On the Generalization Capacities of Neural Controlled Differential Equations

Linus Bleistein [1] [2] **Agathe Guilloux** [1]

Abstract

We consider a supervised learning setup in which the goal is to predicts an outcome from a sample of irregularly sampled time series using Neural Controlled Differential Equations (Kidger et al., 2020). In our framework, the time series is a discretization of an unobserved continuous path, and the outcome depends on this path through a controlled differential equation with unknown vector field. Learning with discrete data thus induces a discretization bias, which we precisely quantify. Using theoretical results on the continuity of the flow of controlled differential equations, we show that the approximation bias is directly related to the approximation error of a Lipschitz function defining the generative model by a shallow neural network. By combining these result with recent work linking the Lipschitz constant of neural networks to their generalization capacities, we upper bound the generalization gap between the expected loss attained by the empirical risk minimizer and the expected loss of the true predictor.

1. Introduction

Time series are ubiquitous in many domains such as finance, agriculture, economics and healthcare. A common set of tasks consists in predicting an outcome $y \in \mathcal{Y}$, such as a scalar or a label, from a time-evolving set of features. This problem has been addressed with a great variety of methods, ranging from auto-regressive models, such as VAR, to deep learning models, such as Recurrent Neural Networks (RNN), Long-Short-Term-Memory Networks (LSTM) and many others. It has also been thoroughly studied through the lenses of stochastic processes, Gaussian processes and many more. Both in practice and theory, most methods and theoretical setups often only tackle the case of regularly sampled time series. In this setup, a time series is seen as a collection $\mathbf{x} = (x_{t_0}, \dots, x_{t_K})$ of $K+1$ datapoints in $\mathbb{R}^d$ for which $\Delta t := t_j - t_{j-1}$ is constant. In real-world scenarios, however, time series are often irregularly sampled. This is often the case when data collection is difficult or expensive. From a modelling perspective, it is natural to consider that the time series $\mathbf{x}$ is a degraded version - through subsampling or missing values - of an unobserved underlying path $x = (x_t)_{t\in[0,1]}$ which, in turn, determines the outcome $y \in \mathcal{Y} \subset \mathbb{R}$. Informally, we have that $y = F(x) + \varepsilon$, where F is an operator that maps the space of paths $\{x : [0,1] \to \mathbb{R}^d\}$ to $\mathcal{Y}$, and ε is a bounded noise term. For instance, in healthcare, the value of a biomarker of interest of a patient is determined by the continuous and unobserved trajectory of her vitals, rather than by the discrete measurements made by a physician. Similarly, in agriculture, crop growth and yield is not determined by the punctual measurements of soil fertility, but by its continuous value through time.

As a consequence, learning from the discrete time series $\mathbf{x}$ rather than from x introduces a discretization bias. It is of great importance to better understand how this bias degrades the performance of learning algorithms. We tackle this question by studying how irregular sampling affects the generalization capacities of Neural Controlled Differential Equations (NCDE), a popular and state-of-the-art method for learning from irregular time dependant data introduced in the seminal work of Kidger et al. (2020).

Out setup. We restrict our attention to a supervised learning setup in which sampling times are irregularly spaced. Consider a sample $\{(y^i, \mathbf{x}^{D,i})\}$, where y^i is the label of data discretized on $D = \{t_0, \dots, t_K\}$. We will always require that this sampling grid is constant across individuals, and assume that $t_0 = 0$ and $t_K = 1$. Consider a predictor $\hat{f}^D$ obtained by empirical risk minimization on this sample such that

$$\hat{f}^D(\mathbf{x}^D) \approx y.$$

Our goal is to upper bound the generalization gap

$$\mathcal{R}^D(\hat{f}^D) - \mathcal{R}(\mathbf{f}^\star)$$

where $\mathcal{R}^D(\hat{f}^D)$ refers to the generalization error of the learn predictor on discretized data, and $\mathcal{R}(\mathbf{f}^\star)$ is the gener-

[1] Inria Paris, F-75015 Paris, France [2] LaMME, UEVE and UMR 8071, Paris Saclay University, F-91042, Evry, France. Correspondence to: Linus Bleistein <linus.bleistein@inria.fr>.

Workshop on New Frontiers in Learning, Control, and Dynamical Systems at the International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023. Copyright 2023 by the author(s).

alization error of the true predictor defining our generative model on continuous data (our framework is introduced in greater detail later on). Upper bounding this quantity crucially allows to quantify how sampling affects prediction performance. This is a central question both from a theoretical perspective and for practitioners, since it could, for instance, justify increasing sampling frequencies of ICU patients (Johnson et al., 2020) or intensifying monitoring of people suffering from sickle-cell anaemia (Bussy et al., 2019).

Contribution. Our contribution is threefold. First, building on a generative model proposed by Bleistein et al. (2023), we obtain an upper bound of the generalization gap which decomposes into two worst-case bounds, an approximation bias and a discretization bias. In a second time, we leverage the continuity of the flow of a CDE and approximation theorems for shallow neural networks to bound both the approximation and the discretization bias. Finally, by using the Lipschitz continuity of NCDE, we obtain an upper bound on the covering number of continuous and discretized NCDE, which explicitly depends on the discretization for the second class. This allows us to control both worst-case bounds and yields an upper bound on the generalization gap. As a byproduct, we also obtain a generalization bound for NCDEs.

Overview. Section 2 covers related works. Section 3 details our setup and assumptions. In Section 4, we state our main result, namely a generalization bound for NCDE and an upper bound on the generalization gap for a very general class of losses. Sections 5 and 6 give the two building blocks of this proof, namely upper bounds on the discretization and approximation biases, and upper bounds on the Rademacher complexity of our model.

2. Related works

Hybridization of deep learning and differential equations. The idea of combining differential systems and deep learning algorithms has been around for many years (Rico-Martinez et al., 1992; 1994). The recent work of Chen et al. (2018) has revived this idea by proposing a model which learns a representation of $x \in \mathbb{R}^d$ by using it to set the initial condition $z_0 = \varphi_\xi(x)$ of the ODE

$$dz_t = \mathbf{G}_\psi(z_t)dt$$

or equivalently in integral form

$$z_t = \varphi_\xi(x) + \int_0^t \mathbf{G}_\psi(z_s)ds$$

where $\mathbf{G}_\psi$ and φ_ξ are neural networks. The terminal value z_1 of this ODE is then used as input of any classical machine learning algorithm. Numeral contributions have build upon this idea and studied the connection between deep learning and differential equations in the recent years (Dupont et al., 2019; Chen et al., 2019b; 2020; Finlay et al., 2020). Massaroli et al. (2020) and Kidger (2022) offer comprehensive introductions to this topic. Links between deep learning and stochastic differential equations (SDE) have also been recently studied (Cohen et al., 2021; Marion et al., 2022; Hayou, 2022).

Learning with irregular time series. In the recent years, numerous competitive models have been proposed to handle time-dependant irregular data. Che et al. (2018) introduce a modified GRU with learn exponential decay of the hidden state between sampling times. Other models hybridizing classical deep learning architectures and ODE include GRU-ODE (De Brouwer et al., 2019) and RNN-ODE (Rubanova et al., 2019). Controlled Differential Equations have been introduced through two distinct frameworks, namely through Neural Rough Differential Equations (Morrill et al., 2021) and NCDEs (Kidger et al., 2020) . The latter extends neural ODEs to sequential data by first interpolating a time series $\mathbf{x}^D$ with cubic splines. The first value of the time series and the interpolated path $\tilde{\mathbf{x}}$ are then used as initial condition and driving signal of the CDE

$$dz_t = \mathbf{G}_\psi(z_t)d\tilde{\mathbf{x}}_t$$

or equivalently

$$z_t = \varphi_\xi(x_{t_0}) + \int \mathbf{G}_\psi(z_s)d\tilde{\mathbf{x}}_s,$$

where the integral is to be understood as the Riemann-Stieltjes integral. As for neural ODEs, the terminal value of this NCDE is then used for classification or regression. Other methods for learning from irregular time series include Gaussian Processes (Li & Marlin, 2016) and the signature transform (Chevyrev & Kormilitzin, 2016; Kidger et al., 2019; Fermanian, 2021; Bleistein et al., 2023).

Generalization bounds. The generalization capacities of recurrent models have been studied since the end of the 1990s, mainly through their VC dimension (Dasgupta & Sontag, 1995; Koiran & Sontag, 1998). Bartlett et al. (2017) sparked a line of research connecting the Lipschitz constant of feedforward neural networks to their generalization capacities. Our central reference is the work of Che et al. (2018), which leverage the Lipschitz continuity of RNN, LSTM and GRU to derive generalization bounds of these models. This work assumes that the time series $\mathbf{x}^D$ is bounded and regularly sampled, which are unecessary assumptions in our work. Fermanian et al. (2021) also obtain generalization bounds for RNN and a particular class of NCDE. However, they crucially impose restrictive regularity conditions which then allow for linearization of the NCDE in the signature space. Generalization bounds are then derived using kernel learning theory. Hanson & Raginsky (2022) considers an encoder-decoder setup involving excess risk bounds on neural ODEs. Closest to our work is the recent work by Marion (2023), which leverages similar tools to study the generalization capacities of a particular class of neural ODE and deep ResNets. However, our work considers NCDE, which learn with time dependant data, and focuses on the effects of data irregularity on the generalization gap.

3. Setup

We first detail our generative model, before introducting NCDE and our learning problem.

3.1. A CDE-based generative model

In order to fully track the approximation bias, we introduce a model that links the underlying path x to the outcome y.

Assumption 3.1. There exists a vector field $\mathbf{G}^* : \mathbb{R}^p \to \mathbb{R}^{p\times d}$, a function $\varphi : \mathbb{R}^d \to \mathbb{R}^p$ and a vector $\alpha_\star \in \mathbb{R}^p$ that governs a latent state $(z_t^\star)_{t\in[0,1]}$ through the CDE

$$dz_t^\star = \mathbf{G}^\star(z_t^\star)dx_t \qquad (1)$$

with initial value $z_0^\star = \varphi(x_0)$ such that the outcome y is given by

$$y = \alpha_\star^\top z_1^\star + \varepsilon,$$

where ε is a noise term bounded by $M_\varepsilon > 0$.

Put otherwise, the outcome y associated to a path x is the value of a linear transformation applied to the final value of a CDE with unknown vector field $\mathbf{G}^\star$. This is a very general model, which encompasses ODE-based models used for instance in biology or physics. Indeed, by setting $x_t = t$ for all t, one recovers a generic ODE.

We call the map $\mathbf{f}^\star : x \mapsto \alpha_\star^\top z_1^\star$ the true predictor. Uniqueness and existence to the CDE (1) is given by the Picard-Lindelöf theorem given in Appendix A.2. In line with this theorem, we make the following assumptions on the generative model and on the path $(x_t)_{t\in[0,1]}$.

Assumption 3.2. The parameters of the generative model satisfy $\|\alpha_\star\| \leq B_\alpha$, $\varphi \in C(\mathbb{R}^d, \mathbb{R}^p)$ and $\mathbf{G}^\star \in \mathrm{Lip}(\mathbb{R}^p, \mathbb{R}^{p\times d})$ with Lipschitz constant $L_{\mathbf{G}^\star}$.

Assumption 3.3. The path $(x_t)_{t\in[0,1]}$ is L_x-Lipschitz continuous, that is $\|x_t - x_s\| \leq L_x|t-s|$ for all $t, s \in [0,1]$.

Assumption 3.3 crucially implies that for all $t \in [0,1]$,

$$\|x\|_{\text{1-var},[0,t]} := \sup_D \sum \|x_{t_{i+1}} - x_{t_i}\| \leq L_x t \leq L_x,$$

where the supremum is taken on all finite discretizations $D = \{0 = t_0 < t_1 < \cdots < t_N = 1\}$ of $[0,t]$, for $N \in \mathbb{N}^*$.

We make a last assumption on the starting point of the considered time series, which we will always assume to be true when speaking of paths in the following.

Assumption 3.4. There exists a constant $B_x > 0$ such that $\|x_0\| \leq B_x$.

3.2. Neural Controlled Differential Equations

Neural Controlled Differential Equations, introduced by Kidger et al. (2020), are a particular form of CDE in which the vector field is chosen as a neural network. They can be seen as a natural extension of neural ODEs (Chen et al., 2018) to sequential data. Informally, NCDEs learn representations of time series by using them as drivers of controlled differential equations, in combination with a neural vector field. This vector field is optimized with respect to the considered learning task as in Neural ODEs. They bear close resemblance with ResNets and Recurrent Neural Networks (Fermanian et al., 2021).

In our setup, the outcome y is determined by the endpoint of an unknown CDE. Our goal is to approximate the dynamics of this CDE using a NCDE, such that the endpoint of this neural model matches the outcome.

