ROUGH TRANSFORMERS FOR CONTINUOUS AND EFFICIENT TIME-SERIES MODELLING

Fernando Moreno-Pino[1,*,†], **Álvaro Arroyo**[1,*], **Harrison Waldon**[1,*], **Xiaowen Dong**[1,2], **Álvaro Cartea**[1,3]

[1] Oxford-Man Institute of Quantitative Finance, University of Oxford

[2] Machine Learning Research Group, University of Oxford

[3] Mathematical Institute, University of Oxford

ABSTRACT

Time-series data in real-world medical settings typically exhibit long-range dependencies and are observed at non-uniform intervals. In such contexts, traditional sequence-based recurrent models struggle. To overcome this, researchers replace recurrent architectures with Neural ODE-based models to model irregularly sampled data and use Transformer-based architectures to account for long-range dependencies. Despite the success of these two approaches, both incur very high computational costs for input sequences of moderate lengths and greater. To mitigate this, we introduce the Rough Transformer, a variation of the Transformer model which operates on continuous-time representations of input sequences and incurs significantly reduced computational costs, critical for addressing long-range dependencies common in medical contexts. In particular, we propose *multi-view signature attention*, which uses path signatures to augment vanilla attention and to capture both local and global dependencies in input data, while remaining robust to changes in the sequence length and sampling frequency. We find that Rough Transformers consistently outperform their vanilla attention counterparts while obtaining the benefits of Neural ODE-based models using a fraction of the computational time and memory resources on synthetic and real-world time-series tasks.

1 INTRODUCTION

Real-world sequential medical data in areas such as Human Activity Recognition (Chen et al., 2021), suicide risk monitoring (Porras-Segovia et al., 2022), or activity detection in atrial fibrillation electrograms (Ríos-Muñoz et al., 2020) often are irregularly sampled, of variable length, and exhibit long-range dependencies. Furthermore, these data, which may be drawn from EEG readings (Vahid et al., 2020) or wearable devices (Moreno-Pino et al., 2022), are often sampled at high frequency for long periods of time, yielding long sequences of data. Hence, many popular machine learning models struggle to model real-world medical sequential data, due to input dimension inflexibility, memory constraints, and computational bottlenecks. Rather than treating these data as *discrete* sequences, effective theoretical models often assume data are generated from some underlying *continuous-time* process (Morariu-Patrichi & Pakkanen, 2022; Ratcliff et al., 2016). Hence, there is an increased interest in developing machine learning methods which analyze sequential medical data using *continuous-time* representations.

One recent approach to modeling continuous-time data involves the development of continuous-time analogues of standard deep learning models, such as Neural ODEs (Chen et al., 2018) and Neural CDEs (Kidger et al., 2020), which extend ResNets (He et al., 2016) and RNNs (Funahashi & Nakamura, 1993), respectively, to continuous-time settings.

Instead of processing discrete data directly, these models operate on a latent continuous-time representation of input sequences. This approach is successful in continuous-time modeling tasks where standard deep recurrent models fail. In particular, extensions of vanilla Neural ODEs to the time-series setting (Rubanova et al., 2019; Kidger et al., 2020) succeed in various domains such as

[*]Equal contribution.

[†]Corresponding author. Email: `fernando.moreno-pino@eng.ox.ac.uk`.

adaptive uncertainty quantification (Norcliffe et al., 2021), counterfactual inference (Seedat et al., 2022), or generative modeling (Calvo-Ordonez et al., 2023).

In many practical medical settings, such as human behaviour changes detection (Berrouiguet et al., 2018), sleep activity recognition (Martínez-García et al., 2023), or heart rate fluctuations (Hausdorff & Peng, 1996), continuous-time data also exhibit long-range dependencies. That is, data from the distant past may impact the system's current behavior. Deep recurrent models struggle in this setting due to vanishing gradients. Several recent works (Melnychuk et al., 2022; Nguyen & Grover, 2022; Moreno-Pino et al., 2023) successfully extract long-range dependencies from sequential data with Transformers (Vaswani et al., 2017), which learn temporal dependencies of a tokenized representation of input sequences. As the attention mechanism is permutation invariant, extracting such temporal dependencies requires a positional embedding of input data which projects data into some latent space. The parallelizable nature of the Transformer allows for rapid training and evaluation on sequences of moderate length and has contributed to its success in fields such as natural language processing (NLP).

