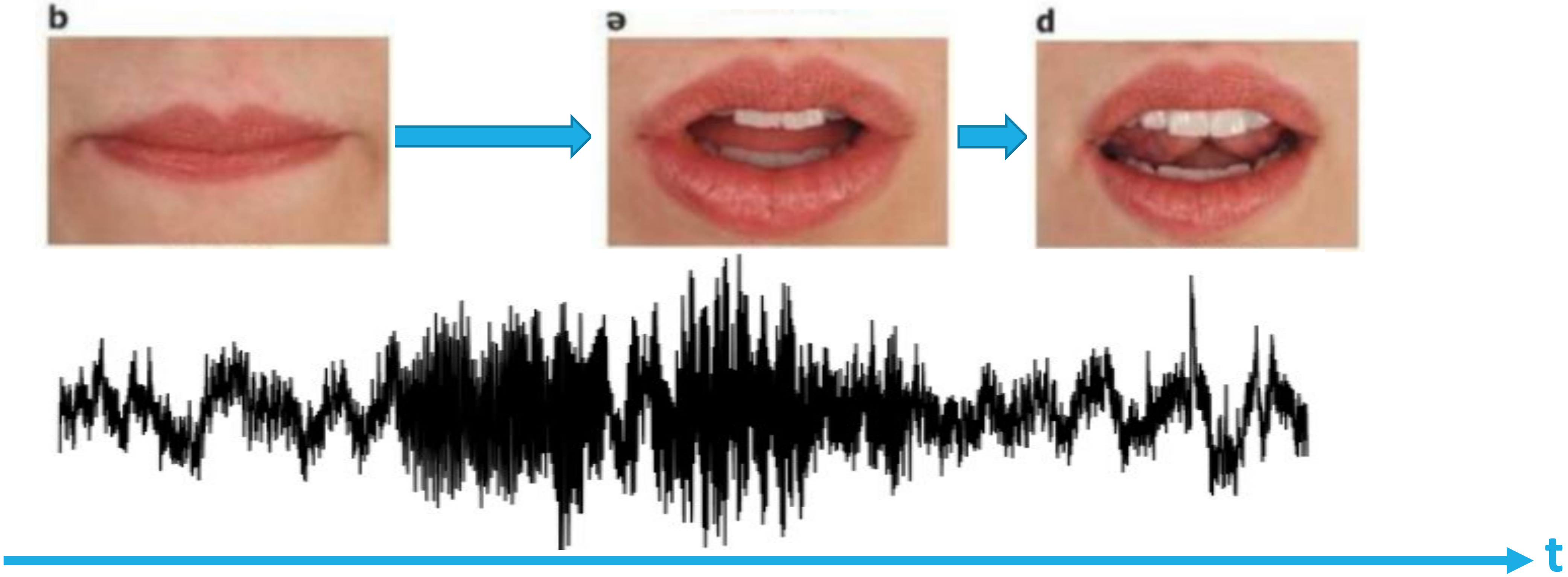
STRODE: Stochastic Boundary Ordinary Differential Equation

HENGGUAN HUANG,

HONGFU LIU, HAO WANG, CHANG XIAO, YE WANG

Time is closely related to what you see and hear

Time is closely related to what you see and hear



Time perception is essential for living organisms

Hunting

Playing

Hearing

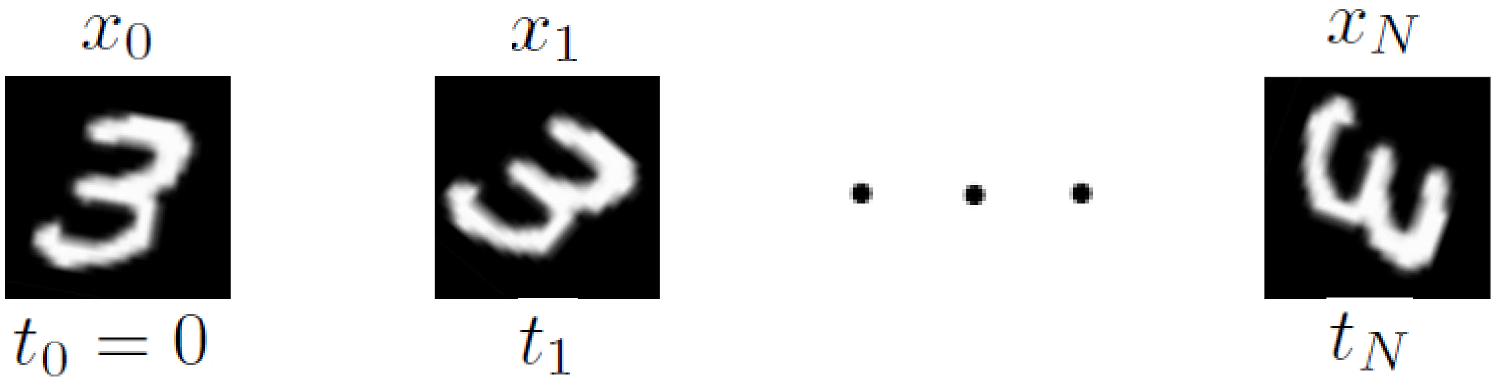
- Mostly, machines fails to perceive time **directly** from visual or audio inputs
- Can gaps between natural and artificial intelligence be bridged further through introducing the "time perception" mechanism?

Existing methods for time-series modeling

- Recurrent neural networks (RNNs) (assume data to be evenly sampled)
- Latent ordinary differential equation (ODE)/ ODE-RNN (for handling irregularly sampled data)
- Jump stochastic differential equation (JSDE) (for modeling marked point process data)
- However, above methods require training data with timing annotations:
  - timing annotations of events contained in regularly-sampled sequence
  - or timing annotation of each data point for irregularly-sampled data

Our Goal: develop time-series models that can jointly infer the timings and the dynamics of time series data without requiring any timing annotations during training.

