Improving Interpretability and Predictive Performance of Neural Temporal Point Processes

Paper ID: 3028

Abstract

Deep learning has become an increasingly popular tool for enriching the traditional point process models, which shows great potential for their high model capacity and flexibility, with applications to event sequence learning and prediction. However, these networks often lack clear interpretability enjoyed by the traditional parametric models such as Hawkes process. This paper aims to fill this gap by devising a novel recurrent network based approach called interpretable neural temporal point process (INTPP). The model enjoys both flexibility (thanks to deep networks) and interpretability in terms of the clear dependency on history. Specifically, we connect INTPP with traditional point processes such as Hawkes process and analyze the meaning of the network layers' output. The results show that the INTPP is really understanding the intrinsic meaning and generation mechanism of the sequence. Every intensity parameter of the INTPP is meaningful in reality. The parameters of Hawkes processes (including multidimensional Hawkes process) can also be evaluated and finally converge by training the recurrent model. Besides, a stable method is devised for neural point process based prediction with uncertainty modeling. Moreover, our analytical method is an order faster than the widely used numerical expectation integration technique. The experimental results verify the efficacy of our techniques.

Introduction and Related Work

Temporal point process (TPP) has been a popular and principled tool for modeling and predicting event data in continuous time space. Traditional methods design and employ different intensity functions or density functions for describing the event occurrence rate over time, with explicit (Xu et al. 2016; Xiao et al. 2016; Zhou, Zha, and Song 2013) or implicit (Eichler, Dahlhaus, and Dueck 2017) parametric forms. Most of these works have a time-varying intensity function, being able to predict futures and analyze the influence among events or their markers e.g. types of events for multi-dimensional TPP (Liniger 2009; Zhou, Zha, and Song 2013; Wu et al. 2018b). Neural point process models (Mei and Eisner 2018) are recently becoming increasingly popular, which can achieve good prediction results (Du et al. 2016; Xiao et al. 2017b; Mei and Eisner 2018). Most of them are based on recurrent neural networks and non-linearity is fulfilled by the hidden layers' output. The paper (Xiao et al. 2017a) proposes an interesting method to model the deep point process with Generative Adversarial Networks (Goodfellow et al. 2014), which is not based on intensity or density for the point process but mainly evaluate events with the distance between event sequences. However, the model cannot give the result in probability view and lacks the analyzability of historical events.

Under TPP's probabilistic framework, (Du et al. 2016; Xiao et al. 2017b) design the intensity or density for the point process and apply maximum likelihood estimation (MLE) to estimate the network's parameters. While (Yan et al. 2018) uses extra techniques such as Generative Adversarial Networks (Goodfellow et al. 2014) and Reinforcement Learning (Upadhyay, De, and Rodriguez 2018; Li et al. 2018; Wu et al. 2019) to improve the robustness. Event imputation has also been recently considered in (Mei, Qin, and Eisner 2019). In these methods, intensity function still play a central role for model learning. Moreover, existing neural point process models lack interpretability, in contrast to the traditional parametric TPP models. Recently (Wang et al. 2017b; Xiao et al. 2019) mitigate such issues with attention model, while the gap for interpretability is still not well filled and remains open. This paper particularly cares: i) can the parameters of the neural point process also have practical meaning as the traditional models e.g. Hawkes process (Hawkes 1971)? ii) can the neural point process models be directly used for some analysis related to causal analysis? We aim to devise a more interpretable neural point process model as such the resulting models can enjoy both flexibility and interpretability.

In this paper, we first identify that most existing neural point process models (Saha et al. 2018; Wu et al. 2018a; Upadhyay, De, and Rodriguez 2018; Yang, Cai, and K.Reddy 2018) follow a recurrent network based design and the intensity function is mostly modeled by a powerful and general form (see Eq. 5) which can be dated back to the seminal work called Recurrent Marked Temporal Point Process (RMTPP) proposed in (Du et al. 2016). Such a formula can also incorporate the popular piecewise Poisson based models (Huang, Wang, and Mak 2019; Li et al. 2018) as a special case. Then we make an impor-

tant observation that Eq. 5 has its fundamental theoretical flaw when $w < 0$, which is not suitable to fit and predict events as in traditional parametric forms e.g. Hawkes processes (Hawkes 1971). In that case, the expectation of next time prediction will be associated with meaningless deviations and the prediction can be earlier than the history events, making the model unrealistic. We solve this issue by devising a principled approach and the important byproduct is that our neural point process model shows strong interpretability with natural connection with the well-established Hawkes process. The paper contributes:

