Journal of Machine Learning Research 26 (2025) 1-67 Submitted 11/24; Revised 7/25; Published 7/25

On Non-asymptotic Theory of Recurrent Neural Networks in Temporal Point Processes

Zhiheng Chen ZHCHEN22@M.FUDAN.EDU.CN
Shanghai Center for Mathematical Sciences
Fudan University
Shanghai, China

Guanhua Fang FANGGH@FUDAN.EDU.CN
Wen Yu WENYU@FUDAN.EDU.CN
Department of Statistics and Data Science
Fudan University
Shanghai, China

Editor: Gabor Lugosi

Abstract

Temporal point process (TPP) is an important tool for modeling and predicting irregularly timed events across various domains. Recently, the recurrent neural network (RNN)-based TPPs have shown practical advantages over traditional parametric TPP models. However, in the current literature, it remains nascent in understanding neural TPPs from theoretical viewpoints. In this paper, we establish the excess risk bounds of RNN-TPPs under many well-known TPP settings. We especially show that an RNN-TPP with no more than four layers can achieve vanishing generalization errors. Our technical contributions include the characterization of the complexity of the multi-layer RNN class, the construction of tanh neural networks for approximating dynamic event intensity functions, and the truncation technique for alleviating the issue of unbounded event sequences. Our results bridge the gap between TPP's application and neural network theory.

Keywords: Recurrent Neural Network, Point Process, Tanh, Excess Risk, Approximation Theory

1. Introduction

Temporal point process (TPP) (Daley and Vere-Jones, 2003, 2008) is an important mathematical framework that provides tools for analyzing and predicting the timing and patterns of events in continuous time. TPP particularly deals with event streaming data where the events occur at irregular time stamps, which is different from classical time series analysis that often assumes a regular time spacing between data points. In real world applications, the events could be anything from transactions in financial markets (Bauwens and Hautsch, 2009; Hawkes, 2018) to user activities in online social network platforms (Farajtabar et al., 2017; Fang et al., 2023), earthquakes in seismology (Wang et al., 2012; Laub et al., 2021), neural spikes in biological experiments (Perkel et al., 1967; Williams et al., 2020), or failure times in survival analysis (Aalen et al., 2008; Fleming and Harrington, 2013).

With the advent of artificial intelligence in last decades, the neural network (McCulloch and Pitts, 1943) has been proved to be a powerful architecture that can be adapted to different applications with distinct purposes. In modern machine learning, researchers have also incorporated deep neural networks into TPPs to handle complex patterns and dependencies in event data, leading to advancements in many areas such as recommendation systems (Du et al., 2015; Hosseini et al., 2017), social network analysis (Du et al., 2016; Zhang et al., 2021), healthcare analytics (Li et al., 2018; Enguehard et al., 2020), etc. Many new TPP models have been proposed in the recent literature, including but not limited to, recurrent temporal point process (Du et al., 2016), fully neural network TPP model (Omi et al., 2019), transformer Hawkes process (Zuo et al., 2020); see Shchur et al. (2021); Lin et al. (2022) and the references therein for a more comprehensive review.

Despite the recent progress in TPP's applications as mentioned above, there is a lack of understanding in neural TPPs from the theoretical perspective. A fundamental question remains: whether the neural network-based TPP can provably have a small generalization error? In this paper, we provide an affirmative answer to this question for *recurrent neural network* (RNN, Medsker and Jain (1999))-based TPPs. To be specific, we establish the non-asymptotic rates of generalization error bounds under mild model assumptions and provide the construction of RNN architectures that could approximate many widely-used TPPs, including homogeneous Poisson process, non-homogeneous Poisson process, self-exciting process, etc.

There are a few challenges in developing the theory of RNN-based TPPs. (a) *Characterization of functional space.* In the machine learning theory, it is necessary to specify the model space to derive any generalization errors. In our setting, the thing becomes more complicated since the model should be data-dependent (i.e., adapts to the past events). Otherwise, the model could not capture the information in event history and fail to provide a good fitting. (b) *Expressive power of RNN architecture.* RNN is the most widely adopted neural architecture in TPP modelling. However, it remains questionable whether the RNNs can approximate most well-known temporal point processes. If the answer is yes, it would be of great interest to know how many hidden layers and how large hidden dimensions will be sufficient for the approximation. (c) *Expressive power of activation function.* In modern neural networks, the activation function is chosen to be a simple non-linear function for the sake of computational feasibility. In RNNs, it is taken as the "tanh" by default. Then it is important to understand the approximability of tanh activation functions. (d) *Variable unbounded length of event sequence.* Unlike the standard RNN's modelling (Tu et al., 2020) where each sample has the same length (or is padded to have the same length), the event sequences in our setting may vary from one to another. In addition, their lengths are potentially unbounded. These add difficulties in computing the complexity of the model space.

To overcome the above challenges, we adopt the following approaches. (a) In TPPs, the intensity function is the core. We recursively construct the multi-layer hidden cells through RNNs to store the event information and adopt the suitable output layer to compute the intensity value. Equipped with suitable input embeddings, our construction can capture the information of event history and adapt to variable lengths of event sequence. (b) For four main categories of TPPs, homogeneous Poisson process, non-homogeneous Poisson process, self-exciting process, and self-correcting process, we carefully study their intensity formula. We can decompose the intensity function into different parts and approximate them component-wisely. Our construction explicitly gives the upper bounds on the model depth, the width of hidden layers, and parameter weights of the RNN architecture to achieve a certain level of approximation accuracy. (c) We use the results in a recent work (De Ryck et al., 2021), where they provide the approximation ability of one- and two-layer tanh neural networks. We adapt such results to our specific RNN structure and give the universal approximation results for each of the intensity components. (d) Thanks to the exponential decay property of the tail probability of the sequence length, we are able to use the truncation technique to decouple the randomness of independent and identically distributed (i.i.d.) samples and the lengths of event sequences. For the space of truncated loss functions, the space complexity can be obtained through calculating the covering number. The classical chaining methods in empirical process theory can hence be applied as well.

Our main technical contributions can be summarized as follows.

(i) We carefully design the functional space of RNN-based TPP models, where the intensity functions are defined through the recursive formula and the interpolation of hidden states. We also choose the proper metric to characterize the distance between two different RNN-based TPPs so that the excess risk analysis becomes possible. We believe our formulation could inspire the theoretical analyses of other continuous-time neural network models.

(ii) In the analysis of the stochastic error in the excess risk of RNN-based TPPs, we provide a truncation technique to decompose the randomness into a bounded component and a tail component. By carefully balancing between the two parts, we establish a nearly optimal stochastic error bound. Additionally, we also derive the complexity of the multi-layer RNN-based TPP class, where we precisely analyze and compute the Lipschitz constant of RNN architecture. This extends the existing result in Chen et al. (2020) where they only give the Lipschitz constant of a single-layer RNN. Therefore, our truncation technique and the Lipschitz result of multi-layer RNNs can be useful and of independent interest for many other related problems.

(iii) We establish the approximation error bounds for the intensity functions of TPPs of four main categories. To the best of our knowledge, there is very few work (De Ryck et al., 2021) on studying the approximation property of tanh activation function. Our work is the first one to provide approximation results for RNN-based statistical models. Our construction procedure largely depends on the Markov nature (Laub et al., 2021) of self-exciting processes so that we can design hidden cells to store sufficient information of past events. Moreover, we decompose the excitation function into different parts. Each of them is a simple smooth function (i.e. either exponential function or trigonometric function) that can be well approximated by a single-layer tanh network. Our construction method can be viewed as a useful tool in analyzing other sequential-type neural networks.

(iv) We illustrate the differences between the architectures of classical RNNs and RNN-based TPPs. Note the fact that the observed events happen at the discrete time grids, while the TPP models should take into account the continuous time domain. Therefore, the interpolation of values in hidden cells at each time point is important and necessary. We show that improper interpolation mechanisms (e.g. constant, linear, exponential decay interpolation) may fail to provide RNN-based TPP with the universal approximation ability.

Our result indicates that the input embedding plays an important role in interpolating the hidden states.

The rest of paper is organized as follows. In Section 2, the background of TPPs, the formulation of RNN-based TPPs, and useful notations are introduced. The main theories along with high-level explanations are given in Section 3. The technical tools for analyzing stochastic errors are provided in Section 4. The construction procedures for approximating different types of intensity functions are listed in Section 5. In Section 6, we provide explanations that the improper interpolation of hidden states in RNN-TPPs may lead to unsatisfactory approximation results. Detailed discussions and concluding remarks are given in Section 7.

2. Preliminaries

2.1 Framework Specification

We observe a set of n irregular event time sequences,

$$\mathbf{D}_{train} := \{S_i; i = 1, ..., n\} = \{(t_{i,1}, ..., t_{i,N_{ei}}); i = 1, ..., n\}, \tag{1}$$

where $0 < t_{i,1} < ... < t_{i,j} < ... < t_{i,N_{ei}} \leq T$ with T being the end time point, and N_{ei} is the number of events in the i-th sequence, S_i. It is assumed that each of S_i's is independently generated from a TPP model with an unknown intensity function $\lambda^*(t)$ defined on $[0, T]$. That is,

$$\lambda^*(t) := \lim_{dt \to 0} \frac{\mathbb{E}[N[t, t+dt)|\mathcal{H}_t]}{dt},$$

where $N[t, t+dt) := N(t+dt) - N(t)$ with $N(t) := \sharp\{j : t_j \leq t\}$ being the number of events observed up to time t, and $\mathcal{H}_t := \sigma(\{N(s); s < t\})$ is the history filtration before time t.

In the literature of TPP's learning (Shchur et al., 2021), the primary goal is to estimate $\lambda^*(t)$ based on $\mathbf{D}_{train}$. Throughout the current work, we adopt the negative log-likelihood function as our objective. To be specific, for any event time sequence $S = (t_1, .., t_{N_e})$, we define

$$\text{loss}(\lambda, S) := -\left\{\sum_{j=1}^{N_e} \log \lambda(t_j) - \int_0^T \lambda(t) \mathrm{d}t\right\}. \tag{2}$$

This loss function is widely used in many literatures (Du et al., 2016; Mei and Eisner, 2017; Zuo et al., 2020). Then the estimator can be defined as

$$\begin{aligned}\hat{\lambda} \;&:=\; \arg\min_{\lambda \in \mathcal{F}} \text{loss}(\lambda) \\ &:=\; \arg\min_{\lambda \in \mathcal{F}} \left\{\frac{1}{n}\sum_{i=1}^{n} \text{loss}(\lambda, S_i)\right\},\end{aligned} \tag{3}$$

where $\mathcal{F}$ is a user-specified functional space. For example, in the existing works, $\mathcal{F}$ can be taken as any space of parametric models (Schoenberg, 2005; Laub et al., 2021), nonparametric models (Cai et al., 2022; Fang et al., 2023), or neural network models (Du et al., 2016; Mei and Eisner, 2017).

In the language of deep learning, $\mathbf{D}_{train}$ is also called a training data set. $\text{loss}(\lambda)$ is known as the loss function of predictor λ. $\hat{\lambda}$ defined in (3) is the empirical risk minimizer (ERM). To evaluate the performance of $\hat{\lambda}$, a common practice in machine (deep) learning is using the excess risk (Hastie et al., 2009; James et al., 2013; Vidyasagar, 2013; Shalev-Shwartz and Ben-David, 2014). To be mathematically formal, we define

$$\text{ER}(\hat{\lambda}) := \mathbb{E}[\text{loss}(\hat{\lambda}, S_{test})] - \mathbb{E}[\text{loss}(\lambda^*, S_{test})], \tag{4}$$

where S_{test} is a testing sample, i.e., a new event time sequence, which is independent of $\mathbf{D}_{train}$ and also follows the intensity $\lambda^*(t)$. The expectation here is taken with respect to the new testing data. We give a proof of $\text{ER}(\hat{\lambda}) \geq 0$ in the supplementary. As a result, (4) is a well-defined excess risk under our model setup.

2.2 RNN Structure

Throughout this paper, we consider $\mathcal{F}$ to be a space of RNN-based TPP models. An arbitrary intensity function λ in $\mathcal{F}$, indexed by the parameter θ, is defined through the following recursive formula,

$$\lambda_\theta(t; S) \quad := \quad f\left(W_x^{(L+1)} h^{(L)}(t; S) + b^{(L+1)}\right) \in \mathbb{R}^1, \tag{5}$$

where the hidden vector function $h^{(L)}(t; S)$ has the following hierarchical form,

$$\begin{aligned}
h^{(1)}(t; S) &= \sigma\left(W_x^{(1)} x(t; S) + W_h^{(1)} h_j^{(1)} + b^{(1)}\right), \\
h^{(2)}(t; S) &= \sigma\left(W_x^{(2)} h^{(1)}(t; S) + W_h^{(2)} h_j^{(2)} + b^{(2)}\right), \\
&\vdots \\
h^{(L)}(t; S) &= \sigma\left(W_x^{(L)} h^{(L-1)}(t; S) + W_h^{(L)} h_j^{(L)} + b^{(L)}\right), \quad \text{for } t \in (t_j, t_{j+1}],
\end{aligned} \tag{6}$$

with

$$\begin{aligned}
h_j^{(1)} &= \sigma\left(W_x^{(1)} x(t_j; S) + W_h^{(1)} h_{j-1}^{(1)} + b^{(1)}\right), \\
h_j^{(2)} &= \sigma\left(W_x^{(2)} h_j^{(1)} + W_h^{(2)} h_{j-1}^{(2)} + b^{(2)}\right), \\
&\vdots \\
h_j^{(L)} &= \sigma\left(W_x^{(L)} h_j^{(L-1)} + W_h^{(L)} h_{j-1}^{(L)} + b^{(L)}\right), \quad \text{for } j \in \{1, ..., N_e\}.
\end{aligned} \tag{7}$$

Here σ, f are two known activation functions of the hidden layers and the output layer, respectively. Both of them are pre-determined by the user. We specifically take $\sigma(x) = \tanh(x) = (\exp(x) - \exp(-x))/(\exp(x) + \exp(-x))$ and $f(x) = \min\{\max\{x, l_f\}, u_f\}$, where l_f and u_f are two fixed positive constants. The input embedding vector function $x(t; S)$ is also known to the user before training. In the current work, we particularly take $x(t; S) = (t, F_S(t))^\top$ where $F_S(t) = t - t_j$ for $t \in (t_j, t_{j+1}]$, $\forall j \in N_e$. The hidden dimension of the l-th layer is denoted by d_l, i.e. $h_j^{(l)} \in \mathbb{R}^{d_l}$, $1 \leq l \leq L$. The model parameters consist of $W_x^{(l)}$, $W_h^{(l)}$, $b^{(l)}$ $(1 \leq l \leq L)$, and $W_x^{(L+1)}$, b^{L+1}, where $W_x^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$, $W_h^{(l)} \in \mathbb{R}^{d_l \times d_l}$ are

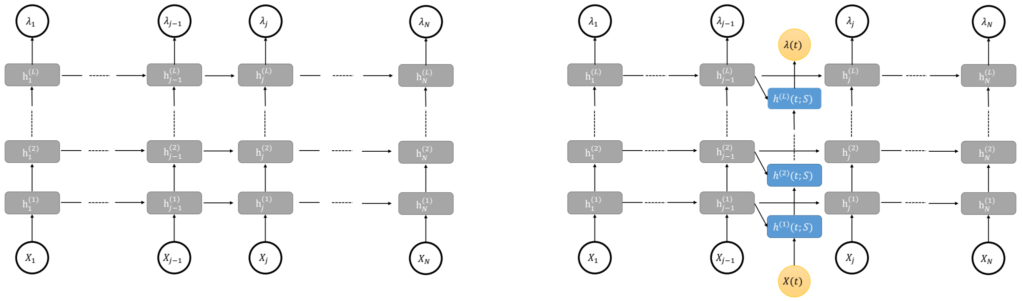


Figure 1: Left: the classical RNN architecture. Right: the RNN-TPP architecture given in (5) - (7). The blue box represents the interpolation of hidden states.

the weight matrices in l-th layer, and $b^{(l)} \in \mathbb{R}^{(l)}$ is the bias vector in l-th layer (here d_0 is the input dimension, i.e., the dimension of $x(t;S)$, and $d_{L+1} = 1$ is the output dimension). For ease of notation, we concatenate all parameter matrices and vectors and write as $\theta = \{W_x^{(l)}, W_h^{(l)}, b^{(l)}; 1 \leq l \leq L+1\}$, where $W_h^{(L+1)} \equiv \mathbf{0}$. By default, we take the initial values $t_0 \equiv 0$ and $h_0^{(l)} \equiv \mathbf{0}$ for $1 \leq l \leq L$. The last time grid $t_{N_e+1} \equiv T$. We call the model defined through equations (5) - (7) as the RNN-TPP. To help readers to gain more intuition, a toy example of RNN-TPP is demonstrated in Appendix A.

Moreover, we define the maximum hidden size $D := \max\{d_1, d_2 \cdots d_L\}$, and the parameter norm

$$\|\theta\| := \max\left\{\|W_x^{(l)}\|_2, \|W_h^{(l)}\|_2, \|b^{(l)}\|_2; 1 \leq l \leq L+1\right\}.$$

Then the RNN-TPP class $\mathcal{F}$ is described by

$$\mathcal{F} = \mathcal{F}_{L,D,B_m,l_f,u_f} := \{\lambda_\theta;\ \|\theta\| \leq B_m\}, \tag{8}$$

where B_m may depend on the hidden size D and the sample size n. To help readers gain more intuitions, a graphical illustration of the network structure is given in Figure 1.

Remark 2.1. *The default choice (De Ryck et al., 2021) of activation function $\sigma(x)$ in RNNs is $\tanh(x)$. In practice, the number of layers L is usually no more than 4.*

Remark 2.2. *By the constructions* (5) - (7)*, it is not hard to see that the intensity $\lambda_\theta(t;S)$ is a left-continuous function of t. In other words, it is a well-defined predictable function with respect to the information filtration generated by event sequence S.*

Remark 2.3. *In the standard application of RNN models, the training data usually consist of discrete-time sequences (e.g., sequences of tokens in natural language processing (NLP) (Yin et al., 2017; Tarwani and Edem, 2017); time series in financial market forecasting (Cao et al., 2019; Chimmula and Zhang, 2020)). Therefore, the classical (single-layer) RNN architecture is defined only through the discrete time grids. That is, the hidden vector at j-th grid is*

$$h_j = \sigma\left(W_x x_j + W_h h_{j-1} + b_h\right),$$

where x_j is the corresponding embedding input. The prediction at time step j is given by $y_j = f(W_y h_j + b_y) \in \mathbb{R}$. In contrast, the RNN-based TPP model should take into account any time point t between grids t_j and t_{j+1}. Hence the interpolation of $h^{(l)}(t; S)$ between $h_j^{(l)}$ and $h_{j+1}^{(l)}$ is heuristically necessary to give reasonable model predictions over the entire time interval $(t_j, t_{j+1}]$.

Remark 2.4. *In the literature, there exist a few methods to interpolate the hidden embedding between $h_j^{(L)}$ and $h_{j+1}^{(L)}$. In Du et al. (2016), a constant embedding mechanism is used, i.e. $h^{(l)}(t; S) \equiv h_j^{(l)}$ for $t \in (t_j, t_{j+1}]$ and any j and l. In Mei and Eisner (2017), the author adopted an exponential decay method to encode the hidden representations under an extended RNN architecture, Long Short Term Memory (LSTM) network. More recently, Rubanova et al. (2019) used the neural ordinary differential equation (ODE) method for solving the intermediate hidden state $h^{(l)}(t; S)$.*

It can be shown that the first two interpolation methods are unable to precisely capture the true intensity in the sense of excess risk. We will give the explanation in Section 6; see Theorem 6.1.

Remark 2.5. *Our result still holds if tanh is replaced with other Sigmoidal-type activation functions (Cybenko, 1989) (e.g., ReLU (Fukushima, 1969)). In the literature of TPP modelling, the most common choice of $f(x)$ is the Softplus function (Dugas et al., 2001; Zhou et al., 2022), $\log(1 + \exp(x))$, which ensures $\lambda_\theta(t; S)$ to be positive and differentiable. Our result also holds if we take $f(x)$ to be $\min\{\max\{\log(1 + \exp(x)), l_f\}, u_f\}$ with $0 < l_f < u_f$. Introducing l_f and u_f only serves the technical purpose, i.e., the predicted intensity value is bounded from above and below.*

2.3 Classical TPPs

In the statistical literature, TPPs can be categorized into several types based on the nature of the intensity functions. Four main categories are summarized as follows.

Homogeneous Poisson process (Kingman, 1992). It is the simplest type where events occur completely independently of one another, and the intensity function is constant, i.e., $\lambda^*(t) \equiv \lambda$, where λ is unknown and needs to be estimated.

Non-homogeneous Poisson process (Kingman, 1992; Daley and Vere-Jones, 2003). In this model, the intensity function varies over time but is still independent of past events. That is, $\lambda^*(t)$ is a non-constant unknown function that is usually estimated via certain nonparametric methods.

Self-exciting process (Hawkes and Oakes, 1974). Future events are influenced by past events, which can lead to clustering of events in time. A well-known example is the Hawkes process (Hawkes, 1971; Hawkes and Oakes, 1974), where the intensity function takes form,

$$\lambda^*(t) = \lambda_0(t) + \sum_{j: t_j < t} \mu(t - t_j), \tag{9}$$

where $\lambda_0(t)$ and $\mu(t)$ are some positive functions which are called the background intensity and excitation/impact function, respectively. In many applications (Laub et al., 2021), the excitation function takes the exponential form that $\mu(t) = \alpha \exp(-\beta t)$, which allows the efficient computation. The model defined in (9) is also known as the *linear* self-exciting process since the intensity is in an additive form of different components. More generally, the non-linear self-exciting process (Brémaud and Massoulié, 1996)

$$\lambda^*(t) = \Psi\left(\lambda_0(t) + \sum_{j:t_j<t} \mu(t-t_j)\right), \tag{10}$$

is also considered in the literature, where Ψ is a non-linear function.
Self-correcting process (Isham and Westcott, 1979; Ogata and Vere-Jones, 1984). The occurrence of an event decreases the likelihood of future events for some time period. To be mathematically formal, the intensity postulates the formula,

$$\lambda^*(t) = \Psi\left(\mu t - \sum_{j:t_j<t} \alpha\right), \tag{11}$$

where both μ and α are positive and Ψ may be a non-linear function.

2.4 Notations

Let $a \wedge b = \min\{a,b\}$ and $a \vee b = \max\{a,b\}$. We use $\mathbb{N}$ and $\mathbb{Z}$ to denote the set of nonnegative integers and all integers, respectively. Denote $[n] = \{1, 2\cdots, n\}$ for a positive integer n. Let $\lceil a \rceil = \min\{b \in \mathbb{Z}, b \geq a\}$. For a set A, denote $\#(A)$ to be its cardinality. For a vector $x = (x_1, \cdots, x_d)^\top \in \mathbb{R}^d$, denote its Euclidean norm as $\|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$. Write $a_N \lesssim b_N$ if there exists some constant $C > 0$ such that $a_N \leq C b_N$ for all index N, and the range of N may be defined case by case. For a function f defined on some domain, denote $\|f\|_{L^\infty}$ as its essential upper bound. For $s \in \mathbb{N}$, the Sobolev norm $\|f\|_{W^{s,\infty}([0,T])}$ is defined as $\|f\|_{W^{s,\infty}([0,T])} = \max_{0 \leq |\alpha| \leq s} \|D^\alpha f\|_{L^\infty([0,T])}$. For a constant $B_0 > 0$, the B_0-ball of Sobolev space $W^{s,\infty}([0,T])$ is defined as

$$W^{s,\infty}([0,T], B_0) := \left\{f \in W^{s,\infty}([0,T]), \|f\|_{W^{s,\infty}([0,T])} \leq B_0\right\}.$$

For constant $C_0 > 0$, the ball $C^{s,\infty}([0,T], C_0)$ is a subset of $W^{s,\infty}([0,T], C_0)$ which contains all s-order smooth functions. In this work, our results are non-asymptotic. For notational brevity, we use $O(\cdot)$ to hide all constants and use $\tilde{O}(\cdot)$ to denote $O(\cdot)$ with hidden log factors. Throughout this paper, α, β, γ, $\mathcal{C}$, and $\mathcal{C}_1$ are positive real numbers and may be defined case by case. In the remaining part of the article, λ^* denotes the ground truth of the intensity function.

3. Main Results

Recent applications in event stream analyses have witnessed the usefulness of TPPs with the incorporation of RNNs. However, there is no study in the existing literature to explain why the RNN structure in TPP modeling is so useful from the theoretical perspective. We attempt to answer the question of whether the RNN-TPPs can *provably* have small generalization error or excess risk. Our answer is **positive**! When the event data are generated according to the classical models described in Section 2.3, we show that the RNN-TPPs can perfectly generalize such data.

To make our presentation easier, we only need to focus on the self-exciting processes. [1] To start with, we first consider the linear case (9).

Some regularity assumptions should be stated before we present the main theorem.

(A1) There exists a constant $B_0 > 0$ such that $\lambda_0 \in W^{s,\infty}([0,T], B_0)$, where $s \geq 1$, $s \in \mathbb{N}$.

(A2) $\int_0^T \mu(t)\mathrm{d}t := c_\mu < 1$.

(A3) There exists a positive constant B_1 such that $\inf_{t\in[0,T]} \lambda_0(t) \geq B_1$.

Assumption (A1) assumes the boundedness of the background intensity, which is also common in neural network approximation studies. Assumption (A2) is standard in the literature of Hawkes process, which guarantees the existence of a stationary version of the process when $\lambda_0(t)$ is constant. Assumption (A3) is an informative lower bound assumption, which ensures that sufficient intensity exists in any subdomain of $[0,T]$.

Now we can present the results on the non-asymptotic bound of excess risk (4) under model (9).

Theorem 3.1. *Under model* (9) *and RNN-TPP class* $\mathcal{F} = \mathcal{F}_{L,D,B_m,l_f,u_f}$ *defined as* (8)*, suppose that assumptions (A1)-(A3) hold, then for* n *i.i.d. sample series* $\{S_i, i \in [n]\}$*, with probability at least* $1-\delta$*, the excess risk* (4) *of ERM* (3) *satisfies:*

(i) (Poisson case) If $\mu \equiv 0$*, for* $L = 2$*,* $D = \tilde{O}(n^{\frac{1}{2(s+1)}})$*,* $B_m = \tilde{O}(n^{\frac{s+1}{4}})$*,* $l_f = B_1 \wedge 1$*, and* $u_f = B_0$*,*

$$\mathbb{E}[loss(\hat{\lambda}, S_{test})] - \mathbb{E}[loss(\lambda^*, S_{test})] \leq \tilde{O}\left(n^{-\frac{s}{2(s+1)}}\right); \tag{12}$$

(ii) (Vanilla Hawkes case) If $\mu(t) = \alpha\exp(-\beta t)$*, for* $L = 2$*,* $D = \tilde{O}(n^{\frac{1}{2(s+1)}})$*,* $B_m = \tilde{O}((\log n)^{3s^2\log^2 n})$*,* $l_f = B_1 \wedge 1$*, and* $u_f = B_0 + O(\log n)$*,*

$$\mathbb{E}[loss(\hat{\lambda}, S_{test})] - \mathbb{E}[loss(\lambda^*, S_{test})] \leq \tilde{O}\left(n^{-\frac{s}{2(s+1)}}\right); \tag{13}$$

(iii) (General case) If $\mu \in C^{k,\infty}([0,T], C_0)$*,* $k \geq 2$*,* $k \in \mathbb{N}$*, for* $L = 2$*,* $D = \tilde{O}(n^{\frac{1}{2}\left(\frac{1}{s+1}\vee\frac{5}{k+4}\right)})$*,* $B_m = \tilde{O}((\log n)^{3s^2\log^2 n})$*,* $l_f = B_1 \wedge 1$*, and* $u_f = B_0 + O(\log n)$*,*

$$\mathbb{E}[loss(\hat{\lambda}, S_{test})] - \mathbb{E}[loss(\lambda^*, S_{test})] \leq \tilde{O}\left(n^{-\frac{1}{2}\left(\frac{s}{s+1}\wedge\frac{k-1}{k+4}\right)}\right). \tag{14}$$

As suggested in Theorem 3.1, there exists a two-layer RNN-TPP model whose excess risk becomes vanishing when the size of the training set goes to infinity. The width of such network grows with the sample size, while the depth remains two.

[1] Homogeneous Poisson, non-homogeneous Poisson, and self-correcting process can be treated similarly due to the following reasons. If we take $\mu(t) \equiv 0$ in (9), the linear self-exciting process reduces to the homogeneous Poisson or non-homogeneous Poisson process. In the RNN-TPP architecture, we can take the input embedding function $x(t;S) = (t, t - F_S(t), N(t-))$, i.e., using an additional input dimension to store the number of past events. Then establishing the excess risk of self-correcting process is technically equivalent to that of non-homogeneous Poisson process.

Remark 3.2. *Here we require the depth of RNN-TPP $L = 2$ due the fact that $\lambda_0 \in W^{s,\infty}([0,T], B_0)$. However, if we allow λ_0 to be sufficiently smooth (i.e., $\lambda_0 \in C^{\infty}([0,T])$), we only need one-layer* tanh *neural network to approximate λ_0. As a result, the number of layers of RNN-TPP can be reduced to one.*

Now we consider the true model to be a non-linear Hawkes process, which is given in (10). For simplicity, we only consider the case $\mu(t) = \alpha \exp(-\beta t)$, which is

$$\lambda^*(t) = \Psi\left(\lambda_0(t) + \sum_{t_j<t} \alpha \exp(-\beta(t-t_j))\right). \tag{15}$$

The regularity of Ψ is presented as Assumption (A4).

(A4) Function Ψ is L-Lipschitz, positive and bounded. In other words, there exist $\tilde{B_1}, \tilde{B_0} > 0$ such that $\tilde{B_1} \leq \Psi \leq \tilde{B_0}$ and $|\Psi(x_1) - \Psi(x_2)| \leq L|x_1 - x_2|$ for any x_1, x_2.

We have a similar bound of excess risk (4) under model (15).

Theorem 3.3. *(Nonlinear Hawkes Case) Under model* (15) *and RNN-TPP class $\mathcal{F} = \mathcal{F}_{L,D,B_m,l_f,u_f}$ defined as* (8)*, suppose that assumptions (A1) and (A4) hold, then for n i.i.d. sample series $\{S_i, i \in [n]\}$, with probability at least $1-\delta$, for $L = 4$, $D = \tilde{O}(n^{\frac{1}{4}})$, $B_m = \tilde{O}((\log n)^{3s^2 \log^2 n})$, $l_f = \tilde{B}_1 \wedge 1$, and $u_f = \tilde{B}_0$, the excess risk* (4) *of ERM* (3) *satisfies:*

$$\mathbb{E}[\mathit{loss}(\hat{\lambda}, S_{\mathit{test}})] - \mathbb{E}[\mathit{loss}(\lambda^*, S_{\mathit{test}})] \leq \tilde{O}\left(n^{-\frac{1}{4}}\right). \tag{16}$$

For the non-linear case, as indicated by Theorem 3.3, we require a deeper RNN-TPP with four layers to achieve the vanishing excess risk. Under the Lipschitz assumption of Ψ, the width of the hidden layers is of order $n^{1/4}$. When Ψ is allowed to have higher-order smoothness, the width can reduce to that of the vanilla Hawkes case.

Remark 3.4. *(i) Two additional layers of RNN are required for the approximation of the arbitrary non-linear Lipschitz continuous function Ψ. (ii) For the model $\lambda^*(t) = \Psi\left(\lambda_0(t) + \sum_{t_j<t} \mu(t-t_j)\right)$ with general excitation function μ, we can obtain the similar excess risk bound using the same technique in the proof of Theorem 3.1.*

To better explain the excess risks that obtained in Theorems 3.1-3.3, we depend on the following decomposition lemma.

