# Supplementary Materials:
# STRODE: Stochastic Boundary Ordinary Differential Equation

## 1 Proof of Lemma 1

**Lemma 1.** *Let $\epsilon$ be a positive real constant. Let $U \subset R^n$ be an open set. Let $f_1, f_2 : [a-2\epsilon, a) \to R^n$ be a continuously differentiable function and $\left\|{f_1}'\right\| \leqslant M$ where $M$ is a positive constant. Let $y_1, y_2 : [a-\epsilon, a) \to U$ satisfy the initial value problem:*

$$y_1' = f_1(t),\ \ y_1(a-\epsilon) = x_1$$
$$y_2' = f_2(t) = f_1(t-\epsilon),\ \ y_2(a-\epsilon) = x_2$$

*Suppose both $x_1$ and $x_2$ depend on $\epsilon$. As $\epsilon \to 0$, we have:*

$$\lim_{\epsilon\to 0}\left(\lim_{t\to a}\|y_1(t) - y_2(t)\|\right) \leqslant \lim_{\epsilon\to 0}\|x_1 - x_2\|$$

*Proof.* For any constant $C \geqslant 0$, we have:

$$\|f_1(t) - f_1(t)\| \leqslant C\,\|x_1 - x_2\| \tag{1}$$

Let constant $K \geqslant 1$, we have:

$$\begin{aligned}\|f_1(t) - f_2(t)\| &= \|f_1(t) - f_1(t-\epsilon)\| \\ &\leqslant K\,\|f_1(t) - f_1(t-\epsilon)\|\end{aligned} \tag{2}$$

Due to Eq.(1) and Eq.(2), two assumptions of Gronwall's Inequality (Theorem 2.1) [Howard, 1998] are met. Then let $C = 0$, for any $t \in [a-\epsilon, a)$, we have:

$$\|y_1(t) - y_2(t)\| \leqslant \|x_1 - x_2\| + \underbrace{\int_{a-\epsilon}^{t} K\,\|f_1(s) - f_1(s-\epsilon)\|\,ds}_{(a)}$$

We then take part (a). There exists a $\theta_s \in (0, \epsilon)$ such that as $\epsilon \to 0$, we have:

$$\begin{aligned}\lim_{\epsilon\to 0}\left(\lim_{t\to a}\int_{a-\epsilon}^{t} K\,\|f_1(s) - f_1(s-\epsilon)\|\,ds\right) &= \lim_{\epsilon\to 0}\int_{a-\epsilon}^{a} K\,\|f_1(s) - f_1(s-\epsilon)\|\,ds \\ &= \lim_{\epsilon\to 0}\int_{a-\epsilon}^{a} K\epsilon\,\left\|{f_1}'(s-\theta_s)\right\|\,ds \\ &\leqslant \lim_{\epsilon\to 0} K\epsilon^2 M = 0\end{aligned}$$

Then we have:

$$\lim_{\epsilon\to 0}\left(\lim_{t\to a}\|y_1(t) - y_2(t)\|\right) \leqslant \lim_{\epsilon\to 0}\left(\lim_{t\to a}\left(\|x_1 - x_2\| + \int_{a-\epsilon}^{t} K\,\|f_1(s) - f_1(s-\epsilon)\|\,ds\right)\right) \tag{3}$$

$$\leqslant \lim_{\epsilon\to 0}\|x_1 - x_2\| \tag{4}$$

□

## 2 Proof of Theorem 1

**Theorem 1.** *Suppose we are given two arbitrary distributions, $q(t)$ and $p(t)$ with $t \in [0, +\infty)$. Let $m = -e^{-t}$. Let $\epsilon$ be a positive real constant and let $g : [-1, 0) \to R$ be a continuous function. There exists a $G : [-1, 0) \to [0, +\infty)$ that satisfies an initial value problem:*

$$\begin{aligned}
&G'(m) = g(m), \ \ G(-1) = 0 \\
&where \ g(m) = \frac{-q(-\log(-m))}{m} \log \frac{q(-\log(-m))}{p(-\log(-m))}
\end{aligned}$$

*Such that as $\epsilon \to 0$, we have:*

$$\lim_{\epsilon \to 0} (\mathrm{KL}(q(t)||p(t))) \leqslant \lim_{\epsilon \to 0} (G(-\epsilon) + \|G(-2\epsilon) - G(-\epsilon)\|)$$

*Proof.* Let $m \in [-1, 0)$ and let $t = -\log(-m)$, we have:

$$\begin{aligned}
\mathrm{KL}(q(t)||p(t)) &= \int_0^{+\infty} q(t) \log(\frac{q(t)}{p(t)}) dt \\
&= \lim_{l \to 0} \int_{-1}^{l} \underbrace{\frac{-q(-\log(-m))}{m} \log \frac{q(-\log(-m))}{p(-\log(-m))}}_{g(m)} dm
\end{aligned}$$