1) Improving the interpretability of neural point process models, specifically:

i) We propose INTPP, an interpretable neural point process, whose intensity function satisfies the theoretical soundness for TPP (Rasmussen 2018). This also shows and corrects the flaws in existing recurrent network based TPP models e.g. RMTPP (Du et al. 2016). Interestingly given a specif form of the intensity, our INTPP can be exactly interpreted as a Hawkes process (HP) such that the corresponding HP's coefficients can be estimated during the learning of INTPP.

ii) We extend the above model to the Multi-dimensional INTPP (M-INTPP). Its log-likelihood is derived in Eq. 28, being slightly different from uni-dimensional INTPP. M-INTPP can also find its natural connection with multi-dimensional Hawkes process, whereby an infectivity matrix can be modeled and estimated to evaluate the influence of different dimensions, showing INTPP's interpretability.

2) Improving the prediction flexibility of point process models, specifically: a sampling based prediction method is devised for intensity-based RNN model such as RMTPP and INTPP, which is more stable and faster compared with expectation integration by Eq. 8. With the sampling method, Median Prediction, Exponential Expectation prediction (EEP), Reference Sampling and Interval Prediction can be used to predict next events with a flexibility of uncertainty modeling. In contrast, existing RNN based models only provide point estimations and has difficulty in uncertainty modeling.

3) Competitive performance. In the experiments, INTPP performs the best compared to RMTPP and other traditional models by different metrics such as RMSE and interval test. Besides, our sampling method, especially Reference Sampling, is more stable and faster than numerical integration based expectation.

Preliminaries and Analysis

In this section, we first introduce the basics for TPP and its neuralized version e.g. RMTPP without loss of generality. Then we show the theoretical flaw of existing neural TPP models, which lies the technical foundation of our approach.

Temporal Point process. We give background to ease the later presentation. Temporal point process is a principled framework for modeling events in continuous time space. It captures the temporal dynamics by the conditional intensity $\lambda^*(t)$. Within a short time window $[t, t+dt)$, $\lambda^*(t)$ represents the occurrence rate of a new event given history $\mathcal{H}_t$: $\lambda^*(t) = \frac{P(N(t+dt)-N(t)=1|\mathcal{H}_t)}{dt}$ where $N(t)$ is the counting process and $*$ reminds it is history dependent. Given the conditional intensity $\lambda^*(t)$, the cumulative probability for next event $j+1$ can be specified:

$$F^*(t) = 1 - S^*(t) = 1 - \exp\left(\int_{t_j}^{t} -\lambda^*(\tau)d\tau\right), \quad (1)$$

where $S^*(t)$ is the survival function, for the probability that no new event has occurred up to time t since t_j. With the cumulative probability, the density function can be specified:

$$f^*(t) = \lambda^*(t)\exp\left(\int_{t_j}^{t} -\lambda^*(\tau)d\tau\right), \quad (2)$$

By using the density of point process, one can get the likelihood to estimate the parameters (model learning). In particular, for the density function, cumulative probability function and survival function, they share a common term which is an important function called **intensity measure**:

$$\Lambda(t_j, t) = \int_{t_j}^{t} -\lambda^*(\tau)d\tau. \quad (3)$$

This is also exactly the cumulative function of intensity. The intensity measure has its physical meaning which presents the expectation of event number between $[0, t]$ and it is used in objective for model learning as will be shown later.

In TPP models, various forms of conditional intensity $\lambda^*(t)$ and density function $f^*(t)$ are modeled to capture different temporal features, which play a central role.

Recurrent Poisson Process (RPP). A simple way to design the recurrent intensity is given a history-based constant before next event's occurrence: $\lambda^*(t) = \sigma(\phi(\mathbf{h_j}))$, which is equivalent to exponential distribution assumption for density. Here σ is non-linear transfer function and ϕ is layer projection for RNN outputs $\mathbf{h_j}$. Some works (Li et al. 2018; Huang, Wang, and Mak 2019) use RPP to model the intensity, which can get a good result. However, the assumptions of RPP do not meet people intuition that the intensity changes over time and the model lacks the value of analysis. For example, it is not known whether the event is motivating or suppressing future events.