Lemma 3.5. *Let $\check{\lambda}^* \in \arg\min_{\lambda\in\mathcal{F}} \mathbb{E}[\mathit{loss}(\lambda, S_{\mathit{test}})]$, for any random sample $\{S_i, i \in [n]\}$, the excess risk of ERM* (3) *satisfies*

$$\begin{aligned}\mathbb{E}[\mathit{loss}(\hat{\lambda}, S_{\mathit{test}})] - \mathbb{E}[\mathit{loss}(\lambda^*, S_{\mathit{test}})] \leq\ & \underbrace{2\sup_{\lambda\in\mathcal{F}} \left|\mathbb{E}[\mathit{loss}(\lambda, S_{\mathit{test}})] - \frac{1}{n}\sum_{i\in[n]} \mathit{loss}(\lambda, S_i)\right|}_{\textit{stochastic error}} \\ & + \underbrace{\mathbb{E}[\mathit{loss}(\check{\lambda}^*, S_{\mathit{test}})] - \mathbb{E}[\mathit{loss}(\lambda^*, S_{\mathit{test}})]}_{\textit{approximation error}}.\end{aligned} \tag{17}$$

By Lemma 3.5, the excess risk of ERM is bounded by the sum of two terms, the stochastic error $2\sup_{\lambda\in\mathcal{F}}|\mathbb{E}[\text{loss}(\lambda, S_{test})] - n^{-1}\sum_{i\in[n]}\text{loss}(\lambda, S_i)|$ and the approximation error $\mathbb{E}[\text{loss}(\check{\lambda}^*, S_{test})] - \mathbb{E}[\text{loss}(\lambda^*, S_{test})]$. The first term can be bounded by the complexity of the function class $\mathcal{F}$ using the empirical process theory, where the unboundedness of the loss function needs to be handled carefully; we present the details in section 4. The second term characterizes the approximation ability of the RNN function class $\mathcal{F}$ to the true intensity λ^* under the measure of the expectation of the negative log-likelihood loss function. In order to bound this term, we need to carefully construct a suitable RNN which can approximate λ^* well. This has not been studied yet in the literature; see section 5 for the details.

Based on Lemma 3.5, the results in Theorem 3.1 admit the following form,

$$\mathbb{E}[\text{loss}(\hat{\lambda}, S_{test})] - \mathbb{E}[\text{loss}(\lambda^*, S_{test})] \leq O\left(\frac{C(N)}{\sqrt{n}} + \frac{1}{R(N)}\right),$$

where $C(N)/\sqrt{n}$ is the stochastic error and $1/R(N)$ is the approximation error. $C(N)$ is the complexity of RNN function class $\mathcal{F}$ and $R(N)$ is the corresponding approximation rate, where N is a tuning parameter. For the Poisson case, we can construct a two-layer RNN-TPP with $O(N)$ width to achieve $O(N^{-s})$ approximation error. Hence $C(N) = O(N)$, $R(N) = O(N^s)$, and the final excess risk bound is $\tilde{O}(n^{-\frac{s}{2(s+1)}})$ in (12). For the vanilla Hawkes case, since the exponential function is C^∞-smooth, we only need extra $O(\text{Poly}(\log N))$ hidden cells in each layer to obtain $\tilde{O}(N^{-s})$ approximation error, and then we have the same order excess risk bound. For the general case, motivated by the vanilla Hawkes case, we decompose $\mu \in C^{k,\infty}([0,T], C_0)$ into two parts. One part is a polynomial of exponential functions which can be well approximated by $O(\text{Poly}(\log N))$-width tanh neural network. The other part is a function $\tilde{\mu} \in C^{k,\infty}([0,T], \tilde{C_0})$ satisfying $\tilde{\mu}^{(j)}(0+) = \tilde{\mu}^{(j)}(T-)$, $j = 0, 1, \cdots, k-1$. It is easy to check that the r-th Fourier coefficients of $\tilde{\mu}$, $\hat{\mu}_r$, decay at the rate of r^{-k}. Then it is sufficient to approximate the first N functions in the Fourier expansion of $\tilde{\mu}$ to get $\tilde{O}(N^{-(k-1)})$ approximation error, which additionally costs $\tilde{O}\left(N^5\right)$ complexity (see section 5.3 for details). Combining this with the approximation result of λ_0, we get the final bound $\tilde{O}(n^{-\frac{1}{2}\left(\frac{s}{s+1}\wedge\frac{k-1}{k+4}\right)})$. Similarly, for the nonlinear Hawkes case, we need $\tilde{O}(N)$ complexity to obtain $\tilde{O}(N^{-1})$ approximation error, which leads to $\tilde{O}(n^{-\frac{1}{4}})$ excess risk bound.

As we emphasize in the above remarks, the number of layers depends on the smoothness of λ_0. If $\lambda_0 \in C^\infty([0,T])$ and $\|\lambda_0\|_{W^{s,\infty}} \leq C^s$, we only need one-layer tanh neural network to approximate λ_0, hence the number of layers in RNN-TPP can be reduced to one.

Remark 3.6. *The main goal of the current paper is to provide a tool for theoretically analyzing the behaviors of RNN-based TPPs. The discussions on the sharpness of our analyses and the optimality of the excess risk bound can be found in Section 7.*

4. Stochastic Error

In this section, we focus on the stochastic error in (17). This type of stochastic error for the RNN function class has been studied in the recent literature, such as Chen et al. (2020)

and Tu et al. (2020). However, they only consider the case where the lengths of the input sequences are bounded, which is not applicable under the TPP setting. Here we establish an upper bound of the stochastic error in (17) by a novel decoupling technique to make the classical results applicable. This technique can be used in many other related problems.

4.1 Main Variance Term

We first give out some mild assumptions for the RNN-TPP function class $\mathcal{F}$ under a more general framework.

(B1) The embedding function $x(\cdot)$ is bounded by a constant $B_{in}(T)$ on the time domain $[0, T]$, i.e. $\|x(\cdot)\|_2 \le B_{in}(T)$.

(B2) The parameter θ lies in a bounded domain Θ. More precisely, we assume that the spectral norms of weight matrices (vectors) and other parameters are bounded respectively, i.e., $\|W_x^{(l)}\|_2 \le B_x$, $\|W_h^{(l)}\|_2 \le B_h$, $\|b^{(l)}\|_2 \le B_b$, $1 \le l \le L+1$, and $B_m = \max\{B_b, B_h, B_x\}$.

(B3) Activation functions σ and f are Lipschitz continuous with parameters ρ_σ and ρ_f respectively, $\sigma(0) = 0$, and there exists $|b_0| \le B_b$ such that $f(b_0) = 1$. Additionally, σ is entrywise bounded by B_σ, and f satisfies $l_f \le \|f\|_{L^\infty} \le u_f$.

Now we consider the first term of (17). For convenience, we denote $X_\theta = \mathbb{E}[\mathrm{loss}(\lambda_\theta, S_{test})] - n^{-1}\sum_{i=1}^n \mathrm{loss}(\lambda_\theta, S_i)$.

Theorem 4.1. *Under assumptions (B1)-(B3) and suppose the event number N_e satisfies the tail condition*

$$\mathbb{P}(N_e \ge s) \le a_N \exp(-c_N s), \;\; s \in \mathbb{N},$$

with probability at least $1-\delta$, we have

$$\begin{aligned}\sup_{\theta\in\Theta} |X_\theta| \le & \frac{500}{\sqrt{n}}\left(T + \frac{1}{l_f}\right)(s_0+1)u_f\left(\sqrt{\log\left(\frac{4}{\delta}\right)} + D\sqrt{(3L+2)}\left(\sqrt{\log\left(1+M(s_0)\right)}+1\right)\right.\\ & \left. + \frac{1}{(1-\exp(-c_N))^2}\right).\end{aligned}$$

Thus

$$\sup_{\theta\in\Theta} |X_\theta| \le \tilde{O}\left(\sqrt{\frac{D^2L^2}{n}}\right)\;, \tag{18}$$

where $s_0 = \lceil {c_N}^{-1}\log\left(2a_N n/\delta\right)\rceil$, $M(s) = \rho_f B_m\sqrt{D}(B_\sigma\sqrt{D} \vee B_{in}(T) \vee 1)(\gamma^L \vee 1)(s+1)^{L-1}(\beta^{s+1}-1)/(\beta-1)$, $\gamma = \rho_\sigma B_x$, $\beta = \rho_\sigma B_h$.

Remark 4.2. *There exist constants a_N, c_N so that the tail condition $\mathbb{P}(N_e \ge s) \le a_N \exp(-c_N s), s \in \mathbb{N}$ always holds for (non) homogeneous Poisson processes, linear and nonlinear Hawkes processes, and self-correcting processes under weak assumptions. To be more concrete, Lemma 4.4 in the following section gives a result for the linear case.*

Remark 4.3. *For one-layer RNN with width D and bounded sequence length T, Chen et al. (2020) gives a $\tilde{O}\big(\sqrt{D^3T/n}\big)$ type stochastic error bound. Our bound reduces the term D^3 to D^2, thanks to the bounded output layer, i.e., $f(x) = \min\{\max\{x, l_f\}, u_f\}$. The term D^2 is also order-optimal by noticing that the number of free parameters in a single-layer RNN is at least D^2.*

The stochastic error in (4.1) is mainly determined by the complexity of the RNN function class $\mathcal{F}$, which will be discussed in the following section. To obtain this bound, we need to handle the unboundedness of the event number. We use a truncation technique to decouple the randomness of the tail of N_e, which allows us to use classical empirical process theory to derive the upper bound. Our computation is motivated by Chen et al. (2020), which gives the generalization error bound of a single-layer RNN function class.

4.2 Key Techniques

To be reader-friendly, the main techniques for proving Theorem 4.1 are summarized as follows.

4.2.1 Probability Bound of Events Number

Define $N_{e(n)} := \max\{N_{ei}, 1 \le i \le n\}$. The following lemma characterizes the tails of event number N_e and $N_{e(n)}$ under model (9) and assumptions (A1) and (A2) (For assumption (A1), we only need $\lambda_0 \le B_0$ in this section). The proof is similar to Proposition 2 in Hansen et al. (2015); see supplementary for the details.

Lemma 4.4. *For model* (9)*, under assumptions (A1) and (A2), with probability at least* $1-\delta$*, we have*

$$N_{e(n)} < \frac{1}{1-c_\mu\eta}\left(\frac{2}{\log(\eta)}\log\left(\frac{2n\sqrt{B_0T}}{\delta(1-c_\mu)}\right)+\eta(B_0T)\right).$$

Hence

$$\mathbb{P}\left(N_e = s\right) \le \mathbb{P}\left(N_e \ge s\right) \le \frac{2\sqrt{B_0T}}{1-c_\mu}\exp\left(\frac{\log(\eta)}{2}\left[\eta(B_0T)-(1-c_\mu\eta)s\right]\right),$$

where $\eta \in \left(1, c_\mu^{-1}\right)$ *is a tuning parameter. Let* $a_N = 2\sqrt{B_0T}\exp(\log(\eta_0)\eta_0(B_0T)/2)/(1-c_\mu)$ *and* $c_N = \log(\eta_0)(1-c_\mu\eta_0)/2$ *with* $\eta_0 \in \left(1, c_\mu^{-1}\right)$ *being fixed. Then*

$$\mathbb{P}\left(N_e = s\right) \le \mathbb{P}\left(N_e \ge s\right) \le a_N\exp(-c_N s). \tag{19}$$

Our result is more refined than Proposition 2 in Hansen et al. (2015), with computing all the constants and giving a tuning parameter to control the probability bound.

For the nonlinear case (10), under Assumption (A4), we can obtain results similar to the non-homogeneous Poisson case, which are included in the above Lemma.

4.2.2 From Unboundedness to Boundedness

The following lemma is the key to handling the unboundedness of X_θ, i.e., the unboundedness of the loss function. For any $s \in \mathbb{N}$, we let $X_\theta(s) = \mathbb{E}\left[\text{loss}(\lambda_\theta, S_{test})\mathbb{1}_{\{N_e \le s\}}\right] - n^{-1}\sum_{i=1}^n \text{loss}(\lambda_\theta, S_i)\mathbb{1}_{\{N_{ei}\le s\}}$ and $E_\theta(s) = \mathbb{E}\left[\text{loss}(\lambda_\theta, S_{test})\mathbb{1}_{\{N_e > s\}}\right]$.

Lemma 4.5. *For any* $s \in \mathbb{N}$ *and nonempty parameter set* Θ*, we have*

$$\mathbb{P}\left(\sup_{\theta\in\Theta}|X_\theta| > t\right) \le \mathbb{P}\left(\sup_{\theta\in\Theta}|X_\theta(s)| + \sup_{\theta\in\Theta}|E_\theta(s)| > t\right) + \mathbb{P}(N_{e(n)} > s). \tag{20}$$

The consequence of this lemma is to decompose $\mathbb{P}\left(\sup_{\theta\in\Theta}|X_\theta| > t\right)$ into two parts. The first part $\mathbb{P}\left(\sup_{\theta\in\Theta}|X_\theta(s)| + \sup_{\theta\in\Theta}|E_\theta(s)| > t\right)$ is the tail probability of the supremum of a set of bounded variables, and can therefore be handled by standard empirical process theory. The second part $\mathbb{P}(N_{e(n)} > s)$ is the tail probability of $N_{e(n)}$. Thanks to Lemma 4.4, this term can be controlled by the exponential decay property of the sub-critical point process. By choosing suitable s, we can make (20) sharper. This result plays a key role in stochastic error calculations.

4.2.3 Complexity of the RNN-TPP Class

To get the result in Theorem 4.1, we need to compute the complexity of the RNN function class which is specified in section 2.2. There are many possible complexity measures in deep learning theory (Suh and Cheng, 2024), and here we choose *covering number* which can be well computed for the RNN function class. In our setup, the key to the computation of the covering number is finding the Lipschitz continuity constant of RNN-TPPs, which separates the spectral norms of weight matrices and the total number of parameters (Chen et al., 2020).

Consider two different sets of parameters $\theta_1 = \{W_{x,1}^{(l)}, W_{h,1}^{(l)}, b_1^{(l)}; 1 \le l \le L+1\}$, $\theta_2 = \{W_{x,2}^{(l)}, W_{h,2}^{(l)}, b_2^{(l)}; 1 \le l \le L+1\}$. Denote $\Delta_b^l = \|b_1^{(l)} - b_2^{(l)}\|_2$, $\Delta_h^l = \|W_{h,1}^{(l)} - W_{h,2}^{(l)}\|_2$, $\Delta_x^l = \|W_{x,1}^{(l)} - W_{x,2}^{(l)}\|_2$, $1 \le l \le L+1$ ($\Delta_h^{L+1} \equiv 0$). The following lemma characterizes the Lipschitz constant of λ_θ.

Lemma 4.6. *Under Assumptions (B1)-(B3), given an input sequence of length N_S, $S = \{t_j\}_{j=1}^{N_S} \subset [0,T]$ (here we set $t_{N_S+1} = T$), for $t \in (t_j, t_{j+1}]$, $1 \le j \le N_S$, and $\theta_1, \theta_2 \in \Theta$, we have*

$$|\lambda_{\theta_1}(t;S) - \lambda_{\theta_2}(t;S)| \le \rho_f \gamma \left(\sum_{l=0}^{L-1} \gamma^l S_j^l \Delta_b^{L-l} + B_\sigma \sqrt{D} \sum_{l=0}^{L-2} \gamma^l S_j^l \Delta_x^{L-l} + B_{in}(T) \gamma^{L-1} S_j^{L-1} \Delta_x^1 \right.$$
$$\left. + B_\sigma \sqrt{D} \sum_{l=0}^{L-1} \gamma^l S_{j-1}^l \Delta_h^{L-l} \right) + \rho_f \Delta_b^{L+1} + \rho_f B_\sigma \sqrt{D} \Delta_x^{L+1}, \quad (21)$$

where $\beta = \rho_\sigma B_h$, $\gamma = \rho_\sigma B_x$, $S_j^l = \sum_{q=0}^{j} \binom{q+l}{l} \beta^q$ ($S_{-1}^l = 0$), and $d = \max\{d_l | 1 \le l \le L+1\}$. We set $\sum_{l=a}^{b} A_l = 0$ if $a > b$.

The proof of Lemma 4.6 is based on the induction. The full proof is given in the supplementary. Our result is an extension of Lemma 2 in Chen et al. (2020), where they only consider the family of **one-layer** RNN models. Lemma 4.6 is of independent interest and can be useful in any other problems regarding RNN-based modeling. Using Lemma 4.6, we can establish a covering number bound for $\mathcal{F}$ under a "truncated" distance.

Denote $\mathcal{N}(\mathcal{F}, \epsilon, d(\cdot,\cdot))$ as the covering number of metric space $\mathcal{F}$, i.e., the minimal cardinality of a subset $\mathcal{C} \subset \mathcal{F}$ that covers $\mathcal{F}$ in scale ϵ with respect to the metric $d(\cdot,\cdot)$. Given a fixed integer N_0, We define a truncated distance,

$$d_{N_0}(\lambda_{\theta_1}, \lambda_{\theta_2}) = \sup_{\#(S) \le N_0} \|\lambda_{\theta_1}(t;S) - \lambda_{\theta_2}(t;S)\|_{L^\infty[0,T]} .$$

The following lemma gives an upper bound of $\mathcal{N}(\mathcal{F}, \epsilon, d_{N_0}(\cdot,\cdot))$.

Lemma 4.7. *Under assumptions (B1)-(B3), for any $\epsilon > 0$ and $\mathcal{F} = \mathcal{F}_{L,D,B_m,l_f,u_f}$ defined as* (8)*, the covering number $\mathcal{N}(\mathcal{F}, \epsilon, d_{N_0}(\cdot,\cdot))$ is bounded by*

$$\mathcal{N}(\mathcal{F}, \epsilon, d_{N_0}(\cdot,\cdot)) \leq \left(1 + \frac{C(N_0)(3L+2)B_m\sqrt{D}}{\epsilon}\right)^{D^2(3L+2)},$$

where $C(N_0) = \rho_f(B_\sigma\sqrt{D} \vee B_{in}(T) \vee 1)(\gamma^L \vee 1)(N_0+1)^{L-1}(\beta^{N_0+1} - 1)/(\beta - 1)$, $\gamma = \rho_\sigma B_x$, and $\beta = \rho_\sigma B_h$.

By Lemma 4.7, taking $N_0 = s$, we can get the non-asymptotic bound of $X_\theta(s)$, which is an important step to obtain the first part of (20).

5. Approximation Error

In this section, we focus on the approximation error, i.e., the second part of (17). The approximation error of deep neural networks has been broadly studied in the literature (Schmidt-Hieber, 2020; Shen et al., 2019; Jiao et al., 2023; Lu et al., 2021). However, most of them only consider the ReLU activation case, which is different from tanh, the activation function usually chosen for RNNs. Recently, De Ryck et al. (2021) studied the approximation properties of shallow tanh neural networks, which provides a technical tool for our analysis. To the best of our knowledge, the approximation ability of RNN-type networks has not been fully studied in the literature. Here we propose a family of approximation results for the intensities of various TPP models stated in section 2.3.

5.1 Poisson Case

We start with the the approximation of (non-homogeneous) Poisson process, whose intensity is independent of the event history, i.e. $\lambda^*(t) = \lambda_0(t)$, where $\lambda_0(t)$ is an unknown function. In this case, we do not need to take into account the transfer of information in the time domain. To be precise, we can take $W_h^l = 0$ for $l \in [L]$. Then the problem degenerates to a standard neural network approximation problem. Using the approximation results for tanh neural networks in De Ryck et al. (2021), we can get the following approximation result.

Theorem 5.1. *(Approximation for Poisson process) Under model $\lambda^*(t) = \lambda_0(t)$ and assumptions (A1) and (A3), for $N \geq 5$, $N \in \mathbb{N}$, there exists an RNN-TPP $\hat{\lambda}^N$ as stated in section 2.2 with $L = 2$, $l_f = B_1$, $u_f = B_0$, and input function $x(t;S) = t$ such that*

$$|\mathbb{E}[loss(\hat{\lambda}^N, S_{test})] - \mathbb{E}[loss(\lambda^*, S_{test})]| \leq 15\exp(2B_0T)(T + 2B_1^{-1})\frac{\mathcal{C}T^s}{N^s}, \tag{22}$$

where $\mathcal{C} = \sqrt{2s}5^s/(s-1)!$. Moreover, the width of $\hat{\lambda}^N$ satisfies $D \leq 3\lceil s/2 \rceil + 6N$ and the weights of $\hat{\lambda}^N$ are less than

$$\mathcal{C}_1\left[\frac{\sqrt{2s}5^s}{(s-1)!}B_0T^s\right]^{-\frac{s}{2}} N^{\frac{1+s^2}{2}}(s(s+2))^{3s(s+2)},$$

where $\mathcal{C}_1$ is an universal constant.

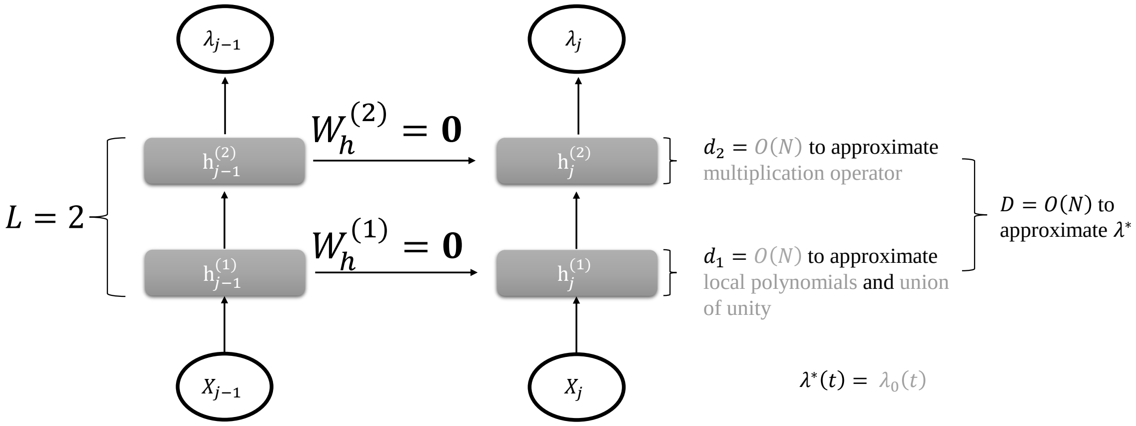

λ_{j-1}
λ_j
$W_h^{(2)} = \mathbf{0}$
$W_h^{(1)} = \mathbf{0}$
$L = 2$
$h_{j-1}^{(2)}$
$h_j^{(2)}$
$h_{j-1}^{(1)}$
$h_j^{(1)}$
$d_2 = O(N)$ to approximate multiplication operator
$D = O(N)$ to approximate λ^*
$d_1 = O(N)$ to approximate local polynomials and union of unity
X_{j-1}
X_j
$\lambda^*(t) = \lambda_0(t)$

Figure 2: The construction of RNN-TPP for the case of Poisson processes.

A graphical representation of RNN approximation is given in Figure 2. For the non-homogeneous Poisson models, the RNN-TPP $\hat{\lambda}^N$ in Theorem 5.1 is indeed a two-layer neural network. From Theorem 5.1, we need an RNN-TPP with $O(N)$ width and $B_m = O(N^{\frac{s^2+2}{2}})$ to obtain $O(N^{-s})$ approximation error. Combining with Theorem 4.1, we can get the part (i) of Theorem 3.1.

5.2 Vanilla Hawkes Case

Recall that the intensity of the vanilla Hawkes process has the form

$$\lambda^*(t) = \lambda_0(t) + \sum_{j:t_j<t} \alpha \exp\{-\beta(t - t_j)\}. \tag{23}$$

Different from Poisson process, the intensity of the vanilla Hawkes process depends on historical events. Hence it can not be approximated by a simple neural network and needs the recurrent structure. We construct an RNN-TPP to approximate the intensity using the Markov property of (23). Specifically, note that if we have observed the first k event times $\{t_1, \cdots, t_k\}$, then for any t satisfying $t_k < t \leq t_{k+1}$, we have

$$\begin{aligned}
\lambda^*(t) - \lambda_0(t) &= \sum_{j:t_j<t} \alpha \exp\{-\beta(t - t_j)\} \\
&= \exp(-\beta(t - t_k)) \sum_{j:t_j \leq t_k} \alpha \exp\{-\beta(t_k - t_j)\} \\
&= (\lambda^*(t_k) - \lambda_0(t_k) + \alpha) \exp(-\beta(t - t_k)).
\end{aligned}$$

Therefore, we can use the hidden layers in RNN-TPP to store the information of $\lambda^*(t_k) - \lambda_0(t_k)$ and then compute $\lambda^*(t) - \lambda_0(t)$ with the help of input $t - t_k$. Together with the approximation of λ_0, we can obtain the final approximation result. A graphical illustration of the above construction procedures are given in Figure 3.

hawkes.pdf

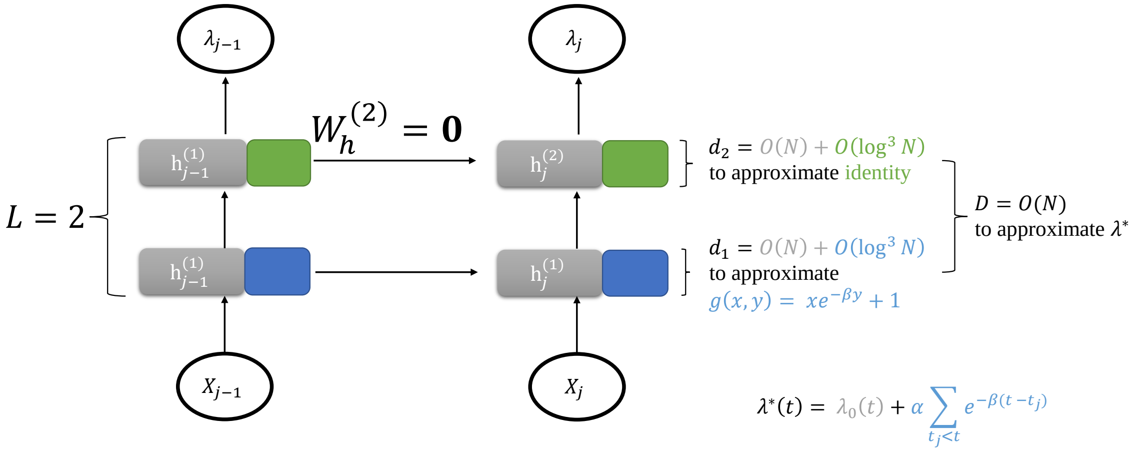

λ_{j-1}
λ_j
$W_h^{(2)} = \mathbf{0}$
$L = 2$
$h_{j-1}^{(1)}$
$h_j^{(2)}$
$h_{j-1}^{(1)}$
$h_j^{(1)}$
$d_2 = O(N) + O(\log^3 N)$
to approximate identity
$D = O(N)$
to approximate λ^*
$d_1 = O(N) + O(\log^3 N)$
to approximate
$g(x, y) = xe^{-\beta y} + 1$
X_{j-1}
X_j
$\lambda^*(t) = \lambda_0(t) + \alpha \sum_{t_j<t} e^{-\beta(t-t_j)}$

Figure 3: The construction of RNN-TPP for the case of the vanilla Hawkes process.

Theorem 5.2. *(Approximation for Vanilla Hawkes process) Under model* (23)*, assumptions (A1), (A3), and $\alpha/\beta < 1$, for $N \geq 5$, $N \in \mathbb{N}$, there exists an RNN-TPP $\hat{\lambda}^N$ as stated in section 2.2 with $L = 2$, $l_f = B_1$, $u_f = B_0 + O(\log N)$, and input function $x(t; S) = (t, t - F_S(t))^\top$ such that*

$$|\mathbb{E}[loss(\hat{\lambda}^N, S_{test})] - \mathbb{E}[loss(\lambda^*, S_{test})]| \lesssim \frac{(\log N)^2}{N^s}. \tag{24}$$

Moreover, the width of $\hat{\lambda}^N$ satisfies $D = O(N)$ and the weights of $\hat{\lambda}^N$ are less than

$$\mathcal{C}_1(\log(N))^{12s^2(\log(N))^2},$$

where $\mathcal{C}_1$ is a constant related to s, B_0, β, and T.

Due to the smoothness of the exponential function, the approximation rate in Theorem 5.2 only adds the $\log(N)$ term compared with the results in Theorem 5.1. Similarly, combining with Theorem 4.1, we can easily get the part (ii) of Theorem 3.1.

5.3 Linear Hawkes Case

Now we consider the general linear Hawkes process, i.e., (9) in section 2.3. Motivated by the approximation construction of the Vanilla Hawkes process, we want to find a decomposition for the general μ where each term has the 'Markov property' so that we can construct the corresponding RNN structure. Precisely, for $\mu \in C^{k,\infty}([0, T], C_0)$, $k \geq 2$, $k \in \mathbb{N}$, we can decompose μ into two parts,

$$\mu(t) = \underbrace{\tilde{\mu}(t)}_{\text{part}_1} + \underbrace{\sum_{j=1}^{k} \alpha_j \exp(-\beta_j t)}_{\text{part}_2}, \quad t \in [0, T],$$

hawkes.pdf

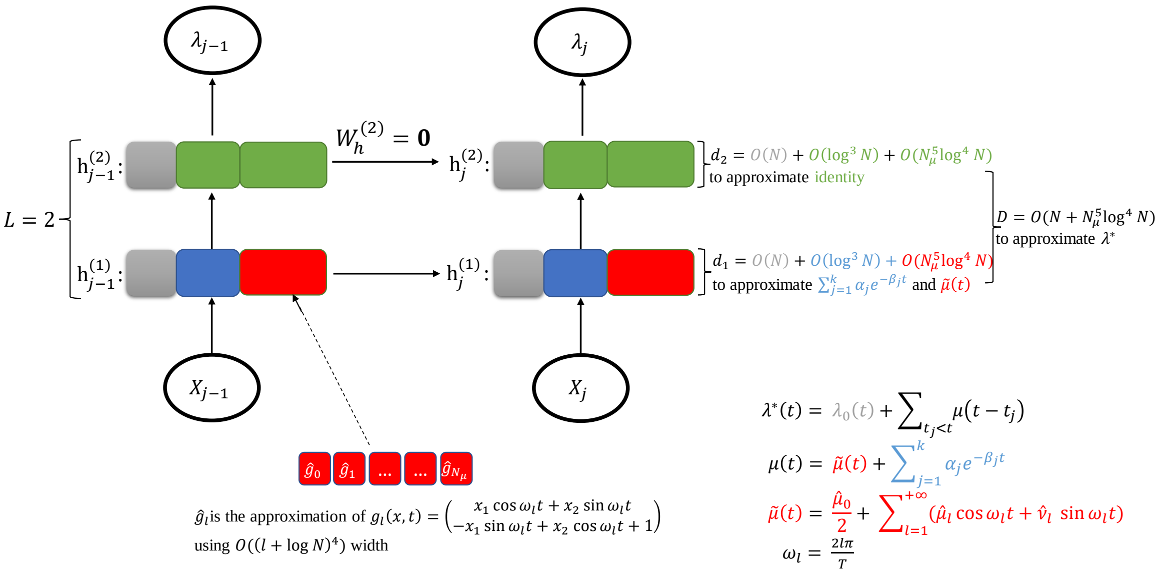

λ_{j-1}
λ_j
$W_h^{(2)} = \mathbf{0}$
$L = 2$
$h_{j-1}^{(2)}$:
$h_j^{(2)}$:
$h_{j-1}^{(1)}$:
$h_j^{(1)}$:
$d_2 = O(N) + O(\log^3 N) + O(N_\mu^5 \log^4 N)$
to approximate identity
$D = O(N + N_\mu^5 \log^4 N)$
to approximate λ^*
$d_1 = O(N) + O(\log^3 N) + O(N_\mu^5 \log^4 N)$
to approximate $\sum_{j=1}^k \alpha_j e^{-\beta_j t}$ and $\tilde{\mu}(t)$
X_{j-1}
X_j
$\hat{g}_0$ $\hat{g}_1$ ... ... $\hat{g}_{N_\mu}$
$\hat{g}_l$is the approximation of $g_l(x,t) = \begin{pmatrix} x_1 \cos\omega_l t + x_2 \sin\omega_l t \\ -x_1 \sin\omega_l t + x_2 \cos\omega_l t + 1 \end{pmatrix}$
using $O((l + \log N)^4)$ width
$\lambda^*(t) = \lambda_0(t) + \sum_{t_j<t} \mu(t - t_j)$
$\mu(t) = \tilde{\mu}(t) + \sum_{j=1}^{k} \alpha_j e^{-\beta_j t}$
$\tilde{\mu}(t) = \frac{\hat{\mu}_0}{2} + \sum_{l=1}^{+\infty} (\hat{\mu}_l \cos\omega_l t + \hat{\nu}_l \sin\omega_l t)$
$\omega_l = \frac{2l\pi}{T}$

Figure 4: The construction of RNN-TPP for the case of general linear Hawkes processes.

where $\tilde{\mu}$ satisfies the *boundary condition*, $\tilde{\mu}^j(0+) = \tilde{\mu}^j(T-)$, $0 \leq j \leq k-1$, $j \in \mathbb{N}$, and $\beta_j = j/k$, $j \in [k]$. The term $\sum_{j=1}^k \alpha_j \exp(-\beta_j t)$ can be handled similarly to the vanilla Hawkes process. For $\tilde{\mu}$, we consider its Fourier expansion,

$$\tilde{\mu}(t) = \frac{\hat{\mu}_0}{2} + \sum_{l=1}^{\infty} \left(\hat{\mu}_l \cos\left(\frac{2l\pi}{T} t\right) + \hat{\nu}_l \sin\left(\frac{2l\pi}{T} t\right) \right).$$

Thanks to the boundary condition, $\tilde{\mu}(t)$ can be well approximated by the **finite** sum of Fourier series. Then we can use the "Markov property" of the trigonometric function pairs $\cos(2l\pi t/T)$ and $\sin(2l\pi t/T)$ to construct the RNN-TPP. The construction is similar to that of the exponential function case but needs more thorny calculations. Combining all the approximation parts, we can get the approximation theorem for (9). The above ideas are visualized in Figure 4.