Consider an autoregressive task for irregularly sampled MNIST rotating digits



Boundary value problem (BVP):

$$h'(t) = f_{\theta_1}(h(t), t)$$

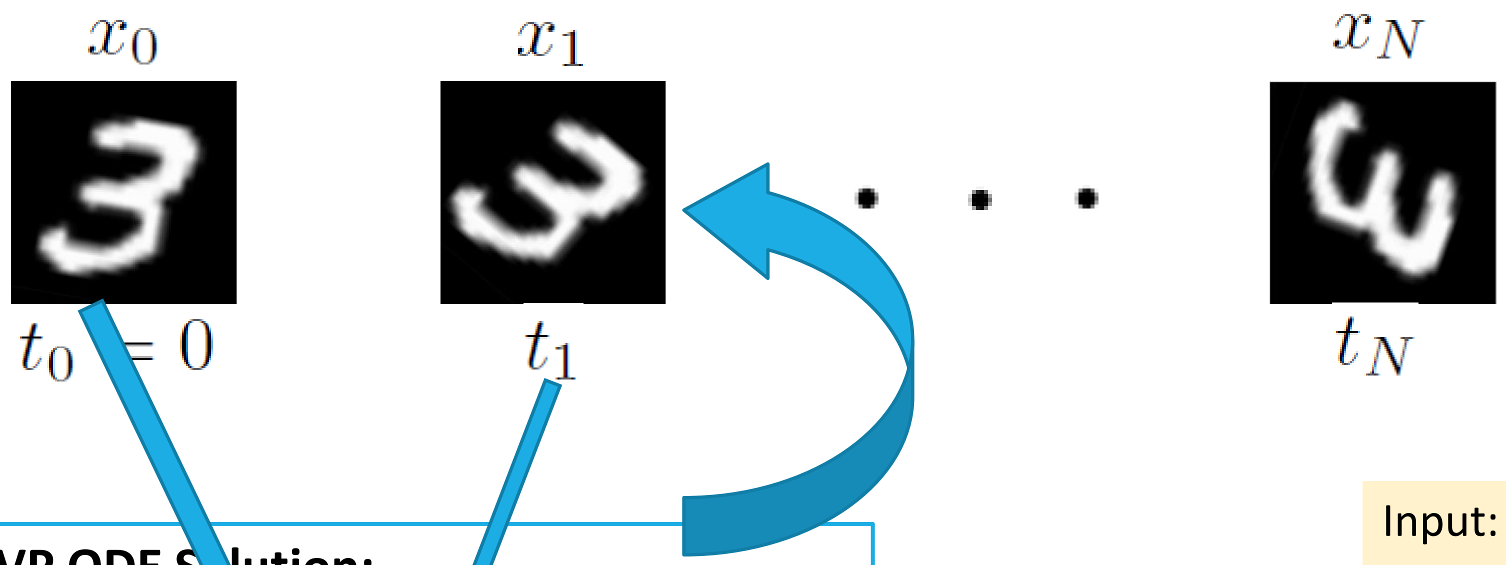
Boundary conditions:

$$\{h(t_0) = x_0, h(t_1) = x_1, ..., h(t_N) = x_N\}$$

Input: $x_0, x_1, \cdots, x_{N-1}$
$t_1, t_2, \cdots, t_N$

Output: $\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_N$

Predicting next frame by solving BVP ODE



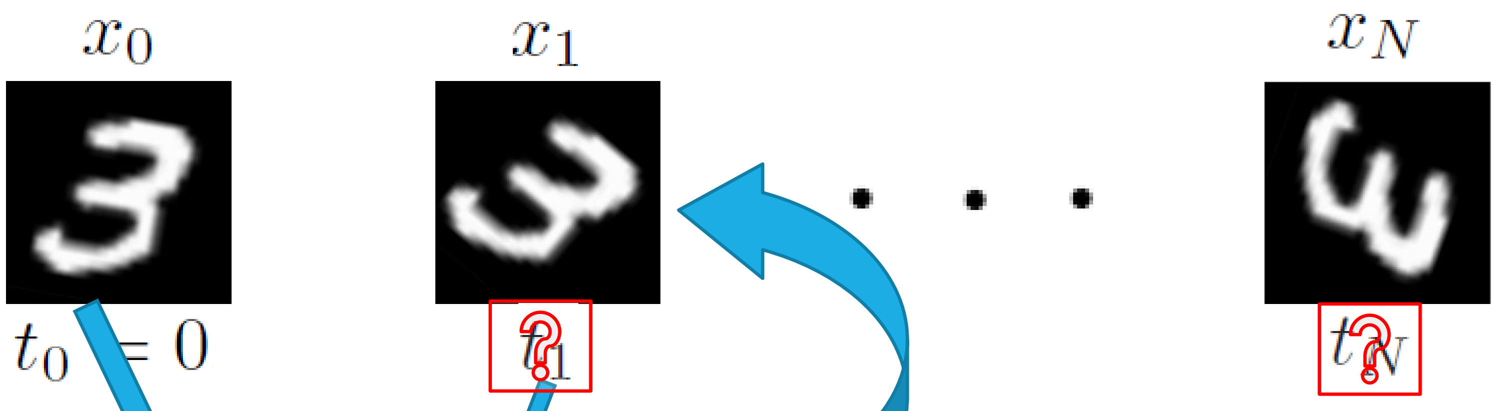
BVP ODE Solution:

$$h(t_{i+1}) = h(t_i) + \int_{t_i}^{t_{i+1}} f_{\theta_1}(h(t), t)dt$$

Input: $x_0, x_1, \cdots, x_{N-1}$
$t_1, t_2, \cdots, t_N$

Output: $\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_N$

However, the timing of each data point is usually unknow (or not exact in many realistic tasks)



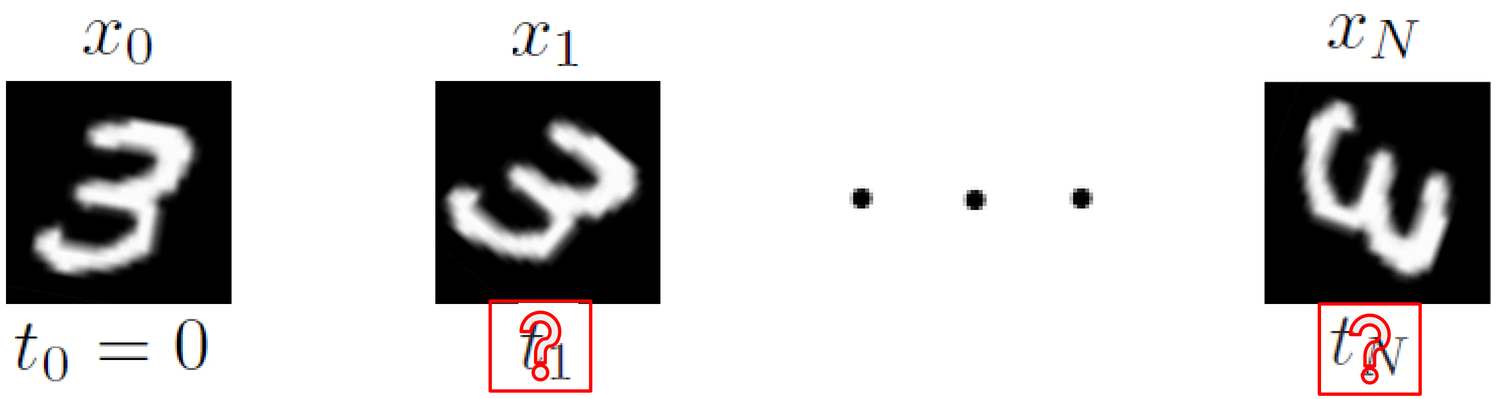
BVP ODE Solution:

$$h(t_{i+1}) = h(t_i) + \int_{t_i}^{t_{i+1}} f_{\theta_1}(h(t), t)\,dt$$

Input: $x_0, x_1, \cdots, x_{N-1}$
$\cancel{t_1, t_2, \cdots, t_N}$

Output: $\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_N$

We propose a stochastic boundary value problem



Stochastic boundary value problem (SBVP):

$$h'(t) = f_{\theta_o}(h(t), t)$$

Boundary conditions:

$$\boxed{\{\tilde{t}_i\}_{i=1}^{N} \sim T(\{p_i(t|\mathbf{x}_i)\}_{i=1}^{N})}$$

$$\{h(0) = x_0, h(\tilde{t}_1) = x_1, ..., h(\tilde{t}_N) = x_N\}$$

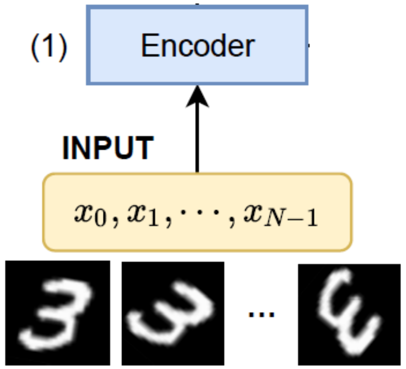
Input: $x_0, x_1, \cdots, x_{N-1}$

$\cancel{t_1, t_2, \cdots, t_N}$

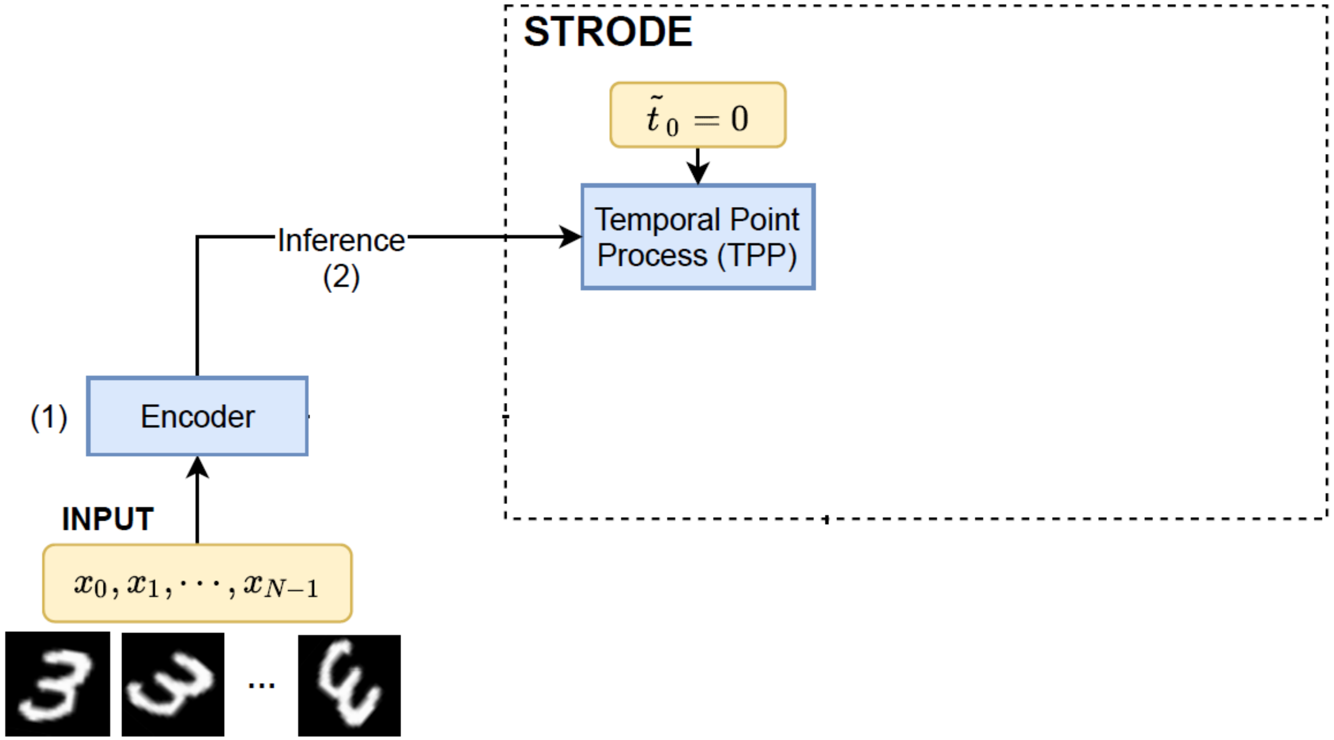
Output: $\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_N$