Recurrent Time-Varying Point Process. Different with a piecewise constant intensity for RPP, many works focus on time-varying deep point process whose intensity changes over time. We generalize to the time-varying intensity:

$$\lambda^*(t) = g\big(\phi(\mathbf{h_j}(t), t)\big), \quad (4)$$

where g is non-linear transfer function such as $max\{0, \cdot\}$, $exp(\cdot)$ and ϕ is a projection for RNN outputs and t.

We begin with a brief retrospection on the Recurrent Marked Temporal Point Process (RMTPP) (Du et al. 2016), which in fact adopts a common intensity form in neural point process literature (Upadhyay, De, and Rodriguez 2018; Saha et al. 2018). Hence our following analysis and results are also general, to a large extent which is similiar to another concurrent work (Wang et al. 2017a).

In (Du et al. 2016), recurrent neural networks (RNNs) (Elman 1990) are used to model the point process with concatenating the time information $W^t t_j$ and marker embedding $\mathbf{y}_j = \mathrm{W}_{em} y_j + b_{em}$ as input vector. Then for output of

RNN model i.e. $\mathbf{h}$, the intensity is specified as:

$$\lambda^*(t) = \exp(\mathbf{v}^\top \cdot \mathbf{h}_j + w(t - t_j) + b) \tag{5}$$

where $\mathbf{v}^\top \cdot \mathbf{h}_j$ denotes the accumulative past influence, $w(t - t_j)$ emphasizes the influence of current event j and b gives a base level. The intensity of RMTPP exactly is a special case for Eq. 4. For conciseness, we set $l_j = \mathbf{v}^\top \cdot \mathbf{h}_j + b$ to represent past influence and call it **deep outputs** in the rest of this paper. Then the conditional density function $f^*(t)$ for $t > t_j$ can be analytically written by:

$$f^*(t) = \lambda^*(t)\exp\left(-\int_{t_j}^{t} \lambda^*(\tau)d\tau\right) \tag{6}$$
$$= \exp\left(l_j + w(t - t_j) + \frac{\exp(l_j) - \exp\left(l_j + w(t - t_j)\right)}{w}\right)$$

An important advantage of RMTPP compared with other deep point process is that it can simply analyze whether the events trigger other events or suppress other events in overall, which has a certain explanatory value.

Besides, there exists another kind of recurrent point processes(Mei and Eisner 2018) which extends RNN output $\mathbf{h}(t)$ to be time-varying and the corresponding intensity is given as:

$$\lambda^*(t) = g(\mathbf{v}^\top \cdot \mathbf{h}_j(t)). \tag{7}$$

However, the parameters related to time t of this model are not analytic to analyze the influence between events.

Then the prediction can be estimated via its expectation:

$$\hat{t}_{j+1} = \int_{t_j}^{+\infty} t \cdot f^*(t)dt \tag{8}$$

However, the integration in Eq. 8 has no analytic solution for expectation of t_{j+1} and thus (Du et al. 2016) has to resort numerical integration which can be slow and inaccurate.

Intensity Constraints. In fact, the parameterized conditional intensity cannot be in arbitrary form. It need to satisfy some constraints to guarantee the mathematical properties as indicated in the following theorem.

Proposition 1. ***(Rasmussen 2018)*** *A conditional intensity $\lambda^*(t)$ uniquely defines a non-terminating point process if it satisfies the following conditions for any $\{t_1, t_2, \ldots, t_j\}$ and $t > t_j$: 1) $\lambda^*(t)$ is non-negative and integrable in the interval $[t_j, \infty)$; 2) $\int_{t_j}^{t} \lambda^*(\tau)d\tau \to \infty$ for $t \to \infty$.*

The point process is non-terminating, which means that the events will occur continuously. The condition 2 in above proposition is necessary to guarantee the probability property that $F^*(t) \to 1$ for $t \to \infty$. In other words, if the condition 2 is not satisfied for a TPP with intensity $\lambda^*(t)$, then its cumulative probability will not converge to 1 (i.e. $\lim_{t\to\infty} F^*(t) < 1$) and some values such as expectation and variance of time t_{j+1} will lose their theoretical basis.