Theorem 5.3. *(Approximation for linear Hawkes process) Under model* (9)*, assumptions (A1)-(A3), and $\mu \in C^{k,\infty}([0,T], C_0)$, $k \geq 2$, $k \in \mathbb{N}$, for $N \geq 5$, $N \in \mathbb{N}$, there exists an RNN-TPP $\hat{\lambda}^{N,N_\mu}$ as stated in section 2.2 with $L = 2$, $l_f = B_1$, $u_f = B_0 + O(\log N)$, and input function $x(t;S) = (t, t - F_S(t))^\top$ such that*

$$|\mathbb{E}[loss(\hat{\lambda}^{N,N_\mu}, S_{test})] - \mathbb{E}[loss(\lambda^*, S_{test})]| \lesssim \frac{(\log N)^2}{N^s} + \frac{\log N}{N_\mu^{k-1}}. \qquad (25)$$

Moreover, the width of $\hat{\lambda}^{N,N_\mu}$ satisfies $D = O(N + N_\mu^5 (\log N)^4)$ and the weights of $\hat{\lambda}^{N,N_\mu}$ are less than

$$\mathcal{C}_1 (\log(N N_\mu))^{12 s^2 (\log(N N_\mu))^2},$$

where $\mathcal{C}_1$ is a constant related to s, k, B_0, C_0, c_μ, and T.

We make a few explanations on Theorem 5.3. There are two tuning parameters in $\hat{\lambda}^{N,N_\mu}$, where N is the tuning parameter to control the approximation error of λ_0, $\sum_{j=1}^k \alpha_j \exp(-\beta_j t)$, and the finite sum of the Fourier series, and N_μ is the tuning parameter to control the number of terms in the Fourier series entering the RNN-TPP. The term $(\log N)^2/N^s$ is obtained similarly to that in the vanilla Hawkes process case, and the term $\log N/N_\mu^{k-1}$ is the error caused by the finite sum approximation for the Fourier series. Moreover, the $O(N_\mu^5(\log N)^4)$ term in the width of RNN-TPP is caused by the approximation construction of the first N_μ terms of the Fourier series. Finally, combining with Theorem 4.1, we can obtain the part (iii) of Theorem 3.1.

5.4 Nonlinear Hawkes Case

Finally, we consider the nonlinear Hawkes process, which is defined in (10) in section 2.3. To make the statement simpler, we only consider the simple case, i.e., $\mu(t) = \alpha \exp(-\beta t)$. The results for the general μ can be obtained similarly. Compared to the vanilla Hawkes case, the additional challenge here is the existence of a nonlinear function Φ. With two additional layers, we can approximate Φ well. Together with the results for the case of the vanilla Hawkes process, we can obtain the desired RNN-TPP architecture. To be clearer, we also provide the graphical illustration in Figure 5.

Theorem 5.4. *(Approximation for nonlinear Hawkes process) Under model* (15)*, assumptions (A1) and (A4), for* $N \geq \max\{5, (2\mathcal{C}B_0T^s+1)^{\frac{1}{s}}\}$ *with* $\mathcal{C} = \sqrt{2s}5^s/(s-1)!$*, there exists an RNN-TPP* $\hat{\lambda}^N$ *as stated in section 2.2 with* $L = 4$*,* $l_f = \tilde{B}_1$*,* $u_f = \tilde{B}_0$*, and input function* $x(t; S) = (t, t - F_S(t))^\top$ *such that*

$$|\mathbb{E}[loss(\hat{\lambda}^N, S_{test})] - \mathbb{E}[loss(\lambda^*, S_{test})]| \lesssim \frac{\log N}{N}. \tag{26}$$

Moreover, the width of $\hat{\lambda}^N$ *satisfies* $D = O(N)$ *and the weights of* $\hat{\lambda}^N$ *are less than*

$$\mathcal{C}_1(\log N)^{12s^2(\log N)^2},$$

where $\mathcal{C}_1$ *is a constant related to* $s, \tilde{B}_0, \alpha, \beta, T,$ *and* L*.*

Since we assume Φ to be Lipschitz continuous, we can only get $\tilde{O}(N^{-1})$ approximation error. The rate can be improved if Φ is allowed to have better smoothness properties. Again, combining with Theorem 4.1, we arrive at Theorem 3.3.

Remark 5.5. *The universal approximation properties of one-layer RNNs are studied in Schäfer and Zimmermann (2007). Our current results are different from theirs in the following sense. (i) RNN-TPP is defined over the continuous time domain* $[0, T]$*, while the standard RNN only considers the discrete points. In other words, our approximation results hold uniformly over all* $t \in [0, T]$*. (The details can be found in the proofs in Appendix D). (ii) In Schäfer and Zimmermann (2007), they do not give the explicit formula of the widths of hidden layers or parameter weights in the construction of RNN approximator. Therefore, their results cannot be directly used in computing the approximation error.*

hawkes.pdf

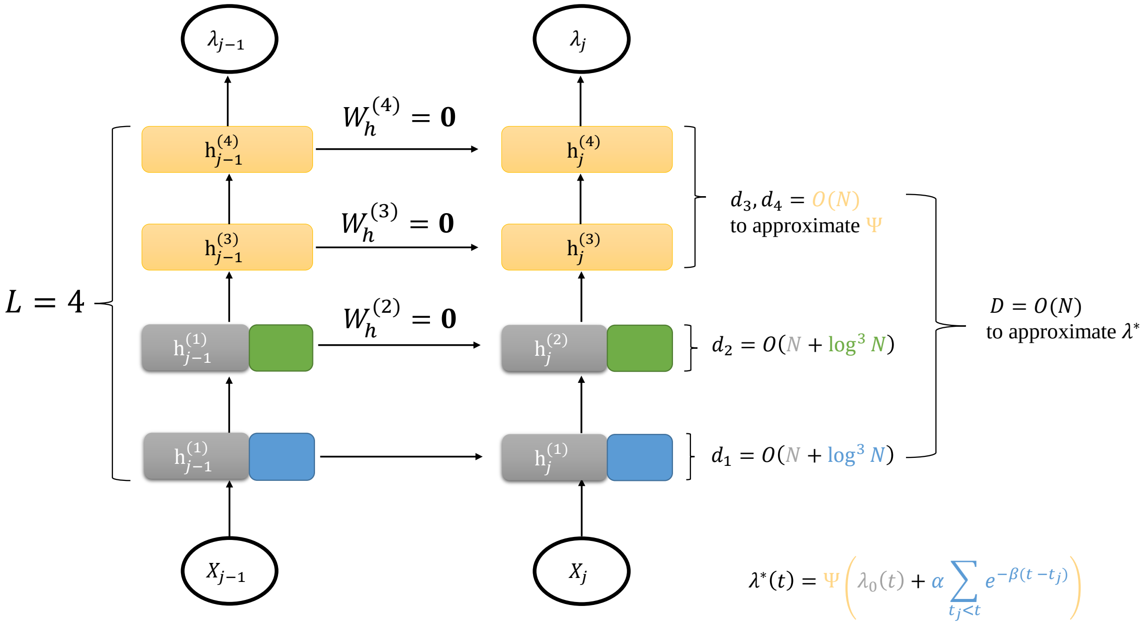


Figure 5: The construction of RNN-TPP for the case of nonlinear Hawkes process.

6. Usefulness of Interpolation of Hidden States

As mentioned in Remark 3, the RNN-TPP needs to take into account any continuous time point t between observed time grids t_j and t_{j+1}. The interpolation of hidden state $h^{(l)}(t;S)$ between $h_j^{(l)}$ and $h_{j+1}^{(l)}$ is essential and important during the construction of RNN-TPPs.

In this section, we give a counter-example to illustrate that an RNN-TPP model without the interpolation of hidden states is unable to precisely capture the true intensity in terms of excess risk (4). For simplicity, we only consider the single-layer RNN-TPP and the argument is the same for multi-layer RNN-TPPs.

We consider a (single-layer) RNN-TPP which admits the following model structure,

$$\begin{aligned} h_j &= \sigma(W_x x(t_j;S) + W_h h_{j-1}), \\ \hat{\lambda}_{ne}(t) &= f(\alpha(t-t_j) + W_y h_j + b) \in \mathbb{R}, \quad t \in (t_j, t_{j+1}], \end{aligned} \tag{27}$$

where $x(t_j;S)$ is the embedding for the j-th event, $h_0 = \mathbf{0}$, $\sigma(x) = \tanh(x)$, $f(x) = (x \vee l_f) \wedge u_f$, and l_f and u_f will be determined from the true intensity. The intensity formula in (27) is well used in existing literature, including Du et al. (2016) and Upadhyay et al. (2018). When $\alpha = 0$, $h^{(1)}(t;S) \equiv h_j$ for all t satisfying $t_j \le t < t_{j+1}$, i.e., it becomes the constant interpolation.

Theorem 6.1. *Suppose the true model intensity on $[0, T]$ has the following form,*

$$\lambda^*(t) \quad = \quad \begin{cases} T & , \ t \in [0, T/3] \\ \dfrac{9}{T}t^2 & , \ t \in (T/3, 2T/3) \\ 4T & , \ t \in [2T/3, T] \end{cases} \quad .$$

Hence we can take $l_f = T$ and $u_f = 4T$, and then there exists a constant $C > 0$ such that

$$\min_{\hat{\lambda}_{ne} \ as \ (27)} \mathbb{E}[loss(\hat{\lambda}_{ne}, S_{test})] - \mathbb{E}[loss(\lambda^*, S_{test})] \geq C > 0. \tag{28}$$

Theorem 6.1 tells us that the RNN-TPPs without a proper hidden state interpolation may fail to offer a good approximation, even under a very simple non-homogeneous Poisson model. Therefore, the user-determined input embedding vector function $x(t; S)$ plays an important role in interpolating the hidden states. It should be carefully chosen so that $x(t; S)$ can summarize the information of past event history to some extent.

Remark 6.2. *The term $\alpha(t - t_j)$ in (27) can be replaced by any monotonic function $\kappa(t - t_j)$, while maintaining the validity of Theorem 6.1. For example, the exponential decay mechanism, i.e. $\kappa(t - t_j) = \exp\{-\delta(t - t_j)\}$, shares the similarity to the memory cell $c(t)$ proposed in Mei and Eisner (2017).*

Remark 6.3. *For other different types of f (e.g. Softplus) in the output layer, the failure of the linear interpolation mechanism can be obtained similarly.*

Remark 6.4. *The parameter α in (27) may be generalized to $\alpha := u(h_j, w)$ for $t \in (t_j, t_{j+1}]$, where w denotes trainable parameters, h_j the previous hidden state, and u a fixed function. Theorem 6.1 remains valid in this case. See Appendix E for details.*

7. Discussion

In this paper, we give a positive answer to the question "whether the RNN-TPPs can provably have small excess risks in the estimation of the well-known TPPs". We establish the excess risk bounds under homogeneous Poisson process, non-homogeneous Poisson process, self-exciting process, and self-correcting process framework. Our analysis focuses on two parts, the stochastic error and the approximation error. For the stochastic error, we use a novel truncation technique to decouple the randomness and make the classical empirical process theory applicable. We carefully compute the Lipschitz constant of multi-layer RNNs, which is a useful intermediate result for future RNN-related work. For approximation error, we construct a series of RNNs to approximate the intensities of different TPPs by providing the explicit network depth, width, and parameter weights. To the best of our knowledge, our work is the first one to study the approximation ability of the multi-layer RNNs over the continuous time domain. We believe the results in the current work add values to both learning theory and neural network fields.

Curious readers may wonder how sharp our theoretical analysis is. In the following, we provide some in-depth discussions. To begin with, we first state the following lemma to characterize the relationship between our excess risk (4) and other L^2-type distances.

Lemma 7.1. *Under assumptions (A1)-(A3), for any predictable $\tilde{\lambda}$ satisfying $B_1 \leq \tilde{\lambda} < +\infty$, we have*

$$2H_2^2(\tilde{\lambda}, \lambda) \leq \mathbb{E}[loss(\tilde{\lambda}, S_{test})] - \mathbb{E}[loss(\lambda^*, S_{test})] \leq \frac{1}{B_1}\mathbb{E}\left[\int_0^T \left(\tilde{\lambda}(t) - \lambda^*(t)\right)^2 \mathrm{d}t\right],$$

where $H_2^2(\tilde{\lambda}, \lambda) := \mathbb{E}\big[\int_0^T \big(\sqrt{\tilde{\lambda}(t)} - \sqrt{\lambda^*(t)}\big)^2 \mathrm{d}t\big]/2$ is the Hellinger distance. All expectations are taken for the test sequence S_{test}.

Lemma 7.1 reveals that our results in Theorem 3.1 and Theorem 3.3 also hold for $H_2^2(\hat{\lambda}, \lambda)$.

To get the faster stochastic error, we need the following error decomposition lemma, which is different from Lemma 3.5.

Lemma 7.2. *Let $\check{\lambda}^* \in \arg\min_{\lambda \in \mathcal{F}} \mathbb{E}[loss(\lambda, S_{test})]$, the excess risk of ERM (3) satisfies*

$$\begin{aligned} &\mathbb{E}_{\mathbf{D}_{train}}\Big[\mathbb{E}[loss(\hat{\lambda}, S_{test})] - \mathbb{E}[loss(\lambda^*, S_{test})]\Big] \\ &\leq \underbrace{\mathbb{E}_{\mathbf{D}_{train}}\left[\mathbb{E}[loss(\hat{\lambda}, S_{test}) - loss(\lambda^*, S_{test})] - \frac{2}{n}\sum_{i\in[n]}\Big(loss(\hat{\lambda}, S_i) - loss(\lambda^*, S_i)\Big)\right]}_{stochastic\ error} \\ &\quad + \underbrace{2\left(\mathbb{E}[loss(\check{\lambda}^*, S_{test})] - \mathbb{E}[loss(\lambda^*, S_{test})]\right)}_{approximation\ error}, \end{aligned}$$

where the expectation $\mathbb{E}_{\mathbf{D}_{train}}$ is taken for the observed samples D_{train}.

It should be noted that Lemma 7.2 is the key to get faster stochastic error (Györfi et al. (2002), Section 11.3). Thus, we can get the faster stochastic error $\tilde{O}(D^2L^2/n)$ using the same technique in Györfi et al. (2002), Section 11.3, see also Jiao et al. (2023). Hence we can square the error rates we established in Theorem 3.1 and Theorem 3.3.

One may be concerned about the optimality of the error bounds in our results. In fact, due to the limitations in the use of non-optimal approximation results in De Ryck et al. (2021), our bound in non-homogeneous Poisson case $O(n^{-2s/(2s+2)})$ does not match the nonparametric minimax optimal rates $O(n^{-2s/(2s+1)})$. Here the main problem is that the approximation error for shallow networks (the depth L is small) is not optimal. There are many works on the approximation rates of ReLU neural networks (Kohler and Langer, 2021; Shen et al., 2022; Lu et al., 2021; Jiao et al., 2023) and neural networks with other activation functions (Zhang et al., 2024) that could lead to the optimal approximation rates. However, to the best of our knowledge, all of them need the network depth L to be greater than a pretty large constant (for example, $L \geq 18$ in Shen et al. (2022) and $L \geq 21$ in Jiao et al. (2023)), which are not satisfied in the practical usage, i.e. L is typically smaller than 4 for RNN-type models. Whether shallow neural networks (for example, $L \leq 4$) can attain the optimal approximation rates is a challenging open problem and still needs further research.

Moreover, the reasons why our rate $\tilde{O}(n^{-2(k-1)/2(k+4)})$ for the general excitation function μ does not match the optimal rate $O(n^{-2k/(2k+1)})$ can be explain as follows. We use the Fourier expansion in the approximation of RNN system so that we can only get $O(N_\mu^{-(k-1)})$

error bound for the first N_μ term in the expansion, which is different from the result of Taylor expansion. This makes the numerator of the exponent take the value from $2k$ to $2(k-1)$. Additionally, to well approximate the intensity function, we need to combine the information from the hidden layer $h^{(l)}(t;S)$ and the input function $x(t;S)$, which makes us require more 'dimensions'. As a result, the denominator of the exponent increases from $2k+1$ to $2(k+4)$. In summary, for a general excitation function μ, the RNN-TPP cannot be simply treated as a uni-variate nonparametric model. Hence, it remains an open question whether the approximation error could be further improved for the general RNN-TPPs. In practice, the excitation function μ is assumed to be sufficiently smooth so that $(k-1)/(k+4)$ is negligible compared with $s/(s+1)$ in (14).

In addition to the above discussions on the optimal rates, there are several other possible extensions along the research line of neural network-based TPPs. First, it is not clear whether the approximation rate can be improved by a more refined RNN structure construction (with possible fewer layers and smaller width) or other possible approaches. Second, we here only consider the "large n" setting where the event sequences are observed in a bounded time domain $[0,T]$ with n repeated samples. It is interesting to extend our results to "large T" setting where the end time T goes to infinity but the number of event sequences, n, remains fixed. Third, in the current work, we do not take into account the different event types. It may be useful to extend our results to the marked TPP settings. Moreover, it is also worth investigating the theoretical performances of other neural network architectures (e.g. Transformer-TPPs) that have performed well in recent empirical applications.

Acknowledgments

We gratefully thank the editor and the anonymous reviewer for their constructive comments. Guanhua Fang is partly supported by the National Natural Science Foundation of China (nos. 12301376) and Shanghai Educational Development Foundation (23CGA02). Wen Yu's research is supported by the National Natural Science Foundation of China (nos. 12071088).

Additional Notations in the Appendix: For two random variables X and Y, we write $X \leq_{s.t.} Y$ if $\mathbb{P}(X > t) \leq \mathbb{P}(Y > t)$ for any $t \in \mathbb{R}$. We use $\mathbb{N}_+$ to denote the set of positive integers. For $d \in \mathbb{N}_+$, define $\mathbf{1}_d := (1, \cdots, 1)^\top \in \mathbb{R}^d$.

Appendix A. A toy example of RNN-TPP in Section 2.2

Let $L = 2$ and $N_e = 2$, the sequence $S = \{t_1, t_2\}$. In this case, the imbedding function $x(t; S)$ is

$$x(t;S) \quad = \quad \begin{cases} (t,t)^\top & , \ t \in [0, t_1] \\ (t, t-t_1)^\top & , \ t \in (t_1, t_2] \\ (t, t-t_2)^\top & , \ t \in (t_2, T] \end{cases} .$$

The hidden neurons are

$$h_1^{(1)} = \sigma\left(W_x^{(1)} x(t_1;S) + b^{(1)}\right), \qquad h_2^{(1)} = \sigma\left(W_x^{(1)} x(t_2;S) + W_h^{(1)} h_1^{(1)} + b^{(1)}\right),$$
$$h_1^{(2)} = \sigma\left(W_x^{(2)} h_1^{(1)} + b^{(2)}\right), \qquad h_2^{(2)} = \sigma\left(W_x^{(2)} h_2^{(1)} + W_h^{(2)} h_1^{(2)} + b^{(2)}\right).$$

The hidden functions are

$$h^{(1)}(t;S) \quad = \quad \begin{cases} \sigma\left(W_x^{(1)} x(t;S) + b^{(1)}\right) & , \ t \in [0, t_1] \\ \sigma\left(W_x^{(1)} x(t;S) + W_h^{(1)} h_1^{(1)} + b^{(1)}\right) & , \ t \in (t_1, t_2] \\ \sigma\left(W_x^{(1)} x(t;S) + W_h^{(1)} h_2^{(1)} + b^{(1)}\right) & , \ t \in (t_2, T] \end{cases} ,$$

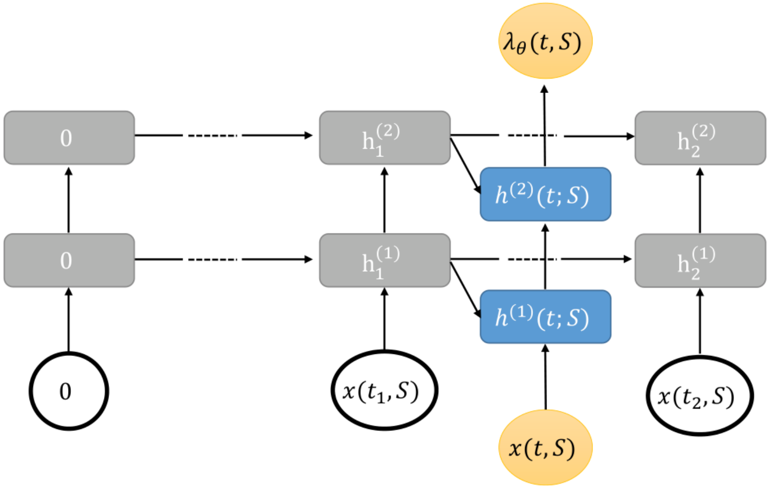
and

$$h^{(2)}(t;S) \quad = \quad \begin{cases} \sigma\left(W_x^{(2)} h^{(1)}(t;S) + b^{(2)}\right) & , \ t \in [0, t_1] \\ \sigma\left(W_x^{(2)} h^{(1)}(t;S) + W_h^{(2)} h_1^{(2)} + b^{(2)}\right) & , \ t \in (t_1, t_2] \\ \sigma\left(W_x^{(2)} h^{(1)}(t;S) + W_h^{(2)} h_2^{(2)} + b^{(2)}\right) & , \ t \in (t_2, T] \end{cases} .$$

The output intensity of RNN-TPP is

$$\lambda_\theta(t;S) \quad := \quad f\left(W_x^{(3)} h^{(2)}(t;S) + b^{(3)}\right) \in \mathbb{R}, \tag{29}$$

where $\theta = \{W_x^{(1)}, W_h^{(1)}, b^{(1)}, W_x^{(2)}, W_h^{(2)}, b^{(2)}, W_x^{(3)}, b^{(3)}\}$. This example is illustrated in Figure 6.



example.jpg

Figure 6: An illustration of (29) on $(t_1, t_2]$.

Appendix B. Proofs in Section 3

B.1 Proof of Lemma 3.5

Proof By the definition of $\check{\lambda}^*$ and $\hat{\lambda}$, we have

$$\begin{aligned}
&\mathbb{E}[\text{loss}(\hat{\lambda}, S_{test})] - \mathbb{E}[\text{loss}(\lambda^*, S_{test})]] \\
&= \mathbb{E}[\text{loss}(\hat{\lambda}, S_{test})] - \mathbb{E}[\text{loss}(\check{\lambda}^*, S_{test})] + \mathbb{E}[\text{loss}(\check{\lambda}^*, S_{test})] - \mathbb{E}[\text{loss}(\lambda^*, S_{test})] \\
&\le \mathbb{E}[\text{loss}(\hat{\lambda}, S_{test})] \underbrace{- \frac{1}{n}\sum_{i\in[n]} \text{loss}(\hat{\lambda}, S_i) + \frac{1}{n}\sum_{i\in[n]} \text{loss}(\check{\lambda}^*, S_i)}_{\ge 0} - \mathbb{E}[\text{loss}(\check{\lambda}^*, S_{test})] \\
&\quad + \mathbb{E}[\text{loss}(\check{\lambda}^*, S_{test})] - \mathbb{E}[\text{loss}(\lambda^*, S_{test})] \\
&\le 2 \sup_{\lambda\in\mathcal{F}} \left| \mathbb{E}[\text{loss}(\lambda, S_{test})] - \frac{1}{n}\sum_{i\in[n]} \text{loss}(\lambda, S_i) \right| \\
&\quad + \mathbb{E}[\text{loss}(\check{\lambda}^*, S_{test})] - \mathbb{E}[\text{loss}(\lambda^*, S_{test})].
\end{aligned}$$

■

Appendix C. Proofs in Section 4

C.1 Proof of Lemma 4.4

Proof From (9) and model assumptions (A1)-(A2), we have $\lambda^*(t) = \lambda_0(t) + \sum_{j:t_j<t} \mu(t-t_j)$, $\int_0^T \mu(t)\mathrm{d}t \le c_\mu < 1$, $\lambda_0(t) \le B_0$. Following the notations in the paper, we denote N_e as the number of event time of λ^* in $[0, T]$. Consider another density $\overline{\lambda}(t) = B_0 + \sum_{j:t_j<t} \mu(t-t_j)$ and similarly denote $\overline{N}_e$ as the number of event time of $\overline{\lambda}$ in $[0, T]$. Then for any fixed event

structure of Hawkes.jpg

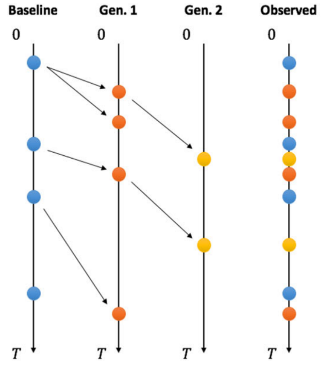


Figure 7: Branching structures of the Hawkes process: parent events (blue circle) and two generations of offspring events.

sequence $S = \{t_j\}$, $\lambda^*(t;S) \leq \overline{\lambda}(t;S)$, and thus $N_e \leq_{s.t.} \overline{N}_e$. Following Daley and Vere-Jones (2003) Example 6.3(c) or Cheysson and Lang (2022) Section 2.2, the point process with intensity $\overline{\lambda}$ is equivalent to a birth-immigration process with immigration intensity B_0 and birth intensity $\mu(t)$. Hence

$$\overline{N}_e = \overline{N}_0 + \sum_{k=1}^{\infty} \overline{N}_k,$$

where $\overline{N}_0 \sim \text{Poisson}(B_0 T)$ is the number of parent events and $\overline{N}_k$ is the number of offspring events in generation k, which are children of generation $k-1$. The branching structure is illustrated Figure 7 (Figure 2 in Fang et al. (2023)). For $t_1 < t_2$, let $\mu_{t_1}^{t_2} = \int_{t_1}^{t_2} \mu(t-t_1)\mathrm{d}t$. We have

$$\mathbb{E}\left[\exp\left(s\overline{N}_0\right)\right] = \exp\left(B_0 T\left(\exp(s)-1\right)\right),$$

and

$$\begin{aligned}
\mathbb{E}\left[\exp\left(s\overline{N}_{k+1}\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(s\overline{N}_{k+1}\right)\,\middle|\,\left\{t_j^{(k)}\right\}_{j=1}^{\overline{N}_k}\right]\right] \\
&= \mathbb{E}\left[\prod_{j=1}^{\overline{N}_k} \exp\left(\mu_{t_j^{(k)}}^{T}\left(\exp(s)-1\right)\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(c_\mu \overline{N}_k\left(\exp(s)-1\right)\right)\right],
\end{aligned}$$

for any $s > 0$. Since $c_\mu < 1$, for any fixed $c_1 \in (c_\mu, 1]$ and any $s \in (0, \log(c_1/c_\mu)]$, we have

$$\mathbb{E}\left[\exp\left(s\overline{N}_k\right)\right] \leq \mathbb{E}\left[\exp\left(c_\mu \overline{N}_{k-1}\left(\exp(s)-1\right)\right)\right] \leq \mathbb{E}\left[\exp\left(c_1 s\overline{N}_{k-1}\right)\right] \leq \cdots \leq \mathbb{E}\left[\exp\left(c_1^k s\overline{N}_0\right)\right],$$

i.e.

$$\mathbb{E}\left[\exp\left(s\overline{N}_k\right)\right] \leq \mathbb{E}\left[\exp\left(c_1^k s\overline{N}_0\right)\right] = \exp\left(B_0 T\left(\exp\left(c_1^k s\right)-1\right)\right) \leq \exp\left(\frac{c_1^{k+1}}{c_\mu}(B_0 T)s\right)$$

for any $k \in \mathbb{N}$.

