1}\right)\varphi(s)\,ds + \exp\left(x\sqrt{L^2+1}\right)|Y_0 - Z_0|_2$$

for all $k = 0, ..., m-1$ *and all* $x \in \left[a_0,\, a_0 + \min\left\{a,\, \frac{b}{M}\right\}\right]$ *where*

$$M = \max\left\{\max_{(x,\, Y)\in\bar{\Gamma}} |f(x,\, Y)|\,,\ \max_{(x,\, Z)\in\bar{\Gamma}} |g(x,\, Z)|\right\}.$$

Proof. The following arguments are based on and extend those in [4]. Let $W = (w_j)_{j=0}^{m-1}$ and

$$F(x,\, W) := \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{m-1} \\ f(x,\, W) \end{bmatrix}.$$

We transform (87) and (88) into

$$\begin{aligned} Y^{'}(x) &= F(x,\, Y(x))\,, \\ Z^{'}(x) &= G(x,\, Z(x))\,, \end{aligned}$$

where

$$F(x,\, Y(x)) := \begin{bmatrix} y^{'}(x) \\ y^{(2)}(x) \\ \vdots \\ y^{(m-1)}(x) \\ f\left(x,\, y(x)\,,\, y^{'}(x)\,,\, ...,\, y^{(m-1)}(x)\right) \end{bmatrix}$$

and

$$G(x,\, Z(x)) := \begin{bmatrix} z^{'}(x) \\ z^{(2)}(x) \\ \vdots \\ z^{(m-1)}(x) \\ g\left(x,\, z(x)\,,\, z^{'}(x)\,,\, ...,\, z^{(m-1)}(x)\right) \end{bmatrix}.$$

The inequality $\frac{d}{ds}\left|Y(s)\right|_2 \le \left|Y'(s)\right|_2$, (89) and (90) yield

$$\begin{aligned}
\frac{d}{ds}\left|Y(s)-Z(s)\right|_2 &\le \left|Y'(s)-Z'(s)\right|_2 \\
&= \left|F(s,\, Y(s))-G(s,\, Z(s))\right|_2 \\
&\le \left|G(s,\, Z(s))-F(s,\, Z(s))\right|_2+\left|F(s,\, Z(s))-F(s,\, Y(s))\right|_2 \\
&\le \varphi(s)+\left[L^2\left|Y(s)-Z(s)\right|_2^2+\sum_{k=1}^{m-1}\left(y^{(k)}(s)-z^{(k)}(s)\right)^2\right]^{\frac{1}{2}} \\
&\le \varphi(s)+\sqrt{L^2+1}\left|Y(s)-Z(s)\right|_2,
\end{aligned}$$

which is equivalent to

$$\frac{d}{ds}\left|Y(s)-Z(s)\right|_2-\sqrt{L^2+1}\left|Y(s)-Z(s)\right|_2 \le \varphi(s).$$

Multiplying both sides above by $\exp\left(-s\sqrt{L^2+1}\right)$ gives

$$\frac{d}{ds}\left[\exp\left(-s\sqrt{L^2+1}\right)\left|Y(s)-Z(s)\right|_2\right] \le \exp\left(-s\sqrt{L^2+1}\right)\varphi(s).$$

By the Cauchy-Peano existence theorem, there exist solutions y and z to (87) and (88), respectively, on $\left[a_0,\, a_0+\min\left\{a,\, \frac{b}{M}\right\}\right]$. Then integrating both sides of the inequality above from a_0 to x gives

$$\exp\left(-x\sqrt{L^2+1}\right)\left|Y(x)-Z(x)\right|_2-\left|Y_0-Z_0\right|_2 \le \int_{a_0}^{x}\exp\left(-s\sqrt{L^2+1}\right)\varphi(s)\,ds$$

for all $x \in \left[a_0,\, a_0+\min\left\{a,\, \frac{b}{M}\right\}\right]$. The above implies that

$$\left|y^{(k)}(x)-z^{(k)}(x)\right| \le \exp\left(x\sqrt{L^2+1}\right)\int_{a_0}^{x}\exp\left(-s\sqrt{L^2+1}\right)\varphi(s)\,ds+\exp\left(x\sqrt{L^2+1}\right)\left|Y_0-Z_0\right|_2$$

for all $x \in \left[a_0,\, a_0+\min\left\{a,\, \frac{b}{M}\right\}\right]$ and all $k=0,...,m-1$. □