Fundamental flaw. To analyze the intensity in Eq. 5, one can find that when $w > 0$, the intensity increases until the next event occurs, whose form is similar to self-correcting point process (Isham and Westcott 1979) and satisfies Pro 1 as non-terminating point process.

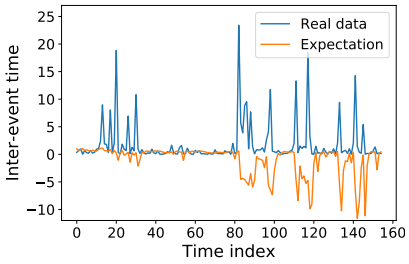


Figure 1: Inter-event time for real testing data and prediction with expectation by Eq. 8 when fitting Hawkes process by RMTPP. Note many points in orange is below zero which means their prediction is earlier than the current time and meaningless, in contrast to the real interval.

However, when $w < 0$, there is an issue with Eq. 5. Consider its intensity measure defined in Eq. 3 for the next event:

$$\Lambda(t_j, t) = \frac{\exp\left(l_j + w(t - t_j)\right) - \exp(l_j)}{w} \tag{9}$$

when $t \to \infty$, we obtain $\lim_{t\to\infty} \Lambda(t_j, t) = -\frac{1}{w}\exp(l_j) < \infty$ which means condition 2 of Theorem 1 is not satisfied. Thus when $w < 0$, the point process with intensity Eq. 5 may terminate at the current event for $\int_{t_j}^{\infty} f^*(t)dt < 1$, which does not satisfy the implied assumptions in RNN model. The limit of Eq. 1 can be written as

$$\overline{p} = \int_{t_j}^{\infty} f^*(t)dt = 1 - \exp\left(-\frac{1}{w}e^{l_j}\right) < 1, \tag{10}$$

which means that the point process continues with probability $\overline{p}$, and terminates with probability $1 - \overline{p}$.

When predicting next event time, the density by Eq. is not strictly defined, which implies the deviation of next time. For the expectation of t_{j+1} in Eq. 8, it can be analyzed with the following form:

$$\mathrm{E}[t_{j+1}] = \underbrace{\int_{t_j}^{+\infty} t \cdot f^*(t)dt}_{\text{with probability } \overline{p}} + \underbrace{0 \cdot (1 - \overline{p})}_{\text{with rest probability } 1-\overline{p}}, \tag{11}$$

which means the expectation is not only for $[t_j, \infty)$, but also for 0 with the probability $1 - \overline{p}$. It really causes the meaningless deviation for next time prediction. Fig. 1 shows the inter-event time ($\Delta_i = t_{i+1} - t_i$) between current event time t_i to the next one t_{i+1} either by real data and by integration based prediction (Eq.8) in the experiments of Hawkes process data with RMTPP model. Interestingly when the ground truth Δ_i is big, the predicted one can also be more smaller than zero, indicating the consequence of the violation of Proposition 1.

Besides, we also cannot guarantee that intensity with Eq. 7 satisfies Prop. 1 and the corresponding expectation is unbiased for another time-varying neural point process.

Modeling Interpretable Neural TPP

Now we propose the interpretive model INTPP which also involves popular techniques to enhance its practical utility.

Intensity Formulation for INTPP

In contrast to Eq. 5, we propose our theoretically-sound intensity in its general form as follows:

$$\lambda^*(t) = g(\phi(\mathbf{h_j}, t)) + c, \tag{12}$$

where $c > 0$ as the base intensity. Without loss of generality, in this paper we set $g = \exp(\cdot)$ and $\phi(\mathbf{h}_j, t) = (\mathbf{v}^\top \cdot \mathbf{h}_j + w(t - t_j) + b$ to make an embodiment in the following form:

$$\lambda^*(t) = \exp\left(l_j + w(t - t_j)\right) + c \tag{13}$$

where $l_j = \mathbf{v}^\top \cdot \mathbf{h}_j + b$ is the deep output to represent accumulative past influence based on RNN. This leads to the following intensity measure,

$$\Lambda(t_j, t) = \frac{1}{w}\exp\big(l_j + w(t - t_j)\big) - \frac{1}{w}\exp(l_j) + c(t - t_j). \tag{14}$$

When $t \to \infty$, $\Lambda(t_j, t) \to \infty$ for all values of w. Also the conditional density function for next event can be written by:

$$\begin{aligned} f^*(t) = &\exp\left(l_j + w(t - t_j) + c\right) \cdot \exp\Bigg(\frac{1}{w}\exp(l_j) \\ &- \frac{1}{w}\exp\big(l_j + w(t - t_j)\big) - c(t - t_j)\Bigg). \end{aligned} \tag{15}$$

Then we can compute the expectation of next time with Eq. 8 in a closed form that can be more stable and without bias.