$\tilde{t}_1, \tilde{t}_2, \cdots, \tilde{t}_N$

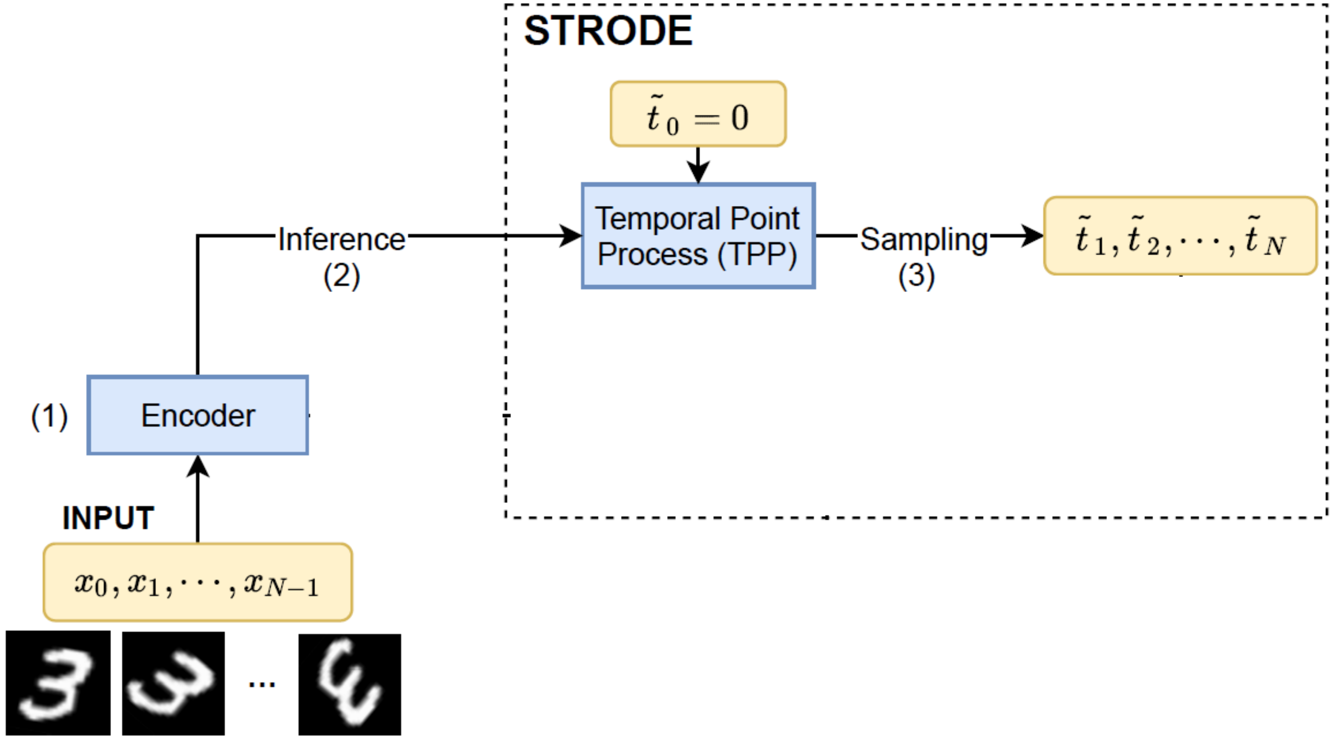
Our model: Stochastic Boundary Ordinary Differential Equation (STRODE)



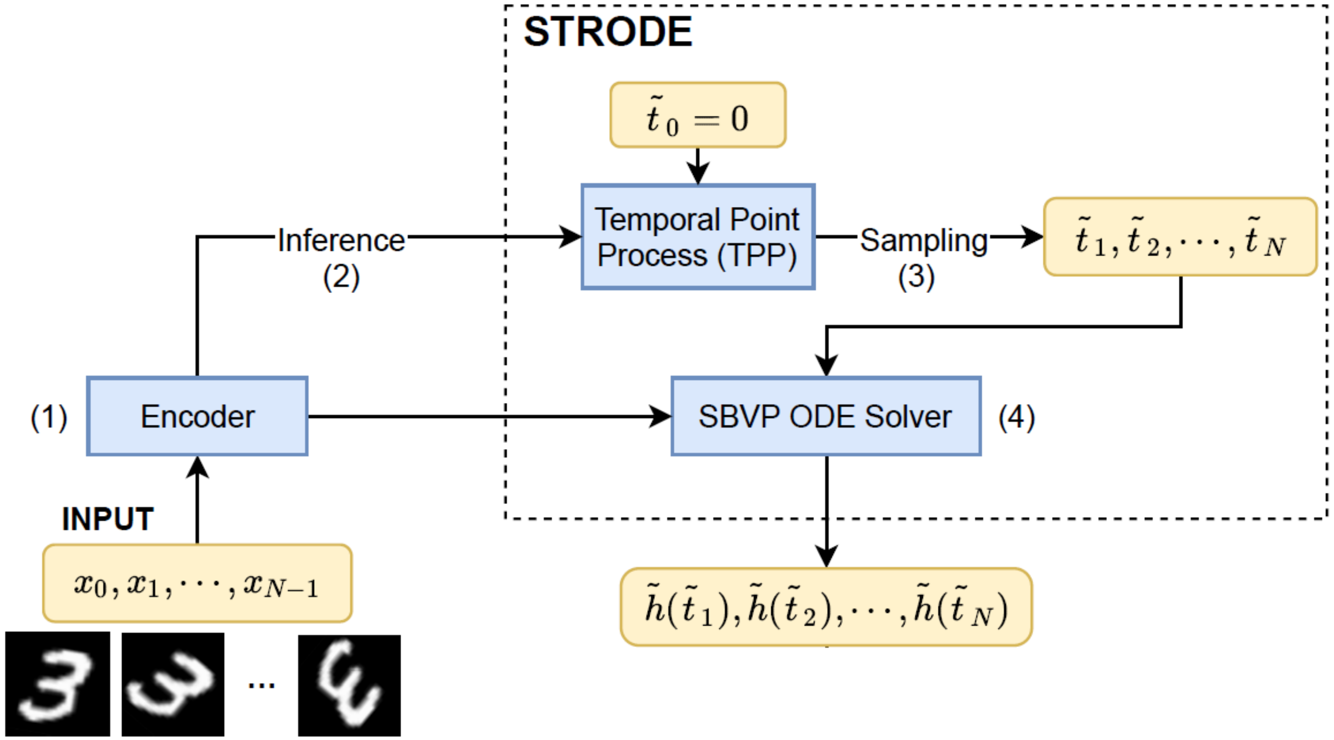
Our model: Stochastic Boundary Ordinary Differential Equation (STRODE)