While the above approaches succeed in certain settings, several limitations hinder their wider applications. On the one hand, Neural ODEs and their analogues (Rubanova et al., 2019; Kidger et al., 2020) bear substantial computational costs when modeling long sequences of high dimension (see Morrill et al., 2021). On the other hand, Transformers operate on discrete-time representations of input sequences, whose relative ordering is represented by the positional embedding. This representation may inhibit their expressivity in continuous-time data modeling tasks (Zeng et al., 2023). Moreover, Transformer-based models suffer from a number of difficulties, including (i) input sequences must be sampled at the same times, (ii) the sequence length must be fixed, and (iii) the computational cost scales quadratically in the length of the input sequence. These difficulties severely limit the application of Transformers to continuous-time data modeling.

2 METHODS

Rough Transformer. We propose the Rough Transformer, a Transformer-based architecture that operates on a continuous-time representation of sequential data. We make use of the the path signature from rough path theory. For a smooth function $f : \mathbb{R}^+ \to \mathbb{R}^d$, the signature $S(f)_{s,t}$ is defined as the infinite sequence of iterated integrals

$$S(f)_{s,t} := \left(1, \int_s^t f'(u)\, du, \iint_{s<u_1<u_2<t} f'(u_1) \otimes f'(u_2)\, du_1\, du_2, ...\right) , \tag{1}$$

for any $s, t \in \mathbb{R}^+$, where $\otimes$ denotes the tensor product; see Appendix A for details on the path signature.

Let $\mathcal{D}$ be a dataset of irregularly sampled time-series. To project a time-series $\mathbf{X} \in \mathcal{D}$ to a continuous-time object, we first let $\tilde{X}$ denote its piecewise-linear interpolation. Fixing a set of times $\mathcal{T} = \{t_1, ..., t_{\overline{L}}\}$, we define the *multi-view signature* for any $t_k \in \mathcal{T}$

$$M(\mathbf{X})_k := \left(S(\tilde{X})_{0,t_k}, S(\tilde{X})_{t_{k-1},t_k}\right) . \tag{2}$$

In what follows, we refer to the components of $\left(S(\tilde{X})_{0,t_k}, S(\tilde{X})_{t_{k-1},t_k}\right)$ as *global* and *local*, respectively. Intuitively, one can interpret the global component as an efficient representation of long-term information (see Theorem A.2 in Appendix A), and the local component as a type of convolutional filter that is invariant to the sampling rate of the signal. Now, define the *multi-view signature transform*

$$M(\mathbf{X}) = (M(\mathbf{X})_1, ..., M(\mathbf{X})_{\overline{L}}) . \tag{3}$$

For a truncation level n, we define $M(\mathbf{X})^{\leq n}$ by truncating each signature component to contain terms consisting of at most n integration operations. Next, define the *multi-view attention mechanism*, which uses the multi-view signature transform to extend the standard attention mechanism to the space of continuous functions Lyons et al. (2007). First, fix a truncation level $n \in \mathbb{N}$, and let $\bar{d} \in \mathbb{N}$ be such that $M(\mathbf{X})_k^{\leq n} \in \mathbb{R}^{\bar{d}}$. For $h = 1, ..., H$ let $W_h^{\tilde{Q},\tilde{K},\tilde{V}} \in \mathbb{R}^{\bar{d}\times\bar{d}'}$ for some $\bar{d}' \in \mathbb{N}$. Then, let

$$\tilde{Q}_h = M(\mathbf{X})^{\leq n} W_h^{\tilde{Q}}, \quad \tilde{K}_h = M(\mathbf{X})^{\leq n} W_h^{\tilde{K}}, \quad \tilde{V}_h = M(\mathbf{X})^{\leq n} W_h^{\tilde{V}} . \tag{4}$$