Since $\overline{N}_k$ can only take integer values, we can get $\mathbb{P}(\overline{N}_k = 0) + e^s \mathbb{P}(\overline{N}_k \neq 0) \leq \mathbb{E}\left[\exp(s\overline{N}_k)\right]$. Thus

$$\mathbb{P}\left(\overline{N}_k \neq 0\right) \leq \frac{\mathbb{E}\left[\exp\left(s\overline{N}_k\right)\right] - 1}{\exp(s) - 1} \leq \frac{c_1^{k+2}}{c_\mu^2}(B_0 T),\ \forall s \in \left(0, \min\left\{\frac{c_\mu}{c_1^{k+1}(B_0T)}, 1\right\} \log\left(\frac{c_1}{c_\mu}\right)\right].$$

Setting $c_1 \searrow c_\mu$, we get

$$\mathbb{P}\left(\overline{N}_k \neq 0\right) \leq c_\mu^k (B_0 T). \tag{30}$$

Now take $c_1 \in (c_\mu, 1)$, and then $c_1^{-1}(1-c_1)\sum_{k=1}^\infty c_1^k = 1$. By Bool's inequality, we have

$$\begin{aligned}\mathbb{P}\left(\overline{N}_e \geq N\right) = \mathbb{P}\left(\sum_{k=0}^\infty \overline{N}_k \geq N\right) &\leq \sum_{k=0}^\infty \mathbb{P}\left(\overline{N}_k \geq \frac{1-c_1}{c_1} c_1^{k+1} N\right) \\ &\leq \sum_{k=0}^{K_0-1} \mathbb{P}\left(\overline{N}_k \geq \frac{1-c_1}{c_1} c_1^{k+1} N\right) + \sum_{k=K_0}^\infty \mathbb{P}\left(\overline{N}_k \neq 0\right).\end{aligned} \tag{31}$$

Now we bound (31). For the second term, by (30), we have

$$\sum_{k=K_0}^\infty \mathbb{P}\left(\overline{N}_k \neq 0\right) \leq \sum_{k=K_0}^\infty c_\mu^k (B_0 T) = c_\mu^{K_0}(B_0T)/(1-c_\mu). \tag{32}$$

Let

$$K_0 \geq \frac{\log\left(2nB_0T/[\delta(1-c_\mu)]\right)}{\log\left(1/c_\mu\right)}, \tag{33}$$

it can be showed that

$$\sum_{k=K_0}^\infty \mathbb{P}\left(\overline{N}_k \neq 0\right) \leq c_\mu^{K_0}(B_0T)/(1-c_\mu) \leq \delta/2n.$$

For the first term, we have

$$\begin{aligned}\sum_{k=0}^{K_0-1} \mathbb{P}\left(\overline{N}_k \geq \frac{1-c_1}{c_1} c_1^{k+1} N\right) &\leq \sum_{k=0}^{K_0-1} \exp\left(-s\left(\frac{1-c_1}{c_1} c_1^{k+1} N\right)\right) \mathbb{E}\left[\exp(s\overline{N}_k)\right] \\ &\leq \sum_{k=0}^{K_0-1} \exp\left(c_1^{k+1} s\left(\frac{B_0T}{c_\mu} - \frac{1-c_1}{c_1} N\right)\right),\end{aligned}$$

where $s \in (0, \log(c_1/c_\mu)]$. We can take

$$c_1 s\left(B_0T/c_\mu - (1-c_1)N/c_1\right) \leq \log(\delta/(2nK_0)), \tag{34}$$

so that

$$\sum_{k=0}^{K_0-1} \mathbb{P}\left(\overline{N}_k \geq \frac{1-c_1}{c_1} c_1^{k+1} N\right) \leq \sum_{k=0}^{K_0-1} \exp\left(c_1^{k+1} s \left(B_0 T/c_\mu - (1-c_1)N/c_1\right)\right) \leq \delta/(2n). \tag{35}$$

To obtain (34), we need

$$N \geq \frac{1}{1-c_1}\left(\frac{1}{s}\log\left(\frac{2nK_0}{\delta}\right) + \frac{c_1}{c_\mu}(B_0 T)\right).$$

From now on, let $\eta = c_1/c_\mu \in (1, 1/c_\mu)$, $s = \log(c_1/c_\mu) = \log(\eta)$. Taking $K_0 = \lceil \log\left(2nB_0T/[\delta(1-c_\mu)]\right)/\log\left(1/c_\mu\right) \rceil$ and $N = \left[\log\left(2nK_0/\delta\right)/\log(\eta) + \eta(B_0T)\right]/(1-c_\mu\eta)$ to obtain (33) and (34). Then by (31), (32) and (35), we have

$$\mathbb{P}\left(N_e \geq N\right) \leq \mathbb{P}\left(\overline{N}_e \geq N\right) \leq \delta/n.$$

Since

$$\mathbb{P}(N_{e(n)} \geq N) = 1 - \mathbb{P}(N_{e(n)} < N) = 1 - \prod_{i=1}^{n} \mathbb{P}\left(N_e < N\right) \leq 1 - \left(1 - \frac{\delta}{n}\right)^n \leq \delta,$$

we get that with probability at least $1-\delta$,

$$N_{e(n)} < N \leq \frac{1}{1-c_\mu\eta}\left[\frac{1}{\log(\eta)}\log\left(\frac{2nK_{n,\delta}}{\delta}\right) + \eta(B_0T)\right],$$

where $\eta \in (1, 1/c_\mu)$, and $K_{n,\delta} = \log\left(2nB_0T/\delta(1-c_\mu)\right)/\log\left(1/c_\mu\right) + 1$. Since $1 - 1/x \leq \log(x) \leq x - 1$, we have $K_{n,\delta} \leq 2nB_0T/[\delta(1-c_\mu)^2]$. Thus with probability at least $1-\delta$,

$$N_{e(n)} < \frac{1}{1-c_\mu\eta}\left[\frac{2}{\log(\eta)}\log\left(\frac{2n\sqrt{B_0T}}{\delta(1-c_\mu)}\right) + \eta(B_0T)\right].$$

Taking $n = 1$ and $2\log\left(2\sqrt{B_0T}/[\delta(1-c_\mu)]\right)/\log(\eta) + \eta(B_0T)/(1-c_\mu\eta) = s$, we have $\delta = 2\sqrt{B_0T}\exp\left(\log(\eta)\left[\eta(B_0T) - (1-c_\mu\eta)s\right]/2\right)/(1-c_\mu)$. Then

$$\mathbb{P}\left(N_e = s\right) \leq \mathbb{P}\left(N_e \geq s\right) \leq \frac{2\sqrt{B_0T}}{1-c_\mu}\exp\left(\frac{\log(\eta)}{2}\left[\eta(B_0T) - (1-c_\mu\eta)s\right]\right).$$

■

C.2 Proof of Lemma 4.5

Proof Conditioning on the event $\{N_{e(n)} \leq s\}$, we have

$$\begin{aligned}
X_\theta &= \mathbb{E}[\text{loss}(\lambda_\theta, S_{test})] - \frac{1}{n}\sum_{i=1}^{n}\text{loss}(\lambda_\theta, S_i) \\
&= \mathbb{E}[\text{loss}(\lambda_\theta, S_{test})] - \frac{1}{n}\sum_{k=1}^{n}\text{loss}(\lambda_\theta, S_i)\mathbb{1}_{\{N_{ei} \leq s\}} \\
&= X_\theta(s) + E_\theta(s).
\end{aligned}$$

Hence, under the condition $N_{e(n)} \leq s$, we have $|X_\theta| \leq |X_\theta(s)| + |E_\theta(s)|$, thus $\sup_{\theta\in\Theta} |X_\theta| \leq \sup_{\theta\in\Theta} |X_\theta(s)| + \sup_{\theta\in\Theta} |E_\theta(s)|$. Then

$$\begin{aligned}\mathbb{P}\left(\sup_{\theta\in\Theta} |X_\theta| > t\right) &= \mathbb{P}\left(\sup_{\theta\in\Theta} |X_\theta| > t,\ N_{e(n)} \leq s\right) + \mathbb{P}\left(\sup_{\theta\in\Theta} |X_\theta| > t,\ N_{e(n)} > s\right) \\ &\leq \mathbb{P}\left(\sup_{\theta\in\Theta} |X_\theta(s)| + \sup_{\theta\in\Theta} |E_\theta(s)| > t,\ N_{e(n)} \leq s\right) \\ &\quad + \mathbb{P}\left(\sup_{\theta\in\Theta} |X_\theta| > t,\ N_{e(n)} > s\right) \\ &\leq \mathbb{P}\left(\sup_{\theta\in\Theta} |X_\theta(s)| + \sup_{\theta\in\Theta} |E_\theta(s)| > t\right) + \mathbb{P}\left(N_{e(n)} > s\right).\end{aligned}$$

■

C.3 Proof of Lemma 4.6

The proof is based on induction. Using the same notation, we give two claims.

Claim C.1. *For $\forall 1 \leq l \leq L$, $1 \leq j \leq N$, $\|h_{j,1}^{(l)} - h_{j,2}^{(l)}\|_2$ is bounded by*

$$\begin{aligned}&\left\|h_{j,1}^{(l)} - h_{j,2}^{(l)}\right\|_2 \\ \leq &\rho_\sigma \left(\sum_{r=0}^{l-1} \gamma^r S_{j-1}^r \Delta_b^{l-r} + B_\sigma\sqrt{D} \sum_{r=0}^{l-2} \gamma^r S_{j-1}^r \Delta_x^{l-r} + B_{in}(T)\gamma^{l-1} S_{j-1}^{l-1} \Delta_x^1 + B_\sigma \sqrt{D} \sum_{r=0}^{l-1} \gamma^r S_{j-2}^r \Delta_h^{l-r}\right).\end{aligned} \tag{36}$$

Proof [Proof of Claim C.1] When $i = 1$, we have

$$\begin{aligned}\left\|h_{1,1}^{(l)} - h_{1,2}^{(l)}\right\|_2 &= \left\|\sigma\left(W_{x,1}^{(l)} h_{1,1}^{(l-1)} + b_1^{(l)}\right) - \sigma\left(W_{x,2}^{(l)} h_{1,2}^{(l-1)} + b_2^{(l)}\right)\right\|_2 \\ &\overset{(i)}{\leq} \rho_\sigma \left(\left\|W_{x,1}^{(l)} h_{1,1}^{(l-1)} - W_{x,2}^{(l)} h_{1,2}^{(l-1)}\right\|_2 + \left\|b_1^{(l)} - b_2^{(l)}\right\|_2\right) \\ &\overset{(ii)}{\leq} \rho_\sigma \left(\left\|W_{x,1}^{(l)} h_{1,1}^{(l-1)} - W_{x,2}^{(l)} h_{1,1}^{(l-1)}\right\|_2 + \left\|W_{x,2}^{(l)} h_{1,1}^{(l-1)} - W_{x,2}^{(l)} h_{1,2}^{(l-1)}\right\|_2 + \left\|b_1^{(l)} - b_2^{(l)}\right\|_2\right) \\ &\overset{(iii)}{\leq} \rho_\sigma \left(B_\sigma \sqrt{D} \Delta_x^l + B_x \left\|h_{1,1}^{(l-1)} - h_{1,2}^{(l-1)}\right\|_2 + \Delta_b^l\right),\end{aligned}$$

where (i) follows from the Lipscitz continuous of σ, and (ii) follows from triangle inequality and (iii) follows from the boundedness of hidden layers and parameter space. Repeat this derivation recursively, we get

$$\begin{aligned}\left\|h_{1,1}^{(l)} - h_{1,2}^{(l)}\right\|_2 &\leq \rho_\sigma \left(B_\sigma\sqrt{D}\Delta_x^l + B_x \left\|h_{1,1}^{(l-1)} - h_{1,2}^{(l-1)}\right\|_2 + \Delta_b^l\right) \\ &\leq \rho_\sigma \Delta_b^l + \rho_\sigma B_\sigma \sqrt{D} \Delta_x^l + \gamma\left(\rho_\sigma \Delta_b^l + \rho_\sigma B_\sigma\sqrt{D}\Delta_x^l + \gamma \left\|h_{1,1}^{(l-2)} - h_{1,2}^{(l-2)}\right\|_2\right) \\ &\leq \cdots\cdots \\ &\leq \rho_\sigma \left(\sum_{r=0}^{l-1} \gamma^r \Delta_b^{l-r} + B_\sigma \sqrt{D} \sum_{r=0}^{l-2} \gamma^r \Delta_x^{l-r} + B_{in}(T)\gamma^{l-1}\Delta_x^1\right).\end{aligned}$$

When $l = 1$, we have

$$\begin{aligned}
&\left\| h_{j,1}^{(1)} - h_{j,2}^{(1)} \right\|_2 \\
= &\left\| \sigma\left(W_{x,1}^{(1)} x(t_j; S) + W_{h,1}^{(1)} h_{j-1,1}^{(1)} + b_1^{(1)} \right) - \sigma\left(W_{x,2}^{(1)} x(t_j; S) + W_{h,2}^{(1)} h_{j-1,2}^{(1)} + b_2^{(1)} \right) \right\|_2 \\
\overset{(i)}{\leq} &\rho_\sigma \left(B_{in}(T) \left\| W_{x,1}^{(1)} - W_{x,2}^{(1)} \right\|_2 + \left\| W_{h,1}^{(1)} h_{j-1,1}^{(1)} - W_{h,2}^{(1)} h_{j-1,2}^{(1)} \right\|_2 + \left\| b_1^{(1)} - b_2^{(1)} \right\|_2 \right) \\
\overset{(ii)}{\leq} &\rho_\sigma \left(B_{in}(T) \Delta_x^1 + B_\sigma \sqrt{D} \Delta_h^1 + B_h \left\| h_{j-1,1}^{(1)} - h_{j-1,2}^{(1)} \right\|_2 + \Delta_b^1 \right),
\end{aligned}$$

where (i), (ii) follows from the triangle inequality and boundedness of embedding function, hidden layers and parameter space. Again repeat it recursively, we can get

$$\begin{aligned}
&\left\| h_{j,1}^{(1)} - h_{j,2}^{(1)} \right\|_2 \\
\leq &\rho_\sigma \left(B_{in}(T) \Delta_x^1 + B_\sigma \sqrt{D} \Delta_h^1 + B_h \left\| h_{j-1,1}^{(1)} - h_{j-1,2}^{(1)} \right\|_2 + \Delta_b^1 \right) \\
\leq &\rho_\sigma \left(B_{in}(T) \Delta_x^1 + B_\sigma \sqrt{D} \Delta_h^1 + \Delta_b^1 \right) + \beta \left(\rho_\sigma \left(B_{in}(T) \Delta_x^1 + B_\sigma \sqrt{D} \Delta_h^1 + \Delta_b^1 \right) + \beta \left\| h_{j-2,1}^{(1)} - h_{j-2,2}^{(1)} \right\|_2 \right) \\
\leq &\cdots\cdots \\
\leq &\rho_\sigma \left(S_{j-1}^0 \Delta_b^1 + B_{in}(T) S_{j-1}^0 \Delta_x^1 + B_\sigma \sqrt{D} S_{j-2}^0 \Delta_h^1 \right).
\end{aligned}$$

Now suppose for all $j < j_0$, $l < l_0$, (36) is true. Consider the case $j = j_0$, $l = l_0$, we have

$$\begin{aligned}
&\left\| h_{j_0,1}^{(l_0)} - h_{j_0,2}^{(l_0)} \right\|_2 \\
= &\left\| \sigma\left(W_{x,1}^{(l_0)} h_{j_0,1}^{(l_0-1)} + W_{h,1}^{(l_0)} h_{j_0-1,1}^{(l_0)} + b_1^{(l_0)} \right) - \sigma\left(W_{x,2}^{(l_0)} h_{j_0,2}^{(l_0-1)} + W_{h,2}^{(l_0)} h_{j_0-1,2}^{(l_0)} + b_2^{(l_0)} \right) \right\|_2 \\
\overset{(i)}{\leq} &\rho_\sigma \left(\left\| W_{x,1}^{(l_0)} h_{j_0,1}^{(l_0-1)} - W_{x,2}^{(l_0)} h_{j_0,2}^{(l_0-1)} \right\|_2 + \left\| W_{h,1}^{(l_0)} h_{j_0-1,1}^{(l_0)} - W_{h,2}^{(l_0)} h_{j_0-1,2}^{(l_0)} \right\|_2 + \left\| b_1^{(l_0)} - b_2^{(l_0)} \right\|_2 \right) \\
\overset{(ii)}{\leq} &\rho_\sigma \left(B_x \left\| h_{j_0,1}^{(l_0-1)} - h_{j_0,2}^{(l_0-1)} \right\|_2 + B_\sigma \sqrt{D} \Delta_x^{l_0} + B_h \left\| h_{j_0-1,1}^{(l_0)} - h_{j_0-1,2}^{(l_0)} \right\|_2 + B_\sigma \sqrt{D} \Delta_h^{l_0} + \Delta_b^{l_0} \right) \\
\overset{(iii)}{\leq} &\rho_\sigma \beta \left(\sum_{r=0}^{l_0-1} \gamma^r S_{j_0-2}^r \Delta_b^{l_0-r} + B_\sigma \sqrt{D} \sum_{r=0}^{l_0-2} \gamma^r S_{j_0-2}^r \Delta_x^{l_0-r} + B_{in}(T) \gamma^{l_0-1} S_{j_0-2}^{l_0-1} \Delta_x^1 \right. \\
&\left. \quad + B_\sigma \sqrt{D} \sum_{r=0}^{l_0-1} \gamma^r S_{j_0-3}^r \Delta_h^{l_0-r} \right) + \rho_\sigma \gamma \left(\sum_{r=0}^{l_0-2} \gamma^r S_{j_0-1}^r \Delta_b^{l_0-1-r} + B_\sigma \sqrt{D} \sum_{r=0}^{l_0-3} \gamma^r S_{j_0-1}^r \Delta_x^{l_0-1-r} \right. \\
&\left. \quad B_{in}(T) \gamma^{l_0-2} S_{j_0-1}^{l_0-2} \Delta_x^1 + B_\sigma \sqrt{D} \sum_{r=0}^{l_0-2} \gamma^r S_{j_0-2}^r \Delta_h^{l_0-1-r} \right) + \rho_\sigma \left(B_\sigma \sqrt{D} \Delta_x^{l_0} + B_\sigma \sqrt{D} \Delta_h^{l_0} + \Delta_b^{l_0} \right) \\
= &\rho_\sigma \left(\sum_{r=1}^{l_0-1} \gamma^r (\beta S_{j_0-2}^r + S_{j_0-1}^{r-1}) \Delta_b^{l_0-r} + B_\sigma \sqrt{D} \sum_{r=1}^{l_0-2} \gamma^r (\beta S_{j_0-2}^r + S_{j_0-1}^{r-1}) \Delta_x^{l_0-r} \right. \\
&\left. \quad + B_{in}(T) \gamma^{l_0-1} (\beta S_{j_0-2}^{l_0-1} + S_{j_0-1}^{l_0-2}) \Delta_x^1 + B_\sigma \sqrt{D} \sum_{r=1}^{l_0-1} \gamma^r (\beta S_{j_0-3}^r + S_{j_0-2}^{r-1}) \Delta_h^{l_0-r} \right)
\end{aligned}$$

$$+ \rho_\sigma \left((1+\beta S^0_{j_0-2}) \left(\Delta_b^{l_0} + B_\sigma \sqrt{D} \Delta_x^{l_0} \right) + (1+\beta S^0_{j_0-3}) B_\sigma \sqrt{D} \Delta_h^{l_0} \right),$$

where (i) follows from the Lipscitz continuous of σ and triangle inequality, (ii) follows from triangle inequality and the boundedness of hidden layers and parameter space, and (iii) follows from the induction hypothesis. Using the fact that $1 + \beta S^0_{j-1} = S^0_j$ and

$$\begin{aligned}\beta S^r_{j-1} + S^{r-1}_j &= \beta \sum_{q=0}^{j-1} \binom{q+r}{r} \beta^q + \sum_{q=0}^{j} \binom{q+r-1}{r-1} \beta^q = 1 + \sum_{q=1}^{j} \left(\binom{q+r-1}{r} + \binom{q+r-1}{r-1} \right) \beta^q \\ &= \sum_{q=0}^{j} \binom{q+r}{r} \beta^q = S^r_j,\end{aligned}$$

(36) is proved. ■

Claim C.2. *For $\forall 1 \le l \le L$, $1 \le j \le N$ and $t \in (t_j, t_{j+1}]$, $\|h_1^{(l)}(t;S) - h_2^{(l)}(t;S)\|_2$ is bounded by*

$$\left\| h_1^{(l)}(t;S) - h_2^{(l)}(t;S) \right\|_2 \le \rho_\sigma \left(\sum_{r=0}^{l-1} \gamma^r S^r_j \Delta_b^{l-r} + B_\sigma \sqrt{D} \sum_{r=0}^{l-2} \gamma^r S^r_j \Delta_x^{l-r} \right. \tag{37}$$

$$\left. + B_{in}(T) \gamma^{l-1} S^{l-1}_j \Delta^1_x + B_\sigma \sqrt{D} \sum_{r=0}^{l-1} \gamma^r S^r_{j-1} \Delta_h^{l-r} \right). \tag{38}$$

Proof [Proof of Claim C.2] It is similar to the proof of Claim C.1 and we will use the same approach. When $l = 1$, by the definition of $h^{(1)}(t;S)$ and (36), for any $1 \le j \le N$ and $t \in (t_j, t_{j+1}]$, we have

$$\begin{aligned}&\left\| h_1^{(1)}(t;S) - h_2^{(1)}(t;S) \right\|_2 \\ &= \left\| \sigma\left(W^{(1)}_{x,1} x(t;t_j) + W^{(1)}_{h,1} h^{(1)}_{j,1} + b^{(1)}_1 \right) - \sigma\left(W^{(1)}_{x,2} x(t;t_j) + W^{(1)}_{h,2} h^{(1)}_{j,2} + b^{(1)}_2 \right) \right\|_2 \\ &\le \rho_\sigma \left(B_{in}(T) \left\| W^{(1)}_{x,1} - W^{(1)}_{x,2} \right\|_2 + \left\| W^{(1)}_{h,1} h^{(1)}_{j,1} - W^{(1)}_{h,2} h^{(1)}_{j,2} \right\|_2 + \left\| b^{(1)}_1 - b^{(1)}_2 \right\|_2 \right) \\ &\le \rho_\sigma \left(B_{in}(T) \Delta^1_x + B_\sigma \sqrt{D} \Delta^1_h + B_h \left\| h^{(1)}_{j,1} - h^{(1)}_{j,2} \right\|_2 + \Delta^1_b \right) \\ &\le \rho_\sigma \beta \left(S^0_{j-1} \Delta^1_b + B_{in}(T) S^0_{j-1} \Delta^1_x + B_\sigma \sqrt{D} S^0_{j-2} \Delta^1_h \right) \\ &\quad + \rho_\sigma \left(B_{in}(T) \Delta^1_x + B_\sigma \sqrt{D} \Delta^1_h + \Delta^1_b \right) \\ &\le \rho_\sigma \left(S^0_j \Delta^1_b + B_{in}(T) S^0_j \Delta^1_x + B_\sigma \sqrt{D} S^0_{j-1} \Delta^1_h \right).\end{aligned}$$

Now suppose for all $l < l_0$, (36) is true for any $1 \le j \le N$ and $t \in (t_j, t_{j+1}]$. Considering the case $l = l_0$, for any $1 \le j \le N$ and $t \in (t_j, t_{j+1}]$, we have

$$\left\| h_1^{(l_0)}(t;S) - h_2^{(l_0)}(t;S) \right\|_2$$

$$
\begin{aligned}
&= \left\|\sigma\left(W_{x,1}^{(l_0)} h_1^{(l_0-1)}(t;S) + W_{h,1}^{(l_0)} h_{j,1}^{(l_0)} + b_1^{(l_0)}\right) - \sigma\left(W_{x,2}^{(l_0)} h_2^{(l_0-1)}(t;S) + W_{h,2}^{(l_0)} h_{j,2}^{(l_0)} + b_2^{(l_0)}\right)\right\|_2 \\
&\le \rho_\sigma\left(\left\|W_{x,1}^{(l_0)} h_1^{(l_0-1)}(t;S) - W_{x,2}^{(l_0)} h_2^{(l_0-1)}(t;S)\right\|_2 + \left\|W_{h,1}^{(l_0)} h_{j,1}^{(l_0)} - W_{h,2}^{(l_0)} h_{j,2}^{(l_0)}\right\|_2 + \left\|b_1^{(l_0)} - b_2^{(l_0)}\right\|_2\right) \\
&\le \rho_\sigma\left(\Delta_b^{l_0} + B_\sigma\sqrt{D}\Delta_x^{l_0} + B_x\left\|h_1^{(l_0-1)}(t;S) - h_2^{(l_0-1)}(t;S)\right\|_2 + B_\sigma\sqrt{D}\Delta_h^{l_0} + B_h\left\|h_{j,1}^{(l_0)} - h_{j,2}^{(l_0)}\right\|_2\right) \\
&\le \rho_\sigma\gamma\left(\sum_{r=0}^{l_0-2}\gamma^r S_j^r \Delta_b^{l_0-1-r} + B_\sigma\sqrt{D}\sum_{r=0}^{l_0-3}\gamma^r S_j^r\Delta_x^{l_0-1-r} + B_{in}(T)\gamma^{l_0-2}S_j^{l_0-2}\Delta_x^1\right. \\
&\qquad\left. + B_\sigma\sqrt{D}\sum_{r=0}^{l_0-2}\gamma^r S_{j-1}^r\Delta_h^{l_0-1-r}\right) + \rho_\sigma\beta\left(\sum_{r=0}^{l_0-1}\gamma^r S_{j-1}^r\Delta_b^{l_0-r} + B_\sigma\sqrt{D}\sum_{r=0}^{l_0-2}\gamma^r S_{j-1}^r\Delta_x^{l_0-r}\right. \\
&\qquad\left. + B_{in}(T)\gamma^{l_0-1}S_{j-1}^{l_0-1}\Delta_x^1 + B_\sigma\sqrt{D}\sum_{l=0}^{l_0-1}\gamma^r S_{j-2}^r\Delta_h^{l_0-r}\right) + \rho_\sigma\left(\Delta_b^{l_0} + B_\sigma\sqrt{D}\Delta_x^{l_0} + B_\sigma\sqrt{D}\Delta_h^{l_0}\right) \\
&\le \rho_\sigma\left(\sum_{r=1}^{l_0-1}\gamma^r(\beta S_{j-1}^r + S_j^{r-1})\Delta_b^{l_0-r} + B_\sigma\sqrt{D}\sum_{r=1}^{l_0-2}\gamma^r(\beta S_{j-1}^r + S_j^{r-1})\Delta_x^{l_0-r}\right. \\
&\qquad\left. + B_{in}(T)\gamma^{l_0-1}(\beta S_{j-1}^{l_0-1} + S_j^{l_0-2})\Delta_x^1 + B_\sigma\sqrt{D}\sum_{r=1}^{l_0-1}\gamma^r(\beta S_{j-2}^r + S_{j-1}^{r-1})\Delta_h^{l_0-r}\right) \\
&\quad + \rho_\sigma\left((1+\beta S_{j-2}^0)\left(\Delta_b^{l_0} + (B_\sigma\sqrt{D}\vee B_{in}(T))\Delta_x^{l_0}\right) + (1+\beta S_{j-3}^0)B_\sigma\sqrt{D}\Delta_h^{l_0}\right) \\
&\le \rho_\sigma\left(\sum_{r=0}^{l_0-1}\gamma^r S_j^r\Delta_b^{l_0-r} + B_\sigma\sqrt{D}\sum_{r=0}^{l_0-2}\gamma^r S_j^r\Delta_x^{l_0-r} + B_{in}(T)\gamma^{l_0-1}S_j^{l_0-1}\Delta_x^1 + B_\sigma\sqrt{D}\sum_{r=0}^{l_0-1}\gamma^r S_{j-1}^r\Delta_h^{l_0-r}\right).
\end{aligned}
$$

Hence (38) is proved. ■

Now we prove Lemma 4.

Proof For $t \in (t_j, t_{j+1}]$, we have

$$
\begin{aligned}
|\lambda_{\theta_1}(t;S) - \lambda_{\theta_2}(t;S)| &= \left|f\left(W_{x,1}^{(L+1)}h^{(L)}(t;S) + b_1^{(L+1)}\right) - f\left(W_{x,2}^{(L+1)}h^{(L)}(t;S) + b_2^{(L+1)}\right)\right| \\
&\le \rho_f\left(\left\|b_1^{(L+1)} - b_2^{(L+1)}\right\|_2 + \left\|W_{x,1}^{(L+1)}h^{(L)}(t;S) - W_{x,2}^{(L+1)}h^{(L)}(t;S)\right\|_2\right) \\
&\le \rho_f\left(\Delta_b^{L+1} + B_\sigma\sqrt{D}\Delta_x^{L+1} + B_x\left\|h_1^{(L)}(t;S) - h_2^{(L)}(t;S)\right\|_2\right) \\
&\le \rho_f\gamma\left(\sum_{l=0}^{L-1}\gamma^l S_j^l\Delta_b^{L-l} + B_\sigma\sqrt{D}\sum_{l=0}^{L-2}\gamma^l S_j^l\Delta_x^{L-l} + B_{in}(T)\gamma^{L-1}S_j^{L-1}\Delta_x^1\right. \\
&\qquad\left. + B_\sigma\sqrt{D}\sum_{l=0}^{L-1}\gamma^l S_{j-1}^l\Delta_h^{L-l}\right) + \rho_f\Delta_b^{L+1} + \rho_f B_\sigma\sqrt{D}\Delta_x^{L+1}.
\end{aligned}
$$

■

C.4 Proof of Lemma 4.7

Proof From Lemma 4, for $\forall\ \lambda_{\theta_1}, \lambda_{\theta_2} \in \mathcal{F}$, we have

$$\begin{aligned}
&d_N(\lambda_{\theta_1}, \lambda_{\theta_2})\\
\leq\ &\rho_f \gamma \left(\sum_{l=0}^{L-1} \gamma^l S_N^l \Delta_b^{L-l} + B_\sigma \sqrt{D} \sum_{l=0}^{L-2} \gamma^l S_N^l \Delta_x^{L-l} + B_{in}(T) \gamma^{L-1} S_N^{L-1} \Delta_x^1 + B_\sigma \sqrt{D} \sum_{l=0}^{L-1} \gamma^l S_{N-1}^l \Delta_h^{L-l} \right)\\
&+ \rho_f \Delta_b^{L+1} + \rho_f B_\sigma \sqrt{D} \Delta_x^{L+1}\\
\leq\ &\rho_f \left(B_\sigma \sqrt{D} \vee B_{in}(T) \vee 1 \right) \left(\gamma^L \vee 1 \right) S_N^{L-1} \Delta_\theta\\
\leq\ &\rho_f \left(B_\sigma \sqrt{D} \vee B_{in}(T) \vee 1 \right) \left(\gamma^L \vee 1 \right) (N+1)^{L-1} \frac{\beta^{N+1} - 1}{\beta - 1} \Delta_\theta,
\end{aligned}$$

where $\Delta_\theta \triangleq \sum_{l=0}^{L+1} \left(\Delta_b^l + \Delta_x^l + \Delta_h^l \right)$.

Define $C(N) \triangleq \rho_f \left(B_\sigma \sqrt{D} \vee B_{in}(T) \vee 1 \right) \left(\gamma^L \vee 1 \right) (N+1)^{L-1} (\beta^{N+1} - 1)/(\beta - 1)$, using Lemma G.1 and $\| \cdot \|_2 \leq \| \cdot \|_F$, we can get

$$\begin{aligned}
\mathcal{N}\left(\mathcal{F}, \epsilon, d_N(\cdot, \cdot)\right) \leq\ &\prod_{l=1}^{L+1} \mathcal{N}\left(W_x^{(l)}, \frac{\epsilon}{C(N)(3L+2)}, \| \cdot \|_F \right) \prod_{l=1}^{L} \mathcal{N}\left(W_h^{(l)}, \frac{\epsilon}{C(N)(3L+2)}, \| \cdot \|_F \right)\\
&\prod_{l=1}^{L+1} \mathcal{N}\left(b^{(l)}, \frac{\epsilon}{C(N)(3L+2)}, \| \cdot \|_2 \right)\\
\leq\ &\left(1 + \frac{C(N)(3L+2) B_m \sqrt{D}}{\epsilon} \right)^{D^2(3L+2)},
\end{aligned}$$

where $B_m = \max\{B_b, B_h, B_x\}$. ■

C.5 Proof of Theorem 4.1

Lemma C.3. *Under assumptions (B1)-(B3), for fixed $s \in \mathbb{N}$, with probability at least $1 - \delta$, we have*

$$\begin{aligned}
&\sup_{\theta \in \Theta} |X_\theta(s)|\\
\leq\ &\frac{50}{\sqrt{n}} \left(T + \frac{1}{l_f} \right) (s+1) \left\{ 8 u_f \left(\sqrt{\log\left(\frac{2}{\delta}\right)} + D\sqrt{(3L+2)\log\left(1 + M(s)\right)} \right) + D\sqrt{3L+2} \right\}.
\end{aligned}$$

Hence

$$\sup_{\|\theta\| \leq B_m} |X_\theta(s)| \leq \tilde{O}\left(\sqrt{\frac{D^2 L^2 s^3}{n}} \right),$$

where $M(s) = \rho_f B_m \sqrt{D} \left(B_\sigma \sqrt{D} \vee B_{in}(T) \vee 1 \right) \left(\gamma^L \vee 1 \right) (s+1)^{L-1} (\beta^{s+1} - 1)/(\beta - 1)$, $B_m = \max\{B_b, B_h, B_x\}$, $\gamma = \rho_\sigma B_x$, and $\beta = \rho_\sigma B_h$.