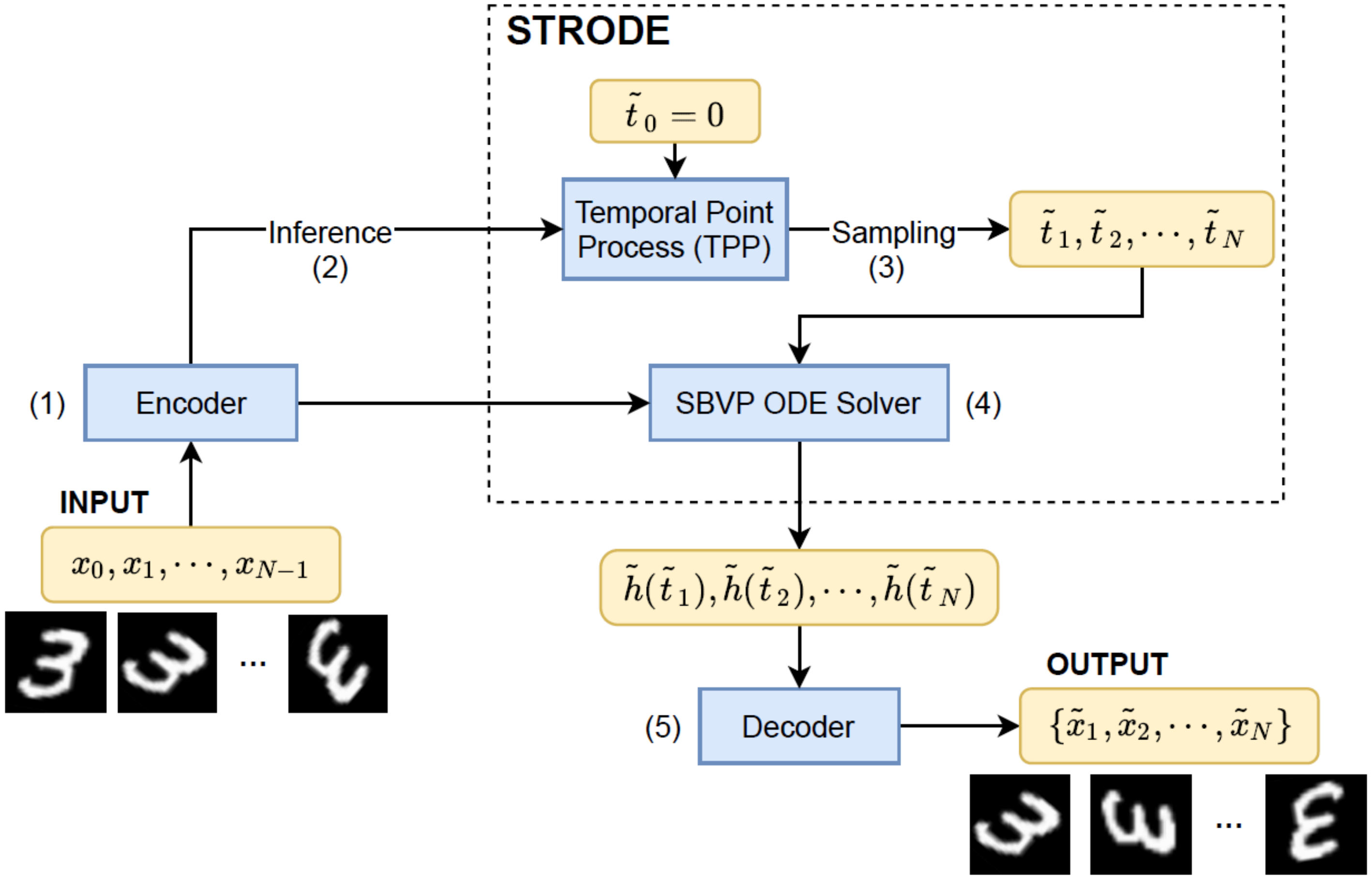
Our model: Stochastic Boundary Ordinary Differential Equation (STRODE)



Our model: Stochastic Boundary Ordinary Differential Equation (STRODE)



Our model: Stochastic Boundary Ordinary Differential Equation (STRODE)



Technical challenges in learning the STRODE

Learning the STRODE is, equivalently, solving the SBVP:

- The inference of the temporal point process (TPP) in SBVP is difficult as timing annotation is unavailable during training
- We, therefore, adopt variational inference to optimize the STRODE:

$$\log P(\mathbf{X}) \geqslant \sum_{i=1}^{N} \{\mathbb{E}_{\tilde{t}_i \sim q_i(t|x_i)} \log p(x_i|\tilde{t}_i) \\ -\mathrm{KL}(q_i(t|x_i)||p_i(t))\}$$

Technical challenges in learning the STRODE

Learning the STRODE is, equivalently, solving the SBVP:

- The inference of the temporal point process (TPP) in SBVP is difficult as timing annotation is unavailable during training
- We, therefore, adopt variational inference to optimize the STRODE:

$$\log P(\mathbf{X}) \geqslant \sum_{i=1}^{N} \{\mathbb{E}_{\tilde{t}_i \sim q_i(t|x_i)} \log \boxed{p(x_i|\tilde{t}_i)} - \mathrm{KL}(q_i(t|x_i)||p_i(t))\}$$

Posterior distribution conditioned on boundary time

Technical challenges in learning the STRODE

Learning the STRODE is, equivalently, solving the SBVP:

- The inference of the temporal point process (TPP) in SBVP is difficult as timing annotation is unavailable during training
- We, therefore, adopt variational inference to optimize the STRODE:

$$\log P(\mathbf{X}) \geqslant \sum_{i=1}^{N} \{ \mathbb{E}_{\tilde{t}_i \sim q_i(t|x_i)} \log p(x_i|\tilde{t}_i) - \mathrm{KL}(\boxed{q_i(t|x_i)} || \boxed{p_i(t)}) \}$$

Approximate posterior of TPP

Prior of TPP

Technical challenges in learning the STRODE

- TPP with general form is usually more powerful than the ones with fixed parametric form
- But KL-divergence between two TPPs with general forms in the evidence lower bound (ELBO) could be computationally intractable

$$\log P(\mathbf{X}) \geqslant \sum_{i=1}^{N} \{\mathbb{E}_{\tilde{t}_i \sim q_i(t|x_i)} \log p(x_i|\tilde{t}_i) \\ - \boxed{\mathrm{KL}(q_i(t|x_i)||p_i(t))}\}$$

If a general form of TPPs are assumed, it will be difficult to calculate the KL term

Technical challenges in learning the STRODE

- Traditional sampling methods (e.g., thinning algorithm) of TPP lead to convergence issues when optimizing STRODE with variational inference

$$\log P(\mathbf{X}) \geqslant \sum_{i=1}^{N} \{\mathbb{E}_{\boxed{\tilde{t}_i \sim q_i(t|x_i)}} \log p(x_i|\tilde{t}_i) - \mathrm{KL}(q_i(t|x_i)||p_i(t))\}$$

Traditional sampling methods lead to convergence issues

Our solution: ODE-based Sampling and Inference of TPP

- We propose an initial value problem (IVP) ODE to describe the dynamics of boundary times

Initial value problem (IVP):

$$\Phi_i{}'(t) = -t q_i(t|x_i)$$

Initial condition:

$$\Phi_i(0) = \int_0^{+\infty} t q_i(t|x_i) dt$$

Our solution: ODE-based Sampling and Inference of TPP

- We propose an initial value problem (IVP) ODE to describe the dynamics of boundary times
- **Differentiable sampling** of the approximate posterior of TPP is achieved by solving such IVP

Initial value problem (IVP):

$$\Phi_i{}'(t) = -tq_i(t|x_i)$$

Initial condition:

$$\Phi_i(0) = \int_0^{+\infty} tq_i(t|x_i)dt$$

General solution:

$$\tilde{t}_i = \Phi_i(t_{i-1}) = \Phi_i(0) + \int_0^{\tilde{t}_{i-1}} -tq_i(t|x_i)dt$$