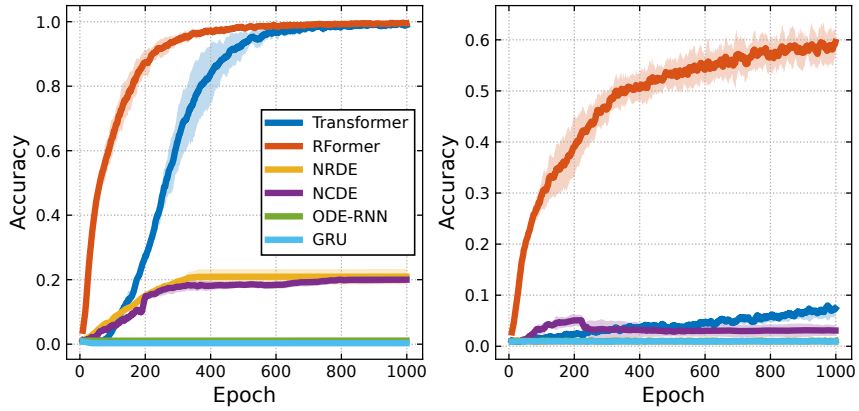

Figure 1: Test accuracy per epoch for the frequency classification task across three random seeds. **Left:** Performance for the full time-series. **Right:** Performance when randomly dropping half of the datapoints every epoch.

Then, the attention calculation is given by $\mathrm{softmax}\left(\frac{\tilde{Q}_h \tilde{K}_h^\intercal}{\sqrt{\bar{d}'}}\right)\tilde{V}_h$. Notice that the attention calculation is similar to vanilla attention, however, we stress that the multi-view attention is built on *continuous-time* objects, the signatures, while the standard attention mechanism acts on discrete objects.

Advantages of Rough Transformers. The Rough Transformer (i) is independent of the sequence length of input data, because one projects input data to its continuous-time interpolation and samples this interpolation at a fixed number of points (ii) is robust to irregular sampling, because the the iterated integrals in equation 1 are invariant to time-reparameterization; see Appendix A (iii) decreases the memory and computational bottleneck inherent to the vanilla Transformer, because we fix the signature sequence length to be $\overline{L} << L$. Furthermore, note that the multi-view attention mechanism does not require backpropagation through the signature calculation and can be computed *in parallel*. This is significantly more computationally efficient when compared to computing signatures batch-wise in every training step.

3 Experiments

Frequency Classification. Our first experiment is based on a set of synthetically generated time-series from continuous-paths of the form $\widehat{X}(t) = g(t)\sin(\omega\, t + \nu) + \eta(t)$, where $g(x)$ is a non-linear trend component, ν and η are two noise terms, and ω is the frequency. Here, the task of the model is to classify the time-series according to its frequency ω. We consider 1000 samples in 100 classes with ω evenly distributed from 10 to 500. Each time-series is regularly sampled with 2000 times-steps on the interval $[0, 6]$. This synthetic experiment is similar to others in recent work on time-series modeling (Li et al., 2019; Yoon et al., 2019; Moreno-Pino et al., 2023). To check the ability of the model to infer the true continuous-time function from irregular observations we drop 50% of the data points at every epoch and report results for both the full and reduced time-series.

Figure 1 shows that the Rough Transformer, hereafter denoted `RFormer`, improves the performance of the Transformer in two ways. First, the inclusion of both local and global information through the multi-view signature enhances the sample efficiency of the model, even though the attention mechanism is now operating on a much shorter sequence. Second, the model is robust to changes in the sampling frequency at inference time, despite operating with half of the context. Therefore, in this task, unlike the vanilla Transformer, `RFormer` adapts to changes in both the length and the sampling frequency of the input stream.[1]

When compared with other models, we see that `GRU` and `ODE-RNN` fail to capture the information in the signal, and are not able to obtain any meaningful performance improvement throughout the training period. This highlights the shortcomings of most RNN-based models when processing sequences

[1]We highlight that the drop in performance is likely due to aliasing, i.e., the sampling rate is below the Nyquist rate, meaning that higher frequency signals cannot be appropriately represented with the number of data points available.

Table 1: Test RMSE (mean $\pm$ std), computed across five seeds on the Heart Rate (HR) dataset. The results correspond to both our reproduction and those presented in (Morrill et al., 2021).