Proof [Proof of Lemma C.3] For $1 \le k \le n$, denote $X_{\theta,k}(s) = \mathbb{E}\left[\mathrm{loss}(\lambda_\theta, S_{test})\mathbb{1}_{\{N_e \le s\}}\right] - \mathrm{loss}(\lambda_\theta, S_k)\mathbb{1}_{\{N_{ek} \le s\}}$. Then $X_\theta(s) = n^{-1}\sum_{k=1}^n X_{\theta,k}(s)$. For two parameters θ_1 and θ_2, we have

$$\begin{aligned}
&\left|\mathrm{loss}(\lambda_{\theta_1}, S_k)\mathbb{1}_{\{N_{ek} \le s\}} - \mathrm{loss}(\lambda_{\theta_2}, S_k)\mathbb{1}_{\{N_{ek} \le s\}}\right| \\
\le{}& \left[\Big|\sum_{j=1}^{N_{ek}}(\log\lambda_{\theta_1}(t_j; S_k) - \log\lambda_{\theta_2}(t_j; S_k))\Big| + \Big|\int_0^T (\lambda_{\theta_1}(t; S_k) - \lambda_{\theta_2}(t; S_k))\,\mathrm{dt}\Big|\right]\mathbb{1}_{\{N_{ek} \le s\}} \\
\le{}& \left[\frac{1}{l_f}\sum_{j=1}^{N_{ek}}|\lambda_{\theta_1}(t_j; S_k) - \lambda_{\theta_2}(t_j; S_k))| + \int_0^T |\lambda_{\theta_1}(t; S_k) - \lambda_{\theta_2}(t; S_k)|\,\mathrm{dt}\right]\mathbb{1}_{\{N_{ek} \le s\}} \\
\le{}& \left(T + \frac{N_{ek}}{l_f}\right) d_{N_{ek}}(\lambda_{\theta_1}, \lambda_{\theta_2})\mathbb{1}_{\{N_{ek} \le s\}} \\
\le{}& \left(T + \frac{1}{l_f}\right)(s+1)d_s(\lambda_{\theta_1}, \lambda_{\theta_2}),
\end{aligned}$$

and similarly,

$$\left|\mathbb{E}\left[\mathrm{loss}(\lambda_{\theta_1}, S_{test})\mathbb{1}_{\{N_e \le s\}}\right] - \mathbb{E}\left[\mathrm{loss}(\lambda_{\theta_2}, S_{test})\mathbb{1}_{\{N_e \le s\}}\right]\right| \le \left(T + \frac{1}{l_f}\right)(s+1)d_s(\lambda_{\theta_1}, \lambda_{\theta_2}).$$

Hence

$$|X_{\theta_1,k}(s) - X_{\theta_2,k}(s)| \le 2\left(T + \frac{1}{l_f}\right)(s+1)d_s(\lambda_{\theta_1}, \lambda_{\theta_2}).$$

By the property of bounded variable, $X_{\theta_1,k}(s) - X_{\theta_2,k}(s)$ is $2\left(T + 1/l_f\right)(s+1)d_s(\lambda_{\theta_1}, \lambda_{\theta_2})$-sub-gaussian. Since $\{X_{\theta_1,k}(s) - X_{\theta_2,k}(s)\}_{k=1}^n$ is mutually independent, $X_{\theta_1}(s) - X_{\theta_2}(s)$ is $2\left(T + 1/l_f\right)(s+1)d_s(\lambda_{\theta_1}, \lambda_{\theta_2})/\sqrt{n}$-sub-gaussian. From assumptions B2 and B3, there exists $\|\theta_0\| \le B_m$ such that $\lambda_{\theta_0} \equiv 1$, implying $X_{\theta_0}(s) = n^{-1}\sum_{k=1}^n T\left(\mathbb{E}\left[\mathbb{1}_{\{N_e \le s\}}\right] - \mathbb{1}_{\{N_{ek} \le s\}}\right)$, hence by a standard concentration inequality of bounded variables, we have with probability at least $1 - \delta$,

$$|X_{\theta_0}(s)| \le T\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{39}$$

The diameter of $\mathcal{F}$ under the distance $d_s(\cdot,\cdot)$ can be bounded by

$$\begin{aligned}
\mathrm{diam}\left(\mathcal{F}|d_s\right) &\le \sup_{\theta_1,\theta_2\in\Theta} d_s(\lambda_{\theta_1}, \lambda_{\theta_2}) \le \sup_{\theta_1,\theta_2\in\Theta}\sup_{\#S\le s}\|\lambda_{\theta_1}(t; S) - \lambda_{\theta_2}(t; S)\|_{L^\infty} \\
&\le 2u_f\ .
\end{aligned} \tag{40}$$

By Lemma 5, we get

$$\log\mathcal{N}\left(\mathcal{F}, \epsilon, d_s(\cdot,\cdot)\right) \le D^2(3L+2)\log\left(1 + \frac{C(s)(3L+2)B_m\sqrt{D}}{\epsilon}\right)\ , \tag{41}$$

where $C(s) = \rho_f \left(B_\sigma\sqrt{D} \vee B_{in}(T) \vee 1\right) \left(\gamma^L \vee 1\right) (s+1)^{L-1}(\beta^{s+1}-1)/(\beta-1)$, $B_m = \max\{B_b, B_h, B_x\}$. Denote $M(s) = C(s)(3L+2)B_m\sqrt{D}$, $\mathcal{D} = \operatorname{diam}(\mathcal{F}|d_s)$. We have

$$\begin{aligned}
\int_0^{2\mathcal{D}} \sqrt{\log\left(1+\frac{M(s)}{\epsilon}\right)}\mathrm{d}\epsilon &= \left(\int_0^a + \int_a^{2\mathcal{D}}\right)\sqrt{\log\left(1+\frac{M(s)}{\epsilon}\right)}\mathrm{d}\epsilon \quad (\forall 0 \le a \le 2\mathcal{D}) \\
&\le \inf_{0\le a\le 2\mathcal{D}}\left\{\int_0^a \sqrt{\frac{M(s)}{\epsilon}}\mathrm{d}\epsilon + \int_a^{2\mathcal{D}}\sqrt{\log\left(1+\frac{M(s)}{\epsilon}\right)}\mathrm{d}\epsilon\right\} \\
&\le \inf_{0\le a\le 2\mathcal{D}}\left\{2\sqrt{M(s)a} + 2\mathcal{D}\sqrt{\log\left(1+\frac{M(s)}{a}\right)}\right\} \\
&\le 2 + 2\mathcal{D}\sqrt{\log\left(1+M(s)^2\right)} \quad (\text{take } a = M(s)^{-1}) \\
&\le 2 + 4\mathcal{D}\sqrt{\log\left(1+M(s)\right)}, \qquad (42)
\end{aligned}$$

where we need $2\mathcal{D}M(s) \ge 1$. If $2\mathcal{D}M(s) < 1$, (42) is obvious since the integral is less than 2.

Combining (40), (41), (42) and using Lemma G.3, we have that with probability at least $1-\delta$,

$$\begin{aligned}
&\sup_{\theta\in\Theta}|X_\theta(s) - X_{\theta_0}(s)| \\
\le&\frac{24}{\sqrt{n}}\left(T+\frac{1}{l_f}\right)(s+1)\left(\mathcal{D}\left(4\sqrt{\log\left(\frac{2}{\delta}\right)} + 4D\sqrt{(3L+2)\log(1+M(s))}\right) + 2D\sqrt{3L+2}\right) \\
\le&\frac{24}{\sqrt{n}}\left(T+\frac{1}{l_f}\right)(s+1)\left\{8u_f\left(\sqrt{\log\left(\frac{2}{\delta}\right)} + D\sqrt{(3L+2)\log\left(1+M(s)\right)}\right) + 2D\sqrt{3L+2}\right\}.
\end{aligned} \qquad (43)$$

Now combining (39) and(43), by the union bound argument, we have that with probability at least $1-\delta$,

$$\begin{aligned}
&\sup_{\theta\in\Theta}|X_\theta(s)| \\
\le&\sup_{\theta\in\Theta}|X_\theta(s) - X_{\theta_0}(s)| + |X_{\theta_0}(s)| \\
\le&\frac{24}{\sqrt{n}}\left(T+\frac{1}{l_f}\right)(s+1)\left\{8u_f\left(\sqrt{\log\left(\frac{4}{\delta}\right)} + D\sqrt{(3L+2)\log\left(1+M(s)\right)}\right) + 2D\sqrt{3L+2}\right\} + T\sqrt{\frac{2\log(4/\delta)}{n}} \\
\le&\frac{50}{\sqrt{n}}\left(T+\frac{1}{l_f}\right)(s+1)\left\{8u_f\left(\sqrt{\log\left(\frac{2}{\delta}\right)} + D\sqrt{(3L+2)\log\left(1+M(s)\right)}\right) + D\sqrt{3L+2}\right\}.
\end{aligned}$$

■

Lemma C.4. *Suppose the event number N_e satisfies the tail condition*

$$\mathbb{P}(N_e \geq s) \leq a_N \exp(-c_N s), \;\; s \in \mathbb{N}.$$

Under assumptions (B1)-(B3), for fixed $s \in \mathbb{N}$, we have

$$\sup_{\theta\in\Theta} |E_\theta(s)| \leq \left(T + \frac{1}{l_f}\right)(u_f+2)\frac{a_N(s+2)}{(1-\exp(-c_N))^2}\exp(-c_N(s+1)) \;.$$

Proof [Proof of Lemma C.4] From assumptions (B2) and (B3), there exists $\theta_0 \in \Theta$ such that $\lambda_{\theta_0} \equiv 1$. Then

$$\begin{aligned} |E_\theta(s)| &= \left|\mathbb{E}\left[\text{loss}(\lambda_\theta, S_{test})\mathbb{1}_{\{N_e>s\}}\right]\right| \leq \mathbb{E}\left[|\text{loss}(\lambda_\theta, S_{test})|\,\mathbb{1}_{\{N_e>s\}}\right] \\ &\leq \mathbb{E}\left[|\text{loss}(\lambda_\theta, S_{test}) - \text{loss}(\lambda_{\theta_0}, S_{test})|\,\mathbb{1}_{\{N_e>s\}}\right] + \mathbb{E}\left[|\text{loss}(\lambda_{\theta_0}, S_{test})|\,\mathbb{1}_{\{N_e>s\}}\right] \\ &\leq \mathbb{E}\left\{\left[\left(T+\frac{1}{l_f}\right)(N_e+1)d_{N_e}(\lambda_\theta,\lambda_{\theta_0})\right]\mathbb{1}_{\{N_e>s\}}\right\} + T\mathbb{P}(N_e>s) \\ &\leq \left(T+\frac{1}{l_f}\right)(u_f+1)\mathbb{E}[(N_e+1)\mathbb{1}_{\{N_e>s\}}] + T\mathbb{P}(N_e>s) \end{aligned}$$

By the tail condition $\mathbb{P}(N_e \geq s) \leq a_N \exp(-c_N s), \;\; s \in \mathbb{N}$, we have

$$\begin{aligned} |E_\theta(s)| &\leq \left(T+\frac{1}{l_f}\right)(u_f+1)\frac{a_N(s+1)}{(1-\exp(-c_N))^2}\exp(-c_N(s+1)) \\ &\quad + \left(T+\frac{1}{l_f}\right)(u_f+2)a_N\exp(-c_N(s+1)) \\ &\leq \left(T+\frac{1}{l_f}\right)(u_f+2)\frac{a_N(s+2)}{(1-\exp(-c_N))^2}\exp(-c_N(s+1)). \end{aligned}$$

■

Now we prove Theorem 4.1.

Proof From Lemma 4.5, we have

$$\mathbb{P}\left(\sup_{\theta\in\Theta}|X_\theta| > t\right) \leq \mathbb{P}\left(\sup_{\theta\in\Theta}|X_\theta(s)| + \sup_{\theta\in\Theta}|E_\theta(s)| > t\right) + \mathbb{P}(N_{e(n)} > s).$$

Since

$$\mathbb{P}(N_{e(n)} > s) \leq n\mathbb{P}(N_e > s) \leq na_N\exp(-c_N s),$$

we can take $s_0 = \lceil \log(2a_N n/\delta)/c_N \rceil$ such that $na_N \exp(-c_N s_0) \leq \delta/2$, so we only need solve $t > 0$ such that

$$\mathbb{P}\left(\sup_{\theta\in\Theta}|X_\theta(s_0)| + \sup_{\theta\in\Theta}|E_\theta(s_0)| > t\right) \leq \frac{\delta}{2} \;.$$

From Lemma C.4, we have

$$\sup_{\theta\in\Theta}|E_\theta(s_0)| \le \left(T+\frac{1}{l_f}\right)(u_f+2)\frac{a_N(s_0+2)}{(1-\exp(-c_N))^2}\exp(-c_N(s_0+1)) := B(s_0).$$

By the definition of s_0, $B(s_0) \le (T+1/l_f)\,(u_f+2)(s_0+2)\delta/[2n(1-\exp(-c_N))^2]$. Thus we only need to solve $t>0$ such that

$$\mathbb{P}\left(\sup_{\theta\in\Theta}|X_\theta(s_0)| + \sup_{\theta\in\Theta}|E_\theta(s_0)| > t\right) \le \mathbb{P}\left(\sup_{\theta\in\Theta}|X_\theta(s_0)| > t - B(s_0)\right) \le \frac{\delta}{2}\ .$$

From Lemma C.3, we can choose

$$\begin{aligned}
t_0 &= \frac{50}{\sqrt{n}}\left(T+\frac{1}{l_f}\right)(s_0+1)\left\{8u_f\left(\sqrt{\log\left(\frac{4}{\delta}\right)} + D\sqrt{(3L+2)\log\left(1+M(s_0)\right)}\right) + D\sqrt{3L+2}\right\} + B(s_0)\\
&\le \frac{50}{\sqrt{n}}\left(T+\frac{1}{l_f}\right)(s_0+1)\left\{8u_f\left(\sqrt{\log\left(\frac{4}{\delta}\right)} + D\sqrt{(3L+2)\log\left(1+M(s_0)\right)}\ \right) + D\sqrt{3L+2}\right\}\\
&\quad + \left(T+\frac{1}{l_f}\right)(u_f+2)\frac{s_0+2}{(1-\exp(-c_N))^2}\frac{\delta}{2n}\\
&\le \frac{400}{\sqrt{n}}\left(T+\frac{1}{l_f}\right)(s_0+1)u_f\left(\sqrt{\log\left(\frac{4}{\delta}\right)} + D\sqrt{(3L+2)}(\sqrt{\log\left(1+M(s_0)\right)}+1) + \frac{1}{(1-\exp(-c_N))^2}\ \right)\ .
\end{aligned}$$

such that $\mathbb{P}\left(\sup_{\theta\in\Theta}|X_\theta| > t\right) \le \delta$. Hence the theorem is proved. ■

Appendix D. Proofs in Section 5

D.1 Proof of Theorem 5.1

Proof For $\lambda^*(t) = \lambda_0(t) \in W^{s,\infty}([0,T],B_0)$, $\delta = 1/2$, and $N \ge 5$, by Lemma G.4, there exists a two-layer NN $\hat{f}^N$ such that

$$\left|\hat{f}^N(x) - \lambda^*(Tx)\right| \le \frac{3\mathcal{C}B_0T^s}{2N^s},\ 0 \le x \le 1,$$

where $\mathcal{C} = \sqrt{2s}5^s/(s-1)!$.

Then we have

$$\left|\hat{f}^N(\frac{t}{T}) - \lambda^*(t)\right| \le \frac{3\mathcal{C}B_0T^s}{2N^s},\ 0 \le t \le T.$$

Since $B_1 \le \lambda^*(t) \le B_0$, taking $l_f = B_1$, $u_f = B_0$ and $\hat{\lambda}^N(t) = f(\hat{f}^N(t/T))$, we have

$$\left|\hat{\lambda}^N(t) - \lambda^*(t)\right| \le \left|\hat{f}^N(\frac{t}{T}) - \lambda^*(t)\right| \le \frac{3\mathcal{C}B_0T^s}{2N^s},\ \forall 0 \le t \le T.$$

Then

$$\begin{aligned}
&|\mathbb{E}[\mathrm{loss}(\hat{\lambda}^N, S_{test})] - \mathbb{E}[\mathrm{loss}(\lambda^*, S_{test})]| \\
&\leq \mathbb{E}\left|\mathrm{loss}(\hat{\lambda}^N, S_{test}) - \mathrm{loss}(\lambda^*, S_{test})\right| \\
&\leq \mathbb{E}\left(\Big|\sum_{j=1}^{N_e}(\log \hat{\lambda}^N(t_j) - \log \lambda^*(t_j))\Big| + \Big|\int_0^T \left(\hat{\lambda}^N(t) - \lambda^*(t)\right) \mathrm{dt}\Big|\right) \\
&\leq \mathbb{E}\left(T + \frac{N_e}{B_1}\right)\left\|\hat{\lambda}^N - \lambda^*\right\|_{L^\infty[0,T]} \\
&\leq \left(T + \frac{1}{B_1}\right)\frac{3\mathcal{C}B_0T^s}{2N^s}\mathbb{E}(N_e+1).
\end{aligned} \tag{44}$$

Since $\lambda_0 \leq B_0$, $\mu \equiv 0$, taking $c = B_0$, $c_0 = 0$, and $\eta = e$ in Lemma 2 , we have

$$\mathbb{P}(N_e \geq s) \leq 2\sqrt{B_0T}\exp\left(\frac{eB_0T - s}{2}\right).$$

Thus

$$\begin{aligned}
\mathbb{E}(N_e+1) &\leq 1 + \sum_{s=1}^{\infty}\mathbb{P}(N_e \geq s) \\
&\leq 1 + \frac{2\sqrt{B_0T}}{1-\exp(-1/2)}\exp\left(\frac{eB_0T-1}{2}\right) \leq 5\sqrt{B_0T+1}\exp\left(\frac{3B_0T}{2}\right).
\end{aligned} \tag{45}$$

Combining (44) and (45), we get

$$\begin{aligned}
|\mathbb{E}[\mathrm{loss}(\hat{\lambda}^N, S_{test})] - \mathbb{E}[\mathrm{loss}(\lambda^*, S_{test})]| &\leq 5\sqrt{B_0T+1}\exp\left(\frac{3B_0T}{2}\right)(T+\frac{1}{B_1})\frac{3\mathcal{C}B_0T^s}{2N^s} \\
&\leq 15\exp\left(2B_0T\right)(T+\frac{1}{B_1})\frac{\mathcal{C}B_0T^s}{N^s},
\end{aligned}$$

where $\mathcal{C} = \sqrt{2s}5^s/(s-1)!$.

$\hat{\lambda}^N$ can be naturally seen as an RNN by taking $W_h^l = 0$, $l = 1, 2$. The width and weights bound can be directly obtained by Lemma G.4 and Remark G.5. ■

D.2 Proof of Theorem 5.2

Proof The proof is divided into several steps. Let $S = \{t_j\}_{j=1}^{N_e}$. Here we agree on $t_0 = 0$, $t_{N_e+1} = T$. To be concise, we denote $S(t) = \sum_{t_j<t}\exp(-\beta(t-t_j)) + 1$, $S_j = \sum_{0<q<j}\exp(-\beta(t_j - t_q)) + 1$, $i \in \mathbb{N}_+$, hence $\lambda^*(t) = \lambda_0(t) + \alpha(S(t) - 1)$, $S_{j+1} = S_j\exp(-\beta(t_{j+1} - t_j)) + 1$, $S(t) = S_j\exp(-\beta(t - t_j)) + 1$ when $t \in (t_j, t_{j+1}]$, where we take $S_0 = 0$ by default.

We first fix $s_0 \in \mathbb{N}_+$.

Step 1. Construct the approximation of $g(x, y) = x\exp(-\beta y) + 1$, where $g \in C^\infty\left([-(s_0+1), 2(s_0+1)] \times [0, T]\right)$.

Let $\tilde{g}(x,y) = g((3x-1)(s_0+1), Ty)$, then $\tilde{g} \in C^\infty\left([0,1]^2\right)$. By simple computation, we have

$$\|\tilde{g}\|_{W^{k,\infty}([0,1]^2)} \le 3(s_0+1)(\beta T \vee 1)^k.$$

Applying Lemma G.6 to $\tilde{g}/[3(s_0+1)]$, for any $\mathcal{N} \in \mathbb{N}_+$, there exists a tanh neural network $\tilde{g}^{\mathcal{N}}$ with only one hidden layer and width $3\lceil\frac{\mathcal{N}+10(\beta T\vee 1)}{2}\rceil\binom{\mathcal{N}+10(\beta T\vee 1)+2}{2}$ such that

$$\left|\tilde{g}(x,y) - \tilde{g}^{\mathcal{N}}(x,y)\right| \le 3(s_0+1)\exp(-\mathcal{N}),\ (x,y) \in [0,1]^2.$$

By coordinate transformation, we get

$$\left|g(x,y) - \tilde{g}^{\mathcal{N}}(\frac{1}{3(s_0+1)}x + \frac{1}{3}, \frac{1}{T}y)\right| \le 3(s_0+1)\exp(-\mathcal{N}),\ (x,y) \in [-(s_0+1), 2(s_0+1)] \times [0,T].$$

Define $\hat{g}^{\mathcal{N}}(x,y) = \tilde{g}^{\mathcal{N}}(x/[3(s_0+1)] + 1/3, y/T)$. Then

$$\left|g(x,y) - \hat{g}^{\mathcal{N}}(x,y)\right| \le 3(s_0+1)\exp(-\mathcal{N}),\ (x,y) \in [-(s_0+1), 2(s_0+1)] \times [0,T].$$

From Lemma G.6 and Remark G.7, the weights of $\hat{g}^{\mathcal{N}}$ are bounded by

$$O\left((s_0+1)\exp(\frac{\mathcal{N}'^2 + \mathcal{N}' - 3Cd\mathcal{N}'}{2})(\mathcal{N}'(\mathcal{N}'+2))^{3\mathcal{N}'(\mathcal{N}'+2)}\right), \tag{46}$$

where $\mathcal{N}' = \mathcal{N} + 10(\beta T \vee 1)$. Taking $\mathcal{N} \leftarrow \mathcal{N} + \lceil\log(3(s_0+1))\rceil$, we have

$$\left|g(x,y) - \hat{g}^{\mathcal{N}}(x,y)\right| \le \exp(-\mathcal{N}),\ (x,y) \in [-(s_0+1), 2(s_0+1)] \times [0,T]. \tag{47}$$

Especially, $\left|g(x,y) - \hat{g}^{\mathcal{N}}(x,y)\right| \le 1$. Since $\hat{g}^{\mathcal{N}} \in \mathbb{R}$, by a small tuning (precisely, width plus 1), we can assume $\hat{g}^{\mathcal{N}}$ has the following structure:

$$\hat{g}^{\mathcal{N}}(x,y) = V_1\sigma\left(\begin{pmatrix} W & B\end{pmatrix}\begin{pmatrix} x \\ y\end{pmatrix} + b_0\right).$$

Step 2. Construct the approximation of S_j and $S(t)$ under the event $\{N_e \le s_0\}$.

Let $h_0 = 0$, $\overline{S}_0 = 0$, for $1 \le j \le s_0$. We construct $h_j^{\mathcal{N}}$ and $\overline{S}_j^{\mathcal{N}}$ recursively by

$$\begin{cases} h_j^{\mathcal{N}} = \sigma\left(\begin{pmatrix} W & B\end{pmatrix}\begin{pmatrix} V_1 h_{j-1}^{\mathcal{N}} \\ t_j - t_{j-1}\end{pmatrix} + b_0\right), \\ \overline{S}_j^{\mathcal{N}} = V_1 h_j^{\mathcal{N}}. \end{cases}$$

Hence $\overline{S}_j^{\mathcal{N}} = \hat{g}^{\mathcal{N}}(\overline{S}_{j-1}^{\mathcal{N}}, t_j - t_{j-1}), 1 \le j \le s_0$, here $t_0 = 0$.

Similarly, we can define $\overline{S}^{\mathcal{N}}(t), t \in (t_{j-1}, t_j]$ by

$$\begin{cases} h^{\mathcal{N}}(t) = \sigma\left(\begin{pmatrix} W & B\end{pmatrix}\begin{pmatrix} V_1 h_{j-1}^{\mathcal{N}} \\ t - t_{j-1}\end{pmatrix} + b_0\right), \\ \overline{S}^{\mathcal{N}}(t) = V_1 h^{\mathcal{N}}(t). \end{cases}$$

Hence $\overline{S}^{\mathcal{N}}(t) = \hat{g}^{\mathcal{N}}(\overline{S}^{\mathcal{N}}_{j-1}, t - t_{j-1}), t \in (t_{j-1}, t_j]$. The approximation error can be bounded by

$$\begin{aligned}
&|S(t) - \overline{S}^{\mathcal{N}}(t)| \\
=& \left|g(S_{j-1}, t_j - t_{j-1}) - \hat{g}^{N}(\overline{S}^{\mathcal{N}}_{j-1}, t_j - t_{j-1})\right| \\
\leq& \left|g(S_{j-1}, t_j - t_{j-1}) - g(\overline{S}^{\mathcal{N}}_{j-1}, t_j - t_{j-1})\right| + \left|g(\overline{S}^{\mathcal{N}}_{j-1}, t_j - t_{j-1}) - \hat{g}^{N}(\overline{S}^{\mathcal{N}}_{j-1}, t_j - t_{j-1})\right| \\
\leq& \left|S_{j-1} - \overline{S}^{\mathcal{N}}_{j-1}\right| + \left\|g - \hat{g}^{\mathcal{N}}\right\|_{\infty} \\
\leq& \cdots \\
\leq& j \left\|g - \hat{g}^{\mathcal{N}}\right\|_{\infty}, \ t \in (t_{j-1}, t_j].
\end{aligned}$$

Under the event $\{N_e \leq s_0\}$, we have

$$\left|S(t) - \overline{S}^{\mathcal{N}}(t)\right| \leq (s_0 + 1) \left\|g - \hat{g}^{\mathcal{N}}\right\|_{\infty}. \tag{48}$$

Step 3. Construct the approximation of identity.

By Lemma 3.1 of De Ryck et al. (2021), for any $h > 0$, there exists a one-layer tanh neural network ψ_h such that

$$|x - \psi_h(x)| \leq (6M)^4 h^2, \ x \in [-M, M]. \tag{49}$$

Actually, ψ_h can be represented as

$$\psi_h(x) = \frac{1}{\sigma'(0)h}\left[\sigma\left(\frac{hx}{2}\right) - \sigma\left(-\frac{hx}{2}\right)\right] = \frac{2}{\sigma'(0)h}\sigma\left(\frac{hx}{2}\right).$$

Step 4. Construct the approximation of $\lambda^*(t)$ under the event $\{N_e \leq s_0\}$.

Since $\lambda_0 \in W^{s,\infty}([0,T], B_0)$, from the proof of Theorem 4 , there exists a two-layer tanh neural network $\overline{\lambda}_0^N$ with width less than $3\lceil s/2 \rceil + 6N$ such that

$$\left|\overline{\lambda}_0^N(t) - \lambda_0(t)\right| \leq \frac{3\mathcal{C}T^s}{2N^s}, \ t \in [0, T]. \tag{50}$$

Moreover, the weights of $\overline{\lambda}_0^N$ can be bounded by

$$O\left(\left[\frac{\sqrt{2s}5^s}{(s-1)!}B_0 T^s\right]^{-s/2} N^{(1+s^2)/2}(s(s+2))^{3s(s+2)}\right).$$

Here we assume $\overline{\lambda}_0^N(t)$ have the following structure

$$\overline{\lambda}_0^N(t) = V_2'\sigma\left(V_1'\sigma\left(B't + b_0'\right) + b_1'\right) + b_2'.$$

Since $\lambda(t) = \lambda_0(t) + \alpha(S(t) - 1)$, we can construct its approximation by

$$h_j^{(1)} \quad = \quad \sigma\left(\begin{pmatrix} WV_1 & 0 \\ 0 & 0 \end{pmatrix} h_{j-1} + \begin{pmatrix} B & 0 \\ 0 & B' \end{pmatrix}\begin{pmatrix} t_j - t_{j-1} \\ t_j \end{pmatrix} + \begin{pmatrix} b_0 \\ b_0' \end{pmatrix}\right), \ 1 \leq j \leq s_0,$$

and

$$\begin{aligned} h^{(1)}(t;S) &= \sigma\left(\begin{pmatrix} WV_1 & 0 \\ 0 & 0 \end{pmatrix} h_j^{(1)} + \begin{pmatrix} B & 0 \\ 0 & B^{'} \end{pmatrix}\begin{pmatrix} t-t_j \\ t \end{pmatrix} + \begin{pmatrix} b_0 \\ b_0^{'} \end{pmatrix}\right), \\ h^{(2)}(t;S) &= \sigma\left(\begin{pmatrix} \frac{h}{2}V_1 & 0 \\ 0 & V_1^{'} \end{pmatrix} h^{(1)}(t;S) + \begin{pmatrix} 0 \\ b_1^{'} \end{pmatrix}\right), \\ \hat{\lambda}(t;S) &= f\left(\begin{pmatrix} \frac{2\alpha}{\sigma^{'}(0)h} & V_2^{'} \end{pmatrix} h^{(2)}(t;S) + \begin{pmatrix} b_2^{'}-\alpha \end{pmatrix}\right) \in \mathbb{R}^1, \ t \in (t_j, t_{j+1}]. \end{aligned} \tag{51}$$

Under the event $\{N_e \leq s_0\}$, we have $B_1 \leq \lambda(t) \leq B_0 + \alpha s_0$. Recall that $f(x) = \min\{\max\{x, l_f\}, u_f\}$. Here we can take $l_f = B_1$, $u_f = B_0 + \alpha s_0$.

Step 5. Estimate the approximation error under the event $\{N_e \leq s_0\}$.

We rewrite (51) as $\hat{\lambda}(t;S) = f(\overline{\lambda}(t;S))$. Under the event $\{N_e \leq s_0\}$ and the construction of f, we have

$$\|\lambda^* - \hat{\lambda}\|_{L^\infty} \leq \|\lambda^* - \overline{\lambda}\|_{L^\infty} \ .$$

From the constuction of $\overline{\lambda}$, we get

$$\overline{\lambda}(t) = \overline{\lambda}_0^N(t) + \alpha\psi_h(\overline{S}^{\mathcal{N}}(t)) - \alpha \ , \tag{52}$$

then

$$\left|\lambda^*(t) - \overline{\lambda}(t)\right| \leq \left|\lambda_0(t) - \overline{\lambda}_0^N(t)\right| + \alpha\left|S(t) - \psi_h(\overline{S}^{\mathcal{N}}(t))\right| \ . \tag{53}$$

From (48) (49) and (47), under the event $\{N_e \leq s_0\}$, we have

$$\begin{aligned} \left|S(t) - \psi_h(\overline{S}^{\mathcal{N}}(t))\right| &\leq \left|S(t) - \overline{S}^{\mathcal{N}}(t)\right| + \left|\overline{S}^{\mathcal{N}}(t) - \psi_h(\overline{S}^{\mathcal{N}}(t))\right| \\ &\leq (s_0+1)\left\|g - \hat{g}^{\mathcal{N}}\right\|_\infty + (6M)^4h^2 \\ &\leq (s_0+1)\exp(-\mathcal{N}) + (12(s_0+1))^4h^2, \ t \in [0,T], \end{aligned}$$

where we take $M = 2(s_0+1)$ to ensure that $\overline{S}^{\mathcal{N}}(t)$ can be well approximated by $\psi_h(\overline{S}^{\mathcal{N}}(t))$. On the other hand, (50) shows that

$$\left|\overline{\lambda}_0^N(t) - \lambda_0(t)\right| \leq \frac{3\mathcal{C}B_0T^s}{2N^s}, \ t \in [0,T].$$

To trade off the two error terms in (53), let $\exp(-\mathcal{N}) \asymp N^{-s}$, and then we can take $\mathcal{N} = \lceil s\log(N)\rceil$. Moreover, take $\mathcal{N} \leftarrow \mathcal{N} + \lceil\log(s_0+1)\rceil$ and $h = (12(s_0+1))^{-2}N^{-s/2}$. Hence, under $\{N_e \leq s_0\}$, we have

$$\left|\lambda^*(t) - \hat{\lambda}(t)\right| \leq \left|\lambda^*(t) - \overline{\lambda}(t)\right| \leq \frac{3\mathcal{C}B_0T^s + 4}{2N^s}, \ t \in [0,T]. \tag{54}$$

Step 6. Estimate the final approximation error.