Parameter Learning

For the joint density for both time and marker, we assume that the next marker is conditioned on next event time, which can be written as

$$f^*(t, k) = f^*_{\text{mark}}(y|t) \cdot f^*_{\text{time}}(t), \tag{16}$$

It means that the marker is not only dependent on the history, but also related to the time. Then for marker prediction, given the learned representation $\mathbf{h}_j$, the marker is learned with a multinomial distribution:

$$f^*_{\text{mark}}(y_{j+1} = k) = \frac{\exp\left(\mathbf{V}^y_{k,:}\mathbf{h}'_j + b^y_k\right)}{\sum_{k=1}^{K}\exp\left(\mathbf{V}^y_{k,:}\mathbf{h}'_j + b^y_k\right)} \tag{17}$$

where K is the number of markers, $\mathbf{V}^y_{k,:}$ is the k-th row of matrix $\mathbf{V}^y$ and $\mathbf{h}'_j = Concat([\mathbf{h}_j, t_{j+1} - t_j])$. when training INTPP, we use real next event in $\mathbf{h}'_j$ to improve marker training and when predicting markers, we predict the event times first and put it in the model to get a better marker prediction.

Then we exploit MLE to learn the parameters.

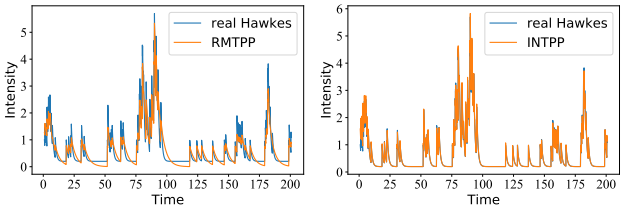
$$L_{time} = \sum_i \sum_j \log(f^*_{\text{time}}(t^i_{j+1})), \tag{18}$$

where $\mathcal{C} = \{\mathcal{S}^i\}_{i=1}^m$ is the sequence set for $\mathcal{S}^i = \{(t^i_j, y^i_j)_{j=0}^n\}$. For the estimation of markers, its loss is:

$$L_{mark} = \sum_i \sum_j \log(f^*_{\text{mark}}(y^i_{j+1}|\mathbf{h}_j, t^i_{j+1})) \tag{19}$$

where $P(y^i_{j+1}|\mathbf{h}_j)$ is the probability that next marker occur mentioned in Eq. 17. At last, we can get the final loss

$$L(\mathcal{C}) = L_{\text{mark}} + L_{\text{time}} \tag{20}$$



(a) RMTPP (b) INTPP

Figure 2: Intensity function computed for testing data.

Interpretation of INTPP

INTPP: a more general intensity. Different from other neural point process models (Du et al. 2016; Mei and Eisner 2018; Xiao et al. 2019), INTPP can be interpreted as a general form of point process and specifically we explore how the history affects the intensity in the form of Eq. 13.

For $t > t_j$, the intensity in Eq. 13 can be transformed to:

$$\lambda^*(t) = c + L_j(\Theta|\mathcal{H}_{t_j}) \cdot \exp(wt) \tag{21}$$

where w is the decaying or increasing parameter of exponential kernel and c is the base intensity which is independent of history $\mathcal{H}_{t_j}$. The weight of exponential kernel satisfies that $L_j(\Theta|\mathcal{H}_{t_j}) = \exp(l_j - w \cdot t_j)$, which is based on event history $\mathcal{H}_{t_j}$(including t_j) and Θ denotes the model parameters.

It can be found that the deep outputs l_j actually affects the weight of exponential kernel and the history does not affect the exponential parameter w and base intensity c while predicting the next events.