Our solution: ODE-based Sampling and Inference of TPP

- We propose an initial value problem (IVP) ODE to describe the dynamics of boundary times
- **Differentiable sampling** of the approximate posterior of TPP is achieved by solving such IVP

Initial value problem (IVP):

$$\Phi_i{}'(t) = -tq_i(t|x_i)$$

Initial condition:

$$\Phi_i(0) = \int_0^{+\infty} \boxed{tq_i(t|x_i)}\,dt$$

General solution:

$$\tilde{t}_i = \Phi_i(t_{i-1}) = \Phi_i(0) + \int_0^{\tilde{t}_{i-1}} \boxed{-tq_i(t|x_i)}\,dt$$

Our solution: ODE-based Sampling and Inference of TPP

- We propose an initial value problem (IVP) ODE to describe the dynamics of boundary times
- **Differentiable sampling** of the approximate posterior of TPP is achieved by solving such IVP

Initial value problem (IVP):

$$\Phi_i{}'(t) = -tq_i(t|x_i)$$

Initial condition:

$$\Phi_i(0) = \int_0^{+\infty} \boxed{tq_i(t|x_i)}\, dt$$

General solution:

$$\tilde{t}_i = \Phi_i(t_{i-1}) = \Phi_i(0) + \int_0^{\tilde{t}_{i-1}} \boxed{-tq_i(t|x_i)}\, dt$$

Approximate solution:

$$\tilde{t}_i = \Phi_i(\tilde{t}_{i-1}) = \boxed{f_{\theta_\Phi}(\tilde{t}_{i-1}, x_i)}$$

A neural network

Our solution: ODE-based Sampling and Inference of TPP

- We propose an initial value problem (IVP) ODE to describe the dynamics of boundary times
- **Differentiable sampling** of the approximate posterior of TPP is achieved by solving such IVP
- Posterior of the TPP can be written as an differential equation accordingly

Initial value problem (IVP):

$$\Phi_i{}'(t) = -t q_i(t|x_i)$$

Initial condition:

$$\Phi_i(0) = \int_0^{+\infty} \boxed{t q_i(t|x_i)} dt$$

Approximate posterior of TPP:

$$q_i(t|x_i) = \frac{-\Phi_i{}'(t)}{t}$$

General solution:

$$\tilde{t}_i = \Phi_i(t_{i-1}) = \Phi_i(0) + \int_0^{\tilde{t}_{i-1}} \boxed{-t q_i(t|x_i)} dt$$

Approximate solution:

$$\tilde{t}_i = \Phi_i(\tilde{t}_{i-1}) = \boxed{f_{\theta_\Phi}(\tilde{t}_{i-1}, x_i)}$$

A neural network

Our solution: ODE-based Sampling and Inference of TPP

- We propose an initial value problem (IVP) ODE to describe the dynamics of boundary times
- **Differentiable sampling** of the approximate posterior of TPP is achieved by solving such IVP
- Posterior of the TPP can be written as an differential equation accordingly

Initial value problem (IVP):

$$\Phi_i{}'(t) = -tq_i(t|x_i)$$

Initial condition:

$$\Phi_i(0) = \int_0^{+\infty} \boxed{tq_i(t|x_i)}dt$$

Approximate posterior of TPP:

$$q_i(t|x_i) = \frac{-\Phi_i{}'(t)}{t}$$

Auto-differentiation

General solution:

$$\tilde{t}_i = \Phi_i(t_{i-1}) = \Phi_i(0) + \int_0^{\tilde{t}_{i-1}} \boxed{-tq_i(t|x_i)}dt$$

Approximate solution:

$$\tilde{t}_i = \Phi_i(\tilde{t}_{i-1}) = \boxed{f_{\theta_\Phi}(\tilde{t}_{i-1}, x_i)}$$

A neural network

Theoretical results: ODE-based Kullback–Leibler (KL) Divergence with analytical upper bound

- Similar to Omi et al.'s work, our prior of TPP in STRODE is a differential equation

Prior of TPP:

$$p_i(t) = \phi_i'(t) e^{-\phi_i(t)}$$

Omi, T., Ueda, N., and Aihara, K. Fully neural network based model for general temporal point processes. Conference on Neural Information Processing Systems (NeurIPS), 2019.

Theoretical results: ODE-based Kullback–Leibler (KL) Divergence with analytical upper bound

- Similar to Omi et al.'s work, our prior of TPP in STRODE is a differential equation

A neural network

Prior of TPP:

$$p_i(t) = \phi_i'(t)e^{-\boxed{\phi_i(t)}}$$

Theoretical results: ODE-based Kullback–Leibler (KL) Divergence with analytical upper bound

- Similar to Omi et al.'s work, our prior of TPP in STRODE is a differential equation
- KL-divergence between two differential equations in the evidence lower bound (ELBO) is computationally intractable

Prior of TPP:

$$p_i(t) = \phi_i'(t)e^{-\phi_i(t)}$$

Approximate posterior of TPP:

$$q_i(t|x_i) = \frac{-\Phi_i{}'(t)}{t}$$

KL

$$\mathrm{KL}(q_i(t|x_i)||p_i(t)) = \int_0^{+\infty} \frac{-\Phi_i{}'(t)}{t} \log \frac{-\Phi_i{}'(t)}{t\phi_i'(t)e^{-\phi_i(t)}}$$

Theoretical results: ODE-based Kullback–Leibler (KL) Divergence with analytical upper bound

- Similar to Omi et al.'s work, our prior of TPP in STRODE is a differential equation
- KL-divergence between two differential equations in the evidence lower bound (ELBO) is computationally intractable
  - upper limit of the integration approaches infinity when calculating the KL

Prior of TPP:

$$p_i(t) = \phi_i'(t)e^{-\phi_i(t)}$$

Approximate posterior of TPP:

$$q_i(t|x_i) = \frac{-\Phi_i{}'(t)}{t}$$

KL

$$\mathrm{KL}(q_i(t|x_i)||p_i(t)) = \int_0^{\boxed{+\infty}} \frac{-\Phi_i{}'(t)}{t} \log \frac{-\Phi_i{}'(t)}{t\phi_i'(t)e^{-\phi_i(t)}}$$