With the integrand $g(m)$, we then construct a $G : [-1, 0) \to [0, +\infty)$ that satisfies an initial value problem:

$$G'(m) = g(m), \ \ G(-1) = 0$$

Then we have:

$$\mathrm{KL}(q(t)||p(t)) = \lim_{l \to 0} G(l) = \lim_{l \to 0} \int_{-1}^{l} g(m) dm$$

Since $g(m)$ is not analytic at the point 0, we separate the solution $\lim_{l \to 0} G(l)$ into two parts:

$$\begin{aligned}
\mathrm{KL}(q(t)||p(t)) &= \int_{-1}^{-\epsilon} g(m) dm + \lim_{l \to 0} \int_{-\epsilon}^{l} g(m) dm \\
&= G(-\epsilon) + \lim_{l \to 0} \int_{-\epsilon}^{l} g(m) dm \qquad (5)
\end{aligned}$$

Let $G_1, G_2 : [-\epsilon, 0) \to [0, +\infty)$ satisfy the initial value problem:

$$\begin{aligned}
&G_1' = g(m), \ \ G_1(-\epsilon) = G(-\epsilon) \\
&G_2' = g(m - \epsilon), \ \ G_2(-\epsilon) = G(-2\epsilon)
\end{aligned}$$

By Lemma 1, as $\epsilon \to 0$ we have:

$$\begin{aligned}
\lim_{\epsilon \to 0} \left( \lim_{l \to 0} \int_{-\epsilon}^{l} g(m) dm \right) &= \lim_{\epsilon \to 0} \left( \lim_{m \to 0} \|G_1(m) - G_2(m)\| \right) \\
&\leqslant \lim_{\epsilon \to 0} \|G(-\epsilon) - G(-2\epsilon)\| \qquad (6)
\end{aligned}$$

Combining Eq.(5) and Eq.(6), we have:

$$\begin{aligned}
\lim_{\epsilon \to 0} (\mathrm{KL}(q(t)||p(t))) &= \lim_{\epsilon \to 0} (G(-\epsilon) + \lim_{l \to 0} \int_{-\epsilon}^{l} g(m) dm) \qquad (7) \\
&\leqslant \lim_{\epsilon \to 0} (G(-\epsilon) + \|G(-\epsilon) - G(-2\epsilon)\|) \qquad (8)
\end{aligned}$$

□

# 3 Experiments

In this section, we provide more details for experiments. Notably, any ODE involved in this experiment are solved by the Euler method with step size 0.1 (such hyperparameter can be tuned to further improve performance).

## 3.1 Training Procedure of STRODEs on Toy Dataset

Our STRODE on toy dataset is trained by maximizing the ELBO, in which the likelihood term is simplified by an MSE term. We adopt the Adam optimizer with a learning rate of $4 \times 10^{-4}$. We apply a dropout rate of 0.1 to the connections between neural network layers except that of ODE solvers.

## 3.2 Detailed Implementation of STRODE on Rotating MNIST Thumbnail

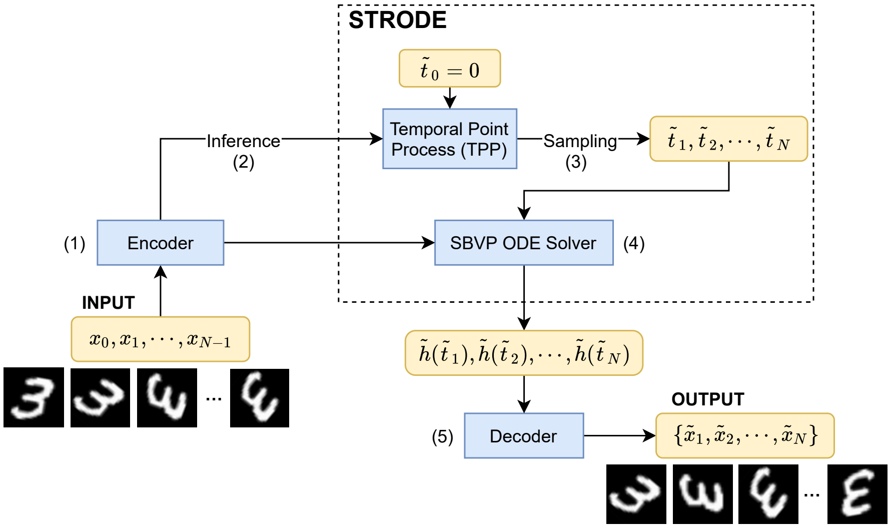
STRODE
$\tilde{t}_0 = 0$
Temporal Point Process (TPP)
Inference (2)
Sampling (3)
$\tilde{t}_1, \tilde{t}_2, \cdots, \tilde{t}_N$
(1)
Encoder
SBVP ODE Solver
(4)
INPUT
$x_0, x_1, \cdots, x_{N-1}$
...
$\tilde{h}(\tilde{t}_1), \tilde{h}(\tilde{t}_2), \cdots, \tilde{h}(\tilde{t}_N)$
OUTPUT
(5)
Decoder
$\{\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_N\}$
...