Model	Test RMSE	
	Full	50% Drop
ODE-RNN$^\diamond$	13.06 ± 0.00	-
Neural CDE$^\diamond$	9.82 ± 0.34	-
Neural RDE$^\diamond$	2.97 ± 0.45	-
GRU†	13.06 ± 0.00	13.06 ± 0.00
ODE-RNN†	13.06 ± 0.00	13.06 ± 0.00
Neural RDE†	$\underline{4.04} \pm 0.11$	$\underline{4.67} \pm 0.40$
Transformer	8.24 ± 2.24	21.01 ± 3.81
RFormer	$\mathbf{3.04 \pm 0.03}$	$\mathbf{3.31 \pm 0.05}$

of moderate length, which are very common in real-world applications. In turn, `Neural-CDE` and `Neural-RDE` (Morrill et al., 2021), which is tailor-made for long-range time series modeling, can capture some useful dependencies in the time series but falls short when compared with both vanilla `Transformer` and `RFormer`. When data are irregularly sampled, we see a very sharp drop in performance for the `Transformer` model. Both `Neural-CDE` and `Neural-RDE` also exhibit a significant drop in performance despite being continuous-time.[2]

HR dataset. Next, we consider the Heart Rate dataset from the TSR archive (Tan et al., 2020), originally sourced from Beth Israel Deaconess Medical Center (BIDMC). This dataset consists of time-series sampled from patient ECG readings, and each model is tasked with forecasting the patient's heart rate (HR) at the sample's conclusion. The data, sampled at 125Hz, consists of three-channel time-series (including time), each spanning 4000 time steps. We used the L2 loss metric to assess the performance. Table 1 shows the results.

The sequences in the HR dataset are sufficiently short to remain within memory when running the `Transformer` model. The baseline `Transformer` model improves over `GRU` and `ODE-RNN`, however it is less competitive when compared with `Neural-RDE`, suggesting that the Transformer is not particularly well-suited for this type of task. However, the `RFormer` model improves over the baseline `Transformer` by 171% and yields the best performance overall, experiencing a very small fall in test loss when it comes to the random drop, similar to other continuous-time models.

Training Efficiency. Attention-based architectures are highly parallelizable on modern GPUs, as opposed to traditional RNN models which require sequential updating. However, vanilla attention experiences a bottleneck in memory and time complexity as the sequence length L grows. As covered above in Section 2, variations of the signature transform allow the model to operate on a reduced sequence length $\bar{L}$ without increasing the dimensionality in a way that would become problematic for the model. This allows us to bypass the quadratic complexity of the model without resorting to sparsity techniques commonly used in the literature (Li et al., 2019; Feng et al., 2023). The previous experiment demonstrates empirically that operating on this continuous-time representation does not hinder (and in fact improves) model performance, with the added benefit of robustness to irregularly-sampled time-series. Table 2 shows the empirical running times for the optimal signature hyperparameters (which can be found in Appendix B), with all models run on the same machine and using the same batch size. We also report the speedup of the `RFormer` model when compared to the vanilla `Transformer`.

The computational efficiency gains of `RFormer` are attained due to the multi-view signature transform reducing the length of the time-series with minimal information loss. This contrasts with NRDEs, which augment NCDEs with local signatures of input data, and find that smaller windows often perform better. Hence, NRDEs do not experience the same computational gains as `RFormer`, as they must perform many costly ODE integration steps.

[2]In the case of NRDEs, this phenomenon was observed consistently for both shorter and longer windows.

Table 2: Seconds per epoch for all models considered. See Appendices C and B for experimental details.

Model	Running Time (seconds / epoch)	
	Sine	HR
GRU	**0.12**	1.07
ODE-RNN	5.39	50.71
Neural CDE	9.83	-
Neural RDE	0.85	9.52
Transformer	0.77	11.71
RFormer	0.55	**0.45**
Speedup	$1.4\times$	$26.11\times$

4 Conclusion

In this paper, we introduced the Rough Transformer, a variant of the original Transformer that allows the processing of discrete-time series as continuous-time signals through the use of multi-view signature attention. Empirical comparisons showed that Rough Transformers outperform vanilla Transformers and continuous-time models on a variety of time-series tasks and are robust to the sampling rate of the signal. Finally, we show that RFormer provides significant speedups in training time compared to regular attention and ODE-based methods, without the need for major architectural modifications or sparsity constraints. In future work, it would be interesting to see if Rough Transformers also provide good performance in other application domains such as finance (Scalzo et al., 2021; Moreno-Pino & Zohren, 2022; Arroyo et al., 2022; 2024) or weather forecasting (Nguyen et al., 2023).

References

Imanol Perez Arribas. Derivatives pricing using signature payoffs. *arXiv preprint arXiv:1809.09466*, 2018.