Definition 3.5. Let $(x_t)_{t\in[0,1]}$ be a continuous path of bounded variation and let $\mathbf{G}_\psi : \mathbb{R}^p \to \mathbb{R}^{p\times d}$ be a neural vector field parametrized by ψ. Consider the solution $z := (z_t)_{t\in[0,1]}$ of the controlled differential equation

$$dz_t = \mathbf{G}_\psi(z_t)dx_t,$$

with initial condition $z_0 = \varphi_\xi(x_0)$, where $\varphi_\xi : \mathbb{R}^d \to \mathbb{R}^d$ is a neural network parametrized by ξ. We call z the *latent space trajectory* and

$$\Phi^\top z_1 =: f_\theta(x)$$

the *prediction* of the NCDE with parameters $\theta = (\Phi, \psi, \xi)$.

The vector field $\mathbf{G}_\psi$ can be any common neural network, since these architectures are Lipschitz continuous (Virmaux & Scaman, 2018). This assures that the solution of the NCDE is well defined. We restrict our attention to neural vector fields of the form

$$\mathbf{G}_\psi(z) = \sigma(\mathbf{A}z + \mathbf{b}), \qquad (2)$$

with L_σ-Lipschitz activation function $\sigma : \mathbb{R} \to \mathbb{R}$ that verifies $\sigma(0) = 0$. $\mathbf{A} : \mathbb{R}^p \to \mathbb{R}^{p\times d}$ is a linear operator and $\mathbf{b} \in \mathbb{R}^{p\times d}$ a matrix. The activation function is evaluated entry-wise. We also restrict ourselves to initializations of the form

$$z_0 = \varphi_\xi(x_0) = \sigma(\mathbf{U}x_0 + v), \qquad (3)$$

where $\mathbf{U} \in \mathbb{R}^{p\times d}$ and $v \in \mathbb{R}^p$. We use the notation

$$\mathbf{NN}_{\mathbf{U},v} : u \mapsto \sigma\big(\mathbf{U}u + v\big)$$

to refer to this initialization network. The activation is assumed to be identical for the initialization layer and the neural vector field, but such an assumption can be relaxed. The learnable parameters of the model thus are $\theta = \big(\Phi, \mathbf{A}, \mathbf{b}, \mathbf{U}, v\big)$.

The restriction on the depth of the neural vector field is mainly made for the sake of simplicity. Indeed, our proof only relies on the Lipschitz constant of $\mathbf{G}_\psi$. Such a constant can be upper bounded for deeper neural vector fields at the price of heavier notations. We refer to Bartlett et al. (2017) and Virmaux & Scaman (2018) for an in-depth discussion of the Lipschitz continuity of deeper neural networks. Also note that in practice, the vector fields used in real-world applications of NCDE have typically only a few hidden layers (Kidger et al., 2020).

Learning with time series. If one has access to the continuous path x the time series $\mathbf{x}^D$ is sampled from, one can directly use this path as an input of a NCDE. However, to apply NCDEs to time series $\mathbf{x}^D$, one needs to embed $\mathbf{x}^D$ into the space of paths of bounded variation. This can be done through any reasonable embedding mapping

$$\rho : \mathbf{x}^D \in \big(\mathbb{R}^d\big)^{K+1} \mapsto (\tilde{\mathbf{x}}_t)_{t\in[0,1]}$$

such as splines, polynomials or linear interpolation.

In this work, we focus on the fill-forward embedding, which simply defines the value of $\tilde{\mathbf{x}}$ between two consecutive points of D as the last observed value of $\mathbf{x}^D$.

Definition 3.6. The fill-forward embedding of a time series $\mathbf{x}^D = (x_{t_0}, \dots, x_{t_K})$ sampled on $D = \{0 = t_0 < \dots < t_K = 1\}$ is the piecewise constant path $(\tilde{\mathbf{x}}_t)_{t\in[0,1]}$ defined for every $t \in [t_k, t_{k+1}[$ as $\tilde{\mathbf{x}}_t = \mathbf{x}^D_{t_k}$ for $k \in \{0, \dots, K-1\}$, and $\tilde{\mathbf{x}}_1 = x_{t_K} = x_1$.

Using this embedding in a NCDE with parameters $\theta = (\Phi, \mathbf{A}, \mathbf{b}, \mathbf{U}, v)$ recovers a piecewise constant latent space trajectory $z^D := (z^D_t)_{t\in[0,1]}$ recursively defined for every $t \in [t_k, t_{k+1}[$ by

$$z^D_t = z^D_{t_{k-1}} + \sigma(\mathbf{A} z_{t_{k-1}} + \mathbf{b})(x_{t_k} - x_{t_{k-1}}) \tag{4}$$

for $k = 1, \dots, K-1$ and initialized as $z^D_0 := \sigma(\mathbf{U} x_0 + v)$. The terminal value is equal to

$$z_1 = z_{t_K} = z_{t_{K-1}} + \sigma(\mathbf{A} z_{t_{K-1}} + \mathbf{b})(x_{t_K} - x_{t_{K-1}}).$$

Formally, the prediction $\Phi^\top z^D_1$ is equal to

$$\Phi^\top z^D_1 = f_\theta \circ \rho_{\text{FF}}\left(\mathbf{x}^D\right),$$

where ρ_{FF} is the fill-forward operator. To lighten notations, we simply write $\Phi^\top z^D_1 = f_\theta(\mathbf{x}^D)$.

We highlight that the latent state z is more informative about the outcome y than z^D. Indeed, it embeds the full trajectory of x, which determines y through the CDE (1), while z^D embeds a version of x degraded through sampling.

This recursive architecture has been studied with random $\mathbf{A}, \mathbf{b}$ by Cirone et al. (2023) under the name of *homogenous controlled ResNet* because of its resemblance with the popular ResNet (He et al., 2015).

Restrictions on the parameter space. In order to obtain generalization bounds, we must furthermore restrict the size of parameter space by requiring that the parameters θ lie in a bounded set Θ. This means that there exist $B_{\mathbf{A}}, B_{\mathbf{b}}, B_{\mathbf{U}}, B_v$ such that

$$\|\mathbf{A}\| \le B_{\mathbf{A}}, \|\mathbf{b}\| \le B_{\mathbf{b}}, \|\mathbf{U}\| \le B_{\mathbf{U}}, \|v\| \le B_v \tag{5}$$

and

$$\|\Phi\| \le B_\alpha \tag{6}$$

for all NCDEs considered, where $\|W\|$ is the Frobenius norm. We recall that B_α is defined in Assumption 3.2. Such an restriction is classical for deriving generalization bounds (Bartlett et al., 2017; Bach, 2021; Fermanian et al., 2021). We call

$$\mathcal{F}_\Theta = \Big\{ f_\theta : (x_t)_{t\in[0,1]} \mapsto f_\theta(x) \in \mathbb{R} \;\text{ s.t. }\; \theta \in \Theta \Big\}$$

this class of predictors.

We now state an important lemma. It is a direct consequence of Gronwall's Lemma, stated in Lemma A.4, and of our generative model and the restrictions on the parameter space. First define

$$\|\mathbf{G}^\star(0)\|_{\text{op}} := \max_{\|u\|=1} \|\mathbf{G}^\star(0)u\|,$$

which is finite since $\mathbf{G}^\star$ is continuous.

Lemma 3.7. *Let y be generated from the CDE (1) from a L_x-Lipschitz path x. One has*

$$|y| \le B_\alpha\Big(B_\varphi + \|\mathbf{G}^\star(0)\|_{\text{op}} L_x\Big) \exp\left(L_{\mathbf{G}^\star} L_x\right) + M_\varepsilon,$$

where $B_\varphi := \max_{\|u\|\le B_x} \|\varphi(u)\|$. Also, both $f_\theta(x)$ and $f_\theta(\mathbf{x}^D)$ are upper bounded by

$$M_\Theta := B_\alpha L_\sigma \exp(L_{\mathbf{A}} L_\sigma L_x)\Big(B_{\mathbf{U}} B_x + B_v + B_{\mathbf{b}} L_x\Big)$$

for all $f_\theta \in \mathcal{F}_\Theta$.

This lemma ensures that the predictions as well as the outcome of our generative model remain bounded. This, in turn, means that the loss function is bounded, which is crucial to leverage results from Bartlett et al. (2017).

3.3. The learning problem

We now detail our learning setup. We consider an i.i.d. sample $\{(y^i, \mathbf{x}^{D,i})\}_{i=1}^n \sim y, x$ with given discretization D.

For a given predictor $f_\theta \in \mathcal{F}_\Theta$, define

$$\mathcal{R}_n^D(f_\theta) = \frac{1}{n}\sum_{i=1}^n \ell(y^i, f_\theta(\mathbf{x}^{D,i}))$$

and

$$\mathcal{R}^D(f_\theta) = \mathbb{E}_{x,y}\Big[\ell(y, f_\theta(\mathbf{x}^D))\Big]$$

as the empirical risk and the expected risk on the discretized data. Similarly, define

$$\mathcal{R}_n(f_\theta) = \frac{1}{n}\sum_{i=1}^n \ell(y^i, f_\theta(x^i))$$

and

$$\mathcal{R}(f_\theta) = \mathbb{E}_{x,y}\Big[\ell(y, f_\theta(x))\Big]$$

the empirical risk and expected risk on the continuous data. We stress that $\mathcal{R}_n(f_\theta)$ cannot be optimized, since we do not have access to the continuous data. Let

$$\hat{f}^D \in \underset{\theta\in\Theta}{\arg\min}\, \mathcal{R}_n^D(f_\theta)$$

be an optimal predictor obtained by empirical risk minimization on the discretized data. In order to obtain generalization bounds, the following technical assumption on the loss are necessary (Mohri et al., 2018).

Assumption 3.8. The loss $\ell : \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$ is Lipschitz with respect to its second variable, that is there exists L_ℓ such that for all $u, u' \in \mathcal{Y}$ and $y \in \mathcal{Y}$,

$$|\ell(y,u) - \ell(y,u')| \leq L_\ell |u-u'|.$$

This hypothesis is satisfied for most classical losses, such as the the mean squared error, as long as the outcome and the predictions are bounded. This is the case in our setup, as stated in Lemma 3.7.

4. A bound on the generalization gap

We first decompose the difference between the expected risk of the learnt predictor $\hat{f}^D$ learn from the sample S and the expected risk of the true predictor $\mathbf{f}^*$.

Lemma 4.1. *For all $f_\theta \in \mathcal{F}_\theta$, the generalization gap*

$$\mathcal{R}^D(\hat{f}^D) - \mathcal{R}(\mathbf{f}^*)$$

is almost surely bounded from above by

$$\sup_{g\in\mathcal{F}_\Theta}\Big[\mathcal{R}^D(g) - \mathcal{R}_n^D(g)\Big] + \sup_{g\in\mathcal{F}_\Theta}\Big[\mathcal{R}(g) - \mathcal{R}_n(g)\Big] + \mathcal{R}_n^D(f_\theta) - \mathcal{R}_n(f_\theta) + \mathcal{R}(f_\theta) - \mathcal{R}(\mathbf{f}^*).$$

We briefly comment on this inequality. The two first suprema are common worst-case type bounds and can be controlled using the Rademacher complexity of the class $\mathcal{F}_\Theta$. The third term $\mathcal{R}_n^D(f_\theta) - \mathcal{R}_n(f_\theta)$ corresponds to a discretization bias, which arises because one can only minimize the empirical risk on the discretized data, and not on the continuous data. It will be bounded using the continuity of the flow of a CDE with respect to its driving signal. The last term is an approximation bias.

We let M_ℓ be a bound on the loss function, which is finite since the loss is continuous and the values of y and $f_\theta(x)$ are bounded as stated in Lemma 3.7. We also let

$$|D| := \max_{j=1,\dots,K-1} |t_{j+1} - t_j|$$

be the greatest gap between two sampling times. The following theorems are our central results.

Theorem 4.2. *One has, with probability at least $1-\delta$, that*

$$\mathcal{R}^D(\hat{f}^D) \leq \mathcal{R}_n(\hat{f}^D) + \frac{24 M_\Theta L_\ell}{\sqrt{n}}\sqrt{2pU_1 + dp(p+2)U_2} + M_\ell\sqrt{\frac{\log 1/\delta}{2n}}$$

with $U_1 := \log 20\sqrt{np}K_1$ and $U_2 := \log 20\sqrt{ndp}K_2$ and K_1, K_2 two constants depending on Θ, L_x, B_x and L_σ.

Turning to the generalization gap, we get the following result.