C Additional results and proofs

C.1 Proposition C.1

Proposition C.1. *Suppose f in (13) is continuous on $[0,\,1]\times[-C_0-b,\, C_0+b]$, $|f(x,\, y;\, \theta)| \le 1$ and*

$$|f(x,\, y;\, \theta)-f(x,\, \tilde{y};\, \theta)| \le |y-\tilde{y}| \tag{91}$$

for all $(x,\, y)$ and $(x,\, \tilde{y})$ in $[0,\,1]\times[-C_0-b,\, C_0+b]$, and $\theta \in \mathbb{B}_q(1)$ with $q \in [1,\,\infty]$; moreover,

$$\left|f(x,\, y;\, \theta)-f\left(x,\, y;\, \theta'\right)\right| \le L_K\left|\theta-\theta'\right|_q, \tag{92}$$

for all $(x,\, y) \in [0,\,1]\times[-C_0-b,\, C_0+b]$ and $\theta,\, \theta' \in \mathbb{B}_q(1)$. Let us consider (1) where $y(\cdot)$ is the (unique) solution to (13) on $[0,\,\alpha)$ where $\alpha=\min\{1,\, b\}$. Suppose we have n design points sampled *from the interval* $[0,\, \tilde{\alpha}]$ *with* $\tilde{\alpha} < \alpha$*. Letting* $B_K = (L_K \vee 1)$*, if*

$$K \log\left(1 + \frac{2L_{\max}L_K}{\delta}\right) \gtrsim \log\left(\frac{2C_0 L_{\max}}{\delta} + 1\right),\ \forall \delta \precsim B_K \sigma \sqrt{\frac{K}{n}} \tag{93}$$

$$\max\{\alpha,\, C_0\} \geq c_0 B_K \sigma \sqrt{\frac{K}{n}}, \tag{94}$$

for a sufficiently large positive universal constant c_0*, then we have*

$$\frac{1}{n}\sum_{i=1}^{n}\left(y^*(x_i;\,\hat{\theta},\,\hat{y}_0) - y^*(x_i;\,\theta^*,\,y_0^*)\right)^2 \precsim B_K^2 \frac{\sigma^2 K}{n} \tag{95}$$

with probability at least $1 - c_1 \exp\left(-c_2 B_K^2 K\right)$*, where* $\left(\hat{\theta},\,\hat{y}_0\right)$ *is a solution to (14).*

Remark. Condition (93) simply restricts C_0 from being too large, and as a consequence, (11) implies that $\log N_\infty(\delta,\,\mathcal{Y}) \precsim K\log\left(1 + \frac{2L_{\max}L_K}{\delta}\right)$. This upper bound implies that $\mathcal{Y}$ is no "larger" than the class of fs parameterized by $\theta \in \mathbb{B}_q(1)$.

Remark. Condition (94) in Proposition C.1 simply excludes the trivial case where b and C_0 are "too small". Without such a condition, as long as f is bounded from above, we would simply replace (95) with

$$\left[\frac{1}{n}\sum_{i=1}^{n}\left(y(x_i;\,\hat{\theta},\,\hat{y}_0) - y(x_i;\,\theta^*,\,y_0)\right)^2\right]^{\frac{1}{2}} \leq c_1 \min\left\{B_K\sigma\sqrt{\frac{K}{n}},\, \max\{\alpha,\, C_0\}\right\}.$$

Proof. Let $\mathcal{F} = \mathcal{Y}$ in (21). We have

$$\begin{aligned}
\frac{1}{\sqrt{n}}\int_0^{\tilde{r}_n}\sqrt{\log N_n(\delta,\,\Lambda(\tilde{r}_n;\,\bar{\mathcal{F}}))}d\delta &\leq \frac{1}{\sqrt{n}}\int_0^{\tilde{r}_n}\sqrt{\log N_\infty(\delta,\,\Lambda(\tilde{r}_n;\,\bar{\mathcal{F}}))}d\delta \\
&\leq \sqrt{\frac{K}{n}}\int_0^{\tilde{r}_n}\sqrt{\log\left(1 + \frac{c\tilde{r}_n(L_K \vee 1)}{\delta}\right)}d\delta \\
&= \tilde{r}_n(L_K \vee 1)\sqrt{\frac{K}{n}}\int_0^{\frac{1}{L_K \vee 1}}\sqrt{\log\left(1 + \frac{c}{t}\right)}dt \\
&\leq (L_K \vee 1)\,\tilde{r}_n\sqrt{\frac{K}{n}}
\end{aligned} \tag{96}$$