Alvaro Arroyo, Bruno Scalzo, Ljubiša Stanković, and Danilo P Mandic. Dynamic portfolio cuts: A spectral approach to graph-theoretic diversification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5468–5472. IEEE, 2022.

Alvaro Arroyo, Alvaro Cartea, Fernando Moreno-Pino, and Stefan Zohren. Deep attentive survival analysis in limit order books: Estimating fill probabilities with convolutional-transformers. *Quantitative Finance*, pp. 1–23, 2024.

Sofian Berrouiguet, David Ramírez, María Luisa Barrigón, Pablo Moreno-Muñoz, Rodrigo Carmona Camacho, Enrique Baca-García, Antonio Artés-Rodríguez, et al. Combining continuous smartphone native sensors data capture and unsupervised data mining techniques for behavioral changes detection: a case series of the evidence-based behavior (eb2) study. *JMIR mHealth and uHealth*, 6 (12):e9472, 2018.

Sergio Calvo-Ordonez, Jiahao Huang, Lipei Zhang, Guang Yang, Carola-Bibiane Schonlieb, and Angelica I Aviles-Rivero. Beyond u: Making diffusion models faster & lighter. *arXiv preprint arXiv:2310.20092*, 2023.

Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys (CSUR)*, 54(4):1–40, 2021.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Aosong Feng, Irene Li, Yuang Jiang, and Rex Ying. Diffuser: efficient transformers with multi-hop attention diffusion for long sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12772–12780, 2023.

Ken-ichi Funahashi and Yuichi Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural networks*, 6(6):801–806, 1993.

Ben Hambly and Terry Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics*, pp. 109–167, 2010.

Jeffrey M Hausdorff and C-K Peng. Multiscaled randomness: A possible source of 1/f noise in biology. *Physical review E*, 54(2):2154, 1996.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems*, 32, 2019.

Terry J Lyons, Michael Caruana, and Thierry Lévy. *Differential equations driven by rough paths*. Springer, 2007.

María Martínez-García, Fernando Moreno-Pino, Pablo M Olmos, and Antonio Artés-Rodríguez. Sleep activity recognition and characterization from multi-source passively sensed data. *arXiv preprint arXiv:2301.10156*, 2023.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning*, pp. 15293–15329. PMLR, 2022.

Maxime Morariu-Patrichi and Mikko S Pakkanen. State-dependent hawkes processes and their application to limit order book modelling. *Quantitative Finance*, 22(3):563–583, 2022.

Fernando Moreno-Pino and Stefan Zohren. Deepvol: Volatility forecasting from high-frequency data with dilated causal convolutions. *arXiv preprint arXiv:2210.04797*, 2022.

Fernando Moreno-Pino, María Martínez-García, Pablo M Olmos, and Antonio Artés-Rodríguez. Heterogeneous hidden markov models for sleep activity recognition from multi-source passively sensed data. *arXiv preprint arXiv:2211.10371*, 2022.

Fernando Moreno-Pino, Pablo M Olmos, and Antonio Artés-Rodríguez. Deep autoregressive models with spectral attention. *Pattern Recognition*, 133:109014, 2023.

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

Tung Nguyen and Aditya Grover. Transformer neural processes: Uncertainty-aware meta learning via sequence modeling. *arXiv preprint arXiv:2207.04179*, 2022.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

Alexander Norcliffe, Cristian Bodnar, Ben Day, Jacob Moss, and Pietro Liò. Neural ode processes. *arXiv preprint arXiv:2103.12413*, 2021.

Alejandro Porras-Segovia, Isaac Diaz-Olivan, Maria Luisa Barrigón, Manon Moreno, Antonio Artes-Rodriguez, Maria Mercedes Perez-Rodriguez, and Enrique Baca-Garcia. Real-world feasibility and acceptability of real-time suicide risk monitoring via smartphones: A 6-month follow-up cohort. *Journal of Psychiatric Research*, 149:145–154, 2022.

Roger Ratcliff, Philip L Smith, Scott D Brown, and Gail McKoon. Diffusion decision model: Current issues and history. *Trends in cognitive sciences*, 20(4):260–281, 2016.