Similar to (44), we have

$$\begin{aligned}
&|\mathbb{E}[\mathrm{loss}(\hat{\lambda}, S_{test})] - \mathbb{E}[\mathrm{loss}(\lambda^*, S_{test})]| \\
&\leq \mathbb{E}\left|\mathrm{loss}(\hat{\lambda}, S_{test}) - \mathrm{loss}(\lambda^*, S_{test})\right| \\
&\leq \mathbb{E}\left(\Big|\sum_{j=1}^{N_e}(\log\hat{\lambda}(t_j) - \log\lambda^*(t_j))\Big| + \Big|\int_0^T \left(\hat{\lambda}(t) - \lambda^*(t)\right)\mathrm{dt}\Big|\right) \\
&\leq \mathbb{E}\left[\left(\Big|\sum_{j=1}^{N_e}(\log\hat{\lambda}(t_j) - \log\lambda^*(t_j))\Big| + \Big|\int_0^T \left(\hat{\lambda}(t) - \lambda^*(t)\right)\mathrm{dt}\Big|\right)\mathbb{1}_{\{N_e\leq s_0\}}\right. \\
&\qquad\left. + \left(\Big|\sum_{j=1}^{N_e}(\log\hat{\lambda}(t_j) - \log\lambda^*(t_j))\Big| + \Big|\int_0^T \left(\hat{\lambda}(t) - \lambda^*(t)\right)\mathrm{dt}\Big|\right)\mathbb{1}_{\{N_e> s_0\}}\right] \\
&\leq \mathbb{E}\left[(T + \frac{N_e}{B_1})\left\|\hat{\lambda} - \lambda^*\right\|_{L^\infty}\mathbb{1}_{\{N_e\leq s_0\}}\right] + \mathbb{E}\left[(T + \frac{N_e}{B_1})\left\|\hat{\lambda} - \lambda^*\right\|_{L^\infty}\mathbb{1}_{\{N_e> s_0\}}\right] \\
&:= \mathbb{I}_1 + \mathbb{I}_2\ . \qquad (55)
\end{aligned}$$

Since $\lambda_0(t) \leq B_0$, $\mu(t) = \alpha\exp(-\beta(t))$, taking $c_\mu = \alpha/\beta$, $\eta = (\alpha+\beta)/(2\alpha)$ in Lemma 2 , we have

$$\begin{aligned}
\mathbb{P}\left(N_e \geq s\right) &\leq \frac{2\sqrt{B_0T}}{1-c_\mu}\exp\left(\frac{\log(\eta)}{2}\left[\eta(B_0T) - (1-c_\mu\eta)s\right]\right) \\
&\leq \frac{2\beta\sqrt{B_0T}}{\beta-\alpha}\exp\left(\frac{\log\left(\frac{\alpha+\beta}{2\alpha}\right)}{2}\left[\frac{\alpha+\beta}{2\alpha}(B_0T) - \frac{\beta-\alpha}{2\beta}s\right]\right) \\
&:= a_e\exp\left(-c_e s\right).
\end{aligned}$$

By (54),

$$\begin{aligned}
\mathbb{I}_1 &\leq \left(T + \frac{1}{B_1}\right)\frac{3\mathcal{C}B_0T^s + 2}{2N^s}\mathbb{E}\left[(N_e+1)\mathbb{1}_{\{N_e\leq s_0\}}\right] \\
&\leq \left(T + \frac{1}{B_1}\right)\frac{3\mathcal{C}B_0T^s + 2}{2N^s}\mathbb{E}\left[(N_e+1)\right] \\
&= \left(T + \frac{1}{B_1}\right)\frac{3\mathcal{C}B_0T^s + 2}{2N^s}\left(1 + \sum_{s=1}^{\infty}\mathbb{P}(N_e \geq s)\right) \\
&\leq \left(T + \frac{1}{B_1}\right)\left(1 + \frac{a_e\exp(-c_e)}{1-\exp(-c_e)}\right)\frac{3\mathcal{C}B_0T^s + 4}{2N^s} \qquad (56)
\end{aligned}$$

On the other hand, from $\|\hat{\lambda}\|_{L^\infty} \leq B_0 + \alpha s_0$ and $\|\lambda^*\|_{L^\infty} \leq B_0 + \alpha N_e$, we have

$$
\begin{aligned}
\mathbb{I}_2 &\leq \mathbb{E}\left[\left(T + \frac{N_e}{B_1}\right)\|\hat{\lambda}\|_{L^\infty}\mathbb{1}_{\{N_e > s_0\}}\right] + \mathbb{E}\left[\left(T + \frac{N_e}{B_1}\right)\|\lambda^*\|_{L^\infty}\mathbb{1}_{\{N_e > s_0\}}\right] \\
&\leq \left(T + \frac{1}{B_1}\right)(B_0 + \alpha s_0)\mathbb{E}\left[(N_e + 1)\mathbb{1}_{\{N_e > s_0\}}\right] + \left(T + \frac{1}{B_1}\right)\mathbb{E}\left[(N_e + 1)(B_0 + \alpha N_e)\mathbb{1}_{\{N_e > s_0\}}\right] \\
&\leq \left(T + \frac{1}{B_1}\right)(B_0 + \alpha s_0)\left((s_0 + 1)\mathbb{P}(N_e \geq s_0 + 1) + \sum_{s = s_0 + 1}^{\infty}\mathbb{P}(N_e \geq s)\right) \\
&\quad + \left(T + \frac{1}{B_1}\right)\left((s_0 + 1)(B_0 + \alpha s_0)\mathbb{P}(N_e \geq s_0 + 1) + \sum_{s = s_0 + 1}^{\infty}(2\alpha s + B_0)\mathbb{P}(N_e \geq s)\right) \\
&\leq \left(T + \frac{1}{B_1}\right)(B_0 + \alpha s_0)a_e\exp(-c_e(s_0 + 1))\left((s_0 + 1) + \frac{1}{1 - \exp(-c_e)}\right) \\
&\quad + \left(T + \frac{1}{B_1}\right)a_e\exp(-c_e(s_0 + 1))\left((s_0 + 1)(B_0 + \alpha s_0) + \frac{2\alpha(s_0 + 1) + B_0}{(1 - \exp(-c_e))^2}\right) \\
&\leq \left(T + \frac{1}{B_1}\right)a_e\exp(-c_e(s_0 + 1))\left(2(s_0 + 1)(B_0 + \alpha s_0) + \frac{3\alpha(s_0 + 1) + 2B_0}{(1 - \exp(-c_e))^2}\right). \qquad (57)
\end{aligned}
$$

Combing (55) (56) (57), we have

$$
\begin{aligned}
&|\mathbb{E}[\mathrm{loss}(\hat{\lambda}, S_{test})] - \mathbb{E}[\mathrm{loss}(\lambda^*, S_{test})]| \\
\leq &\left(T + \frac{1}{B_1}\right)a_e\exp(-c_e(s_0 + 1))\left(2(s_0 + 1)(B_0 + \alpha s_0) + \frac{3\alpha(s_0 + 1) + 2B_0}{(1 - \exp(-c_e))^2}\right) \\
&+ \left(T + \frac{1}{B_1}\right)\left(1 + \frac{a_e\exp(-c_e)}{1 - \exp(-c_e)}\right)\frac{3\mathcal{C}B_0T^s + 4}{2N^s}.
\end{aligned}
$$

Let $s_0 = \lceil s\log(N)/c_e\rceil$, and denote $\hat{\lambda}^N = \hat{\lambda}$. We have

$$
|\mathbb{E}[\mathrm{loss}(\hat{\lambda}^N, S_{test})] - \mathbb{E}[\mathrm{loss}(\lambda^*, S_{test})]| \lesssim \frac{(\log N)^2}{N^s} \ .
$$

Step 7. Bound the sizes of the network width and weights.

From step 1-6, the width of the network is less than

$$
3\left\lceil\frac{\tilde{\mathcal{N}}}{2}\right\rceil\binom{\tilde{\mathcal{N}} + 2}{2} + 3\left\lceil\frac{s}{2}\right\rceil + 6N + 2 \ ,
$$

where $\tilde{\mathcal{N}} = \mathcal{N} + 10(\beta T \vee 1) + 2\lceil\log(3(s_0 + 1))\rceil$. Since $s_0 = \lceil\frac{s}{c_e}\log(N)\rceil$ and $\mathcal{N} = \lceil s\log(N)\rceil$, we have $D \leq O(N)$.

From the construction of $\hat{g}^{\mathcal{N}}$, ψ_h and $\overline{\lambda}_0^N$, the weights of the network is less than

$$
\begin{aligned}
O\Bigg(\max\Bigg\{&\frac{2}{\sigma'(0)h}\exp(\frac{\tilde{\mathcal{N}}^2 + \tilde{\mathcal{N}} - 3Cd\tilde{\mathcal{N}}}{2})(\tilde{\mathcal{N}}(\tilde{\mathcal{N}} + 2))^{3\tilde{\mathcal{N}}(\tilde{\mathcal{N}} + 2)}, \\
&\left[\frac{\sqrt{2s}5^s}{(s - 1)!}B_0T^s\right]^{-s/2}N^{(1 + s^2)/2}(s(s + 2))^{3s(s + 2)}\Bigg\}\Bigg),
\end{aligned}
$$

where $\tilde{\mathcal{N}} = \mathcal{N} + 10(\beta T \vee 1) + 2\lceil \log(3(s_0+1)) \rceil$. Since $s_0 = \lceil \frac{s}{c_e} \log(N) \rceil$, $h = (12(s_0+1))^{-2} N^{-s/2}$, $\mathcal{N} = \lceil s \log(N) \rceil$, the weights are less than

$$\mathcal{C}_1 (\log(N))^{12 s^2 (\log(N))^2} ,$$

where $\mathcal{C}_1$ is a constant related to s, B_0, α, β, and T. ■

D.3 Proof of Theorem 5.3

Lemma D.1. *Suppose $\mu \in C^{k,\infty}([0,T], C_0)$, $k \geq 2$, $k \in \mathbb{N}$. The fourier series of μ is given by*

$$S_\infty(t) = \frac{\hat{\mu}_0}{2} + \sum_{l=1}^{\infty} \left(\hat{\mu}_l \cos\left(\frac{2l\pi}{T} t \right) + \hat{\nu}_l \sin\left(\frac{2l\pi}{T} t \right) \right), \tag{58}$$

where $\hat{\mu}_l = 2\int_0^T \mu(t)\cos(2l\pi t/T)\mathrm{d}t/T$, $\hat{\nu}_l = 2\int_0^T \mu(t)\sin(2l\pi t/T)\mathrm{d}t/T$, $l \geq 0$. If $\mu^{(j)}(0+) = \mu^{(j)}(T-)$, $0 \leq j \leq k-1$, then

$$|\hat{\mu}_l| \leq \frac{2C_0 T^k}{(2l\pi)^k}, |\hat{\nu}_l| \leq \frac{2C_0 T^k}{(2l\pi)^k}$$

and $S_\infty(t) = \mu(t)$ on $t \in [0,T]$. Moveover, denote the partial sum of $S_\infty(t)$ as $S_{N_\mu}(t) = \hat{\mu}_0/2 + \sum_{l=1}^{N_\mu} (\hat{\mu}_l \cos(2l\pi t/T) + \hat{\nu}_l \sin(2l\pi t/T))$,

$$\left| \mu(t) - S_{N_\mu}(t) \right| \leq \frac{2C_0 T^{k+1}}{(k-1)(2\pi)^k N_\mu^{k-1}}, \ t \in [0,T].$$

Proof [Proof of Lemma D.1] The proof is a standard Fourier analysis exercise and we omit it. ■

Theorem D.2. *Under model assumptions (A1)-(A3) and $\mu^{(j)}(0+) = \mu^{(j)}(T-)$, $0 \leq j \leq k-1$, for $N \geq 5$, there exists an RNN structure $\hat{\lambda}^{N,N_\mu}$ as stated in section 2.2 with $L = 2$, $l_f = B_1$, $u_f = B_0 + O(\log N)$, and input function $x(t;S) = (t, t - F_S(t))^\top$ such that*

$$|\mathbb{E}[loss(\hat{\lambda}^{N,N_\mu}, S_{test})] - \mathbb{E}[loss(\lambda^*, S_{test})]| \lesssim \frac{1}{1-c_\mu} \exp\left(\frac{2B_0 T}{c_\mu^2} \right) \left(\frac{T^s + \log^2 N}{N^s} + \frac{T^k \log N}{N_\mu^{k-1}} \right) .$$

Moreover, the width of $\tilde{\lambda}^N$ satisfies $D \lesssim N + N_\mu^5 \log^4 N$ and the weights of $\hat{\lambda}^N$ are less than

$$\mathcal{C}_1 (\log(N N_\mu))^{12 s^2 (\log(N N_\mu))^2} ,$$

where $\mathcal{C}_1$ is a constant related to s, B_0, C_0, c_μ, and T.

Proof [Proof of Theorem D.2] Similar to the proof of Theorem 5 , the proof is divided into several steps. Denote $w_l = 2l\pi/T$, $g_{l,1}(x,t) = x_1 \cos w_l t + x_2 \sin w_l t$, $g_{l,2}(x,t) = -x_1 \sin w_l t + x_2 \cos w_l t + 1$, $g_l(x,t) = (g_{l,1}(x,t), g_{l,2}(x,t))^\top \in \mathbb{R}^2$, where $x \in \mathbb{R}^2$, $l \in \mathbb{N}_+$. For $l \in \mathbb{N}_+$, define

$$S_l(t) = \sum_{t_j < t} \begin{pmatrix} \sin w_l(t-t_j) \\ \cos w_l(t-t_j) \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

and

$$S_{l,j} = \sum_{0<q<j} \begin{pmatrix} \sin w_l(t_j-t_q) \\ \cos w_l(t_j-t_q) \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Hence we have

$$S_{l,j+1} = \begin{pmatrix} \cos w_l(t_{j+1}-t_j) & \sin w_l(t_{j+1}-t_j) \\ -\sin w_l(t_{j+1}-t_j) & \cos w_l(t_{j+1}-t_j) \end{pmatrix} S_{l,j} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} = g_l(S_{l,j}, t_{j+1}-t_j)$$

and

$$S_l(t) = g_l(S_{l,j}, t-t_j),\ t \in (t_j, t_{j+1}].$$

where we agree on $S_{l,0} = \mathbf{0}$. Define $S_0(t) = \#\{j : t_j < t\}$. If we assume $t_1 > 0$, the true intensity can be rewritten as

$$\lambda^*(t) = \lambda_0(t) + \frac{\hat{\mu}_0}{2}(S_0(t)-1) + \sum_{l=1}^{\infty} (\hat{\nu}_l, \hat{\mu}_l) \cdot \left(S_l(t) - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right),\ t \in [0,T], \tag{59}$$

where $a \cdot b$ refers to the standard inner product of vectors a and b.

We first fix $s_0 \in \mathbb{N}_+$. Since $\mathbb{P}(t_1 = 0) = 0$. We assume $t_1 > 0$ so that (59) holds.

Step 1. Construct the approximation of $g_l(x,t) = (g_{l,1}(x,t), g_{l,2}(x,t))^\top \in \mathbb{R}^2$, where $g_{l,1}, g_{l,2} \in C^\infty\left([-3(s_0+1), 3(s_0+1)]^2 \times [0,T]\right)$. Here $x \in \mathbb{R}^2$.

Let $\tilde{g}_{l,i}(x,t) = g_{l,i}(3(s_0+1)(2x - \mathbf{1}_2), Tt)$, $i = 1,2$. Then $\tilde{g}_{l,i} \in C^\infty\left([0,1]^3\right)$. By simple computation, we have

$$\|\tilde{g}_{l,i}\|_{W^{k,\infty}([0,1]^3)} \le 6(s_0+1)(w_l T)^k.$$

Applying Lemma G.6 to $\tilde{g}_{l,i}/[6(s_0+1)]$, for any $\mathcal{N} \in \mathbb{N}_+$, there exists a tanh neural network $\tilde{g}_{l,i}^{\mathcal{N}}$ with only one hidden layer and width $3\lceil(\mathcal{N} + 15w_nT)/2\rceil\binom{\mathcal{N}+15w_nT+3}{3}$ such that

$$\left|\tilde{g}_{l,i}(x,t) - \tilde{g}_{l,i}^{\mathcal{N}}(x,t)\right| \le 6(s_0+1)\exp(-\mathcal{N}),\ (x,t) \in [0,1]^3.$$

By coordinate transformation, we get

$$\left| g_{l,i}(x,t) - \tilde{g}_{l,i}^{\mathcal{N}}\left(\frac{x}{6(s_0+1)} + \frac{1}{2}\mathbf{1}_2, \frac{t}{T} \right) \right| \le 6(s_0+1)\exp(-\mathcal{N}),\ (x,t) \in [-3(s_0+1), 3(s_0+1)]^2 \times [0,T].$$

Define $\hat{g}_{l,i}^{\mathcal{N}}(x,t) = \tilde{g}_{l,i}^{\mathcal{N}}(x/[6(s_0+1)] + 1/2\mathbf{1}_2, t/T)$, then

$$\left|g_{l,i}(x,t) - \hat{g}_{l,i}^{\mathcal{N}}(x,t)\right| \le 6(s_0+1)\exp(-\mathcal{N}),\ (x,y) \in [-3(s_0+1), 3(s_0+1)]^2 \times [0,T].$$

Taking $\mathcal{N} \leftarrow \mathcal{N} + \lceil \log(6(s_0+1)) \rceil$, we have

$$\left| g_{l,i}(x,t) - \hat{g}_{l,i}^{\mathcal{N}}(x,t) \right| \leq \exp(-\mathcal{N}),\ (x,y) \in [-3(s_0+1), 3(s_0+1)]^2 \times [0,T].$$

Especially, $\left| g_{l,i}(x,t) - \hat{g}_{l,i}^{\mathcal{N}}(x,t) \right| \leq 1$. The width of this NN is bounded by $3\lceil u/2 \rceil \binom{u+3}{3}$. From Lemma G.6 and Remark G.7, the weights of $\hat{g}_{l,i}^{\mathcal{N}}$ are bounded by

$$O\left((s_0+1) \exp(\frac{\mathcal{N}'^2 + \mathcal{N}' - 3Cd\mathcal{N}'}{2}) (\mathcal{N}'(\mathcal{N}'+2))^{3\mathcal{N}'(\mathcal{N}'+2)} \right),$$

where $\mathcal{N}' = \mathcal{N} + \lceil \log(6(s_0+1)) \rceil + 15 w_l T$. Since $\hat{g}_{l,i}^{\mathcal{N}} \in \mathbb{R}$, by a small tuning(precisely, let width plus 1), we can assume $\hat{g}_{l,i}^{\mathcal{N}}$ has the following structure:

$$\hat{g}_{l,i}^{\mathcal{N}}(x,y) = V_{l,i} \sigma \left(\begin{pmatrix} W_{l,i} & B_{l,i} \end{pmatrix} \begin{pmatrix} x_{l,i} \\ t_{l,i} \end{pmatrix} + b_{l,i} \right).$$

Denote $\hat{g}_l^{\mathcal{N}}(x,t) = \left(\hat{g}_{l,1}^{\mathcal{N}}(x,t), \hat{g}_{l,2}^{\mathcal{N}}(x,t) \right)^\top$.

Step 1′. Construct the approximation of identity and $g_0(x) = x+1$, $x \in [-(s_0+1), 2(s_0+1)]$. Here $x \in \mathbb{R}$.

Similarly to step 3 in the proof of Theorem 5 , taking $\psi_h(x) = 2\sigma\left(hy/2\right)/[\sigma'(0)h]$, we have

$$|x - \psi_h(x)| \leq (6M)^4 h^2,\ x \in [-M, M].$$

For $g_0(x) = x+1$, $x \in [-(s_0+1), 2(s_0+1)]$, we can construct a similar approximation as the proof of of Theorem 5. There exists a tanh neural network $\hat{g}_0^{\mathcal{N}}$ with only one hidden layer and width $3\lceil (\mathcal{N}''+5)/2 \rceil$ such that

$$\left| g_0(x) - \hat{g}_0^{\mathcal{N}}(x) \right| \leq \exp(-\mathcal{N}),\ x \in [-(s_0+1), 2(s_0+1)],$$

where $\mathcal{N}'' = \mathcal{N} + \lceil (s_0+3)\log 2 \rceil$. The weight of $\hat{g}_0^{\mathcal{N}}$ is bounded by

$$O\left((s_0+1) \exp\left(\frac{\mathcal{N}''^2 + \mathcal{N}'' - 3Cd\mathcal{N}''}{2} \right) \left[\mathcal{N}''(\mathcal{N}''+2) \right]^{3\mathcal{N}''(\mathcal{N}''+2)} \right).$$

Step 2. Construct the approximation of $S_{n,i}$ and $S_n(t)$ under the event $\{N_e \leq s_0\}$.

Let $h_{l,0}^{\mathcal{N}} = (h_{l,0,1}^{\mathcal{N}}, h_{l,0,2}^{\mathcal{N}})^\top = \mathbf{0}$, $\overline{S}_{l,0}^{\mathcal{N}} = (\overline{S}_{l,0,1}^{\mathcal{N}}, \overline{S}_{l,0,2}^{\mathcal{N}})^\top = \mathbf{0}$. For $1 \leq j \leq s_0$, we construct $h_{l,j}^{\mathcal{N}}$ and $\overline{S}_{l,j}^{\mathcal{N}}$ recursively by

$$\begin{cases} h_{l,j}^{\mathcal{N}} = \sigma\left(\begin{pmatrix} W_{l,1} \begin{pmatrix} V_{l,1} & 0 \\ 0 & V_{l,2} \end{pmatrix} \\ W_{l,2} \begin{pmatrix} V_{l,1} & 0 \\ 0 & V_{l,2} \end{pmatrix} \end{pmatrix} h_{l,j-1}^{\mathcal{N}} + \begin{pmatrix} B_{l,1} \\ B_{l,2} \end{pmatrix} (t_j - t_{j-1}) + \begin{pmatrix} b_{l,1} \\ b_{l,2} \end{pmatrix} \right), \\ \overline{S}_{l,j}^{\mathcal{N}} = \begin{pmatrix} V_{l,1} & 0 \\ 0 & V_{l,2} \end{pmatrix} h_{l,j}^{\mathcal{N}}. \end{cases}$$

Hence $\overline{S}_{l,j}^{\mathcal{N}} = \hat{g}_l^{\mathcal{N}}(\overline{S}_{l,j-1}^{\mathcal{N}}, t_j - t_{j-1}), 1 \leq j \leq s_0$. Here we agree on $t_0 = 0$.

Similarly, we can define $\overline{S}_l^{\mathcal{N}}(t), t \in (t_{j-1}, t_j]$ by

$$\begin{cases} h_l^{\mathcal{N}}(t) = \sigma\left(\begin{pmatrix} W_{l,1}\begin{pmatrix} V_{l,1} & 0\\ 0 & V_{l,2}\end{pmatrix}\\ W_{l,2}\begin{pmatrix} V_{l,1} & 0\\ 0 & V_{l,2}\end{pmatrix}\end{pmatrix} h_{l,j-1}^{\mathcal{N}} + \begin{pmatrix} B_{l,1}\\ B_{l,2}\end{pmatrix}(t_j - t_{j-1}) + \begin{pmatrix} b_{l,1}\\ b_{l,2}\end{pmatrix}\right), \\ \overline{S}_l^{\mathcal{N}}(t) = \begin{pmatrix} V_{l,1} & 0\\ 0 & V_{l,2}\end{pmatrix} h_l^{\mathcal{N}}(t). \end{cases}$$

Hence $\overline{S}_l^{\mathcal{N}}(t) = \hat{g}_l^{\mathcal{N}}(\overline{S}_{l,j-1}^{\mathcal{N}}, t - t_{j-1}), t \in (t_{j-1}, t_j]$. The approximation error can be bounded by

$$\begin{aligned} & \left\| S_l(t) - \overline{S}_l^{\mathcal{N}}(t) \right\|_2 \\ = & \left\| g_l(S_{l,i-1}, t - t_{j-1}) - \hat{g}_l^{\mathcal{N}}(\overline{S}_{l,j-1}^{\mathcal{N}}, t - t_{j-1}) \right\|_2 \\ \leq & \left\| g_l(S_{l,j-1}, t - t_{j-1}) - g_l(\overline{S}_{l,j-1}^{\mathcal{N}}, t - t_{j-1}) \right\|_2 + \left\| g_l(\overline{S}_{l,j-1}^{\mathcal{N}}, t - t_{j-1}) - \hat{g}_l^{\mathcal{N}}(\overline{S}_{l,j-1}^{\mathcal{N}}, t - t_{j-1}) \right\|_2 \\ \leq & \left\| S_{l,j-1} - \overline{S}_{l,j-1}^{\mathcal{N}} \right\|_2 + \sqrt{2} \max\left\{ \left\| g_{l,1} - \hat{g}_{l,1}^{\mathcal{N}} \right\|_\infty \vee \left\| g_{l,2} - \hat{g}_{l,2}^{\mathcal{N}} \right\|_\infty \right\} \\ \leq & \left\| S_{l,j-1} - \overline{S}_{l,j-1}^{\mathcal{N}} \right\|_2 + \sqrt{2} \exp(-\mathcal{N}) \qquad (60) \\ \leq & \cdots \\ \leq & \sqrt{2} j \exp(-\mathcal{N}), \ t \in (t_{j-1}, t_j]. \end{aligned}$$

Under the event $\{N_e \leq s_0\}$, we have

$$\left\| S_l(t) - \overline{S}_l^{\mathcal{N}}(t) \right\|_2 \leq \sqrt{2}(s_0 + 1) \exp(-\mathcal{N}).$$

Moreover, $\left\| \overline{S}_{l,j}^{\mathcal{N}} \right\|_2 \leq \left\| S_{l,j} - \overline{S}_{l,j}^{\mathcal{N}} \right\|_2 + \| S_{l,j} \|_2 \leq \sqrt{2}(s_0 + 1) + (s_0 + 1) \leq 3(s_0 + 1), \ j \leq s_0$, then $\overline{S}_{l,j}^{\mathcal{N}} \in [-3(s_0 + 1), 3(s_0 + 1)]^2$ and (60) can be verified by induction under the event $\{N_e \leq s_0\}$.

For the approximation of $S_0(t)$, we can similarly construct a simple RNN such that $\overline{S}_{0,j}^{\mathcal{N}} = \hat{g}_0^{\mathcal{N}}(\overline{S}_{0,j-1}^{\mathcal{N}})$ and $\left| S_0(t) - \overline{S}_0^{\mathcal{N}}(t) \right| \leq (s_0 + 1) \exp(-\mathcal{N})$.

Step 3. Construct the approximation of $\lambda^*(t)$ under the event $\{N_e \leq s_0\}$.

Since $\lambda_0 \in W^{s,\infty}([0, T], B_0)$, from the proof of Theorem 4 , there exists a two layer tanh neural network $\overline{\lambda}_0^N$ with width less than $3\lceil s/2 \rceil + 6N$ such that

$$\left| \overline{\lambda}_0^N(t) - \lambda_0(t) \right| \leq \frac{3\mathcal{C} B_0 T^s}{2N^s}, \ \forall 0 \leq t \leq T. \qquad (61)$$

Moreover, the weights of $\overline{\lambda}_0^N$ can be bounded by

$$O\left(\left[\frac{\sqrt{2} s 5^s}{(s-1)!} B_0 T^s \right]^{-\frac{s}{2}} N^{(1+s^2)/2} (s(s+2))^{3s(s+2)} \right) .$$

Here we assume $\overline{\lambda}_0^N(t)$ have the following structure

$$\overline{\lambda}_0^N(t) = V_2^{'}\sigma\left(V_1^{'}\sigma\left(B^{'}t + b_0^{'}\right) + b_1^{'}\right) + b_2^{'} \ .$$

Since $\lambda^*(t) = \lambda_0(t) + \hat{\mu}_0(S_0(t)-1)/2 + \sum_{l=1}^{\infty}(\hat{\nu}_l, \hat{\mu}_l)\cdot\left(S_l(t) - \begin{pmatrix}0\\1\end{pmatrix}\right)$, we can construct its (finite sum) approximation by

$$\overline{\lambda}(t) = \overline{\lambda}_0^N(t) + \frac{\hat{\mu}_0}{2}(\psi_h(\overline{S_0}^{\mathcal{N}}(t)) - 1) + \sum_{l=1}^{N_\mu}(\hat{\nu}_l, \hat{\mu}_l)\cdot\left(\psi_h(\overline{S_l}^{\mathcal{N}}(t)) - \begin{pmatrix}0\\1\end{pmatrix}\right) \ .$$

It can be seen as a parallelism of $(N_\mu + 2)$ RNNs defined before.

Under the event $\{N_e \le s_0\}$, we have $B_1 \le \lambda(t) \le B_0 + C_0 s_0$. Recall that $f(x) = \min\{\max\{x, l_f\}, u_f\}$. Here we can take $l_f = B_1$, $u_f = B_0 + C_0 s_0$. The final output is $\hat{\lambda}(t;S) = f(\overline{\lambda}(t;S))$.

Step 4. Compute the approximation error under the event $\{N_e \le s_0\}$.

Under the event $\{N_e \le s_0\}$ and the construction of f, we have

$$\|\lambda^* - \hat{\lambda}\|_{L^\infty} \le \|\lambda^* - \overline{\lambda}\|_{L^\infty} \ . \tag{62}$$

By the construction of $\overline{\lambda}$,

$$\begin{aligned}\left|\lambda^*(t) - \overline{\lambda}(t)\right| \le &\left|\lambda_0(t) - \overline{\lambda}_0^N(t)\right| + \frac{\hat{\mu}_0}{2}\left|S_0(t) - \psi_h(\overline{S}_0^{\mathcal{N}}(t))\right| + \left|\sum_{l=1}^{N_\mu}(\hat{\nu}_l, \hat{\mu}_l)\cdot\left(S_l(t) - \psi_h(\overline{S_l}^{\mathcal{N}}(t))\right)\right| \\ &+ \left|\sum_{l>N_\mu}(\hat{\nu}_l, \hat{\mu}_l)\cdot\left(S_l(t) - \begin{pmatrix}0\\1\end{pmatrix}\right)\right| \ .\end{aligned} \tag{63}$$

Under the event $\{N_e \le s_0\}$, for the second term, we have

$$\begin{aligned}\left|S_0(t) - \psi_h(\overline{S}_0^{\mathcal{N}}(t))\right| &\le \left|S_0(t) - \overline{S}_0^{\mathcal{N}}(t)\right| + \left|\overline{S}_0^{\mathcal{N}}(t) - \psi_h(\overline{S}_0^{\mathcal{N}}(t))\right| \\ &\le (s_0+1)\left\|g_0 - \hat{g}_0^{\mathcal{N}}\right\|_\infty + (6M)^4h^2 \\ &\le (s_0+1)\exp(-\mathcal{N}) + (18(s_0+1))^4h^2, \ 0 \le t \le T.\end{aligned} \tag{64}$$

For the third term, similarly,

$$\begin{aligned}&\left|\sum_{l=1}^{N_\mu}(\hat{\nu}_l, \hat{\mu}_l)\cdot\left(S_l(t) - \psi_h(\overline{S_l}^{\mathcal{N}}(t))\right)\right| \\ &\le \sum_{l=1}^{N_\mu}\left\|(\hat{\nu}_l, \hat{\mu}_l)^\top\right\|_2\left\|S_l(t) - \psi_h(\overline{S_l}^{\mathcal{N}}(t))\right\|_2 \\ &\le \sum_{l=1}^{N_\mu}\sqrt{2}C_0\left(\left\|S_l(t) - \overline{S_l}^{\mathcal{N}}(t)\right\|_2 + \left\|\overline{S_l}^{\mathcal{N}}(t) - \psi_h(\overline{S_l}^{\mathcal{N}}(t))\right\|_2\right) \\ &\le 2C_0N_\mu\left((s_0+1)\exp(-\mathcal{N}) + (18(s_0+1))^4h^2\right), \ 0 \le t \le T,\end{aligned} \tag{65}$$

where we take $M = 3(s_0+1)$ in (64) and (65) to ensure that $\overline{S}_0^{\mathcal{N}}(t)$ and $\overline{S}_l^{\mathcal{N}}(t)$ can be well approximated by $\psi_h(\overline{S}_0^{\mathcal{N}}(t))$ and $\psi_h(\overline{S}_l^{\mathcal{N}}(t))$.