Theorem 4.3. *The generalization gap*

$$\mathcal{R}^D(\hat{f}^D) - \mathcal{R}(\mathbf{f}^*)$$

is bounded from above for any $f_\theta \in \mathcal{F}_\Theta$ with parameters $\theta = (\Phi, \psi, \mathbf{U}, v)$ such that $\Phi = \alpha$ by

$$\begin{aligned}&\frac{16}{n} + \frac{96 M_\Theta L_\ell}{\sqrt{n}}\sqrt{2pU_1 + dp(p+2)U_2}\\ &+ L_\ell C_2\big(\Theta, L_x, L_\sigma\big)|D|\\ &+ L_\ell B_\alpha \exp(L_{\mathbf{G}^\star} L_x)\Big[L_x \left\|\mathbf{G}_\psi - \mathbf{G}^\star\right\|_{\infty,\Omega(\Theta,L_x,B_x)}\\ &+ \left\|\varphi - \mathbf{NN}_{\mathbf{U},v}\right\|_{\infty,B_x}\Big].\end{aligned}$$

The constant M_Θ is given in Lemma 3.7. The constant $C_2\big(\Theta, L_x, L_\sigma\big)$ is given in Proposition 5.1. The quantity $\left\|\mathbf{G}_\psi - \mathbf{G}^\star\right\|_{\infty,\Omega(\Theta,L_x,B_x)}$ is defined as

$$\max_{u\in\Omega(\Theta,L_x,B_x)} \left\|\mathbf{G}_\psi(u) - \mathbf{G}^\star(u)\right\|,$$

and

$$\left\|\varphi - \mathbf{NN}_{\mathbf{U},v}\right\|_{\infty,B_x} := \max_{\|u\|\leq B_x} \left\|\varphi(u) - \mathbf{NN}_{\mathbf{U},v}(u)\right\|.$$

The diameter of $\Omega(\Theta, L_x, B_x)$ is upper bounded in Proposition 5.2.

Not assuming that $\Phi = \alpha$ introduces a supplementary source of bias directly proportional to the difference $\|\Phi - \alpha\|$. However, since we assume that α and Φ lie in a common ball of radius B_α, we make this assumption which simplifies the presentation of our results.

Let us comment the three terms of the bound of Theorem 4.3.

1. The first term is a common worst case bound when deriving generalization bounds. It depends on the Rademacher complexity of $\mathcal{F}_\Theta$, and not on the discretization. We show in our proofs that an identical upper bounds hold for discrete and continuous input.
2. The second term corresponds to the discretization bias, and is directly proportional to $|D|$. It therefor vanishes at the same speed than $|D|$. For instance, if we consider the sequence D_K of equidistant discretizations of $[0, 1]$ with K points, the discretization bias vanishes at linear speed.
3. Finally, observe that the approximation bias writes as the sum of approximation errors on the vector fields and the initial condition, rather than an general approximation error on the true predictor $\mathbf{f}^\star$.

Outline of the proof. The proof schematically works in two times. We first bound the two sources of bias by leveraging the continuity of the flow of a CDE. Informally, this theorem states that the difference Δ(terminal value) between the terminal value of two CDEs decomposes as

$$\Delta(\text{vector fields}) + \Delta(\text{initial conditions}) + \Delta(\text{driving paths}).$$

This neat decomposition allows us to bound the discretization bias since it depends on the terminal value of two CDEs whose driving paths differ. This is done in Proposition 5.1. It also allows to control the approximation bias, which depends on the terminal value of two CDEs with identical driving path but whose vector fields and initial conditions differ, which we do in Proposition 5.2. Continuity of the flow is stated in full generality in Theorem A.7. This property is illustrated in Figure 1.

In a second time, we proceed to show that NCDEs are Lipschitz with respect to their parameters θ. Building on Bartlett et al. (2017), we first control the covering number of $\mathcal{F}_\Theta$ in Proposition 6.1. This gives us a bound on the Rademacher complexity of class of predictors, stated in 6.2. Theorem 4.2 then follows using standard arguments from Mohri et al. (2018). Combining these results with bounds on both sources of bias gives Theorem 4.3.

5. Bounding the discretization and the approximation bias

We bound the two sources of bias, namely the approximation bias and the discretization bias.

Proposition 5.1. *For any $f_\theta \in \mathcal{F}_\Theta$, one has*

$$\mathcal{R}_n^D(f_\theta) - \mathcal{R}_n(f_\theta) \leq L_\ell C_2\big(\Theta, L_x, L_\sigma\big)|D|,$$

where $C_2\big(\Theta, L_x, L_\sigma\big)$ is equal to

$$B_\alpha \Bigg[L_\sigma B_{\mathbf{A}} \Big(L_\sigma (B_{\mathbf{U}} B_x + B_v) + L_\sigma B_{\mathbf{b}} L_x \Big) \times \exp(L_\sigma B_{\mathbf{A}} L_x) + L_\sigma B_{\mathbf{b}} \Bigg] \exp(L_\sigma B_{\mathbf{A}} L_x).$$

The proof is given in Appendix B.1. Turning to the approximation bias, we get the following bound, using the same technique as for bounding the discretization bias.

Proposition 5.2. *For any $f_\theta \in \mathcal{F}_\Theta$ with parameters $\theta = (\Phi, \psi, \mathbf{U}, v)$ such that $\Phi = \alpha$, the approximation bias $\mathcal{R}(f_\theta) - \mathcal{R}(\mathbf{f}^*)$ is bounded from above by*

$$L_\ell B_\alpha \exp(L_{\mathbf{G}^\star} L_x) L_x \left\| \mathbf{G}_\psi - \mathbf{G}^\star \right\|_{\infty, \Omega(\Theta, L_x, B_x)} + L_\ell B_\alpha \exp(L_{\mathbf{G}^\star} L_x) \left\| \varphi - NN_{\mathbf{U}, v} \right\|_{\infty, B_x},$$

where

$$\Omega(\Theta, L_x, B_x) \subset \big\{ u \in \mathbb{R}^p \mid \|u\| \leq (L_\sigma (B_{\mathbf{U}} B_x + B_v) + L_\sigma B_{\mathbf{b}} L_x) \exp(L_\sigma L_{\mathbf{A}} L_x) \big\}.$$

Notice that this bound has a double dependence on the unknown vector field $\mathbf{G}^\star$. The first one is through its Lipschitz constant $L_{\mathbf{G}^\star}$. The second one comes from the error made when approximating $\mathbf{G}^\star$ by $\mathbf{G}_\psi$ and φ by the shallow initialization neural network. Controlling this second term using approximation results for neural networks is left for future work.

6. Bouding the Rademacher complexity of NCDE

We first upper bound the covering number of $\mathcal{F}_\Theta$. This result is obtained by showing that NCDE are Lipschitz with respect to their parameters, such that covering each parameter class yields a covering of the whole class of predictors.

Proposition 6.1. *The covering number $\mathcal{N}\big(\mathcal{F}_\Theta, \beta\big)$ of $\mathcal{F}_\Theta$ is bounded from above by*

$$\left(1 + \frac{10\sqrt{p}K_1}{\beta}\right)^{2p} \left(1 + \frac{10\sqrt{dp}K_3}{\beta}\right)^{dp(p+2)}$$

where K_1, K_2 are two constants depending on Θ, L_x, B_x and L_σ.

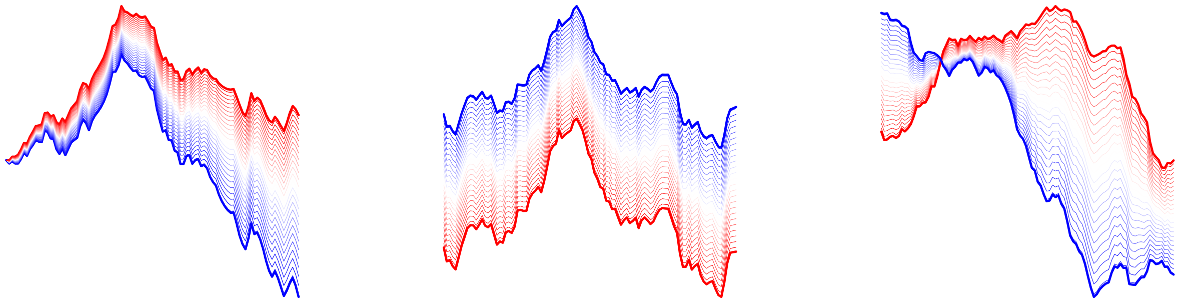

Figure 1: On the **left**, we plot in bold the solutions of two NCDEs who only differ in their vector field $\mathbf{A}_1$ and $\mathbf{A}_2$. We then plot the solutions of the NCDEs with interpolated vector field $\delta\mathbf{A}_1 + (1-\delta)\mathbf{A}_2$ for $\delta \in [0,1]$ in red-blue gradient. In the **middle**, we consider a given NCDE and interpolate linearly between two initial conditions x_0^1 and x_0^2. On the **right**, we consider a given NCDE and drive it with the linear interpolation of two paths $(x_t^1)_t$ and $(x_t^2)_t$. In all three cases, the solutions evolve continuously as we interpolate between the models.

In combination with arguments borrowed from Chen et al. (2019a) and Bartlett et al. (2017), Proposition 6.1 allows to upper bound the empirical Rademacher complexity, which we denote by $\mathrm{Rad}(\mathcal{H})$ for a function class $\mathcal{H}$.

Proposition 6.2. *The empirical Rademacher complexity* $\mathrm{Rad}(\mathcal{F}_\Theta)$ *associated to an i.i.d. sample of size* n *is upper bounded by*

$$\frac{4}{n} + \frac{24M_\Theta}{\sqrt{n}}\sqrt{2pU_1 + dp(p+2)U_2}$$

where $U_1 := \log 20\sqrt{np}K_1$ *and* $U_2 := \log 20\sqrt{ndp}K_2$, *and* K_1, K_2 *are two constants depending on* Θ, L_x, B_x *and* L_σ. *The constant* M_Θ *is given in Lemma 3.7.*

7. Experiments

Our bounds rely crucially on restrictions on our parameter space: the vaster the parameter space we consider - i.e. the bigger the norm of the model's parameters, the looser our bounds.

In a first experiment, we analyse how the norm of the NCDE's parameters evolve during training in a teacher-student setup. Paths are not downsampled in this experiment to neutralize the discretization bias. The target is generated from a teacher NCDE: the approximation bias is thus null. Figure 2 considers a single training run. First, despite only seeing the endpoint of the true red trajectory, the trained model achieves almost perfect interpolation of the whole trajectory on the training set. Secondly, the model's parameters stay within close range of their initialized value. Since our model is initialized with Pytorch's standard scaled Gaussian initialization, this means that the parameter's norms remain well behaved during training and that our bounds do not explode.

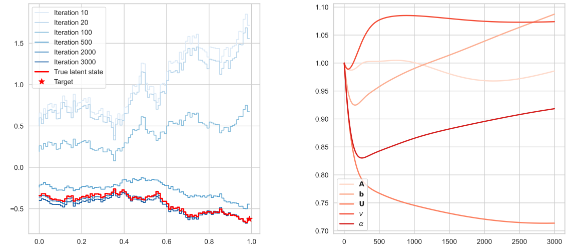

Figure 2: On the **left**, we display the evolution of the latent state trajectory $(\alpha^\top z_t)_t$ and the true latent state trajectory $(\alpha_\star^\top z_t^\star)_t$ through training. On the **right**, we display the evolution of the parameter's norms normalized by their value at initialization. Training is performed with Adam (Kingma & Ba, 2014), standard α, β parameters and a learning rate of 9×10^{-4} for 3000 iterations on a training set of size $n = 100$.

We then iterate this experiment and display the distribution of the parameter's norms - this time in absolute value - after training in Figure 3. Since the norms of the parameters scale with their dimensions, greater values are observed for the parameters $\mathbf{A}$ and $\mathbf{b}$ who are of dimension dp^2 and dp respectively. The parameters norm remain relatively small and do not disperse excessively over training runs.

Finally, we conduct a last experiment to analyse the interaction between training error, generalization error and discretization. We first fix a test and train dataset. For

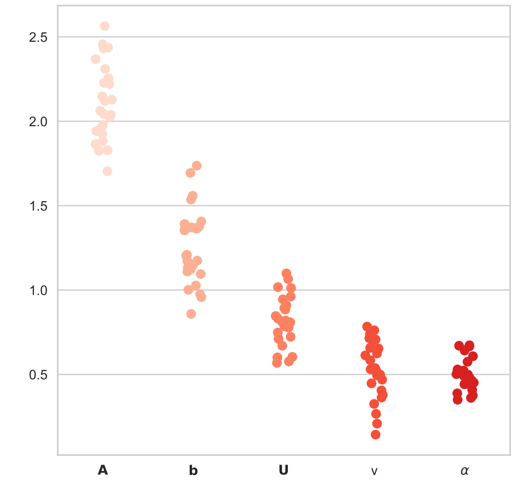

Figure 3: Absolute value of the parameters norm after training for 1000 iterations on a training set of size $n = 100$. Training is performed with Adam (Kingma & Ba, 2014), standard α, β parameters and a learning rate of 9×10^{-4}. Each dot corresponds to one random initialization and training run of an NCDE.