Gonzalo R Ríos-Muñoz, Fernando Moreno-Pino, Nina Soto, Pablo M Olmos, Antonio Artés-Rodríguez, Francisco Ferández-Avilés, and Angel Arenal. Hidden markov models for activity detection in atrial fibrillation electrograms. In *2020 Computing in Cardiology*, pp. 1–4. IEEE, 2020.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Bruno Scalzo, Alvaro Arroyo, Ljubiša Stanković, and Danilo P Mandic. Nonstationary portfolios: Diversification in the spectral domain. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5155–5159. IEEE, 2021.

Nabeel Seedat, Fergus Imrie, Alexis Bellot, Zhaozhi Qian, and Mihaela van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. *arXiv preprint arXiv:2206.08311*, 2022.

Chang Wei Tan, Christoph Bergmeir, Francois Petitjean, and Geoffrey I Webb. Monash university, uea, ucr time series extrinsic regression archive. *arXiv preprint arXiv:2006.10996*, 2020.

Amirali Vahid, Moritz Mückschel, Sebastian Stober, Ann-Kathrin Stock, and Christian Beste. Applying deep learning to single-trial eeg data provides evidence for complementary theories on action control. *Communications biology*, 3(1):112, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

A PROPERTIES OF PATH SIGNATURES

First, we recall that smooth paths are uniquely determined by their signatures, which motivates its use as a feature map.

Proposition A.1. *Given a smooth path $\widehat{X} : [0,T] \to \mathbb{R}^d$, then the map $P_{\widehat{X}} : [0,T] \to \mathbb{R}^{1+d}$ where $P_{\widehat{X}}(t) = (t, \widehat{X}(t))$ is uniquely determined by it's signature $S(P_{\widehat{X}})_{0,T}$.*

The proof of this fact in (Hambly & Lyons, 2010).

For Rough Transformers, several features of path signatures are important. First, linear functionals on path signatures possess universal approximation properties for continuous functionals.

Theorem A.2. *Fix $T > 0$, and let $K \subset C_b^1([0,T];\mathbb{R}^d)$. Let $f : C_b^1([0,T];\mathbb{R}^{d+1}) \to \mathbb{R}$ be continuous with respect to the sup-norm topology on $C_b^1([0,T];\mathbb{R}^{d+1})$. Then for any $\epsilon > 0$, there exists a linear functional ℓ such that*

$$|f(P_{\widehat{X}}) - \langle \ell, S(P_{\widehat{X}})_{0,T} \rangle| \leq \epsilon \,, \tag{5}$$

for any $\widehat{X} \in K$.

For a proof of A.2, see (Arribas, 2018). Even though Theorem equation A.2 guarantees that *linear* functionals are sufficient for universal approximation, linear models are not always sufficient in practice. This motivates the development of nonlinear models built upon the path signature which efficiently extract path behavior.

The second feature is that the terms of the path signature decay factorially, as described by the following proposition.

Proposition A.3. *Given $\widehat{X} \in C_b^1([0,T];\mathbb{R}^d)$, for any $s,t \in [0,T]$, we have that for any $I \in \mathcal{I}_d^n$*

$$|S(\widehat{X})^I_{0,T}| = O\left(1/n!\right) \,. \tag{6}$$

For a proof of Proposition A.3, see (Lyons et al., 2007). Hence, the number of terms in signature grows exponentially in the level of the signature, but the tail of the signature is well-behaved, so only a few levels in a truncated signature are necessary to adequately approximate continuous functionals.

A.1 SIGNATURES OF PIECEWISE LINEAR PATHS.

In the Rough Transformer, we use linear interpolation of input time-series to get a continuous-time representation of the data. As mentioned in Section 2, the signature computation in this case is particularly simple.

Suppose $\widehat{X}_k : [t_k, t_{k+1}] \to \mathbb{R}^d$ is a linear interpolation between two points $X_k, X_{k+1} \in \mathbb{R}^d$. That is,

$$\widehat{X}_k(t) = X_k + \frac{t - t_k}{t_{k+1} - t_k}\left(X_{k+1} - X_k\right) \,. \tag{7}$$