For the fourth term, using Lemma D.1,

$$\begin{aligned}\left|\sum_{l>N_\mu}(\hat{\nu}_l,\hat{\mu}_l)\cdot\left(S_l(t)-\begin{pmatrix}0\\1\end{pmatrix}\right)\right| &= \sum_{t_j<t}\left|\mu(t-t_j)-S_{N_\mu}(t-t_j)\right| \\ &\le s_0\frac{4C_0T^k}{(k-1)(2\pi)^kN_\mu^{k-1}}.\end{aligned}$$

Here $S_{N_\mu}(t)$ is the finite sum of the fourier series defined in Lemma D.1 .

Finally, by (61), we have $|\overline{\lambda}_0^N(t)-\lambda_0(t)| \le 3\mathcal{C}B_0T^s/(2N^s),\ 0\le t\le T$. To trade off the error terms in (63), take $\mathcal{N} = \lceil\log((s_0+1)N^sN_\mu)\rceil$ and $h = (18(s_0+1))^{-2}N^{-s/2}N_\mu^{-1/2}$. Then under the event $\{N_e\le s_0\}$, we have

$$\begin{aligned}\left|\lambda^*(t)-\overline{\lambda}(t)\right| &\le \frac{2C_0N_\mu+1}{N^sN_\mu}+s_0\frac{4C_0T^k}{(k-1)(2\pi)^kN_\mu^{k-1}}+\frac{3\mathcal{C}B_0T^s}{2N^s} \\ &\le \frac{3\mathcal{C}B_0T^s+4C_0+2}{2N^s}+s_0\frac{4C_0T^k}{(k-1)(2\pi)^kN_\mu^{k-1}},\ t\in[0,T].\end{aligned}\tag{66}$$

Step 5. Compute the final approximation error.

By (55),

$$\begin{aligned}&|\mathbb{E}[\text{loss}(\hat{\lambda},S_{test})]-\mathbb{E}[\text{loss}(\lambda^*,S_{test})]| \\ &\le \mathbb{E}\left[(T+\frac{N_e}{B_1})\left\|\hat{\lambda}-\lambda^*\right\|_{L^\infty}\mathbb{1}_{\{N_e\le s_0\}}\right]+\mathbb{E}\left[(T+\frac{N_e}{B_1})\left\|\hat{\lambda}-\lambda^*\right\|_{L^\infty}\mathbb{1}_{\{N_e>s_0\}}\right] \\ &:= \mathbb{I}_1+\mathbb{I}_2\ .\end{aligned}\tag{67}$$

Taking $\eta = (c_\mu+1)/(2c_\mu)$ in Lemma 2 , we have

$$\begin{aligned}\mathbb{P}\left(N_e\ge s\right) &\le \frac{2\sqrt{B_0T}}{1-c_\mu}\exp\left(\frac{\log(\eta)}{2}\left[\eta(B_0T)-(1-c_\mu\eta)s\right]\right) \\ &\le \frac{2\sqrt{B_0T}}{1-c_\mu}\exp\left(\frac{\log\left(\frac{c_\mu+1}{2c_\mu}\right)}{2}\left[\frac{c_\mu+1}{2c_\mu}(B_0T)-\frac{1-c_\mu}{2}s\right]\right) \\ &:= a_e\exp\left(-c_es\right)\ .\end{aligned}$$

By (62) and (66),

$$\begin{aligned}\mathbb{I}_1 &\leq \left(T + \frac{1}{B_1}\right) \|\lambda^* - \hat{\lambda}\|_{L^\infty} \mathbb{E}\left[(N_e+1)\mathbb{1}_{\{N_e \leq s_0\}}\right] \\ &\leq \left(T + \frac{1}{B_1}\right) \|\lambda^* - \overline{\lambda}\|_{L^\infty} \mathbb{E}\left[(N_e+1)\right] \\ &= \left(T + \frac{1}{B_1}\right) \|\lambda^* - \overline{\lambda}\|_{L^\infty} \left(1 + \sum_{s=1}^{\infty} \mathbb{P}(N_e \geq s)\right) \\ &\leq \left(T + \frac{1}{B_1}\right) \left(1 + \frac{a_e \exp(-c_e)}{1-\exp(-c_e)}\right) \left(\frac{3\mathcal{C}B_0T^s + 4C_0 + 2}{2N^s} + s_0\frac{4C_0T^k}{(k-1)(2\pi)^k N_\mu^{k-1}}\right). \quad (68)\end{aligned}$$

Since $\|\hat{\lambda}\|_{L^\infty} \leq B_0 + C_0 s_0$ and $\|\lambda^*\|_{L^\infty} \leq B_0 + C_0 N_e$, similar to (57), we have

$$\begin{aligned}\mathbb{I}_2 &\leq \mathbb{E}\left[\left(T + \frac{N_e}{B_1}\right)\|\hat{\lambda}\|_{L^\infty}\mathbb{1}_{\{N_e > s_0\}}\right] + \mathbb{E}\left[\left(T + \frac{N_e}{B_1}\right)\|\lambda^*\|_{L^\infty}\,\mathbb{1}_{\{N_e > s_0\}}\right] \\ &\leq \left(T + \frac{1}{B_1}\right)(B_0 + C_0 s_0)\mathbb{E}\left[(N_e+1)\mathbb{1}_{\{N_e > s_0\}}\right] + \left(T + \frac{1}{B_1}\right) E\left[(N_e+1)(B_0 + C_0N_e)\mathbb{1}_{\{N_e > s_0\}}\right] \\ &\leq \left(T + \frac{1}{B_1}\right) a_e \exp(-c_e(s_0+1)) \left(2(s_0+1)(B_0 + C_0s_0) + \frac{3C_0(s_0+1) + 2B_0}{(1-\exp(-c_e))^2}\right). \quad (69)\end{aligned}$$

Combining (67), (68), and (69), we have

$$\begin{aligned}&|\mathbb{E}[\text{loss}(\hat{\lambda}, S_{test})] - \mathbb{E}[\text{loss}(\lambda^*, S_{test})]| \\ &\leq \left(T + \frac{1}{B_1}\right) a_e \exp(-c_e(s_0+1)) \left(2(s_0+1)(B_0 + C_0s_0) + \frac{3C_0(s_0+1) + 2B_0}{(1-\exp(-c_e))^2}\right) \\ &\quad + \left(T + \frac{1}{B_1}\right) \left(1 + \frac{a_e \exp(-c_e)}{1-\exp(-c_e)}\right) \left(\frac{3\mathcal{C}B_0T^s + 4C_0 + 2}{2N^s} + s_0\frac{4C_0T^k}{(k-1)(2\pi)^k N_\mu^{k-1}}\right).\end{aligned}$$

Let $s_0 = \lceil s\log(N)/c_e \rceil$ and denote $\hat{\lambda}^{N,N_\mu} = \hat{\lambda}$. We have

$$|\mathbb{E}[\text{loss}(\hat{\lambda}^{N,N_\mu}, S_{test})] - \mathbb{E}[\text{loss}(\lambda^*, S_{test})]| \lesssim \frac{1}{1-c_\mu}\exp\left(\frac{2B_0T}{c_\mu^2}\right)\left(\frac{T^s + \log^2 N}{N^s} + \frac{T^k \log N}{N_\mu^{k-1}}\right).$$

Step 6. Bound the sizes of the network width and weights.

From step 1-5, we have the width of the network being less than

$$\left(3\lceil\frac{\mathcal{N}'}{2}\rceil\binom{\mathcal{N}'+3}{3}\right)2N_\mu + 3\left\lceil\frac{s}{2}\right\rceil + 6N + 3\lceil\frac{\mathcal{N}''+5}{2}\rceil$$

where $\mathcal{N}' = \mathcal{N} + \lceil\log(6(s_0+1))\rceil + 15w_{N_\mu}T$, $\mathcal{N}'' = \mathcal{N} + \lceil(s_0+3)\log 2\rceil$, $\mathcal{N} = \lceil\log((s_0+1)N^sN_\mu)\rceil$, $s_0 = \lceil s\log(N)/c_e\rceil$. Hence

$$D \lesssim N + N_\mu^5 \log^4 N\ .$$

From the construction of $\hat{g}^{\mathcal{N}}_{l,i}$, $\hat{g}^{\mathcal{N}}_0$, ψ_h, $\overline{\lambda}^N_0$, the weights of the network is less than

$$\begin{aligned}\mathcal{C}_1' \max\Bigg\{&\left((s_0+1)\exp(\frac{\mathcal{N}'^2+\mathcal{N}'-3Cd\mathcal{N}'}{2})(\mathcal{N}'(\mathcal{N}'+2))^{3\mathcal{N}'(\mathcal{N}'+2)}\right),\\ &\left((s_0+1)\exp(\frac{\mathcal{N}''^2+\mathcal{N}''-3Cd\mathcal{N}''}{2})(\mathcal{N}''(\mathcal{N}''+2))^{3\mathcal{N}''(\mathcal{N}''+2)}\right), \frac{2}{\sigma'(0)h},\\ &\left(\left[\frac{\sqrt{2s}5^s}{(s-1)!}B_0T^s\right]^{-s/2} N^{(1+s^2)/2}(s(s+2))^{3s(s+2)}\right)\Bigg\},\end{aligned}$$

where $h = (18(s_0+1))^{-2}N^{-s/2}N_\mu^{-1/2}$. Hence the weights of the network is less than

$$\mathcal{C}_1(\log(NN_\mu))^{12s^2(\log(NN_\mu))^2} ,$$

where $\mathcal{C}_1$ is a constant related to s, B_0, C_0, c_μ, and T. ■

Lemma D.3. *Let $\delta_j = \frac{j}{k}$, $1 \le j \le k$, $\delta = (\delta_1, \delta_2, \cdots, \delta_k)^\top$ and*

$$V_\delta = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \delta_1 & \delta_2 & \cdots & \delta_k \\ \delta_1^2 & \delta_2^2 & \cdots & \delta_k^2 \\ \vdots & \vdots & \ddots & \vdots \\ \delta_1^{k-1} & \delta_2^{k-1} & \cdots & \delta_k^{k-1} \end{pmatrix}$$

then V_δ is invertible and $\|V_\delta^{-1}\|_\infty \le C * 8^k$, where C is a universal constant.

Proof [Proof of Lemma D.3] See Gautschi (1990). ■

Lemma D.4. *For $\mu \in C^{k,\infty}([0,T], C_0)$ and $\delta = (\delta_1, \delta_2, \cdots, \delta_k)^\top$ defined in Lemma D.3, there exists $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)^\top$ such that*

$$\tilde{\mu}(t) := \mu(t) + \sum_{q=1}^{k} \alpha_q \exp(-\delta_q t) \tag{70}$$

satisfying $\|\alpha\|_\infty \le 2C_0C8^k/(1-\exp(-T))$, $\tilde{\mu} \in C^{k,\infty}([0,T], C_0 + k\|\alpha\|_\infty)$ and $\tilde{\mu}^{(q)}(0+) = \tilde{\mu}^{(q)}(T-)$, $0 \le q \le k-1$, where the constant C is defined in Lemma D.3.

Proof [Proof of Lemma D.4] We only need to solve the following equations:

$$\tilde{\mu}^{(q)}(0+) = \tilde{\mu}^{(q)}(T-),\ 0 \le q \le k-1. \tag{71}$$

In matrix form,

$$\begin{pmatrix} 1-e^{-\delta_1 T} & 1-e^{-\delta_2 T} & \cdots & 1-e^{-\delta_k T} \\ (-\delta_1)(1-e^{-\delta_1 T}) & (-\delta_2)(1-e^{-\delta_2 T}) & \cdots & (-\delta_k)(1-e^{-\delta_k T}) \\ \vdots & \vdots & \ddots & \vdots \\ (-\delta_1)^{k-1}(1-e^{-\delta_1 T}) & (-\delta_2)^{k-1}(1-e^{-\delta_2 T}) & \cdots & (-\delta_k)^{k-1}(1-e^{-\delta_k T}) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix}$$
$$= \begin{pmatrix} \mu(T-)-\mu(0+) \\ \mu^{(1)}(T-)-\mu^{(1)}(0+) \\ \vdots \\ \mu^{(k-1)}(T-)-\mu^{(k-1)}(0+) \end{pmatrix}. \tag{72}$$

Rewrite (72) as

$$DV_\delta \Lambda_\delta \alpha = \Delta_\mu,$$

where $D = \mathrm{diag}\{1, -1, \cdots, (-1)^{k-1}$, $\Lambda_\delta = \mathrm{diag}\{1-e^{-\delta_1 T}, 1-e^{-\delta_2 T}, \cdots, 1-e^{-\delta_k T}\}$, $\Lambda_\mu = (\mu(T-)-\mu(0+), \mu^{(1)}(T-)-\mu^{(1)}(0+), \cdots, \mu^{(k-1)}(T-)-\mu^{(k-1)}(0+))^\top$, and V_δ is defined in Lemma D.3. By Lemma D.3 and $\delta_j = j/k$, $1 \le j \le k$, we have $DV_\delta\Lambda_\delta$ is invertible and

$$\begin{aligned} \|\alpha\|_\infty &\le \|D^{-1}\|_\infty \|V_\delta^{-1}\|_\infty \|\Lambda_\delta^{-1}\|_\infty \|\Delta_\mu\|_\infty \\ &\le (C * 8^k)\frac{1}{(1-\exp(-T))}(2C_0) = \frac{2C_0C8^k}{1-\exp(-T)}, \end{aligned}$$

where the constant C is defined in Lemma D.3. By (70), we have $\tilde{\mu} \in C^{k,\infty}([0,T], C_0 + k\|\alpha\|_\infty)$. ■

Now we prove Theorem 6. The proof is based on Theorem 5, Theorem D.2, and Lemma D.4. From Lemma D.4, for $\mu \in C^{k,\infty}([0,T], C_0)$, there exists $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)^\top \in \mathbb{R}^k$ such that $\tilde{\mu}(t) := \mu(t) + \sum_{j=1}^k \alpha_j \exp(-\delta_j t)$ satisfying the boundary condition $\tilde{\mu}^{(q)}(0+) = \tilde{\mu}^{(q)}(T-)$, $0 \le q \le k-1$, and we have $\tilde{\mu} \in C^{k,\infty}([0,T], C_0 + k\frac{2C_0C8^k}{1-\exp(-T)})$. Define $\tilde{\nu}(t) := \mu(t) - \tilde{\mu}(t) = -\sum_{q=1}^k \alpha_q \exp(-\delta_q t)$. Denote

$$\begin{aligned} \lambda_1^*(t) &:= \lambda_0(t) + \sum_{t_j<t} \tilde{\mu}(t-t_j), \\ \lambda_2^*(t) &:= \sum_{t_j<t} \tilde{\nu}(t-t_j) = \sum_{q=1}^k \sum_{t_j<t} (-\alpha_q)\exp(-\delta_q(t-t_j)) := \sum_{j=1}^k \lambda_{2q}^*(t), \end{aligned}$$

and then $\lambda^*(t) = \lambda_1^*(t) + \lambda_2^*(t)$.

Fix $s_0 \in \mathbb{N}_+$. By the proof of Theorem D.2, under the event $\{N_e \le s_0\}$, there exists an RNN (without the output layer) $\overline{\lambda}_1(t)$ such that

$$\left|\lambda_1^*(t) - \overline{\lambda}_1(t)\right| \le \frac{3\mathcal{C}B_0T^s + 4\tilde{C}_0 + 2}{2N^s} + s_0\frac{4\tilde{C}_0T^k}{(k-1)(2\pi)^kN_\mu^{k-1}},\ t \in [0,T],$$

where $\tilde{C_0} = C_0 + 2kC_0 C 8^k/(1-\exp(-T))$.

By the proof of Theorem 5 , under the event $\{N_e \leq s_0\}$, for$1 \leq j \leq k$, there exists an RNN (without the output layer) $\overline{\lambda}_{2q}(t)$ such that

$$\left|\lambda^*_{2q}(t) - \overline{\lambda}_{2q}(t)\right| \leq \frac{2\alpha_q}{N^s} \leq \frac{4C_0 C 8^k}{(1-\exp(-T))N^s}, t \in [0,T].$$

Let $\overline{\lambda}_2(t) = \sum_{q=1}^k \overline{\lambda}_{2q}(t)$. We have

$$\left|\lambda^*_2(t) - \overline{\lambda}_2(t)\right| \leq \frac{2(\tilde{C}_0 - C_0)}{N^s}, t \in [0,T].$$

Let $\overline{\lambda}(t) = \overline{\lambda}_1(t) + \overline{\lambda}_2(t)$,

$$\left|\lambda^*(t) - \overline{\lambda}(t)\right| \leq \left|\lambda^*_1(t) - \overline{\lambda}_1(t)\right| + \left|\lambda^*_2(t) - \overline{\lambda}_2(t)\right| \leq \frac{3\mathcal{C}B_0T^s + 8\tilde{C}_0 + 2}{2N^s} + s_0 \frac{4\tilde{C}_0 T^k}{(k-1)(2\pi)^k N_\mu^{k-1}}.$$

Under the event $\{N_e \leq s_0\}$, $B_1 \leq \lambda^* \leq B_0 + C_0 s_0$. Hence we can take $l_f = B_1$ and $u_f = B_0 + C_0 s_0$ and denote $\hat{\lambda}(t) = f(\overline{\lambda}(t))$. Then $\|\lambda^* - \hat{\lambda}\|_\infty \leq \|\lambda^* - \overline{\lambda}\|_\infty$. By similar arguments in Theorem D.2, we have

$$\begin{aligned}
&|\mathbb{E}[\text{loss}(\hat{\lambda}, S_{test})] - \mathbb{E}[\text{loss}(\lambda^*, S_{test})]| \\
&\leq \left(T + \frac{1}{B_1}\right) a_e \exp(-c_e(s_0+1)) \left(2(s_0+1)(B_0 + C_0 s_0) + \frac{3C_0(s_0+1) + 2B_0}{(1-\exp(-c_e))^2}\right) \\
&\quad + \left(T + \frac{1}{B_1}\right)\left(1 + \frac{a_e \exp(-c_e)}{1-\exp(-c_e)}\right)\left(\frac{3\mathcal{C}B_0T^s + 8\tilde{C}_0}{2N^s} + s_0 \frac{4\tilde{C}_0 T^k}{(k-1)(2\pi)^k N_\mu^{k-1}}\right).
\end{aligned}$$

Let $s_0 = \lceil s\log(N)/c_e \rceil$ and denote $\hat{\lambda}^{N,N_\mu} = \hat{\lambda}$. We have

$$|\mathbb{E}[\tilde{\text{loss}}(\hat{\lambda}^{N,N_\mu})] - \mathbb{E}[\tilde{\text{loss}}(\lambda^*)]| \lesssim \frac{\log^2 N}{N^s} + \frac{\log N}{N_\mu^{k-1}}$$

The width and elements weights bound can also be obtained similarly to the proof of Theorem D.2.

D.4 Proof of Theorem 5.4

Denote $\lambda^*_1(t) = \lambda_0(t) + \sum_{t_j<t} \alpha \exp(-\beta(t-t_j))$. Then $\lambda^*(t) = \Psi\left(\lambda^*_1(t)\right)$. Fix $s_0 \in \mathbb{N}_+$. From the proof of Theorem 5 , under the event $\{N_e \leq s_0\}$, there exists a 2-layer recurrent neural network $\overline{\lambda}_1(t)$ as (52) such that

$$\left|\overline{\lambda}_1(t) - \lambda^*_1(t)\right| \leq \frac{3\mathcal{C}B_0T^s + 2}{2N_1^s}, \ \forall t \in [0,T]. \tag{73}$$

Moreover, the width of $\overline{\lambda}_1(t)$ satisfies $D \lesssim N_1$ and the weights of $\overline{\lambda}_1(t)$ are bounded by

$$O\left((\log N_1)^{12s^2(\log N_1)^2}\right).$$

Under the event $\{N_e \le s_0\}$, the function $\lambda_1^*(t)$ satisfies $0 \le \lambda_1^* \le B_0 + \alpha s_0$. Using (73) and taking $(3\mathcal{C}B_0T^s + 2)/2N_1^s \le 1$, we have $\overline{\lambda}_1 \in [-1, B_0 + \alpha s_0 + 1]$. Hence we need to construct an approximation of Ψ on $[-1, B_0 + \alpha s_0 + 1]$. Let $\tilde{\Psi}(x) = \Psi(\rho x - 1)$, where $\rho = B_0 + \alpha s_0 + 2$. Then $\Psi(x) = \tilde{\Psi}((x+1)/\rho)$.

Since Ψ is L-lipschitz and $\tilde{\Psi}$ is defined on $[0,1]$ and ρL-Lipschitz, by the Corollary 5.4 of De Ryck et al. (2021), there exists a tanh neural network $\tilde{\Psi}^{N_2}$ with 2 hidden layers such that

$$\left\|\tilde{\Psi} - \tilde{\Psi}^{N_2}\right\|_{L^\infty[0,1]} \le \frac{7(\rho L \vee \tilde{B}_0)}{N_2}.$$

Let $\Psi^{N_2}(x) = \tilde{\Psi}^{N_2}((x+1)/\rho)$. Then

$$\left|\Psi(x) - \Psi^{N_2}(x)\right| \le \frac{7(\rho L \vee \tilde{B}_0)}{N_2}, \ x \in [-1, B_0 + \alpha s_0 + 1].$$

Then under the event $\{N_e \le s_0\}$, we have

$$\begin{aligned}\left|\Psi(\lambda_1^*(t)) - \Psi^{N_2}(\overline{\lambda}_1(t))\right| &\le \left|\Psi(\lambda_1^*(t)) - \Psi(\overline{\lambda}_1(t))\right| + \left|\Psi(\overline{\lambda}_1(t)) - \Psi^{N_2}(\overline{\lambda}_1(t))\right| \\ &\le L\left|\overline{\lambda}_1(t) - \lambda_1^*(t)\right| + \left\|\Psi - \Psi^{N_2}\right\|_{L^\infty} \\ &\le L\frac{3\mathcal{C}B_0T^s + 2}{2N_1^s} + \frac{7(\rho L \vee \tilde{B}_0)}{N_2}. \end{aligned} \tag{74}$$

Recall that $f(x) = \min\{\max\{x, l_f\}, u_f\}$. Since $\tilde{B}_1 \le \Psi \le \tilde{B}_0$, we can take $l_f = \tilde{B}_1$ and $u_f = \tilde{B}_0$. Define $\hat{\lambda}(t) = f\left(\Psi^{N_2}(\overline{\lambda}_1(t))\right)$. We have

$$\left|\lambda^*(t) - \hat{\lambda}(t)\right| \le \left|\Psi(\lambda_1^*(t)) - \Psi^{N_2}(\overline{\lambda}_1(t))\right|, \ \forall t \in [0,T]. \tag{75}$$

Similar to (55), we have

$$\begin{aligned}&\left|\mathbb{E}[\mathrm{loss}(\hat{\lambda}, S_{test})] - \mathbb{E}[\mathrm{loss}(\lambda^*, S_{test})]\right| \\ &\le \mathbb{E}\left|\mathrm{loss}(\hat{\lambda}, S_{test}) - \mathrm{loss}(\lambda^*, S_{test})\right| \\ &\le \mathbb{E}\left(\Big|\sum_{j=1}^{N_e}(\log\hat{\lambda}(t_j) - \log\lambda^*(t_j))\Big| + \Big|\int_0^T\left(\hat{\lambda}(t) - \lambda^*(t)\right)\mathrm{dt}\Big|\right) \\ &\le \mathbb{E}\left[\left(\Big|\sum_{j=1}^{N_e}(\log\hat{\lambda}(t_j) - \log\lambda^*(t_j))\Big| + \Big|\int_0^T\left(\hat{\lambda}(t) - \lambda^*(t)\right)\mathrm{dt}\Big|\right)\mathbb{1}_{\{N_e \le s_0\}}\right. \\ &\qquad \left. + \left(\Big|\sum_{j=1}^{N_e}(\log\hat{\lambda}(t_j) - \log\lambda^*(t_j))\Big| + \Big|\int_0^T\left(\hat{\lambda}(t) - \lambda^*(t)\right)\mathrm{dt}\Big|\right)\mathbb{1}_{\{N_e > s_0\}}\right] \\ &\le \mathbb{E}\left[(T + \frac{N_e}{\tilde{B}_1})\left\|\hat{\lambda} - \lambda^*\right\|_{L^\infty}\mathbb{1}_{\{N_e \le s_0\}}\right] + \mathbb{E}\left[(T + \frac{N_e}{\tilde{B}_1})\left\|\hat{\lambda} - \lambda^*\right\|_{L^\infty}\mathbb{1}_{\{N_e > s_0\}}\right] \\ &:= \mathbb{I}_1 + \mathbb{I}_2 \ . \end{aligned} \tag{76}$$

Since $\Psi \leq \tilde{B}_0$, similar to (55), taking $\eta = e$ in Lemma 2 , we have

$$\mathbb{P}(N_e \geq s) \leq 2\sqrt{\tilde{B}_0 T} \exp\left(\frac{e\tilde{B}_0 T - s}{2}\right),$$

and similar to (45), we have

$$\begin{aligned}\mathbb{E}(N_e + 1) &\leq 1 + \sum_{s=1}^{\infty} \mathbb{P}(N_e \geq s)\\ &\leq 1 + \frac{2\sqrt{\tilde{B}_0 T}}{1 - \exp(-1/2)} \exp\left(\frac{e\tilde{B}_0 T - 1}{2}\right) \leq 5\sqrt{\tilde{B}_0 T + 1} \exp\left(\frac{3\tilde{B}_0 T}{2}\right). \end{aligned} \tag{77}$$

By (74), (75), and (77),

$$\begin{aligned}\mathbb{I}_1 &\leq \left(T + \frac{1}{\tilde{B}_1}\right) \frac{3\mathcal{C}B_0 T^s + 2}{2N^s} \mathbb{E}\left[(N_e + 1)\mathbb{1}_{\{N_e \leq s_0\}}\right]\\ &\leq \left(T + \frac{1}{\tilde{B}_1}\right) \frac{3\mathcal{C}B_0 T^s + 2}{2N^s} \mathbb{E}\left[(N_e + 1)\right]\\ &\leq \left(T + \frac{1}{\tilde{B}_1}\right) 5\sqrt{\tilde{B}_0 T + 1} \exp\left(\frac{3\tilde{B}_0 T}{2}\right) \left(L\frac{3\mathcal{C}B_0 T^s + 2}{2N_1^s} + \frac{7(\rho L \vee \tilde{B}_0)}{N_2}\right)\\ &\leq 5L \exp\left(2\tilde{B}_0 T\right) \left(T + \frac{1}{\tilde{B}_1}\right) \left(\frac{3\mathcal{C}B_0 T^s + 2}{2N_1^s} + \frac{7(\rho \vee (\tilde{B}_0/L))}{N_2}\right). \end{aligned} \tag{78}$$

On the other hand, since $\|\hat{\lambda}\|_{L^\infty} \leq \tilde{B}_0$ and $\|\lambda^*\|_{L^\infty} \leq \tilde{B}_0$, we have

$$\begin{aligned}\mathbb{I}_2 &\leq \mathbb{E}\left[\left(T + \frac{N_e}{\tilde{B}_1}\right) \|\hat{\lambda}\|_{L^\infty} \mathbb{1}_{\{N_e > s_0\}}\right] + \mathbb{E}\left[\left(T + \frac{N_e}{\tilde{B}_1}\right) \|\lambda^*\|_{L^\infty} \mathbb{1}_{\{N_e > s_0\}}\right]\\ &\leq 2\left(T + \frac{1}{\tilde{B}_1}\right) \tilde{B}_0 \mathbb{E}\left[(N_e + 1)\mathbb{1}_{\{N_e > s_0\}}\right]\\ &\leq 2\left(T + \frac{1}{\tilde{B}_1}\right) \tilde{B}_0 \left((s_0 + 1)\mathbb{P}(N_e \geq s_0 + 1) + \sum_{s=s_0+1}^{\infty} \mathbb{P}(N_e \geq s)\right)\\ &\leq 4\left(T + \frac{1}{\tilde{B}_1}\right) \tilde{B}_0 \sqrt{\tilde{B}_0 T} \exp\left(\frac{e\tilde{B}_0 T - (s_0 + 1)}{2}\right) \left((s_0 + 1) + \frac{1}{1 - e^{-\frac{1}{2}}}\right)\\ &\leq 4\left(T + \frac{1}{\tilde{B}_1}\right) \tilde{B}_0 \sqrt{\tilde{B}_0 T} \exp\left(\frac{3\tilde{B}_0 T - (s_0 + 1)}{2}\right) (s_0 + 4)\\ &\leq 4\left(T + \frac{1}{\tilde{B}_1}\right) \tilde{B}_0 \exp\left(2\tilde{B}_0 T\right) (s_0 + 4) \exp\left(-\frac{s_0 + 1}{2}\right). \end{aligned} \tag{79}$$

Combining (76), (78), and (79), we have

$$\begin{aligned}&|\mathbb{E}[\mathrm{loss}(\hat{\lambda}, S_{test})] - \mathbb{E}[\mathrm{loss}(\lambda^*, S_{test})]|\\ \leq\,& 5L \exp\left(2\tilde{B}_0 T\right)\left(T + \frac{1}{\tilde{B_1}}\right)\left(\frac{3\mathcal{C}B_0T^s + 2}{2N_1^s} + \frac{7(\rho \vee (\tilde{B}_0/L))}{N_2}\right)\\ &+ 4\left(T + \frac{1}{\tilde{B_1}}\right)\tilde{B}_0 \exp\left(2\tilde{B}_0T\right)(s_0+4)\exp\left(-\frac{s_0+1}{2}\right).\end{aligned}$$

Let $s_0 = \lceil 2\log N \rceil$, $N_1 = N_2 = N$ and denote $\hat{\lambda}^N = \hat{\lambda}$. We have

$$|\mathbb{E}[\mathrm{loss}(\hat{\lambda}^N, S_{test})] - \mathbb{E}[\mathrm{loss}(\lambda^*, S_{test})]| \lesssim \frac{\log N}{N}.$$

Similar to the proof of Theorem 5, we can bound the width of the network by

$$\max\left\{3\left\lceil\frac{\tilde{\mathcal{N}}}{2}\right\rceil\binom{\tilde{\mathcal{N}}+2}{2} + 3\left\lceil\frac{s}{2}\right\rceil + 6N + 2, 6N\right\},$$

where $\tilde{\mathcal{N}} = \lceil s\log(N)\rceil + 10(\delta T \vee 1) + 2\lceil \log(3(s_0+1))\rceil$. Hence we have $D \lesssim N$.

Moreover, from the construction of $\hat{\lambda}$, the weights of the network is less than

$$\mathcal{C}_1' \max\left\{(\log(N))^{12s^2(\log(N))^2}, \frac{N}{\rho\sqrt{\rho L}}\right\},$$

where $\rho = B_0 + \alpha s_0 + 2 = B_0 + \alpha\lceil 2\log N\rceil + 2$, $\mathcal{C}_1'$ is a constant related to $s, B_0, \alpha, \delta, T, \tilde{B}_0$, and L. Then the weights of the network can be bounded by

$$\mathcal{C}_1(\log(N))^{12s^2(\log(N))^2},$$

where C_1 are constants related to $s, B_0, \alpha, \delta, T, \tilde{B}_0$, and L.