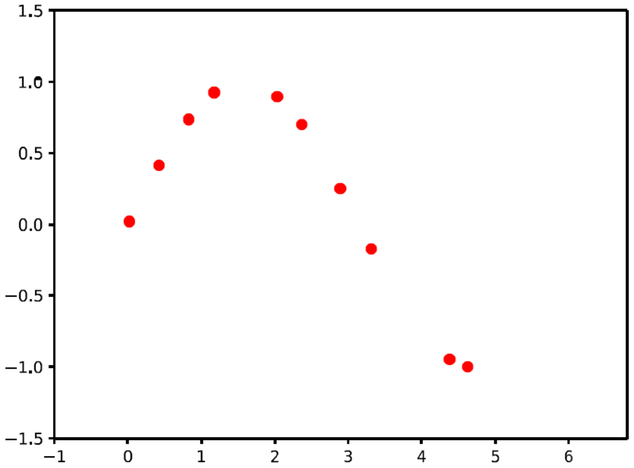
Theoretical results: ODE-based Kullback–Leibler (KL) Divergence with analytical upper bound

We derive **an analytical upper bound** for the KL term in ELBO **(Theorem 1)**

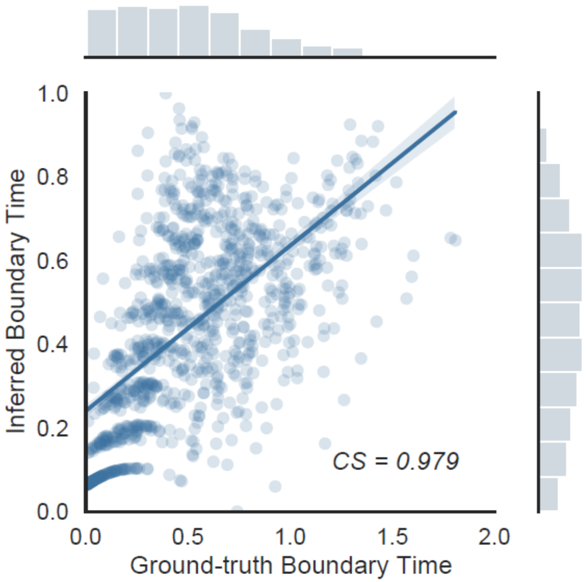
- We first introduce an ODE to to assist the derivation of the upper bound
- Then we separate the KL term into two parts, where one part is computed by solving the IVP, but the other involves an improper integral
- Unlike the well-known Gronwall's Inequality which bound such integral with an unbounded Lipschitz constant
- We derive an computationally tractable upper bound of such integral (Lemma 1)

STRODE is capable of inferring timings of irregularly sampled sine waves

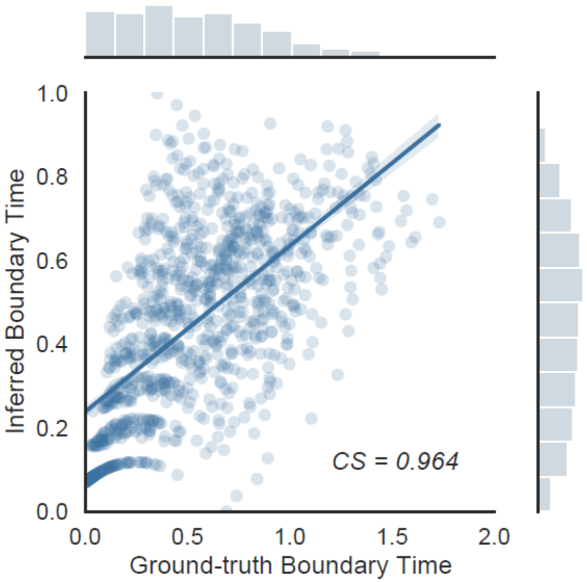
Training data samples



Cosine similarity (CS) between the inferred timings and the ground truth.



Sine waves are sampled with Hawkes process



Sine waves are sampled with Poisson process

STRODE can be generalized to irregularly sampled high dimensional data (*Rotating MNIST Thumbnail*)

Training data samples

Cosine similarity (CS) (mean std) and MSE results on two subsets of Rotating MNIST Thumbnail*

DATASET	Hawkes		Exponential	
	CS	MSE ($\times 10^{-3}$)	CS	MSE ($\times 10^{-3}$)
NODE (Chen et al., 2018)	0.907	6.66±0.03	0.923	7.69±0.02
ODE-RNN (Rubanova et al., 2019)	0.907	6.82±0.01	0.923	6.07±0.10
STRODE (Ours)	0.966±0.007	**6.01±0.11**	0.973±0.003	7.26±0.27
STRODE-RNN (Ours)	**0.967±0.012**	6.35±0.14	**0.974±0.005**	**5.94±0.03**

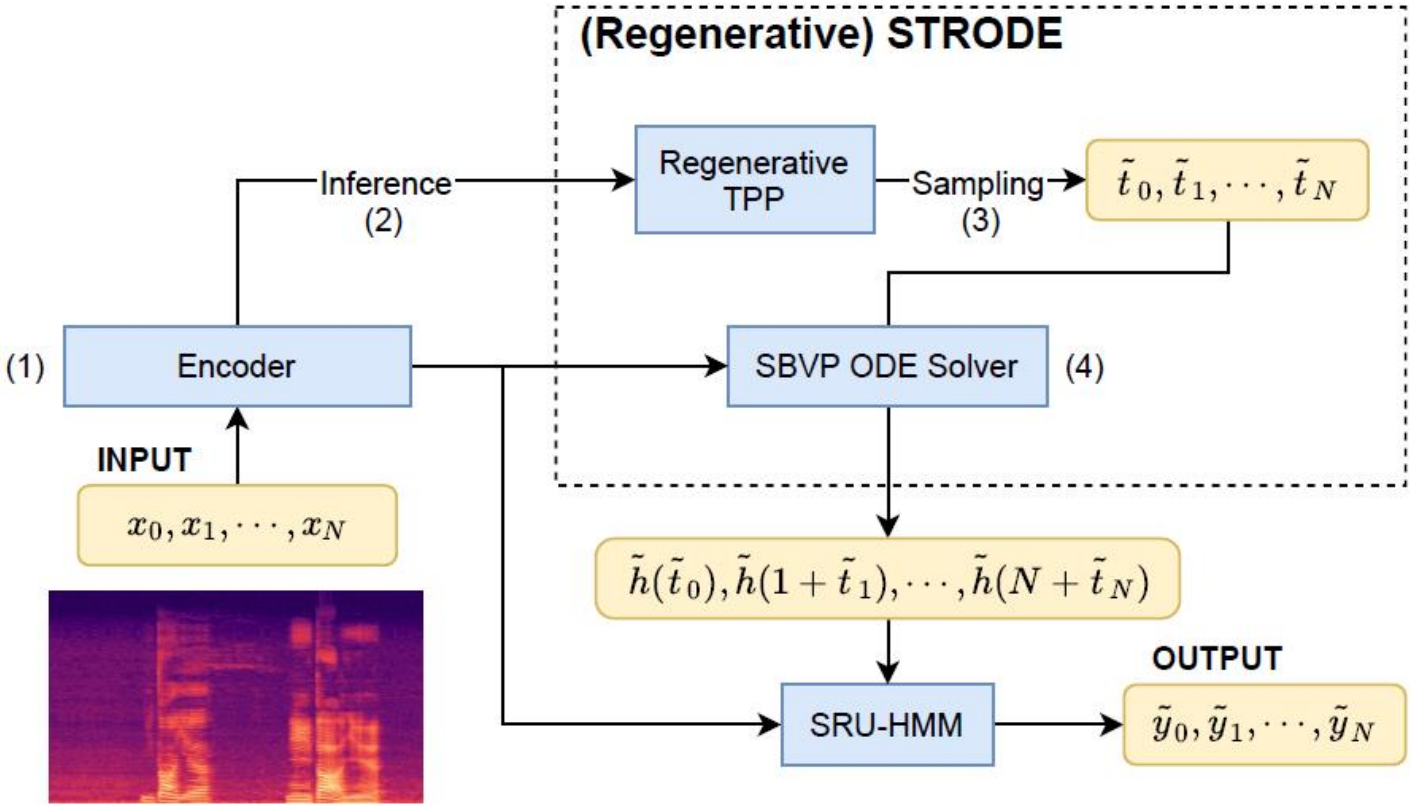
*: We find that results differ when using different GPUs. We, therefore, rerun the experiments with a NVIDIA TESLA V100 GPU