where we have applied a change of variable $t = \frac{\delta}{\tilde{r}_n(L_K \vee 1)}$ in the third line. Setting $B_K\tilde{r}_n\sqrt{\frac{K}{n}} \asymp \frac{\tilde{r}_n^2}{\sigma}$ yields $\tilde{r}_n \asymp B_K\sigma\sqrt{\frac{K}{n}}$, where $B_K = (L_K \vee 1)$. By Theorem 13.5 in [11], we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\left(y(x_i;\,\hat{\theta},\,\hat{y}_0) - y(x_i;\,\theta^*,\,y_0)\right)^2 \precsim B_K^2\frac{\sigma^2 K}{n}.$$

with probability at least $1 - c_1 \exp\left\{-c_2 B_K^2 K\right\}$. $\square$

C.2 Proposition C.2

Proposition C.2. *Suppose the conditions in Proposition C.1 hold. In terms of* $\hat{y}_{R+1}\left(x_i;\, \hat{\theta},\, \hat{y}_0\right)$ *where* $\hat{\theta}$ *is obtained from solving (17), if*

$$\max\{\tilde{\alpha},\, C_0\} \geq c_0 \sigma \tilde{b}\sqrt{\frac{K}{n}} \tag{97}$$

for a sufficiently large positive universal constant c_0*, then we have*

$$\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{y}_{R+1}(x_i;\, \hat{\theta},\, \hat{y}_0) - y(x_i;\, \theta^*,\, y_0)\right]^2\right\}^{\frac{1}{2}} \precsim \sigma\left(\tilde{b}\sqrt{\frac{K}{n}} + \frac{1}{1-\tilde{\alpha}}\left|\hat{y}_0 - y_0\right| + \frac{\tilde{\alpha}^{R+1}}{1-\tilde{\alpha}}\max\{C_0,\, \tilde{\alpha}\}\right) \tag{98}$$

with probability at least $1 - c_1\exp(-c_2 n) - c_3\exp\left(-c_4 K\tilde{b}^2\right)$*, where* $\tilde{b} = \left(\frac{\tilde{\alpha}L_K}{1-\tilde{\alpha}} \vee 1\right)$.

Remark. If the integral in (15) is hard to compute analytically, numerical integration can be used in (16). This would introduce an additional approximation error $\sigma(R+1)\cdot Err$, where Err is an upper bound on the error incurred in each iteration of (16), depending on the smoothness of f and which numerical method is used. For example, if f is twice differentiable with bounded first and second derivatives, and the integral is approximated with the midpoint rule with T slices, then $Err \precsim T^{-2}$.

Remark. The bound (98) reflects three sources of errors: $\tilde{b}\sqrt{\frac{K}{n}}$ is due to the estimation error in $\hat{\theta}$, $\frac{1}{1-\tilde{\alpha}}\left|\hat{y}_0 - y_0\right|$ is due to the estimation error in $\hat{y}_0$, and $\frac{\tilde{\alpha}^{R+1}}{1-\tilde{\alpha}}\max\{C_0,\, \tilde{\alpha}\}$ is due to the error from the finite $(R+1)$ Picard iterations.

Proof. In what follows, we suppress the dependence of $\hat{y}_{R+1}(x_i;\, \hat{\theta},\, \hat{y}_0)$ on $\hat{y}_0$ ($y(x_i;\, \theta^*,\, y_0)$ on y_0) and simply write $\hat{y}_{R+1}(x_i, \hat{\theta})$ (respectively, $y(x_i, \theta^*)$). By (77), we need to bound

$$\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left(\hat{y}_{R+1}(x_i, \hat{\theta}) - y(x_i, \theta^*)\right)\right|.$$