Appendix E. Proof in Section 6

E.1 Proof of Theorem 6.1

Without loss of generality, we denote $t_1 = T/3$, $t_2 = 2T/3$ for simplicity. Since both λ^* and $\hat{\lambda}_{ne}$ are predictable, by (87)(see the following section), we have

$$\begin{aligned}&\mathbb{E}[\mathrm{loss}(\hat{\lambda}_{ne}, S_{test})] - \mathbb{E}[\mathrm{loss}(\lambda^*, S_{test})]\\ =&\mathbb{E}\left[\int_0^T \left(\hat{\lambda}_{ne}(t) - \log\hat{\lambda}_{ne}(t) * \lambda^*(t)\right)\mathrm{d}t\right] - E\left[\int_0^T (\lambda^*(t) - \log\lambda^*(t) * \lambda^*(t))\,\mathrm{d}t\right]\\ :=&\mathbb{E}\left[\int_0^T \left(g(\hat{\lambda}_{ne}(t), \lambda^*(t)) - g(\lambda^*(t), \lambda^*(t))\right)\mathrm{d}t\right],\end{aligned}$$

where $g(x,y) = x - \log x * y \geq y - \log y * y = g(y,y)$, $\forall x, y > 0$, and the equality holds if and only if $x = y$. Thus

$$\mathbb{E}[\mathrm{loss}(\hat{\lambda}_{ne}, S_{test})] - \mathbb{E}[\mathrm{loss}(\lambda^*, S_{test})] = \mathbb{E}\left[\int_0^T \left(g(\hat{\lambda}_{ne}(t), \lambda^*(t)) - g(\lambda^*(t), \lambda^*(t))\right)\mathrm{d}t\right] \geq 0.$$

Denote $\mathcal{E} = \{\text{there is no event in } [0, 2T/3]\}$, and $\mathbb{P}(\mathcal{E}) > 0$. Denote $I_0 = [T/3, 2T/3]$. By a similar argument, we have

$$\mathbb{E}[\text{loss}(\hat{\lambda}_{ne}, S_{test})] - \mathbb{E}[\text{loss}(\lambda^*, S_{test})] \geq \mathbb{E}\left[\left(\int_{I_0}\left(g(\hat{\lambda}_{ne}(t), \lambda^*(t)) - g(\lambda^*(t), \lambda^*(t))\right)\mathrm{d}t\right)\mathbb{1}_{\mathcal{E}}\right]. \tag{80}$$

Under the event $\mathcal{E}$,

$$\hat{\lambda}_{ne}(t) = f(\alpha t + b) = \begin{cases} T & , \ \alpha t + b < T \\ \alpha t + b & , \ T \leq \alpha t + b \leq 4T \\ 4T & , \ \alpha t + b > 4T \end{cases}, \ t \in I_0, \tag{81}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\left(\int_{I_0}\left(g\left(\hat{\lambda}_{ne}(t), \lambda^*(t)\right) - g\left(\lambda^*(t), \lambda^*(t)\right)\right)\mathrm{d}t\right)\mathbb{1}_{\mathcal{E}}\right] \\
=&\left[\int_{I_0}\left(g\left(f(\alpha t + b), \frac{9}{T}t^2\right) - g\left(\frac{9}{T}t^2, \frac{9}{T}t^2\right)\right)\mathrm{d}t\right]\mathbb{P}(\mathcal{E}) \\
:=&F(\alpha, b)P(\mathcal{E}).
\end{aligned} \tag{82}$$

Then we only need to show

$$\inf_{\alpha\in\mathbb{R}, b\in\mathbb{R}} F(\alpha, b) > 0.$$

Case 1. $|\alpha| > 18$, $b \in \mathbb{R}$.

Since $|3T/\alpha| \leq T/6$, $\hat{\lambda}_{ne}(t) \in \{T, 4T\}$ on $I_1 := [T/3, 5T/12]$ or $I_2 := [7T/12, 2T/3]$. From $g(x, y) \geq g(y, y), x, y > 0$,

$$\begin{aligned}
\inf_{|\alpha|>18, b\in\mathbb{R}} F(\alpha, b) \geq \min\Bigg\{&\int_{I_1}\left(g\left(T, \frac{9}{T}t^2\right) - g\left(\frac{9}{T}t^2, \frac{9}{T}t^2\right)\right)\mathrm{d}t, \\
&\int_{I_2}\left(g\left(T, \frac{9}{T}t^2\right) - g\left(\frac{9}{T}t^2, \frac{9}{T}t^2\right)\right)\mathrm{d}t, \\
&\int_{I_1}\left(g\left(4T, \frac{9}{T}t^2\right) - g\left(\frac{9}{T}t^2, \frac{9}{T}t^2\right)\right)\mathrm{d}t, \\
&\int_{I_2}\left(g\left(4T, \frac{9}{T}t^2\right) - g\left(\frac{9}{T}t^2, \frac{9}{T}t^2\right)\right)\mathrm{d}t\Bigg\} \\
:= C_1 > 0.&
\end{aligned} \tag{83}$$

Case 2. $|\alpha| \leq 18$, $|b| > 16T$.

In this case , we can check that $\{t : T \leq \alpha t + b \leq 4T\} \cap I_0 = \emptyset$. Hence

$$\begin{aligned}
\inf_{|\alpha|\leq 18, |b|>16T} F(\alpha, b) \geq \min\Bigg\{&\int_{I_0}\left(g\left(T, \frac{9}{T}t^2\right) - g\left(\frac{9}{T}t^2, \frac{9}{T}t^2\right)\right)\mathrm{d}t, \\
&\int_{I_0}\left(g\left(4T, \frac{9}{T}t^2\right) - g\left(\frac{9}{T}t^2, \frac{9}{T}t^2\right)\right)\mathrm{d}t\Bigg\} \\
:= C_2 > 0.&
\end{aligned} \tag{84}$$

Case 3. $|\alpha| \le 18$, $|b| \le 16T$.

By (82) , F is continuous with respect to (α, b). For fixed (α, b), since $f(\alpha t + b) \not\equiv \frac{9}{T^2}t^2$, $F(\alpha, b) > 0$. Since $\{|\alpha| \le 18, |b| \le 16T\}$ is a compact set in $\mathbb{R}^2$, there exists $C_3 > 0$ such that

$$\inf_{|\alpha|\le 18, |b|\le 16T} F(\alpha, b) \ge C_3 > 0. \tag{85}$$

By (80), (82), (83), (84), and (85),

$$\mathbb{E}[\tilde{\text{loss}}(\hat{\lambda}_{ne})] - \mathbb{E}[\tilde{\text{loss}}(\lambda^*)] \ge \min\{C_1, C_2, C_3\}\mathbb{P}(\mathcal{E}) := C > 0.$$

Hence Theorem 8 is proved.

Remark E.1. *Note that we have proved the excess risk*

$$\mathbb{E}[loss(\hat{\lambda}, S_{test})] - \mathbb{E}[loss(\lambda^*, S_{test})] \tag{86}$$

is always positive if $\hat{\lambda} \ne \lambda^*$ in the proof of Theorem 8. Thus (86) *is a well-defined excess risk.*

E.2 Explanation of Remark 6.4

Notice that the proof of (80) remains valid regardless of the specific form of $\hat{\lambda}_{ne}$. Now consider $\alpha = u(h_j, w)$ for $t \in (t_j, t_{j+1}]$, where w denotes trainable parameters, h_j represents the hidden state at t_j, and u is a fixed function. Under event $\mathcal{E}$, we observe $\alpha = u(h_0, w)$ throughout $[0, 2T/3]$, effectively reducing it to a single training parameter. Consequently, $\hat{\lambda}_{ne}$ simplifies to the form in (81), and the theorem follows through similar arguments.

Appendix F. Proofs in Section 7

F.1 Proof of Lemma 7.1

Proof Since the compensator of $N(t)$ is $\Lambda(t) = \int_0^t \lambda^*(s)\mathrm{d}s$, for a predictable stochastic process $\lambda(t), t \in [0, T]$, we have

$$\begin{aligned}\mathbb{E}[\text{loss}(\lambda, S_{test})] &= \mathbb{E}\left[-\sum_{t_j<T} \log \lambda(t_j) + \int_0^T \lambda(t)\mathrm{d}t\right] \\ &= \mathbb{E}\left[-\int_0^T \log \lambda(t)\mathrm{d}N(t) + \int_0^T \lambda(t)\mathrm{d}t\right] \\ &= \mathbb{E}\left[\int_0^T (\lambda(t) - \log \lambda(t) * \lambda^*(t))\,\mathrm{d}t\right]. \end{aligned} \tag{87}$$

Thus by $\tilde{\lambda}, \lambda^* \geq B_1 > 0$, we have

$$\begin{aligned}
&\mathbb{E}[\text{loss}(\tilde{\lambda}, S_{test})] - \mathbb{E}[\text{loss}(\lambda^*, S_{test})] \\
=&\mathbb{E}\left[\int_0^T \left(\tilde{\lambda}(t) - \log \tilde{\lambda}(t) * \lambda^*(t)\right) \mathrm{d}t\right] - \mathbb{E}\left[\int_0^T \left(\lambda^*(t) - \log \lambda^*(t) * \lambda^*(t)\right) \mathrm{d}t\right] \\
=&\mathbb{E}\left[\int_0^T \left(\frac{\tilde{\lambda}(t)}{\lambda^*(t)} - 1 - \log\left(\frac{\tilde{\lambda}(t)}{\lambda^*(t)}\right)\right) \lambda^*(t)\mathrm{d}t\right] \\
\geq&\mathbb{E}\left[\int_0^T \left(\sqrt{\frac{\tilde{\lambda}(t)}{\lambda^*(t)}} - 1\right)^2 \lambda^*(t)\mathrm{d}t\right] \\
=&\mathbb{E}\left[\int_0^T \left(\sqrt{\tilde{\lambda}(t)} - \sqrt{\lambda^*(t)}\right)^2 \mathrm{d}t\right] \\
=&2H_2^2\left(\tilde{\lambda}, \lambda^*\right),
\end{aligned}$$

where the inequality follows by $x - 1 - \log x \geq (\sqrt{x} - 1)^2$, $x > 0$.

On the other hand, define $g_t(p) := (p\tilde{\lambda}(t) + (1-p)\lambda^*(t)) - \log(p\tilde{\lambda}(t) + (1-p)\lambda^*(t)) * \lambda^*(t)$ $p \in [0,1]$, then we have

$$\begin{aligned}
&\mathbb{E}[\text{loss}(\tilde{\lambda}, S_{test})] - \mathbb{E}[\text{loss}(\lambda^*, S_{test})] \\
=&\mathbb{E}\left[\int_0^T \left(\tilde{\lambda}(t) - \log \tilde{\lambda}(t) * \lambda^*(t)\right) \mathrm{d}t\right] - \mathbb{E}\left[\int_0^T \left(\lambda^*(t) - \log \lambda^*(t) * \lambda^*(t)\right) \mathrm{d}t\right] \\
=&\mathbb{E}\left[\int_0^T (g_t(1) - g_t(0))\mathrm{d}t\right] \\
\leq&\mathbb{E}\left[\int_0^T \sup_{p\in[0,1]} g_t'(p)\mathrm{d}t\right] \\
=&\mathbb{E}\left[\int_0^T \sup_{p\in[0,1]} \frac{p(\tilde{\lambda}(t) - \lambda^*(t))^2}{p\tilde{\lambda}(t) + (1-p)\lambda^*(t)}\mathrm{d}t\right] \\
\leq&\mathbb{E}\left[\int_0^T (\tilde{\lambda}(t) - \lambda^*(t))^2\mathrm{d}t\right],
\end{aligned}$$

where the last inequality follows by $\tilde{\lambda}, \lambda^* \geq B_1$. ■

F.2 Proof of Lemma 7.2

Proof By direct calculation, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{D}_{train}}\left[\mathbb{E}[\text{loss}(\hat{\lambda}, S_{test})] - \mathbb{E}[\text{loss}(\lambda^*, S_{test})]\right] \\
=&\mathbb{E}_{\mathbf{D}_{train}}\left[\mathbb{E}[\text{loss}(\hat{\lambda}, S_{test}) - \text{loss}(\lambda^*, S_{test})] - \frac{2}{n}\sum_{i\in[n]}\left(\text{loss}(\hat{\lambda}, S_i) - \text{loss}(\lambda^*, S_i)\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&+ 2\mathbb{E}_{\mathbf{D}_{train}}\left[\frac{2}{n}\sum_{i\in[n]}\left(\text{loss}(\hat{\lambda}, S_i) - \text{loss}(\lambda^*, S_i)\right)\right], && (88)\\
&\overset{(i)}{\leq}\mathbb{E}_{\mathbf{D}_{train}}\left[\mathbb{E}[\text{loss}(\hat{\lambda}, S_{test}) - \text{loss}(\lambda^*, S_{test})] - \frac{2}{n}\sum_{i\in[n]}\left(\text{loss}(\hat{\lambda}, S_i) - \text{loss}(\lambda^*, S_i)\right)\right]\\
&+ 2\mathbb{E}_{\mathbf{D}_{train}}\left[\frac{2}{n}\sum_{i\in[n]}\left(\text{loss}(\check{\lambda}^*, S_i) - \text{loss}(\lambda^*, S_i)\right)\right], && (89)\\
&=\mathbb{E}_{\mathbf{D}_{train}}\left[\mathbb{E}[\text{loss}(\hat{\lambda}, S_{test}) - \text{loss}(\lambda^*, S_{test})] - \frac{2}{n}\sum_{i\in[n]}\left(\text{loss}(\hat{\lambda}, S_i) - \text{loss}(\lambda^*, S_i)\right)\right]\\
&+ 2\left(\mathbb{E}[\text{loss}(\check{\lambda}^*, S_{test})] - \mathbb{E}[\text{loss}(\lambda^*, S_{test})]\right),
\end{aligned}
$$

where (i) follows from the definition of $\hat{\lambda}$ (see (3)). ■

Appendix G. Supporting Lemmas

Lemma G.1. *(Lemma 8 in Chen et al. (2020)) Let $\mathcal{G} = \{A \in \mathbb{R}^{d_1\times d_2} : \|A\|_2 \leq \lambda\}$ be the set of matrices with bounded spectral norm and $\epsilon > 0$ be given. The covering number $\mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|_F)$ is bounded above by*

$$\mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|_F) \leq \left(1 + \frac{(\sqrt{d_1}\wedge\sqrt{d_2})\lambda}{\epsilon}\right)^{d_1 d_2}.$$

The following lemma is a bridge between the covering number and the upper bound of sub-gaussian process.

Definition G.2. *A stochastic process $\{X_h\}_{h\in H}$ is called a sub-gaussian process for metric $d(\cdot,\cdot)$ on H if*

$$\mathbb{E}\left[\exp\left(\lambda\left(X_{h_1} - X_{h_2}\right)\right)\right] \leq \exp\left(\frac{\lambda^2 d(h_1,h_2)^2}{2}\right) \quad \textit{for } \lambda\in\mathbb{R},\ h_1, h_2\in H.$$

A stochastic process $\{X_h\}_{h\in H}$ is called a centered sub-gaussian process for metric $d(\cdot,\cdot)$ on H if $\{X_h\}_{h\in H}$ is a sub-gaussian process for metric $d(\cdot,\cdot)$ and $\mathbb{E}[X_h] = 0,\ \forall h\in H$.

Lemma G.3. *Suppose $\{X_h\}_{h\in H}$ is a centered sub-gaussian process for metric $K\cdot d(\cdot,\cdot)$ on metric space H, where the diameter of H is finite, i.e. $\operatorname{diam}(H) = \sup_{h_1,h_2\in H} d(h_1,h_2) < +\infty$. Then with probability at least $1-\delta$, for any fixed $h_0\in H$, we have*

$$\sup_{h\in H}|X_h - X_{h_0}| \leq 6K\left(8\operatorname{diam}(H)\sqrt{\log\left(\frac{2}{\delta}\right)} + \sum_{k=-\kappa}^{\infty} 2^{-k}\sqrt{\log\mathcal{N}\left(H, d, 2^{-k}\right)}\right)$$

and

$$\sup_{h\in H}|X_h - X_{h_0}| \leq 12K\left(4\operatorname{diam}(H)\sqrt{\log\left(\frac{2}{\delta}\right)} + \int_0^{2\operatorname{diam}(H)} \sqrt{\log\mathcal{N}\left(H,d,\epsilon\right)}\,\mathrm{d}\epsilon\right),$$

where $\kappa \in \mathbb{Z}_+$ *satisfies* $2^{\kappa-1} < \operatorname{diam}(H) \leq 2^{\kappa}$.

Proof [Proof of Lemma G.3] Let $\kappa \in \mathbb{Z}_+$ satisfy $2^{\kappa-1} < \operatorname{diam}(H) \leq 2^{\kappa}$. Define $\epsilon_k = 2^{-k}, k \in \mathbb{Z}, k \geq -\kappa$. Let H_k be the ϵ_k-net of H with metric $d(\cdot,\cdot)$, i.e., $H_k \subset H$ covers H at scale ϵ_k with respect to the metric $d(\cdot,\cdot)$. Clearly $|H_{-\kappa}| = 1$. We take $H_{-\kappa} = \{h_0\}$. Define $\pi_k(h)$ as the closest element of h in H_k under the metric $d(\cdot,\cdot)$. Then $\forall h \in H, \forall N \geq -\kappa, N \in \mathbb{Z}$, we have

$$X_h - X_{h_0} = \sum_{k=-\kappa+1}^{\infty}\left(X_{\pi_k(h)} - X_{\pi_{k-1}(h)}\right) \quad a.s.\ .$$

Thus

$$\sup_{h\in H}|X_h - X_{h_0}| \leq \sum_{k=-\kappa+1}^{\infty}\sup_{h\in H}\left|X_{\pi_k(h)} - X_{\pi_{k-1}(h)}\right| \quad a.s.\ .$$

Consider $P_k = \{X_{\pi_k(h)} - X_{\pi_{k-1}(h)} | h \in H\}$, $|P_k| \leq |H_{k-1}||H_k| \leq |H_k|^2$ and any element in P_k is $K(\epsilon_k + \epsilon_{k-1})$ sub-gaussian. By Hoeffding's inequality and union bound argument, we have

$$\begin{aligned}\mathbb{P}\left(\sup_{X\in P_k}|X| \geq t\right) &= \mathbb{P}\left(\bigcup_{X\in P_k}\{|X| \geq t\}\right)\\ &\leq \sum_{X\in P_k}\mathbb{P}\left(|X| \geq t\right)\\ &\leq 2|P_k|\exp\left(-\frac{t^2}{2K^2(\epsilon_{k-1}+\epsilon_k)^2}\right)\\ &\leq 2|P_k|\exp\left(-\frac{t^2}{18K^2\epsilon_k^2}\right).\end{aligned}$$

Let $2|P_k|\exp\left(-t^2/18K^2\epsilon_k^2\right) = \delta_k \leq 1/2$,
hence $t = \sqrt{18}K\epsilon_k\sqrt{\log(|P_k|) + \log(2/\delta_k)} \leq 3\sqrt{2}K\epsilon_k(\sqrt{\log(|P_k|)} + \sqrt{\log(2/\delta_k)})$. Then with probability at least $1-\delta_k$, we have

$$\begin{aligned}\sup_{X\in P_k}|X| &\leq 3\sqrt{2}K\epsilon_k\left(\sqrt{\log(|P_k|)} + \sqrt{\log(2/\delta_k)}\right)\\ &\leq 6K\epsilon_k\left(\sqrt{\log(|H_k|)} + \sqrt{\log(1/\delta_k)}\right).\end{aligned}$$

Thus, with probability at least $1 - \sum_{k=-\kappa}^{+\infty}\delta_k$, we get

$$\sup_{h\in H}|X_h - X_{h_0}| \leq 6K\sum_{k=-\kappa}^{\infty}2^{-k}\left(\sqrt{\log\mathcal{N}\left(H,d,2^{-k}\right)} + \sqrt{\log\left(1/\delta_k\right)}\right),$$

Let $\delta_k = \delta/2^{k+\kappa+1}$. Then $\sum_{k=-\kappa}^{\infty} \delta_k = \delta$. We have

$$\begin{aligned}\sum_{k=-\kappa}^{\infty} 2^{-k}\sqrt{\log(1/\delta_k)} &= \sum_{k=-\kappa}^{\infty} 2^{-k}\sqrt{\log(2^{k+\kappa+1}/\delta)} \\ &\le \sum_{k=-\kappa}^{\infty} 2^{-k}\sqrt{k+\kappa+1}\sqrt{\log(2/\delta)} \\ &\le 8\,\mathrm{diam}(H)\sqrt{\log(2/\delta)}\ .\end{aligned}$$

Thus,

$$\sup_{h\in H}|X_h - X_{h_0}| \le 6K\left(8\,\mathrm{diam}(H)\sqrt{\log\left(\frac{2}{\delta}\right)} + \sum_{k=-\kappa}^{\infty} 2^{-k}\sqrt{\log\mathcal{N}(H,d,2^{-k})}\right).$$

Since

$$\sum_{k=-\kappa}^{\infty} 2^{-k}\sqrt{\log\mathcal{N}(H,d,2^{-k})} \le 2\int_0^{2^\kappa}\sqrt{\log\mathcal{N}(H,d,\epsilon)}\,\mathrm{d}\epsilon \le 2\int_0^{2\,\mathrm{diam}(H)}\sqrt{\log\mathcal{N}(H,d,\epsilon)}\,\mathrm{d}\epsilon\ ,$$

the lemma is proved. ■

Lemma G.4. *(Theorem 5.1 in De Ryck et al. (2021)) Let $d, s \in \mathbb{N}_+$, $\delta > 0$ and $f \in W^{s,\infty}([0,1]^d)$. There exist constants $\mathcal{C}(d,s,f)$ and $N_0(d) > 0$ such that for every integer $N > N_0(d)$, there exists a tanh neural network $\hat{f}^N$ with two hidden layers, with one width at most $3\lceil s/2\rceil\binom{s+d-1}{d} + d(N-1)$ and the other width at most $3\lceil(d+2)/2\rceil\binom{2d+1}{d}N^d$ (or $3\lceil s/2\rceil + N - 1$ and $6N$ for $d = 1$), such that*

$$\left\|f - \hat{f}^N\right\|_{L^\infty([0,1]^d)} \le (1+\delta)\frac{\mathcal{C}(d,s,f)}{N^s}\ .$$

If $f \in C^s([0,1]^d)$, then it holds that

$$\mathcal{C}(d,s,f) = \frac{(3d)^s}{s!2^s}\|f\|_{W^{s,\infty}([0,1]^d)}, \quad N_0(d) = \frac{3d}{2},$$

and else, it holds that

$$\mathcal{C}(d,s,f) = \frac{\pi^{1/4}\sqrt{s}(5d)^s}{(s-1)!}\|f\|_{W^{s,\infty}([0,1]^d)}, \quad N_0(d) = 5d^2.$$

Moreover, the weights of $\hat{f}^N$ scale as $O(\mathcal{C}(d,s,f)^{-s/2}N^{d(d+s^2)/2}(s(s+2))^{3s(s+2)})$.

Remark G.5. *By Lemma G.4, there exists a constant $C(\delta)$ which is only dependent with δ, such that*

$$|\text{the weights of } \hat{f}^N| \le C(\delta)\mathcal{C}(d,s,f)^{-s/2}N^{d(d+s^2)/2}(s(s+2))^{3s(s+2)}.$$

Lemma G.6. *(Corollary 5.8 in De Ryck et al. (2021)) Let $d \in \mathbb{N}_+$, $\Omega \subset \mathbb{R}^d$ open with $[0,1]^d \subset \Omega$ and let f be analytic on Ω. If, for some $C > 0$, f satisfies that $\|f\|_{W^{s,\infty}([0,1]^d)} \leq C^s$ for all $s \in \mathbb{N}$, then for any $\mathcal{N} \in \mathbb{N}_+$, there exists a one-layer* tanh *neural network $\hat{f}^{\mathcal{N}}$ of width $3\lceil(\mathcal{N} + 5Cd)/2\rceil\binom{\mathcal{N}+(5C+1)d}{d}$ (or $3\lceil \mathcal{N}/2\rceil$ for $d = 1$) such that*

$$\left\|f - \hat{f}^N\right\|_{L^\infty([0,1]^d)} \leq \exp(-\mathcal{N}) \ .$$

Remark G.7. *In De Ryck et al. (2021), the construction of $\hat{f}^{\mathcal{N}}$ in Lemma G.6 uses Lemma G.4 directly. Hence the weights of $\hat{f}^{\mathcal{N}}$ can be derived from Lemma G.4. Then there exists a constant $\tilde{C}$ such that*

$$|\textit{the weights of } \hat{f}^{\mathcal{N}}| \leq \tilde{C}\exp(\frac{\mathcal{N}'^2 + \mathcal{N}' - 3Cd\mathcal{N}'}{2})(\mathcal{N}'(\mathcal{N}'+2))^{3\mathcal{N}'(\mathcal{N}'+2)},$$

where $\mathcal{N}' = \mathcal{N} + 5Cd$. We emphasize that the original literature (De Ryck et al., 2021) does not give this result, but it can be obtained by simple calculations.

References

Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

Luc Bauwens and Nikolaus Hautsch. Modelling financial high frequency data using point processes. In *Handbook of financial time series*, pages 953–979. Springer, 2009.

Pierre Brémaud and Laurent Massoulié. Stability of nonlinear hawkes processes. *The Annals of Probability*, pages 1563–1588, 1996.

Biao Cai, Jingfei Zhang, and Yongtao Guan. Latent network structure learning from high-dimensional multivariate point processes. *Journal of the American Statistical Association*, pages 1–14, 2022.

Jian Cao, Zhi Li, and Jian Li. Financial time series forecasting model based on ceemdan and lstm. *Physica A: Statistical mechanics and its applications*, 519:127–139, 2019.

Minshuo Chen, Xingguo Li, and Tuo Zhao. On generalization bounds of a family of recurrent neural networks. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 1233–1243. PMLR, 2020.

Felix Cheysson and Gabriel Lang. Spectral estimation of Hawkes processes from count data. *The Annals of Statistics*, 50(3):1722 – 1746, 2022. doi: 10.1214/22-AOS2173. URL `https://doi.org/10.1214/22-AOS2173`.

Vinay Kumar Reddy Chimmula and Lei Zhang. Time series forecasting of covid-19 transmission in canada using lstm networks. *Chaos, solitons & fractals*, 135:109864, 2020.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer, 2003.

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer, 2008.

Tim De Ryck, Samuel Lanthaler, and Siddhartha Mishra. On the approximation of functions by tanh neural networks. *Neural Networks*, 143:732–750, 2021.

Nan Du, Yichen Wang, Niao He, Jimeng Sun, and Le Song. Time-sensitive recommendation from recurrent user activities. *Advances in neural information processing systems*, 28, 2015.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1555–1564, 2016.

Michel J Dugas, Patrick Gosselin, and Robert Ladouceur. Intolerance of uncertainty and worry: Investigating specificity in a nonclinical sample. *Cognitive therapy and Research*, 25:551–558, 2001.

Joseph Enguehard, Dan Busbridge, Adam Bozson, Claire Woodcock, and Nils Hammerla. Neural temporal point processes for modelling electronic health records. In *Machine Learning for Health*, pages 85–113. PMLR, 2020.

Guanhua Fang, Ganggang Xu, Haochen Xu, Xuening Zhu, and Yongtao Guan. Group network hawkes process. *Journal of the American Statistical Association*, pages 1–17, 2023.

Mehrdad Farajtabar, Yichen Wang, Manuel Gomez-Rodriguez, Shuang Li, Hongyuan Zha, and Le Song. Coevolve: A joint point process model for information diffusion and network evolution. *Journal of Machine Learning Research*, 18(41):1–49, 2017.

Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*, volume 625. John Wiley & Sons, 2013.

Kunihiko Fukushima. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):322–333, 1969.

Walter Gautschi. How (un)stable are vandermonde systems? *Asymptotic and Computational Analysis*, 1990.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. A distribution-free theory of nonparametric regression. In *Springer Series in Statistics*, 2002.

Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 2015.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

Alan G Hawkes. Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198, 2018.

Alan G Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of applied probability*, 11(3):493–503, 1974.

Seyed Abbas Hosseini, Keivan Alizadeh, Ali Khodadadi, Ali Arabzadeh, Mehrdad Farajtabar, Hongyuan Zha, and Hamid R Rabiee. Recurrent poisson factorization for temporal recommendation. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 847–855, 2017.

Valerie Isham and Mark Westcott. A self-correcting point process. *Stochastic processes and their applications*, 8(3):335–347, 1979.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics*, 51(2):691–716, 2023.

John Frank Charles Kingman. *Poisson processes*, volume 3. Clarendon Press, 1992.

Michael Kohler and Sophie Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231 – 2249, 2021.

Patrick J Laub, Young Lee, and Thomas Taimre. *The elements of Hawkes processes.* Springer, 2021.

Shuang Li, Shuai Xiao, Shixiang Zhu, Nan Du, Yao Xie, and Le Song. Learning temporal point processes via reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Haitao Lin, Lirong Wu, Guojiang Zhao, Pai Liu, and Stan Z Li. Exploring generative neural temporal point process. *arXiv preprint arXiv:2208.01874*, 2022.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

Larry Medsker and Lakhmi C Jain. *Recurrent neural networks: design and applications.* CRC press, 1999.

Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. *Advances in neural information processing systems*, 30, 2017.

Yosihiko Ogata and David Vere-Jones. Inference for earthquake models: a self-correcting model. *Stochastic processes and their applications*, 17(2):337–347, 1984.

Takahiro Omi, Kazuyuki Aihara, et al. Fully neural network based model for general temporal point processes. *Advances in neural information processing systems*, 32, 2019.

Donald H Perkel, George L Gerstein, and George P Moore. Neuronal spike trains and stochastic point processes: I. the single spike train. *Biophysical journal*, 7(4):391–418, 1967.

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.

Anton Maximilian Schäfer and Hans-Georg Zimmermann. Recurrent neural networks are universal approximators. *International journal of neural systems*, 17(04):253–263, 2007.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4), 2020. doi: 10.1214/19-AOS1875.

Frederic Paik Schoenberg. Consistent parametric estimation of the intensity of a spatial–temporal point process. *Journal of Statistical Planning and Inference*, 128(1):79–93, 2005.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

Oleksandr Shchur, Ali Caner Türkmen, Tim Januschowski, and Stephan Günnemann. Neural temporal point processes: A review. *arXiv preprint arXiv:2104.03528*, 2021.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation characterized by number of neurons. *arXiv preprint arXiv:1906.05497*, 2019.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of relu networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157: 101–135, 2022. ISSN 0021-7824.

Namjoon Suh and Guang Cheng. A survey on statistical theory of deep learning: Approximation, training dynamics, and generative models. *arXiv preprint arXiv:2401.07187*, 2024.

Kanchan M Tarwani and Swathi Edem. Survey on recurrent neural network in natural language processing. *Int. J. Eng. Trends Technol*, 48(6):301–304, 2017.

Zhuozhuo Tu, Fengxiang He, and Dacheng Tao. Understanding generalization in recurrent neural networks. In *International Conference on Learning Representations*, 2020.

Utkarsh Upadhyay, Abir De, and Manuel Gomez Rodriguez. Deep reinforcement learning of marked temporal point processes. *Advances in neural information processing systems*, 31, 2018.

Mathukumalli Vidyasagar. *Learning and generalisation: with applications to neural networks*. Springer Science & Business Media, 2013.

Ting Wang, Mark Bebbington, and David Harte. Markov-modulated hawkes process with stepwise decay. *Annals of the Institute of Statistical Mathematics*, 64:521–544, 2012.

Alex Williams, Anthony Degleris, Yixin Wang, and Scott Linderman. Point process models for sequence detection in high-dimensional neural spike trains. *Advances in neural information processing systems*, 33:14350–14361, 2020.

Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv:1702.01923*, 2017.

Shijun Zhang, Jianfeng Lu, and Hongkai Zhao. Deep network approximation: Beyond relu to diverse activation functions. *Journal of Machine Learning Research*, 25(35):1–39, 2024.

Yizhou Zhang, Karishma Sharma, and Yan Liu. Vigdet: Knowledge informed neural temporal point process for coordination detection on social media. *Advances in Neural Information Processing Systems*, 34:3218–3231, 2021.

Zihao Zhou, Xingyi Yang, Ryan Rossi, Handong Zhao, and Rose Yu. Neural point process for learning spatiotemporal event dynamics. In *Learning for Dynamics and Control Conference*, pages 777–789. PMLR, 2022.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer hawkes process. In *International conference on machine learning*, pages 11692–11702. PMLR, 2020.