Then the signature of $\widehat{X}_k$ is given explicitly:

$$S(\widehat{X}_k)_{t_k,t_{k+1}} = \left(1, X_{k+1} - X_k, \frac{1}{2}(X_{k+1} - X_k)^{\otimes 2}, \frac{1}{3!}(X_{k+1} - X_k)^{\otimes 3}, ..., \frac{1}{n!}(X_{k+1} - X_k)^{\otimes n}, ...\right) \,, \tag{8}$$

where $\otimes$ denotes the tensor product. Let $\widehat{X}_k * \widehat{X}_{k+1}$ denote the *concatenation* of $\widehat{X}_k$ and $\widehat{X}_{k+1}$. That is, $\widehat{X}_k * \widehat{X}_{k+1} : [t_k, t_{k+2}] \to \mathbb{R}^d$ is defined

$$\widehat{X}_k * \widehat{X}_{k+1}(t) = \begin{cases} \widehat{X}_k(t) & t \in [t_k, t_{k+1}] \\ \widehat{X}_{k+1}(t) & t \in (t_2, t_{k+2}] \,. \end{cases} \tag{9}$$

The signature of the concatenation $\widehat{X}_k * \widehat{X}_{k+1}$ is given by *Chen's relation*.

Proposition A.4 (Chen's Relation)**.** *The following identity holds:*

$$S(\widehat{X}_k * \widehat{X}_{k+1})_{t_k,t_{k+2}} = S(\widehat{X}_k)_{t_k,t_{k+1}} \otimes S(\widehat{X}_{k+1})_{t_{k+1},t_{k+2}} \,, \tag{10}$$

where for elements $A, B \in T((\mathbb{R}^d))$ with $A = (A_0, A_1, A_2, ...)$ and $B = (B_0, B_1, B_2, ...)$ the tensor product $\otimes$ is defined

$$A \otimes B = \left(\sum_{j=0}^{k} A_j \otimes B_{k-j} \right)_{k \geq 0} \,. \tag{11}$$

Let $\mathbf{X} = (X_0, ..., X_L)$ be a time-series. Then the linear interpolation $\tilde{X} : [0, T] \to \mathbb{R}^d$ can be represented as the concatenation of a finite number of linear paths:

$$\tilde{X} = \widehat{X}_0 * \cdots * \widehat{X}_{L-1} \,. \tag{12}$$

Hence, the signature can be computed:

$$S(\tilde{X})_{0,T} = S(\widehat{X}_0)_{0,t_1} \otimes \cdots \otimes S(\widehat{X}_{L-1})_{t_{L-1},T} \,. \tag{13}$$

B EXPERIMENTAL DETAILS

All experiments were conducted on an NVIDIA GeForce RTX 3090 GPU with 24,564 MiB of memory, utilizing CUDA version 12.3. The timings presented in Table 2 were obtained by executing each model independently for each dataset and averaging the resulting times across 100 epochs. We take down the hyperparameters used in the `RFormer` model for each of the datasets in the paper, which were chosen in accordance with the performance of the model in the validation set.

Table 3: Validation accuracy on the sinusoidal dataset.

Dataset	**Learning Rate**	**Number of Windows**	**Sig. Depth**	**Sig. Type**
Sinusoidal	1×10^{-3}	75	2	Multi-View
HR	1×10^{-3}	75	4	Local

C MODELS VALIDATION

This section collects the validation of Step and Depth for the neural RDE model in each of the evaluated datasets. Optimal values are selected for evaluation on test-set. Early-stopping was used with same criteria than Morrill et al., 2021.

Table 4: Validation accuracy on the sinusoidal dataset.

Acc. Val	Step	Depth	Memory Usage (Mb)	Elapsed Time (s)
17.26	2	2	778.9	6912.7
12.21	2	3	770.3	1194.43
16.35	4	2	382.2	2702.48
19.27	4	3	386.16	574.97
20.99	8	2	193	1321.36
24.02	**8**	**3**	**194.17**	**332.17**
17.15	16	2	97.13	136.43
21.59	16	3	98.17	156.93
17.46	24	2	65.96	105.94
20.59	24	3	66.68	98.97

Table 5: Validation loss on the HR dataset.

Acc. Val	Step	Depth	Memory Usage (Mb)	Elapsed Time (s)
2.44	**2**	**2**	**5044.44**	**56492.33**
3.03	2	3	5059.28	39855.19
3.67	4	2	2515.40	10765.58
16.04	4	3	2531.44	7157.20
5.35	8	2	1259.30	3723.94
2.70	8	3	1268.60	18682.82
3.58	16	2	632.08	3518.96
3.64	16	3	636.64	7922.96
3.86	24	2	422.74	3710.95
3.55	24	3	426.83	6567.02