We can write $\hat{y}_{R+1}(x_i, \hat{\theta}) - y(x_i, \theta^*) = \sum_{j=1}^{3} T_j(x_i)$ where

$$\begin{aligned}
T_1(x_i) &= \hat{y}_{R+1}(x_i;\, \hat{\theta}) - \hat{y}_{R+1}(x_i;\, \theta^*), \quad \text{estimation error due to } \hat{\theta} \\
T_2(x_i) &= \hat{y}_{R+1}(x_i;\, \theta^*) - y_{R+1}(x_i;\, \theta^*), \quad \text{estimation error due to } \hat{y}_0 \\
T_3(x_i) &= y_{R+1}(x_i;\, \theta^*) - y(x_i;\, \theta^*), \quad \text{estimation error due to the finite iterations,}
\end{aligned}$$

where

$$\begin{aligned}
y_0 &= y_0, \\
y_1(x;\, \theta) &= y_0^* + \int_0^x f(s,\, y_0;\theta)\, ds, \\
y_2(x;\, \theta) &= y_0^* + \int_0^x f(s,\, y_1(s;\, \theta);\theta)\, ds, \\
&\vdots \\
y_{R+1}(x;\, \theta) &= y_0 + \int_0^x f(s,\, y_R(s;\, \theta);\theta)\, ds.
\end{aligned}$$

As a result, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left(\hat{y}_{R+1}(x_i,\hat{\theta})-y(x_i,\theta^*)\right)\right| \precsim \left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i T_1\left(x_i\right)\right| + \sigma\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[T_2\left(x_i\right)\right]^2} + \sigma\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[T_3\left(x_i\right)\right]^2} \quad (99)$$

with probability at least $1-c_1\exp\left(-c_2 n\right)$.

We first analyze $\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i T_1\left(x_i\right)\right|$ in (99). In (21), let

$$\mathcal{F}=\left\{g_\theta\left(x\right)=\hat{y}_{R+1}\left(x;\,\theta\right):\ \theta\in\mathbb{B}_q\left(1\right),\,x\in[0,\,\tilde{\alpha}]\right\} \quad (100)$$

where $\hat{y}_{R+1}\left(s;\,\theta\right)$ is constructed in the following fashion:

$$\begin{aligned}
\hat{y}_0 &= \hat{y}_0,\\
\hat{y}_1\left(x;\,\theta\right) &= \hat{y}_0+\int_0^x f\left(s,\,\hat{y}_0;\theta\right)ds,\\
\hat{y}_2\left(x;\,\theta\right) &= \hat{y}_0+\int_0^x f\left(s,\,\hat{y}_1\left(s;\,\theta\right);\theta\right)ds,\\
&\vdots\\
\hat{y}_{R+1}\left(x;\,\theta\right) &= \hat{y}_0+\int_0^x f\left(s,\,\hat{y}_R\left(s;\,\theta\right);\theta\right)ds.
\end{aligned}$$

For any $\theta,\,\theta^{'}\in\mathbb{B}_q\left(1\right)$, at the beginning, we have

$$\left|\hat{y}_1\left(x;\theta\right)-\hat{y}_1\left(x;\theta^{'}\right)\right|\leq\tilde{\alpha}L_K\left|\theta-\theta^{'}\right|_q \quad \forall x\in[0,\,\tilde{\alpha}] \quad (101)$$

where the inequality follows from (10). For the second iteration, we have

$$\begin{aligned}
\left|\hat{y}_2\left(x;\,\theta\right)-\hat{y}_2\left(x;\,\theta^{'}\right)\right| &\leq \left|\int_0^x f\left(s,\,\hat{y}_1\left(s;\,\theta\right);\theta\right)ds-\int_0^x f\left(s,\,\hat{y}_1\left(s;\,\theta^{'}\right);\theta\right)ds\right|+\\
&\quad \left|\int_0^x f\left(s,\,\hat{y}_1\left(s;\,\theta^{'}\right);\theta\right)ds-\int_0^x f\left(s,\,\hat{y}_1\left(s;\,\theta^{'}\right);\theta^{'}\right)ds\right|\\
&\leq \underbrace{\tilde{\alpha}\left(\tilde{\alpha}L_K\left|\theta-\theta^{'}\right|_q\right)}_{(i)}+\underbrace{\tilde{\alpha}L_K\left|\theta-\theta^{'}\right|_q}_{(ii)}\\
&\leq \tilde{\alpha}^2L_K\left|\theta-\theta^{'}\right|_q+\tilde{\alpha}L_K\left|\theta-\theta^{'}\right|_q \quad \forall x\in[0,\,\tilde{\alpha}]
\end{aligned}$$