Figure 1: Architecture of STRODE for Rotating MNIST Thumbnail

(1) **Encoder**: our encoder for this task contains four convolution layers (kernel size: $5 \times 5$) with the following input-output dimensions: $1 \times 64 \times 64 \to 128 \times 16 \times 16 \to 256 \times 8 \times 8 \to 512 \times 4 \times 4 \to 512 \times 1 \times 1$, where numbers indicate #feature maps$\times$width $\times$height. Batch-normalization is applied to each layer except the first and last ones. Activations are LeakyReLU.

(2) **Inference**: the implementation is similar to what we adopt on the toy dataset, except that the architecture of the neural network for both $\Phi_i(t)$ and $\phi_i(t)$ includes 2 fully connected layers, each with 128 hidden nodes and Tanh, whose outputs are further transformed into a scalar by another fully connected layer with Softplus.

(3) **Sampling**: the implementation is similar to what we adopt on the toy dataset.

(4) **SBVP ODE Solver**: the implementation is similar to what we adopt on the toy dataset.

(5) **Decoder**: our decoder for this task contains four deconvolution layers (kernel size $5 \times 5$ ) with the following input-output dimensions: $512 \times 1 \times 1 \to 512 \times 4 \times 4 \to 256 \times 8 \times 8 \to 128 \times 16 \times 16 \to 1 \times 64 \times 64$, where numbers indicate #feature maps$\times$width $\times$height. We adopt a batch-normalization layer with ReLU activation after each deconvolution layer except the last one. Tanh activation is applied after the last deconvolution layer.

## 3.3 Detailed Implementation of STRODE on CHiME-5

(1) **Encoder**: we adopt a 9-layer SRU to implement the encoder. Each SRU layer contains 900 hidden states.

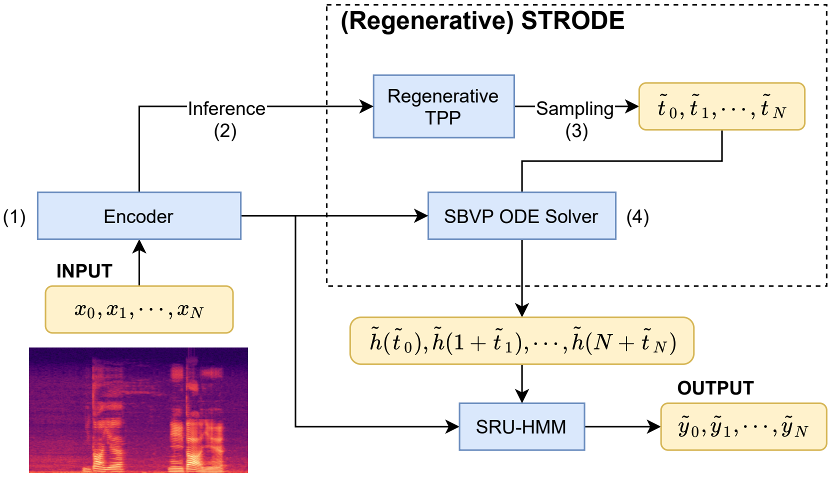
(Regenerative) STRODE
Regenerative TPP
Inference (2)
Sampling (3)
$\tilde{t}_0, \tilde{t}_1, \cdots, \tilde{t}_N$
(1)
Encoder
SBVP ODE Solver
(4)
INPUT
$x_0, x_1, \cdots, x_N$
$\tilde{h}(\tilde{t}_0), \tilde{h}(1+\tilde{t}_1), \cdots, \tilde{h}(N+\tilde{t}_N)$
OUTPUT
SRU-HMM
$\tilde{y}_0, \tilde{y}_1, \cdots, \tilde{y}_N$

Figure 2: Architecture of STRODE for CHiME-5

(2) **Inference**: the implementation is similar to what we adopt on the toy dataset, except that the architecture of the neural network for both $\Phi_i(t)$ and $\phi_i(t)$ includes 2 fully connected layers, each with 128 hidden nodes and Tanh, whose outputs are further transformed into a scalar by another fully connected layer with Softplus.

(3) **Sampling**: the boundary time samples are sequentially generated through adopting Eq. (34) in the main paper.

(4) **SBVP ODE Solver**: we follow Eq. (35) of the main paper to obtain the SBVP ODE solution for each frame, in which the neural network $f_{\theta_o}$ includes 2 fully connected layers, each with 64 hidden nodes and Tanh.

# References

[Howard, 1998] Howard, R. (1998). The gronwall inequality. *lecture notes*.