An extension of STRODE for real application: postdictive acoustic modeling

Postdiction: a phenomena in cognition of human brain, in which accuracy of "prediction" is reassured with sufficient future information to be integrated.

- There are advantages to this process for many real-world tasks.
- For example, understanding a word aids in distinguishing its constituent phonemes from another in human speech processing
- However, such process is difficult to be incorporated into acoustic modeling
- This is because the temporal range of subsequent context is mostly unannotated
- Such process could lead to input latency due to future context required

An extension of STRODE for real application: postdictive acoustic modeling



- We adopt STRODE to infer such temporal ranges
- Our STRODE further produce future acoustic features as additional inputs of the original acoustic model

STRODE outperforms ODE-RNN in realistic conversation speech data (CHiME-5)

N: number of hidden states per layer;
P: number of model parameters;
T: training time per epoch (hrs).

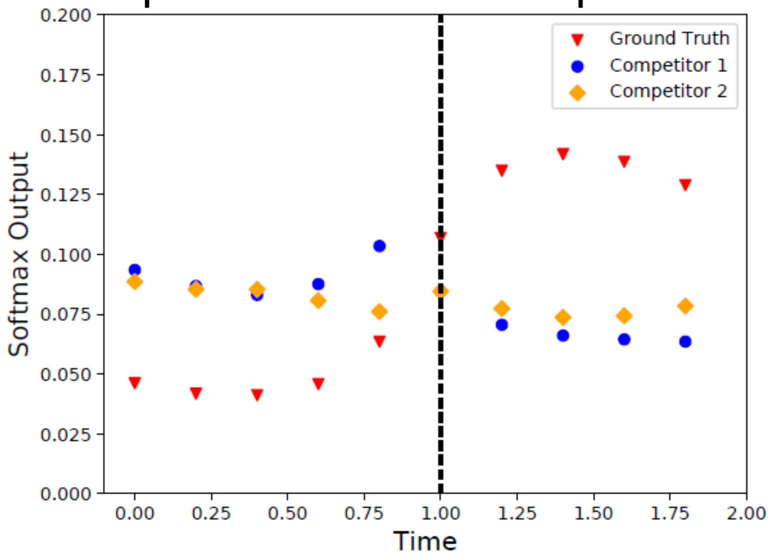
Model	N	P	T
ODE-RNN (Rubanova et al., 2019)	1100	77M	0.6
RTN (Huang et al., 2020)	1024	70M	0.3
STRODE (Ours)	900	76M	0.7

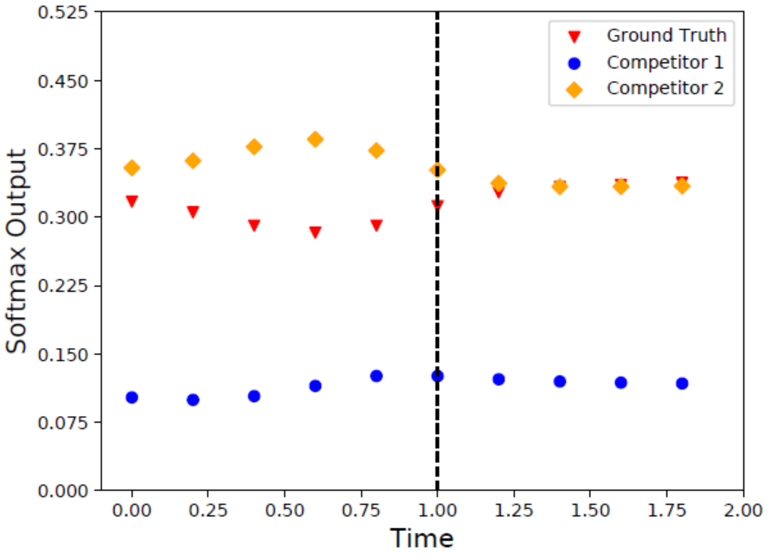
WER (%) on eval of CHiME-5

Model	WER
Kaldi DNN (Povey et al., 2011)	64.5
ODE-RNN (Rubanova et al., 2019)	59.0
RTN (Huang et al., 2020)	57.4
STRODE (Ours)	**56.3**

Biological interpretability of STRODE: it has the potential to model Postdictive mechanisms in neuroscience

The Softmax outputs by taking the ODE solutions at future time points as an extra input of the acoustic model





- The dotted line corresponds to the original Softmax output of STRODE-based acoustic model
- STRODE allows continuous-time evaluation of predictions, whose patterns surprisingly match the postdiction!

Take-away

We generalize neural ODE in handling a special type of boundary value problem with random boundary times, our STRODE

- Infers both the timings and the dynamics of time series without requiring any timing annotations during training
- Can be applied to address real-world problems, e.g., postdictive acoustic modeling
- We give a learning framework of STRODE with theoretical guarantees
- Code: https://github.com/Waffle-Liu/STRODE