where (i) and (ii) in the second inequality follow from (91) with (101) and (10), respectively. Continuing with this pattern until the $(R+1)$th iteration, we obtain

$$\begin{aligned}
\left|\hat{y}_{R+1}\left(x;\,\theta\right)-\hat{y}_{R+1}\left(x;\,\theta^{'}\right)\right| &\leq \left(L_K\left|\theta-\theta^{'}\right|_q\right)\sum_{i=1}^{R+1}\tilde{\alpha}^i\\
&\leq \frac{\tilde{\alpha}L_K}{1-\tilde{\alpha}}\left|\theta-\theta^{'}\right|_q \quad \forall x\in[0,\,\tilde{\alpha}]\,. \quad (102)
\end{aligned}$$

In particular, (102) holds for $x\in\{x_1,\,x_2,\,...,\,x_n\}$. Consequently,

$$\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{y}_{R+1}(x_i;\,\theta)-\hat{y}_{R+1}(x_i;\,\theta^{'})\right]^2\right\}^{\frac{1}{2}}\leq\frac{\tilde{\alpha}L_K}{1-\tilde{\alpha}}\left|\theta-\theta^{'}\right|_q.$$

Let $\tilde{b}=\left(\frac{\tilde{\alpha}L_K}{1-\tilde{\alpha}}\vee 1\right)$. For a given $\delta>0$, let us consider the smallest $\frac{\delta}{2\tilde{b}}-$covering $\left\{\theta^1,...,\theta^N\right\}$ (with respect to the l_q-norm), and by (102), for any $\theta,\,\theta^{'}\in\mathbb{B}_q(1)$, we can find some θ^i and θ^j from the covering set $\left\{\theta^1,...,\theta^N\right\}$ such that

$$\begin{aligned}
& \left|\hat{y}_{R+1}(x;\,\theta)-\hat{y}_{R+1}\left(x;\,\theta^{'}\right)-\left(\hat{y}_{R+1}\left(x;\,\theta^{i}\right)-\hat{y}_{R+1}\left(x;\,\theta^{j}\right)\right)\right| \\
\leq\; & \left|\hat{y}_{R+1}(x;\,\theta)-\hat{y}_{R+1}\left(x;\,\theta^{i}\right)\right|+\left|\hat{y}_{R+1}\left(x;\,\theta^{'}\right)-\hat{y}_{R+1}\left(x;\,\theta^{j}\right)\right| \\
\leq\; & \delta.
\end{aligned}$$

Thus, $\{g_{\theta^1},\,g_{\theta^2},...,\,g_{\theta^N}\}\times\{g_{\theta^1},\,g_{\theta^2},...,\,g_{\theta^N}\}$ forms a $\delta-$cover of $\bar{\mathcal{F}}$ in terms of $\mathcal{F}$ defined in (100). Consequently, we have

$$\begin{aligned}
\frac{1}{\sqrt{n}}\int_0^{\tilde{r}_n}\sqrt{\log N_n(\delta,\,\Lambda(\tilde{r}_n;\,\bar{\mathcal{F}}))}d\delta & \leq \frac{1}{\sqrt{n}}\int_0^{\tilde{r}_n}\sqrt{\log N_\infty(\delta,\,\Lambda(\tilde{r}_n;\,\bar{\mathcal{F}}))}d\delta \\
& \leq \sqrt{\frac{K}{n}}\int_0^{\tilde{r}_n}\sqrt{2\log\left(1+\frac{c\tilde{b}\tilde{r}_n}{\delta}\right)}d\delta \\
& \precsim \tilde{b}\tilde{r}_n\sqrt{\frac{K}{n}}.
\end{aligned}$$

Setting $\tilde{b}\tilde{r}_n\sqrt{\frac{K}{n}}\asymp\frac{\tilde{r}_n^2}{\sigma}$ yields $\tilde{r}_n\asymp\sigma\tilde{b}\sqrt{\frac{K}{n}}$ and therefore,

$$\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_iT_1(x_i)\right|\precsim\sigma^2\tilde{b}^2\frac{K}{n} \tag{103}$$

with probability at least $1-c_1\exp\left(-c_2\tilde{b}^2K\right)$.