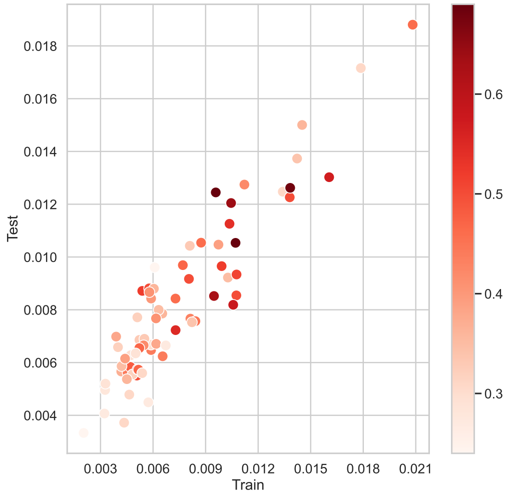

Figure 4: Train and test error with paths downsampled to 5 points. Every point corresponds to a random downsampling of the paths $(\mathbf{x}^i)$, initialization and training of an NCDE model. The colorbar indicates the coarsness of the sampling, that is the biggest time between two sampling points. Training is performed with Adam (Kingma & Ba, 2014), standard α, β parameters and a learning rate of 1×10^{-3} for 1200 iterations.

every random initialization and training of an NCDE, we randomly downsample the paths $(\mathbf{x}^i)$ on a 5 point grid. We then compute the loss on the downsampled training and the downsampled test dataset. The results are displayed in Figure 4. The darker the points, the greater the sampling gap $|D|$. First, one can see that the training error is positively correlated with the generalization error, as predicted by Theorem 4.2. Interestingly, the sampling gap $|D|$ is also positively correlated with the training and generalization error. Indeed, as more information is missed between sampling times when $|D|$ increases, the prediction of the label becomes less precise.

8. Conclusion

Several perspectives are of high interest for extending our work. First, a deeper empirical assessment of our claims is needed. Concerning the approximation bias, we believe that approximation theorems for deeper neural networks can be used for quantifying its dependence in Θ and in the size of the latent space p, leading to a trade-off between complexity of the model and approximation capacities.

Acknowledgements. We thank Pierre Marion for insightful comments and clarifications, and Cristopher Salvi for discussions on Controlled ResNets. LB conducted this research while visiting MBZUAI's Machine Learning Department (UAE).

References

Bach, F. Learning theory from first principles. *Online version*, 2021.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Bleistein, L., Fermanian, A., Jannot, A.-S., and Guilloux, A. Learning the dynamics of sparsely observed interacting systems. *arXiv preprint arXiv:2301.11647*, 2023.

Bussy, S., Veil, R., Looten, V., Burgun, A., Gaïffas, S., Guilloux, A., Ranque, B., and Jannot, A.-S. Comparison of methods for early-readmission prediction in a high-dimensional heterogeneous covariates and time-to-event outcome framework. *BMC medical research methodology*, 19:1–9, 2019.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Chen, M., Li, X., and Zhao, T. On generalization bounds of a family of recurrent neural networks. *arXiv preprint arXiv:1910.12947*, 2019a.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Chen, R. T., Behrmann, J., Duvenaud, D. K., and Jacobsen, J.-H. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019b.

Chen, R. T., Amos, B., and Nickel, M. Learning neural event functions for ordinary differential equations. *arXiv preprint arXiv:2011.03902*, 2020.

Chevyrev, I. and Kormilitzin, A. A primer on the signature method in machine learning. *arXiv preprint arXiv:1603.03788*, 2016.

Cirone, N. M., Lemercier, M., and Salvi, C. Neural signature kernels as infinite-width-depth-limits of controlled resnets. *arXiv preprint arXiv:2303.17671*, 2023.

Clark, D. S. Short proof of a discrete gronwall inequality. *Discrete applied mathematics*, 16(3):279–281, 1987.

Cohen, A.-S., Cont, R., Rossier, A., and Xu, R. Scaling properties of deep residual networks. In *International Conference on Machine Learning*, pp. 2039–2048. PMLR, 2021.

Dasgupta, B. and Sontag, E. Sample complexity for learning recurrent perceptron mappings. *Advances in Neural Information Processing Systems*, 8, 1995.

De Brouwer, E., Simm, J., Arany, A., and Moreau, Y. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. *Advances in neural information processing systems*, 32, 2019.

Dupont, E., Doucet, A., and Teh, Y. W. Augmented neural odes. *Advances in neural information processing systems*, 32, 2019.

Fermanian, A. Embedding and learning with signatures. *Computational Statistics & Data Analysis*, 157:107148, 2021.

Fermanian, A., Marion, P., Vert, J.-P., and Biau, G. Framing rnn as a kernel method: A neural ode approach. *Advances in Neural Information Processing Systems*, 34: 3121–3134, 2021.

Finlay, C., Jacobsen, J.-H., Nurbekyan, L., and Oberman, A. How to train your neural ode: the world of jacobian and kinetic regularization. In *International conference on machine learning*, pp. 3154–3164. PMLR, 2020.

Friz, P. K. and Victoir, N. B. *Multidimensional stochastic processes as rough paths: theory and applications*, volume 120. Cambridge University Press, 2010.

Hanson, J. and Raginsky, M. Fitting an immersed submanifold to data via sussmann's orbit theorem. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 5323–5328. IEEE, 2022.

Hayou, S. On the infinite-depth limit of finite-width neural networks. *arXiv preprint arXiv:2210.00688*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. arxiv 2015. *arXiv preprint arXiv:1512.03385*, 14, 2015.

Holte, J. M. Discrete gronwall lemma and applications. In *MAA-NCS meeting at the University of North Dakota*, volume 24, pp. 1–7, 2009.

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., and Mark, R. Mimic-iv. *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)*, 2020.

Kidger, P. On neural differential equations. *arXiv preprint arXiv:2202.02435*, 2022.

Kidger, P., Bonnier, P., Perez Arribas, I., Salvi, C., and Lyons, T. Deep signature transforms. *Advances in Neural Information Processing Systems*, 32, 2019.

Kidger, P., Morrill, J., Foster, J., and Lyons, T. Neural controlled differential equations for irregular time series. *Advances in Neural Information Processing Systems*, 33: 6696–6707, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Koiran, P. and Sontag, E. D. Vapnik-chervonenkis dimension of recurrent neural networks. *Discrete Applied Mathematics*, 86(1):63–79, 1998.

Li, S. C.-X. and Marlin, B. M. A scalable end-to-end gaussian process adapter for irregularly sampled time series classification. *Advances in neural information processing systems*, 29, 2016.

Marion, P. Generalization bounds for neural ordinary differential equations and deep residual networks, 2023.

Marion, P., Fermanian, A., Biau, G., and Vert, J.-P. Scaling resnets in the large-depth regime. *arXiv preprint arXiv:2206.06929*, 2022.

Massaroli, S., Poli, M., Park, J., Yamashita, A., and Asama, H. Dissecting neural odes. *Advances in Neural Information Processing Systems*, 33:3952–3963, 2020.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Morrill, J., Salvi, C., Kidger, P., and Foster, J. Neural rough differential equations for long time series. In *International Conference on Machine Learning*, pp. 7829–7838. PMLR, 2021.

Rico-Martinez, R., Krischer, K., Kevrekidis, I., Kube, M., and Hudson, J. Discrete-vs. continuous-time nonlinear signal processing of cu electrodissolution data. *Chemical Engineering Communications*, 118(1):25–48, 1992.

Rico-Martinez, R., Anderson, J., and Kevrekidis, I. Continuous-time nonlinear signal processing: a neural network based approach for gray box identification. In *Proceedings of IEEE Workshop on Neural Networks for Signal Processing*, pp. 596–605. IEEE, 1994.

Rubanova, Y., Chen, R. T., and Duvenaud, D. K. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Virmaux, A. and Scaman, K. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

The appendix is structured as follows. Appendix A gives preliminary results on CDEs and covering numbers. Appendix B details the proof of the bounds of the approximation and discretization biases. Appendix C collects the proofs of the generalization bound, and the bound on the generalization gap. Proofs of Appendix A are given in Appendix D for the sake of clarity.

A. Mathematical background

A.1. Riemann-Stieltjes integral

Lemma A.1. *Let $f : [0,1] \to \mathbb{R}^{p\times d}$ be a continuous function, and $g : [0,1] \to \mathbb{R}^d$ be of finite total variation. Then Riemann-Stieljes integral*

$$\int_0^t f_s dg_s \in \mathbb{R}^p,$$

where the product is understood in the sens of matrix-vector multiplication, is well-defined and finite for every $t \in [0,1]$.

We refer to Friz & Victoir (2010) for a thorough introduction to the Riemann-Stieltjes integral.

A.2. Picard-Lindelöf Theorem

Theorem A.2. *Let x be a continuous path of bounded variation, and assume that $\mathbf{G} : \mathbb{R}^p \to \mathbb{R}^{p\times d}$ is Lipschitz continuous. Then the CDE*

$$dy_t = \mathbf{G}(y_t)dx_t$$

with initial condition $y_0 = \varphi(x_0)$ has a unique solution.

A full proof can be found in Fermanian et al. (2021). Remark that since in our setting, NCDE are Lipschitz since the neural vector fields are Lipschitz (Virmaux & Scaman, 2018). This ensures that the solutions to NCDEs are well defined.

We also need the following variation, which ensures that the CDE driven by $\tilde{\mathbf{x}}$ is well-defined.

Lemma A.3. *Let x be a piecewise constant and right-continuous path taking finite values in $\mathbb{R}^d$, with a finite number of discontinuities at $0 < t_1, \dots, t_K = 1$, i.e.*

$$\lim_{t\to t_i^+} x_t = x_{t_i}$$

and

$$x_t = x_{t_i}$$

for all $t \in [t_i, t_{i+1}[$, for all $i = 1, \dots, K-1$. Assume that $\mathbf{G} : \mathbb{R}^p \to \mathbb{R}^{p\times d}$ is Lipschitz continuous. Then the CDE

$$dy_t = \mathbf{G}(y_t)dx_t$$

with initial condition $y_0 = \varphi(x_0)$ has a unique solution.

This result can be obtained by first remarking that since x is piecewise constant, the solution to this CDE will also be piecewise constant, with discontinues at $0 < t_1 < \dots < t_K$: indeed, the variations of x between two points of discontinuity being null, the variations of the solution will also be null between these two points. The solution can then be recursively obtained by seeing that for all $t \in [t_i, t_{i+1}[$

$$y_t = y_{t_i} = y_{t_{i-1}} + \mathbf{G}(y_{t_{i-1}})(x_{t_i} - x_{t_{i-1}})$$

with $y_t = \varphi(x_0)$ for $t \in [t_0, t_1[$, where $t_0 = 0$, and $y_{t_K} = y_1 = y_{t_{K-1}} + \mathbf{G}(y_{t_{K-1}})(x_{t_K} - x_{t_{K-1}})$.

A.3. Gronwall's Lemmas

Lemma A.4 (Gronwall's Lemma for CDEs)**.** *Let $x : [0,1] \to \mathbb{R}^d$ be a continuous path of bounded variations, and $\phi : [0,1] \to \mathbb{R}^d$ be a bounded measurable function. If*

$$\phi(t) \le K_t + L_x \int_0^t \phi(s) \left\| dx_s \right\|$$

for all $t \in [0,1]$, where K_t is a time dependant constant, then

$$\phi(t) \leq K_t \exp\left(L_x \left\|x\right\|_{\text{1-var},[0,t]}\right)$$

for all $t \geq 0$.

See Friz & Victoir (2010) for a proof. Remark that this Lemma does not require $\phi(\cdot)$ to be continuous. We also state a variant of the discrete Gronwall Lemma, which will allow us to obtain the bound for discrete inputs $\mathbf{x}^D$.

Lemma A.5 (Gronwall's Lemma for sequences)**.** *Let $(y_k)_{k\geq 0}, (b_k)_{k\geq 0}$ and $(f_k)_{k\geq 0}$ be positive sequences of real numbers such that*

$$y_n \leq f_n + \sum_{l=0}^{n-1} b_l y_l$$

for all $n \geq 0$. Then

$$y_n \leq f_n + \sum_{l=0}^{n-1} f_l b_l \prod_{j=l+1}^{n-1} (1 + b_j)$$

for all $n \geq 0$.

A proof can be found in Holte (2009) and Clark (1987).

A.4. An upper bound on the total variation of the solution of a CDE

We have the following bound on the total variation of the solution to a CDE.