To analyze $\sqrt{\frac{1}{n}\sum_{i=1}^n[T_2(x_i)]^2}$ in (99), note that (91) implies

$$\begin{aligned}
|\hat{y}_1(x_i;\,\theta^*)-y_1(x_i;\,\theta^*)| & \leq |\hat{y}_0-y_0|+\tilde{\alpha}|\hat{y}_0-y_0|, \\
|\hat{y}_2(x_i;\,\theta^*)-y_2(x_i;\,\theta^*)| & \leq |\hat{y}_0-y_0|+\tilde{\alpha}|\hat{y}_0-y_0|+\tilde{\alpha}^2|\hat{y}_0-y_0|, \\
& \vdots \\
|\hat{y}_{R+1}(x_i;\,\theta^*)-y_{R+1}(x_i;\,\theta^*)| & \leq |\hat{y}_0-y_0|+\tilde{\alpha}|\hat{y}_0-y_0|+\cdots+\tilde{\alpha}^{R+1}|\hat{y}_0-y_0|\leq\frac{1}{1-\tilde{\alpha}}|\hat{y}_0-y_0|
\end{aligned}$$

for all $i=1,...,n$. As a result, we have

$$\sqrt{\frac{1}{n}\sum_{i=1}^n[T_2(x_i)]^2}\leq\frac{1}{1-\tilde{\alpha}}|\hat{y}_0-y_0|.$$

For $\sqrt{\frac{1}{n}\sum_{i=1}^n[T_3(x_i)]^2}$ in (99), standard argument for the Picard-Lindelöf Theorem implies that

$$\sup_{x\in[0,\,\tilde{\alpha}]}\left|y_r(x;\,\theta^*)-y_{r+r^{'}}(x;\,\theta^*)\right|\leq\tilde{\alpha}^r\frac{1-\tilde{\alpha}^{r^{'}}}{1-\tilde{\alpha}}\sup_{x\in[0,\,\tilde{\alpha}]}|y_1(x;\,\theta^*)-y_0|$$

for any non-negative integers r and $r^{'}$; as a result, we have

$$\sqrt{\frac{1}{n}\sum_{i=1}^n[T_3(x_i)]^2}\leq\frac{\tilde{\alpha}^{R+1}}{1-\tilde{\alpha}}\sup_{x\in[0,\,\tilde{\alpha}]}|y_1(x;\,\theta^*)-y_0|\asymp\frac{\tilde{\alpha}^{R+1}}{1-\tilde{\alpha}}\max\{C_0,\,\tilde{\alpha}\}.$$

□

References

[1] P. Bartlett and S. Mendelson (2002). "Gaussian and Rademacher complexities: risk bounds and structural results". *Journal of Machine Learning Research*, 3, 463–482.

[2] E. A. Coddington and N. Levinson (1955). *Theory of ordinary differential equations.* McGraw-Hill, New York, NY.

[3] T. H. Gronwall (1919). *Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. Annals of Mathematics*, 20, 292–296.

[4] R. Howard (1998). "The Gronwall inequality". Unpublished notes. Department of Mathematics, University of South Carolina.

[5] A. Kolmogorov and B. Tikhomirov (1959). "ϵ-entropy and ϵ-capacity of sets in functional spaces". *Uspekhi Mat. Nauk.*, 86, 3-86. Appeared in English as American Mathematical Society Translations, 17, 277–364, 1961.

[6] V. Koltchinskii (2006). "Local Rademacher complexities and oracle inequalities in risk minimization". *Annals of Statistics*, 34, 2593–2656.

[7] M. Ledoux and M. Talagrand (1991). *Probability in Banach spaces: isoperimetry and processes.* Springer, New York, NY.

[8] Mendelson, S. 2002. Geometric parameters of kernel machines. Pages 29–43 of: Proceedings of COLT.

[9] S. van de Geer (2000). *Empirical Processes in M-Estimation.* Cambridge University Press.

[10] A. W. van der Vaart and J. Wellner (1996). *Weak convergence and empirical processes.* Springer-Verlag, New York, NY.

[11] M. J. Wainwright (2019). *High-dimensional statistics: A non-asymptotic viewpoint.* Cambridge University Press.