Proposition A.6. *Let $\mathbf{F} : \mathbb{R}^{p\times d} \to \mathbb{R}^d$ be a Lipschitz vector field with Lipschitz constant $L_{\mathbf{F}}$ and $(x_t)_{t\in[0,1]}$ be a continuous path of total variation bounded by L_x. Let $(z_t)_{t\in[0,1]}$ be the solution of the CDE*

$$dz_t = \mathbf{F}(z_t)dx_t$$

with initial condition $z_0 \in \mathbb{R}^p$. Then for all $t \in [0,1]$, one has

$$\|z\|_{\text{1-var},[0,t]} \leq C_1(L_{\mathbf{F}}, \mathbf{F}, L_x) \left\|x\right\|_{\text{1-var},[0,t]}$$

with

$$C_1(L_{\mathbf{F}}, \mathbf{F}, L_x) := \left[L_{\mathbf{F}} \big(\|z_0\| + \|\mathbf{F}(0)\|_{\text{op}} L_x \big) \exp(L_{\mathbf{F}} L_x) + \|\mathbf{F}(0)\|_{\text{op}} \right] \exp(L_{\mathbf{F}} L_x).$$

For $f \in \mathcal{F}_\Theta$, let z be the trajectory of the latent state associated to an input x, and z^D be the trajectory of the latent state associated to an input $\mathbf{x}^D$. Both $\|z_1\|_{\text{1-var}}$ and $\left\|z_1^D\right\|_{\text{1-var}}$ are upper bounded by

$$C_1(L_{\mathbf{G}_\psi}, \mathbf{G}_\psi, L_x) L_x$$

where $C_1(L_{\mathbf{G}_\psi}, \mathbf{G}_\psi, L_x)$ is defined as

$$\left[L_\sigma B_{\mathbf{A}} L_\sigma \big(B_{\mathbf{U}} B_x + B_v + B_{\mathbf{b}} L_x \big) \exp(L_\sigma B_{\mathbf{A}} L_x) + L_\sigma B_{\mathbf{b}} \right] \exp(L_\sigma B_{\mathbf{A}} L_x).$$

The proof is given in Appendix D.2.

A.5. Continuity of the flow of a CDE

Theorem A.7. *Let $\mathbf{F}, \mathbf{G} : \mathbb{R}^p \to \mathbb{R}^{p\times d}$ be two Lipschitz vector fields with Lipschitz constants $L_{\mathbf{F}}, L_{\mathbf{G}}$. Let x, r be either continuous or piecewise constant paths of total variation bounded by L_x and L_r. Consider the controlled differential equations*

$$dw_t = \mathbf{F}(w_t)dx_t \;\; \textit{and} \;\; dv_t = \mathbf{G}(v_t)dr_t$$

with initial conditions $w_0 \in \mathbb{R}^p$ and $v_0 \in \mathbb{R}^p$ respectively.

One has that

$$\begin{aligned}\|w_t - v_t\| \leq \Bigg(& \|w_0 - v_0\| + \|x_0 - r_0\| \\ & + \|x - r\|_{\infty,[0,t]} \left(1 + L_{\mathbf{F}} L_r C_1(L_{\mathbf{F}}, \mathbf{F}, L_x)\right) + \max_{v \in \Omega(\mathbf{G})} \|\mathbf{F}(v) - \mathbf{G}(v)\|_{\text{op}} L_r \Bigg) \\ & \times \exp(L_{\mathbf{F}} L_x).\end{aligned}$$

and symmetrically that

$$\begin{aligned}\|w_t - v_t\| \leq \Bigg(& \|w_0 - v_0\| + \|x_0 - r_0\| \\ & + \|x - r\|_{\infty,[0,t]} \left(1 + L_{\mathbf{G}} L_x C_1(L_{\mathbf{G}}, \mathbf{G}, L_r)\right) + \max_{v \in \Omega(\mathbf{F})} \|\mathbf{F}(v) - \mathbf{G}(v)\|_{\text{op}} L_x \Bigg) \\ & \times \exp(L_{\mathbf{G}} L_r),\end{aligned}$$

where the constant $C_1(\cdot,\cdot,\cdot)$ is given in Proposition A.6 and the sets $\Omega(\mathbf{G})$ and $\Omega(\mathbf{F})$ are defined as

$$\Omega(\mathbf{G}) = \left\{u \in \mathbb{R}^p \mid \|u\| \leq (\|v_0\| + \|\mathbf{G}(0)\| L_r) \exp(L_{\mathbf{G}} L_r)\right\}.$$

and

$$\Omega(\mathbf{F}) = \left\{u \in \mathbb{R}^p \mid \|u\| \leq (\|w_0\| + \|\mathbf{F}(0)\| L_x) \exp(L_{\mathbf{F}} L_x)\right\}.$$

The proof is given in Appendix D.3. We stress that any combination of continuous and piecewise constant paths can be used.

Using this result, we obtain the following theorem.

Theorem A.8. *Let $f_{\theta_1}, f_{\theta_2} \in \mathcal{F}_\Theta$ two predictors with respective parameters $\theta_1 = (\mathbf{A}_1, \mathbf{b}_1, \mathbf{U}_1, v_1, \Phi_1)$ and $\theta_2 = (\mathbf{A}_2, \mathbf{b}_2, \mathbf{U}_2, v_2, \Phi_2)$.*

Both

$$|f_{\theta_1}(x) - f_{\theta_2}(x)|$$

and

$$|f_{\theta_1}(\mathbf{x}^D) - f_{\theta_2}(\mathbf{x}^D)|$$

are upper bounded by

$$L_1 \|\Phi_1 - \Phi_2\| + L_2 \|\mathbf{A}_1 - \mathbf{A}_2\| + L_3 \|\mathbf{b}_1 - \mathbf{b}_2\| + L_4 \|\mathbf{U}_1 - \mathbf{U}_2\| + L_5 \|v_1 - v_2\|$$

where L_1, L_2, L_3, L_4, L_5 are explicit Lipschitz constants.

The proof is in Appendix D.4.

A.6. Covering numbers

First, we recall the definition of the covering number of a class of functions.

Definition A.9. Let $\mathcal{H} = \{h : \mathcal{U} \subset \mathbb{R}^e \to \mathbb{R}^f\}$ be a class of functions. The covering number $\mathcal{N}\big(\mathcal{H}, \beta\big)$ of $\mathcal{H}$ is the minimal cardinality of a subset $C \subset \mathcal{H}$ such that for all $h \in \mathcal{H}$, there exists $\hat{h} \in C$ such that

$$\sup_{x \in \mathcal{U}} \left\| h(x) - \hat{h}(x) \right\| \leq \beta.$$

We state a Lemma from Chen et al. (2019a).

Lemma A.10. *Let $\mathcal{G} = \{A \in \mathbb{R}^{d_1 \times d_2}$ s.t. $\|A\| \leq \lambda\}$ for a given $\lambda > 0$. The covering number $\mathcal{N}(\mathcal{G}, \beta)$ is upper bounded by*

$$\left(1 + \frac{2\min(d_1^{\frac{1}{2}}, d_2^{\frac{1}{2}})\lambda}{\varepsilon}\right)^{d_1 d_2}.$$

We now proceed to prove the bounds on the approximation and discretization biases, and the generalization inequalities. Let us precisely detail the different steps of our proofs.

1. The bound of the discretization bias, stated in Proposition 5.1), follows from the continuity of the flow given in Theorem A.7. Indeed, the discretization error depends on the difference between z_1 and z_1^D, which are solutions to identical CDE, but with different driving paths x and embedded path $\tilde{\mathbf{x}}$.

2. The bound of the approximation bias is obtained in a similar fashion. Indeed, it directly depends on the difference between $z_1^\star$ and z_1. However, this time, only the initial condition and the vector fields are different, while the driving paths are identical.

3. Turning to the generalization inequalities, the first step consists in showing that NCDEs are Lipschitz with respect to their parameters, which we do in Theorem A.8. This leverages once again the continuity of the flow, since we aim at bounding the difference between two NCDEs with different parameters but identical driving paths.

4. We then use the central idea of Chen et al. (2019a) which consists in obtaining an upper bound on the covering number of a parameterized class by covering each of its parameters in Proposition 6.1. While the authors of this article use such a bound for RNN, the same technique applies for NCDE.

5. We then proceed to connect the covering number of $\mathcal{F}_\Theta$ to its Rademacher complexity building on arguments made by Bartlett et al. (2017) in Proposition 6.2.

6. Once this last result is obtained, we can obtain the generalization bound by resorting to classical techniques (Mohri et al., 2018).

7. The bound on the generalization gap is obtained by combining all the precedent elements and using a symmetrization argument to upper bound the two worst case bounds.

B. Proof of bias bounds

B.1. Proof of Proposition 5.1

Take $f_\theta \in \mathcal{F}_\Theta$. One has

$$\mathcal{R}_n^D(f_\theta) - \mathcal{R}_n(f_\theta) = \frac{1}{n}\sum_{i=1}^{n}\Big[\ell(y^i, f_\theta(\mathbf{x}^{D,i})) - \ell(y^i, f_\theta(x^i))\Big]. \tag{7}$$

Considering a single individual $i \in \{1, \dots, n\}$, one has

$$\ell(y^i, f_\theta(\mathbf{x}^{D,i})) - \ell(y^i, f_\theta(x^i)) \leq L_\ell |f_\theta(\mathbf{x}^{D,i}) - f_\theta(x^i)| \tag{8}$$

using the Lipschitz continuity of the loss function with respect to its second argument. From the Cauchy-Schwarz inequality, it follows that

$$|f_\theta(\mathbf{x}^{D,i}) - f_\theta(x^i)| = |\Phi^\top\big(z_1^D - z_1\big)| \leq B_\alpha \left\| z_1^{D,i} - z_1^i \right\| \tag{9}$$

where $z^{D,i}$ and z_1^i refer to the endpoint of the latent space trajectory of the NCDE, resp. with discrete and continuous input, associated to the predictor f_θ.

z_1^D and z_1 correspond to the endpoint of two CDEs with identical vector field and identical initial condition, but whose driving path differ. Using the continuity of the flow stated in Theorem A.7, and using the fact that the total variation of the piecewise constant path corresponding to the fill-forward embedding of the time series $\mathbf{x}^{D,i}$ is bounded by L_x, one has the inequality

$$\left\|z_1^D - z_1\right\| \leq \left\|x - \tilde{\mathbf{x}}^i\right\|_{\infty,[0,t]} \big(1 + L_\sigma L_{\mathbf{A}} L_x C_1(L_{\mathbf{G}_\psi}, \mathbf{G}_\psi, L_x)\big) \exp(L_\sigma B_{\mathbf{A}} L_x) \tag{10}$$

where $\tilde{\mathbf{x}}^i$ is the piecewise constant path corresponding to the fill-forward embedding of the time series $\mathbf{x}^{D,i}$ and we recall that $C_1(L_{\mathbf{G}_\psi}, \mathbf{G}_\psi, L_x)$ is equal to

$$\Big[L_\sigma B_{\mathbf{A}} L_\sigma \big(B_{\mathbf{U}} B_x + B_v + B_{\mathbf{b}} L_x\big) \exp(L_\sigma B_{\mathbf{A}} L_x) + L_\sigma B_{\mathbf{b}}\Big] \exp(L_\sigma B_{\mathbf{A}} L_x).$$

One also has that

$$\left\|\mathbf{x}^{D,i} - x^i\right\|_{\infty,[0,1]} = \max_{j=1,\dots,K} \max_{s\in[t_j,t_{j+1}]} \left\|x^i_{t_j} - x^i_s\right\| \leq L_x|D| \tag{11}$$

since the discretization is identical between individuals. Putting everything together, this yields for all $i = 1, \dots, n$ that

$$|f_\theta(\mathbf{x}^{D,i}) - f_\theta(x^i)| \leq \underbrace{B_\alpha\big(1 + L_\sigma L_{\mathbf{A}} L_x C_1(L_{\mathbf{G}_\psi}, \mathbf{G}_\psi, L_x)\big)\exp(L_\sigma B_{\mathbf{A}} L_x) L_x}_{C_2\left(\Theta, L_x, L_\sigma\right)}|D| \tag{12}$$

and finally

$$\mathcal{R}^D_n(f_\theta) - \mathcal{R}_n(f_\theta) \leq L_\ell C_2\big(\Theta, L_x, L_\sigma\big)|D|. \tag{13}$$

This concludes the proof.

B.2. Proof of Proposition 5.2

For any $f_\theta \in \mathcal{F}_\Theta$ with continuous input, the approximation bias writes

$$\mathcal{R}(f_\theta) - \mathcal{R}(\mathbf{f}^\star) = \mathbb{E}\Big[\ell(y, f_\theta(x)) - \ell(y, \mathbf{f}^\star(x))\Big]. \tag{14}$$

Using the Lipschitz continuity of ℓ and the Cauchy-Schwarz inequality, one gets

$$\ell(y, f_\theta(x)) - \ell(y, \mathbf{f}^\star(x)) \leq L_\ell \left\|\alpha_\star - \Phi\right\| \left\|z_1\right\| + L_\ell B_\alpha \left\|z^\star_1 - z_1\right\|. \tag{15}$$

Since $z^\star_1$ and z_1 are the solution to two CDEs with different vector fields and different initial conditions, the continuity of the flow stated in Theorem A.7 yields

$$\left\|z^\star_1 - z_1\right\| \leq \exp(L_{\mathbf{G}^\star} L_x)\Big[L \max_{u\in\Omega(\mathbf{G}_\psi)} \left\|\mathbf{G}_\psi(u) - \mathbf{G}^\star(u)\right\| + \max_{\|u\|\leq B_x} \left\|\varphi(u) - \mathbf{NN}_{\mathbf{U},v}(u)\right\|\Big] \tag{16}$$

where

$$\begin{aligned}\Omega(\mathbf{G}_\psi) &= \big\{u \in \mathbb{R}^p \mid \|u\| \leq (\|z_0\| + L_\sigma B_{\mathbf{b}} L_x)\exp(L_\sigma L_{\mathbf{A}} L_x)\big\} \\ &\subset \big\{u \in \mathbb{R}^p \mid \|u\| \leq (L_\sigma(B_{\mathbf{U}} B_x + B_v) + L_\sigma B_{\mathbf{b}} L_x)\exp(L_\sigma L_{\mathbf{A}} L_x)\big\}.\end{aligned}$$

This means that $\operatorname{diam}(\Omega(\mathbf{G}_\psi))$ is a function of Θ, L_x and B_x. To clarify this, we now write $\Omega(\Theta, L_x, B_x)$.

Putting everything together and taking expectations yields

$$\begin{aligned}\mathcal{R}(f_\theta) - \mathcal{R}(\mathbf{f}^\star) \leq L_\ell B_\alpha \exp(L_{\mathbf{G}^\star} L_x)\Bigg[L_x &\max_{u\in\Omega(\Theta,L_x,B_x)} \left\|\mathbf{G}_\psi(u) - \mathbf{G}^\star(u)\right\| & (17)\\ &+ \max_{\|u\|\leq B_x} \left\|\varphi(u) - \mathbf{NN}_{\mathbf{U},v}(u)\right\|\Bigg] & (18)\\ &+ L_\ell \left\|\alpha_\star - \Phi\right\| \left\|z_1\right\|. & (19)\\ && (20)\end{aligned}$$

We now turn to a predictor for which one has $\Phi = \alpha_\star$. This yields

$$\begin{aligned}&\mathcal{R}(f_\theta) - \mathcal{R}(\mathbf{f}^\star) & (21)\\ &\leq L_\ell B_\alpha \exp(L_{\mathbf{G}^\star} L_x)\Big[L_x \left\|\mathbf{G}_\psi - \mathbf{G}^\star\right\|_{\infty,\Omega(\Theta,L_x,B_x)} + \left\|\varphi - \mathbf{NN}_{\mathbf{U},v}\right\|_{\infty,B_x}\Big]. & (22)\end{aligned}$$

This concludes the proof.

C. Proof of generalization bound

C.1. Proof of Lemma 4.1

For all $f \in \mathcal{F}_\Theta$, one has the decomposition

$$\begin{align}
\mathcal{R}^D(\hat{f}^D) - \mathcal{R}(\mathbf{f}^*) &= \mathcal{R}^D(\hat{f}^D) - \mathcal{R}_n^D(\hat{f}^D) \tag{23} \\
&+ \mathcal{R}_n^D(\hat{f}^D) - \mathcal{R}_n^D(f) \tag{24} \\
&+ \mathcal{R}_n^D(f) - \mathcal{R}_n(f) \tag{25} \\
&+ \mathcal{R}_n(f) - \mathcal{R}(f) \tag{26} \\
&+ \mathcal{R}(f) - \mathcal{R}(\mathbf{f}^\star). \tag{27}
\end{align}$$

By optimality of $\hat{f}^D$, one has almost surely

$$\mathcal{R}_n^D(\hat{f}^D) - \mathcal{R}_n^D(f) \leq 0. \tag{28}$$

One is then left with the inequality

$$\begin{align}
\mathcal{R}^D(\hat{f}^D) - \mathcal{R}(\mathbf{f}^*) &\leq \mathcal{R}^D(\hat{f}^D) - \mathcal{R}_n^D(\hat{f}^D) \tag{29} \\
&+ \mathcal{R}_n^D(f) - \mathcal{R}_n(f) \tag{30} \\
&+ \mathcal{R}_n(f) - \mathcal{R}(f) \tag{31} \\
&+ \mathcal{R}(f) - \mathcal{R}(\mathbf{f}^\star). \tag{32}
\end{align}$$

Taking the supremum on the two differences between the empirical and expected risk yields that

$$\begin{align}
\mathcal{R}^D(\hat{f}^D) - \mathcal{R}(\mathbf{f}^*) \leq \sup_{g \in \mathcal{F}_\Theta} \Big[\mathcal{R}^D(g) - \mathcal{R}_n^D(g)\Big] + \sup_{g \in \mathcal{F}_\Theta} \Big[\mathcal{R}(g) - \mathcal{R}_n(g)\Big] \tag{33} \\
+ \mathcal{R}_n^D(f) - \mathcal{R}_n(f) + \mathcal{R}(f) - \mathcal{R}(\mathbf{f}^\star). \tag{34}
\end{align}$$

This concludes the proof.

C.2. Proof of Proposition 6.1

We follow the proof strategy of Chen et al. (2019a). Starting from Theorem A.8, one can see that since

$$\begin{align}
\sup_x \|f_{\theta_1}(x) - f_{\theta_2}(x)\| \leq L_1 \|\Phi_1 - \Phi_2\| + L_2 \|\mathbf{A}_1 - \mathbf{A}_2\| + L_3 \|\mathbf{b}_1 - \mathbf{b}_2\| \tag{35} \\
+ L_4 \|\mathbf{U}_1 - \mathbf{U}_2\| + L_5 \|v_1 - v_2\|, \tag{36}
\end{align}$$

where the supremum is taken on all x such that $\|x\|_{\text{1-var},[0,1]} \leq L_x$ and $\|x_0\| \leq B_x$, it is sufficient to have $\hat{\theta} = (\hat{\Phi}, \hat{\mathbf{A}}, \hat{\mathbf{b}}, \hat{\mathbf{U}}, \hat{v})$ such that

$$\left\|\Phi - \hat{\Phi}\right\| \leq \frac{\beta}{5L_1}, \quad \left\|\mathbf{A} - \hat{\mathbf{A}}\right\| \leq \frac{\beta}{5L_2}, \quad \left\|\mathbf{b} - \hat{\mathbf{b}}\right\| \leq \frac{\beta}{5L_3}, \quad \left\|\mathbf{U} - \hat{\mathbf{U}}\right\| \leq \frac{\beta}{5L_4} \tag{37}$$

and

$$\|v - \hat{v}\| \leq \frac{\beta}{5L_5} \tag{38}$$

to obtain an β covering of $\mathcal{F}_\Theta$, since in this case we get that

$$\sup_x \left\|f_{\theta_1}(x) - f_{\hat{\theta}}(x)\right\| \leq \beta. \tag{39}$$

Using Lemma A.10 and denoting by $\mathcal{G}_\Phi = \{h : z \mapsto \Phi^\top z,\ \|\Phi\| \leq B_\alpha\}$, and using the corresponding definitions for $\mathcal{G}_{\mathbf{A}}$, $\mathcal{G}_{\mathbf{b}}$, $\mathcal{G}_{\mathbf{U}}$ and $\mathcal{G}_v$, we get that

$$\mathcal{N}\Big(\mathcal{G}_\Phi, \frac{\beta}{5L_1}\Big) \leq \left(1 + \frac{10\sqrt{p}B_\alpha L_1}{\beta}\right)^p, \mathcal{N}\Big(\mathcal{G}_{\mathbf{A}}, \frac{\beta}{5L_2}\Big) \leq \left(1 + \frac{10\sqrt{dp}B_{\mathbf{A}} L_2}{\beta}\right)^{dp^2} \tag{40}$$

$$\mathcal{N}\Big(\mathcal{G}_{\mathbf{b}}, \frac{\beta}{5L_3}\Big) \leq \left(1 + \frac{10\min\{\sqrt{p},\sqrt{d}\}B_{\mathbf{b}} L_3}{\beta}\right)^{dp}, \tag{41}$$

$$\mathcal{N}\Big(\mathcal{G}_{\mathbf{U}}, \frac{\beta}{5L_4}\Big) \leq \left(1 + \frac{10\min\{\sqrt{p},\sqrt{d}\}B_{\mathbf{U}} L_4}{\beta}\right)^{dp} \tag{42}$$

and

$$\mathcal{N}\Big(\mathcal{G}_v, \frac{\beta}{5L_5}\Big) \leq \left(1 + \frac{10\sqrt{p}B_v L_5}{\beta}\right)^p. \tag{43}$$

The covering number of $\mathcal{F}_\Theta$ is obtained by multiplying the covering number of each functional class (Chen et al., 2019a). Defining

$$K_1 := \max\{B_\alpha L_1, B_v L_5\}, \tag{44}$$
$$K_2 := \max\{B_{\mathbf{b}} L_3, B_{\mathbf{U}} L_4\}, \tag{45}$$
$$K_3 := \max\{B_{\mathbf{b}} L_3, B_{\mathbf{U}} L_4, B_{\mathbf{A}} L_2\} \tag{46}$$

and using the fact that $\min\{\sqrt{d},\sqrt{p}\} \leq \sqrt{dp}$, we finally get that

$$\mathcal{N}(\mathcal{F}_\Theta, \beta) \leq \left(1 + \frac{10\sqrt{p}K_1}{\beta}\right)^{2p}\left(1 + \frac{10\min\{\sqrt{p},\sqrt{d}\}K_2}{\beta}\right)^{2dp}\left(1 + \frac{10\sqrt{dp}B_{\mathbf{A}} L_2}{\beta}\right)^{dp^2} \tag{47}$$

$$\leq \left(1 + \frac{10\sqrt{p}K_1}{\beta}\right)^{2p}\left(1 + \frac{10\sqrt{dp}K_3}{\beta}\right)^{dp(2+p)}. \tag{48}$$

The proof is identical for inputs $\mathbf{x}^D$. This concludes the proof.

C.3. Proof of Proposition 6.2

We use the following Lemma from Bartlett et al. (2017).

Lemma C.1. *Let $\mathcal{H}$ be a class of real-valued functions taking values in $[-M, M]$, for $M > 0$, and assume that $0 \in \mathcal{H}$. Then the empirical Rademacher complexity associated to a sample of n datapoints verifies*

$$\text{Rad}(\mathcal{H}) \leq \inf_{\beta>0}\left[\frac{4\beta}{\sqrt{n}} + \frac{12}{n}\int_\beta^{2M\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{H}, \beta)}d\beta\right] \tag{49}$$

We apply this Lemma to the class $\mathcal{F}_\Theta$. We trivially have that $\mathbf{0} \in \mathcal{F}$, since this function is recovered by taking $\theta = \mathbf{0}$. By Lemma 3.7, the value of f_θ is bounded by

$$M_\Theta = B_\alpha L_\sigma \exp(L_{\mathbf{A}} L_\sigma L_x)\Big(B_{\mathbf{U}} B_x + B_v + B_{\mathbf{b}} L_x\Big). \tag{50}$$

In our setup, we get, from Proposition 6.1, that

$$\int_{\beta}^{2M_\Theta\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{F}_\Theta, \beta)} d\beta \tag{51}$$

$$\leq \int_{\beta}^{2M_\Theta\sqrt{n}} \sqrt{2p\log\left(1+\frac{10\sqrt{p}K_1}{\beta}\right)+dp(2+p)\log\left(1+\frac{10\sqrt{dp}K_2}{\beta}\right)} d\beta \tag{52}$$

$$\leq 2M_\Theta\sqrt{n}\sqrt{2p\log\left(1+\frac{10\sqrt{p}K_1}{\beta}\right)+dp(2+p)\log\left(1+\frac{10\sqrt{dp}K_2}{\beta}\right)}. \tag{53}$$

Since for $x>1$, $\log(1+x)\leq\log(2x)$, we have, for β small enough to ensure that both

$$\frac{10\sqrt{dp}K_2}{\beta} \geq \frac{10\sqrt{p}K_1}{\beta} > 1$$

so that

$$\int_{\beta}^{2M_\Theta\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{F}_\Theta, \beta)} d\beta \tag{54}$$

$$\leq 2M_\Theta\sqrt{n}\sqrt{2p\log\left(\frac{20\sqrt{p}K_1}{\beta}\right)+dp(p+2)\log\left(\frac{20\sqrt{dp}K_2}{\beta}\right)} \tag{55}$$

Taking $\beta=\frac{1}{\sqrt{n}}$, one gets

$$\mathrm{Rad}(\mathcal{F}_\Theta)\leq\frac{4}{n}+\frac{24M_\Theta}{\sqrt{n}}\sqrt{2p\log 20\sqrt{np}K_1+dp(p+2)\log 20\sqrt{ndp}K_2}. \tag{56}$$

yielding the same upper bound on the complexity.

C.4. Proof of Theorems 4.2 and 4.3

The first part of the theorem is a straightforward application of Mohri et al. (2018), Theorem 11.3. Since our loss is L_ℓ-Lipschitz and bounded by M_ℓ, this gets us immediately that with probability at least $1-\delta$, one has

$$\mathcal{R}^D(\hat{f}^D)\leq\mathcal{R}_n(\hat{f}^D)+\frac{24M_\Theta L_\ell}{\sqrt{n}}\sqrt{2pU_1+dp(p+2)U_2}+M_\ell\sqrt{\frac{\log 1/\delta}{2n}} \tag{57}$$

with $U_1:=\log 20\sqrt{np}K_1$ and $U_2:=\log 20\sqrt{ndp}K_2$.

Note that this also gets us

$$\mathcal{R}(\hat{f})\leq\mathcal{R}_n(\hat{f})+\frac{24M_\Theta L_\ell}{\sqrt{n}}\sqrt{2pU_1+dp(p+2)U_2}+M_\ell\sqrt{\frac{\log 1/\delta}{2n}}, \tag{58}$$

with

$$\hat{f}\in\arg\min_{\theta\in\Theta}\mathcal{R}_n(f_\theta)$$

We now turn to the generalization gap. Using the generalization gap decomposition given in Lemma 4.1, it is clear that for any predictor f_θ parametrized by $\theta=(\Phi,\psi,\mathbf{U},v)$ such that $\Phi=\alpha$,

$$\begin{aligned}
&\mathcal{R}^D(\hat{f}^D) - \mathcal{R}(\mathbf{f}^*) && (59)\\
&\leq \sup_{g \in \mathcal{F}_\Theta} \Big[\mathcal{R}^D(g) - \mathcal{R}_n^D(g)\Big] + \sup_{g \in \mathcal{F}_\Theta} \Big[\mathcal{R}(g) - \mathcal{R}_n(g)\Big] && (60)\\
&\quad + L_\ell C_2\big(\Theta, L_x, L_\sigma\big)|D| && (61)\\
&\quad + L_\ell B_\alpha \exp(L_{\mathbf{G}^\star} L_x)\Big[L_x \left\|\mathbf{G}_\psi - \mathbf{G}^\star\right\|_{\infty, \Omega(\Theta, L_x, B_x)} + \left\|\varphi - \mathbf{NN}_{\mathbf{U},v}\right\|_{\infty, B_x}\Big]. && (62)
\end{aligned}$$

Now, by a classical symmetrization argument - see for instance Bach (2021), Proposition 4.2 - the obtained bounds on the Rademacher complexity imply that

$$\sup_{g \in \mathcal{F}_\Theta} \Big[\mathcal{R}(g) - \mathcal{R}_n(g)\Big] \leq 2\mathrm{Rad}(\mathcal{F}_\Theta) \tag{63}$$

and

$$\sup_{g \in \mathcal{F}_\Theta} \Big[\mathcal{R}^D(g) - \mathcal{R}_n^D(g)\Big] \leq 2\mathrm{Rad}(\mathcal{F}_\Theta). \tag{64}$$

Combining all these elements gets us the final result.

D. Supplementary proofs

D.1. Proof of Lemma 3.7

We recall that $\|W\|_{\text{op}} := \max_{\|x\|=1} \|Wx\|$, and that $\|W\|_{\text{op}} \leq \|W\|$, which means that $\|Wx\| \leq \|W\| \, \|x\|$ for all x. We first prove this result for a general CDE

$$dz_t = \mathbf{F}(z_t)dx_t$$

with initial condition $z_0 \in \mathbb{R}^p$, where the driving path x is supposed to be continuous of bounded variation (or piecewise constant with a finite number of discontinuities.)

By definition,

$$z_t = z_0 + \int_0^t \mathbf{F}(z_t)dx_t. \tag{65}$$

Taking norms, this yields

$$\|z_t\| \leq \|z_0\| + \int_0^t \|\mathbf{F}(z_t)\|_{\text{op}} \|dx_t\| . \tag{66}$$

Notice that since we have assumed $\mathbf{F}$ to be Lipschitz, one has for all $z \in \mathbb{R}^p$

$$\begin{aligned} \|\mathbf{F}(z)\|_{\text{op}} &\leq \|\mathbf{F}(z) - \mathbf{F}(0)\|_{\text{op}} + \|\mathbf{F}(0)\|_{\text{op}} && (67)\\ &\leq \|\mathbf{F}(z) - \mathbf{F}(0)\| + \|\mathbf{F}(0)\|_{\text{op}} && (68)\\ &\leq L_{\mathbf{F}} \|z\| + \|\mathbf{F}(0)\|_{\text{op}} , && (69) \end{aligned}$$

where the last inequality follows from the fact that $\mathbf{F}$ is Lipschitz. It follows that

$$\|z_t\| \leq \|z_0\| + \int_0^t (L_{\mathbf{F}} \|z_t\| + \|\mathbf{F}(0)\|_{\text{op}}) \|dx_t\| . \tag{70}$$

Using the fact that $\int_0^t \|dx_s\| = \|x\|_{\text{1-var},[0,t]}$, one gets

$$\|z_t\| \leq \|z_0\| + \|\mathbf{F}(0)\|_{\text{op}} \|x\|_{\text{1-var},[0,t]} + L_{\mathbf{F}} \int_0^t \|z_s\| \|dx_s\| . \tag{71}$$

Applying Gronwall's Lemma for CDEs yields

$$\|z_t\| \leq \Big(\|z_0\| + \|\mathbf{F}(0)\|_{\text{op}} \|x\|_{\text{1-var},[0,t]} \Big) \exp\big(L_{\mathbf{F}} \|x\|_{\text{1-var},[0,t]}\big). \tag{72}$$

Now, turning to the generative CDE (1), we obtain as a consequence that

$$|y| \leq B_\alpha \Big(B_\varphi + \|\mathbf{G}^\star(0)\|_{\text{op}} L_x \Big) \exp\big(L_{\mathbf{G}^\star} L_x\big) + M_\varepsilon, \tag{73}$$

since y is the sum of a linear transformation of the endpoint of a CDE an a noise term ε bounded by M_ε.

Turning now to $f_\theta \in \mathcal{F}_\Theta$, one has that

$$|f_\theta(x)| = |\Phi^\top z_1| \leq \|\Phi\| \, \|z_1\| \leq B_\alpha \|z_1\| \tag{74}$$

and since z_1 is the solution of a NCDE, we can directly leverage the previous result to bound $\|z_1\|$. From the definition of $\mathcal{F}_\Theta$, it follows that

$$\|z_0\| \leq L_\sigma B_{\mathbf{U}} B_x + L_\sigma B_v, \ \|\mathbf{F}(0)\|_{\text{op}} \leq L_\sigma B_{\mathbf{b}} \text{ and } L_{\mathbf{F}} \leq L_\sigma B_{\mathbf{A}}. \tag{75}$$

As a direct consequence, one has

$$|f_\theta(x)| \leq B_\alpha L_\sigma \exp(L_{\mathbf{A}} L_\sigma L_x)\Big(B_{\mathbf{U}} B_x + B_v + B_{\mathbf{b}} L_x\Big). \tag{76}$$

We now turn to $f_\theta(\mathbf{x}^D)$. Bounding the value of z_1^D can be done by applying this lemma to the values of z^D at points in D, since it is constant between those points. The proof leverages the discrete version of Gronwall's Lemma, stated in Lemma A.5. One has

$$\|z_{t_k}\| \leq \|z_0\| + \sum_{l=0}^{k-1} \left\|\sigma\big(\mathbf{A} z_{t_l} + \mathbf{b}\big)\Delta x_{t_{l+1}}\right\| \leq \sum_{l=0}^{k-1} \left\|\sigma\big(\mathbf{A} z_{t_l} + \mathbf{b}\big)\right\| \left\|\Delta x_{t_{l+1}}\right\|, \tag{77}$$

for all $k = 1, \dots, K$. Now, since

$$\left\|\sigma\big(\mathbf{A} z_{t_l} + \mathbf{b}\big)\right\| \leq L_\sigma B_{\mathbf{A}} \|z_l\| + L_\sigma B_{\mathbf{b}} \tag{78}$$

one has

$$\|z_{t_k}\| \leq \|z_0\| + L_\sigma B_{\mathbf{b}} \sum_{l=0}^{k-1} \left\|\Delta x_{t_{l+1}}\right\| + L_\sigma B_{\mathbf{A}} \sum_{l=0}^{k-1} \|z_l\| \|\Delta x_{l+1}\| \tag{79}$$

$$\leq \|z_0\| + L_\sigma B_{\mathbf{b}} L_x + L_\sigma B_{\mathbf{A}} \sum_{l=0}^{k-1} \|z_l\| \|\Delta x_{l+1}\| \tag{80}$$

and one gets by the discrete Gronwall Lemma that

$$\|z_{t_k}\| \leq \big(\|z_0\| + L_\sigma B_{\mathbf{b}} L_x\big) \prod_{l=0}^{k-1} \Big(1 + L_\sigma B_{\mathbf{A}} \left\|\Delta x_{t_{l+1}}\right\|\Big) \tag{81}$$

$$\leq \big(\|z_0\| + L_\sigma B_{\mathbf{b}} L_x\big) \exp(L_\sigma B_{\mathbf{A}} \|x\|_{\text{1-var},[0,1]}) \tag{82}$$

$$\leq L_\sigma \exp(L_{\mathbf{A}} L_\sigma L_x)\Big(B_{\mathbf{U}} B_x + B_v + B_{\mathbf{b}} L_x\Big). \tag{83}$$

From this we get that

$$|f_\theta(\mathbf{x}^D)| \leq B_\alpha L_\sigma \exp(L_{\mathbf{A}} L_\sigma L_x)\Big(B_{\mathbf{U}} B_x + B_v + B_{\mathbf{b}} L_x\Big) \tag{84}$$

This concludes the proof.

D.2. Proof of Proposition A.6

We first consider a general CDE

$$dz_t = \mathbf{F}(z_t) dx_t \tag{85}$$

with initial condition $z_0 \in \mathbb{R}^p$. Take $s, r \in [0, t]$. By definition,

$$\|z_r - z_s\| = \left\|\int_s^r \mathbf{F}(z_u) dx_u\right\| \tag{86}$$

$$= \left\|\int_s^r \mathbf{F}\big(z_u - z_s + z_s\big) dx_u\right\| \tag{87}$$

$$\leq \int_s^r \big(L_{\mathbf{F}} \|z_u - z_s + z_s\| + \|\mathbf{F}(0)\|_{\text{op}}\big) \|dx_u\| \tag{88}$$

$$\leq \int_s^r \big(L_{\mathbf{F}} \|z_u - z_s\| + L_{\mathbf{F}} \|z_s\| + \|\mathbf{F}(0)\|_{\text{op}}\big) \|dx_u\| \tag{89}$$

$$= \big(L_{\mathbf{F}} \|z_s\| + \|\mathbf{F}(0)\|_{\text{op}}\big) \|x\|_{\text{1-var},[s,r]} + L_{\mathbf{F}} \int_s^r \|z_u - z_s\| \|dx_u\|. \tag{90}$$

Now since z_s is the solution of a CDE evaluated at s, it can be bounded by Lemma 3.7 by

$$\|z_s\| \leq \big(\|z_0\| + \|\mathbf{F}(0)\|_{\text{op}} \|x\|_{\text{1-var},[0,s]}\big) \exp(L_{\mathbf{F}} \|x\|_{\text{1-var},[0,s]}). \tag{91}$$

This means that

$$\|z_r - z_s\| \tag{92}$$

$$\leq \Big[L_{\mathbf{F}}\big(\|z_0\| + \|\mathbf{F}(0)\|_{\text{op}} \|x\|_{\text{1-var},[0,s]}\big) \exp(L_{\mathbf{F}} \|x\|_{\text{1-var},[0,s]}) + \|\mathbf{F}(0)\|_{\text{op}}\Big] \|x\|_{\text{1-var},[s,r]} \tag{93}$$

$$+ L_{\mathbf{F}} \int_s^r \|z_u - z_s\| \|dx_u\|. \tag{94}$$

We may now use Gronwall's Lemma and the fact that $\|x\|_{\text{1-var},[0,s]} \leq L_x$ for all $s \in [0,1]$ to obtain that

$$\|z_r - z_s\| \leq \left[L_{\mathbf{F}}\big(\|z_0\| + \|\mathbf{F}(0)\|_{\text{op}} L_x\big) \exp(L_{\mathbf{F}} L_x) + \|\mathbf{F}(0)\|_{\text{op}}\right] \exp(L_{\mathbf{F}} L_x) \|x\|_{\text{1-var},[s,r]}. \tag{95}$$

and thus

$$\|z_r - z_s\| \leq C_1(L_{\mathbf{F}}, \mathbf{F}, L_x) \|x\|_{\text{1-var},[s,r]}. \tag{96}$$

This means that the variations of z on arbitrary intervals $[r,s]$ are bounded by the variations of x on the same interval, times an interval independent constant.

From this, we may immediately conclude that

$$\|z\|_{\text{1-var},[0,t]} \leq C_1(L_{\mathbf{F}}, \mathbf{F}, L_x) \|x\|_{\text{1-var},[0,t]}, \tag{97}$$

which concludes the proof for a general CDE.

We now turn to $f_\theta \in \mathcal{F}_\Theta$. The previous result allows to bound the total variation of a CDE and thus of the trajectory of the latent state. Using this proposition with

$$C_1(L_{\mathbf{G}_\psi}, \mathbf{G}_\psi, L_x) := \left[L_\sigma B_{\mathbf{A}} L_\sigma \big(B_{\mathbf{U}} B_x + B_v + B_{\mathbf{b}} L_x\big) \exp(L_\sigma B_{\mathbf{A}} L_x) + L_\sigma B_{\mathbf{b}}\right] \exp(L_\sigma B_{\mathbf{A}} L_x)$$

yields

$$\|z\|_{\text{1-var}} \leq C_1(L_{\mathbf{G}_\psi}, \mathbf{G}_\psi, L_x) L_x. \tag{98}$$

This concludes the proof for f_θ with continuous input. For a discrete input, the proof directly transfers by remarking that

$$\left\|z^D\right\|_{\text{1-var},[0,1]} = \sum_{k=1}^{K-1} \left\|z^D_{t_k} - z^D_{t_{k-1}}\right\| \tag{99}$$

and by using the discrete Gronwall Lemma.

D.3. Proof of Theorem A.7

For all $t \in [0, 1]$, one has the decomposition

$$w_t - v_t = w_0 - v_0 + \int_0^t \mathbf{F}(w_s) dx_s - \int_0^t \mathbf{G}(v_s) dr_s \tag{100}$$

$$= w_0 - v_0 + \int_0^t \big(\mathbf{F}(w_s) - \mathbf{F}(v_s)\big) dx_s + \int_0^t \mathbf{F}(v_s) dx_s - \int_0^t \mathbf{G}(v_s) dr_s \tag{101}$$

$$= w_0 - v_0 \tag{102}$$

$$+ \int_0^t \big(\mathbf{F}(w_s) - \mathbf{F}(v_s)\big) dx_s \tag{103}$$

$$+ \int_0^t \mathbf{F}(v_s) d(x_s - r_s) \tag{104}$$

$$+ \int_0^t \big(\mathbf{F}(v_s) - \mathbf{G}(v_s)\big) dr_s. \tag{105}$$

We control every one of these terms separately and conclude by applying Gronwall's Lemma. Writing

$$A := \left\| \int_0^t \big(\mathbf{F}(w_s) - \mathbf{F}(v_s)\big) dx_s \right\|,$$

$$B := \left\| \int_0^t \mathbf{F}(v_s) d(x_s - r_s) \right\|,$$

$$C := \left\| \int_0^t \big(\mathbf{F}(v_s) - \mathbf{G}(v_s)\big) dr_s \right\|$$

it is clear that

$$\|w_t - z_t\| \leq \|w_0 - z_0\| + A + B + C. \tag{106}$$

Control of the term A. One has

$$A \leq \int_0^t \|\mathbf{F}(w_s) - \mathbf{F}(v_s)\|_{\text{op}} \|dx_s\|. \tag{107}$$

Since F is $L_{\mathbf{F}}$-Lipschitz, this gives

$$A \leq L_{\mathbf{F}} \int_0^t \|w_s - v_s\| \|dx_s\|. \tag{108}$$

Control of the term B. One gets using integration by parts

$$\int_0^t \mathbf{F}(v_s) d(x_s - r_s) = (x_t - r_t) - (x_0 - r_0) - \int_0^t (x_s - r_s)^\top d\mathbf{F}(v_s). \tag{109}$$

This gives

$$B \leq \|x_0 - r_0\| + \|x - r\|_{\infty,[0,t]} (1 + \|\mathbf{F}(v)\|_{\text{1-var},[0,t]}). \tag{110}$$

Since $\mathbf{F}$ is $L_{\mathbf{F}}$-Lipschitz, one gets

$$\|\mathbf{F}(v)\|_{\text{1-var},[0,t]} \leq L_{\mathbf{F}} \|v\|_{\text{1-var},[0,t]}. \tag{111}$$

Since v is the solution to a CDE, we can resort to Proposition A.6 to bound its total variation. This yields

$$\|v\|_{\text{1-var},[0,t]} \leq C_1(L_{\mathbf{F}}, \mathbf{F}, L_r) \|r\|_{\text{1-var},[0,t]} \tag{112}$$

and finally

$$B \leq \|x_0 - r_0\| + \|x - r\|_{\infty,[0,t]} (1 + L_{\mathbf{F}} C_1(L_{\mathbf{F}}, \mathbf{F}, L_x) \|r\|_{\text{1-var},[0,t]}). \tag{113}$$

Control of the term C. Finally, we get that

$$C \leq \int_0^t \|\mathbf{F}(v_s) - \mathbf{G}(v_s)\|_{\text{op}} \|dr_s\| \tag{114}$$

$$\leq \max_{v \in \Omega(\mathbf{G})} \|\mathbf{F}(v) - \mathbf{G}(v)\|_{\text{op}} L_r \tag{115}$$

where we recall that $\Omega(\mathbf{G})$ is the closed ball

$$\Omega(\mathbf{G}) = \big\{u \in \mathbb{R}^p \mid \|u\| \leq (\|v_0\| + \|\mathbf{G}(0)\| L_r) \exp(L_{\mathbf{G}} L_r)\big\}.$$

Indeed, since v is the solution of a CDE, its norm at any time $t \in [0,1]$ is bounded as stated in Lemma 3.7. One can therefor bound the difference $\|\mathbf{F}(v_s) - \mathbf{G}(v_s)\|_{\text{op}}$ by considering all possible values of v.

Putting everything together. Combining the obtained bounds on all terms, one is left with

$$\|w_t - v_t\| \leq \|w_0 - v_0\| \tag{116}$$

$$+ \|x_0 - r_0\| + \|x - r\|_{\infty,[0,t]} (1 + L_{\mathbf{F}} L_r C_1(L_{\mathbf{F}}, \mathbf{F}, L_x)) \tag{117}$$

$$+ L_r \max_{v \in \Omega(\mathbf{G})} \|\mathbf{F}(v) - \mathbf{G}(v)\|_{\text{op}} \tag{118}$$

$$+ L_{\mathbf{F}} \int_0^t \|w_s - v_s\| \|dx_s\|. \tag{119}$$

The proof is concluded by using Gronwall's Lemma A.4. If the path (x_t) is continuous, we resort to its continuous version. If it is piecewise constant, we use its discrete version. The path $(\|w_t - v_t\|)_t$ only needs to be measurable and bounded. There are no assumptions on its continuity. This finally yields

$$\|w_t - v_t\| \leq \Bigg(\|w_0 - v_0\| + \|x_0 - r_0\| \tag{120}$$

$$+ \|x - r\|_{\infty,[0,t]} \big(1 + L_{\mathbf{F}} L_r C_1(L_{\mathbf{F}}, \mathbf{F}, L_x)\big) + \max_{v \in \Omega(\mathbf{G})} \|\mathbf{F}(v) - \mathbf{G}(v)\|_{\text{op}} L_r \Bigg) \tag{121}$$

$$\times \exp(L_{\mathbf{F}} L_x). \tag{122}$$

Now, notice that in our proof, the two CDEs are exchangeable. This means that we immediately get the alternative bound

$$\|w_t - v_t\| \leq \Bigg(\|w_0 - v_0\| + \|x_0 - r_0\| \tag{123}$$

$$+ \|x - r\|_{\infty,[0,t]} \big(1 + L_{\mathbf{G}} L_x C_1(L_{\mathbf{G}}, \mathbf{G}, L_r)\big) + \max_{v \in \Omega(\mathbf{F})} \|\mathbf{F}(v) - \mathbf{G}(v)\|_{\text{op}} L_x \Bigg) \tag{124}$$

$$\times \exp(L_{\mathbf{G}} L_r) \tag{125}$$

where we recall that

$$\Omega(\mathbf{F}) = \big\{u \in \mathbb{R}^p \mid \|u\| \leq (\|w_0\| + \|\mathbf{F}(0)\| L_x) \exp(L_{\mathbf{F}} L_x)\big\}.$$

D.4. Proof of Theorem A.8

One has

$$\|f_{\theta_1}(x) - f_{\theta_2}(x)\| = \left\|\Phi_1^\top z_1^1 - \Phi_2^\top z_1^2\right\| \leq \|\Phi_1 - \Phi_2\| \left\|z_1^1\right\| + B_\alpha \left\|z_1^1 - z_1^2\right\|. \tag{126}$$

Using Theorem A.7, one gets that

$$\left\|z_1^1 - z_1^2\right\| \leq \exp(L_\sigma B_{\mathbf{A}} L_x)\Big(\left\|z_0^1 - z_0^2\right\| + L_x \|\mathbf{G_1} - \mathbf{G}_2\|_{\infty,\Omega} \Big), \tag{127}$$

where

$$\Omega \subset \left\{ u \in \mathbb{R}^p \mid \|u\| \leq (L_\sigma(B_{\mathbf{U}} B_x + B_v) + L_\sigma B_{\mathbf{b}} L_x) \exp(L_\sigma L_{\mathbf{A}} L_x) \right\}.$$

We first bound

$$\left\| z_0^1 - z_0^2 \right\| \leq L_\sigma \left\| \mathbf{U}_1 - \mathbf{U}_2 \right\| B_x + L_\sigma \left\| v_1 - v_2 \right\|. \tag{128}$$

One also has that

$$\|\mathbf{G}_1 - \mathbf{G}_2\|_{\infty,\Omega} \leq L_\sigma \left\| \mathbf{A}_1 - \mathbf{A}_2 \right\| \times \max_{u \in \Omega} \|u\| + L_\sigma \left\| \mathbf{b}_1 - \mathbf{b}_2 \right\| \tag{129}$$

It follows from the inclusion of Ω that

$$\max_{u \in \Omega} \|u\| \leq (L_\sigma(B_{\mathbf{U}} B_x + B_v) + L_\sigma B_{\mathbf{b}} L_x) \exp(L_\sigma L_{\mathbf{A}} L_x). \tag{130}$$

Since z_1^1 is bounded as the endpoint of a NCDE, one gets using Theorem A.7 that

$$\|\Phi_1 - \Phi_2\| \left\| z_1^1 \right\| \leq L_\sigma \Big[B_{\mathbf{U}} B_x + B_v + (B_{\mathbf{A}} + B_{\mathbf{b}}) L \Big] \exp(L_\sigma L_x B_{\mathbf{A}}) \left\| \Phi_1 - \Phi_2 \right\| \tag{131}$$

Putting everything together, one gets that

$$\|f_1(x) - f_2(x)\| \leq L_1 \left\| \Phi_1 - \Phi_2 \right\| \tag{132}$$

$$+ L_2 \left\| \mathbf{A}_1 - \mathbf{A}_2 \right\| + L_3 \left\| \mathbf{b}_1 - \mathbf{b}_2 \right\| \tag{133}$$

$$+ L_4 \left\| \mathbf{U}_1 - \mathbf{U}_2 \right\| + L_5 \left\| v_1 - v_2 \right\| \tag{134}$$

with

$$L_1 := L_\sigma \Big[B_{\mathbf{U}} B_x + B_v + (B_{\mathbf{A}} + B_{\mathbf{b}}) L \Big] \exp(L_\sigma L_x B_{\mathbf{A}}), \tag{135}$$

$$L_2 := B_\alpha \exp(2 L_\sigma B_{\mathbf{A}} L_x) L_x L_\sigma (L_\sigma(B_{\mathbf{U}} B_x + B_v) + L_\sigma B_{\mathbf{b}} L_x), \tag{136}$$

$$L_3 := B_\alpha \exp(L_\sigma B_{\mathbf{A}} L_x) L_x L_\sigma \tag{137}$$

$$L_4 := B_\alpha B_x \exp(L_\sigma B_{\mathbf{A}} L_x) L_\sigma \tag{138}$$

$$L_5 := B_\alpha \exp(L_\sigma B_{\mathbf{A}} L_x) L_\sigma. \tag{139}$$

which concludes our proof. The proof for $|f_{\theta_1}(\mathbf{x}^D) - f_{\theta_2}(\mathbf{x}^D)|$ is identical.