Connection to the Hawkes process. Although the intensity seems to be as a quite simple form, while in fact INTPP with its intensity in the form of Eq. 13 can be interpreted in the view of Hawkes process. For a vanilla Hawkes process, the intensity can be written as:

$$\lambda^*(t) = \mu + \alpha \sum_{t_k < t} e^{-\beta(t - t_k)}. \tag{22}$$

Focusing on $t > t_j$, the intensity can be written exactly as

$$\lambda^*(t) = \mu + \big(\alpha \sum_{k=1}^{j} e^{\beta \cdot t_k}\big)e^{-\beta t}, \tag{23}$$

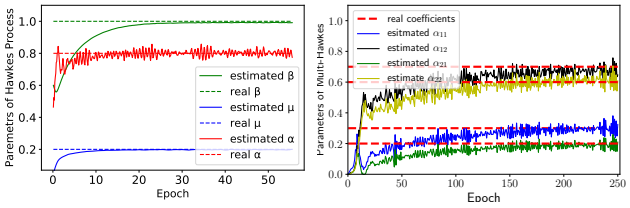
One can find the exact identify between Eq. 23 and Eq. 21. It means that the Hawkes process is also based on three vital terms: base intensity, exponential kernel and the history-dependent weight of the kernel. So if INTPP is perfectly fitted, Hawkes process is a special case of INTPP.

Furthermore, there are three connections between Hawkes process and INTPP's hidden outputs:

$$\begin{cases} c = \mu, w = -\beta, \\ L_j = \exp(l_j - w \cdot t_j) = \alpha \sum_{k=1}^{j} e^{\beta \cdot t_k}, \end{cases} \tag{24}$$

We can find that when fitting Hawkes process, deep output l_j is exactly fitting the following term, which can be parameterized with the parameters w, α, β in Hawkes process:

$$l_j = w \cdot t_j + \log(\alpha \sum_{k=1}^{j} e^{\beta \cdot t_i}). \tag{25}$$



(a) 1-D Hawkes process (b) Muti-Hawkes process

Figure 3: Estimating 1-D Hawkes process and the infectivity matrix of multi-dim Hawkes process with deep outputs l_j.

Hence the output l_j is explanatory in the view of Hawkes process (or more generally traditional parametric models in the form of Eq. 21).

Estimating Hawkes parameters via INTPP Interestingly, based on the above analysis, one can even evaluate the parameters of Hawkes process by training INTPP. According to Eq. 24, one can easily find that

$$\begin{cases} \hat{\mu} = c, \hat{\beta} = -w, \\ \hat{\alpha} = \frac{1}{n}\sum_j \frac{\exp(l_j)}{\sum_{k=1}^{j}\exp(w\cdot(t_j - t_k))}, \end{cases} \tag{26}$$

Fig. 3(a) is the result of evaluating Hawkes parameters by training INTPP. Note that $\hat{\mu}, \hat{\beta}, \hat{\alpha}$ can converge to the ground truth ($\hat{\alpha}$ is oscillatory because of different batch inputs).

Multi-dimensional Extension

INTPP can be extended to a multi-dimensional case. For D dimensional INTPP and its next time $t > t_j$, the intensity is:

$$\lambda_d(t) = c_d + \exp\left(l_j^d + w_{dd_j}(t - t_j)\right) \tag{27}$$

where $d = 1, 2, \ldots, D$ and $l_j = \mathbf{v}_d^\top \cdot \mathbf{h}_j + b$, here $\mathbf{h}_j$ is output of RNN and $\mathbf{v_d}$ is a liner layer projection of $\mathbf{h}_j$, b is a constant. So after the layer of RNN, there exist D dimensional intensity function for INTPP and for its optimization, the log-likelihood can be written as

$$\begin{aligned} \mathcal{L} &= \sum_{j=1}^{n}\log(\lambda_{d_j}(t_j)) - \sum_{d=1}^{D}\int_{t_0}^{t_n}\lambda_d(t)dt \\ &= \sum_{j=1}^{n}\log(\lambda_{d_j}(t_j)) - \sum_{d=1}^{D}\sum_{j=1}^{n}\Lambda_d(t_{j-1}, t_j), \end{aligned} \tag{28}$$

where Λ_d is the intensity measure for d dimension and the equation is equivalent to Eq. 18 when $D = 1$.

The Multi-INTPP can also interpreted by Multi-Hawkes processes. For a Multi-Hawkes process with its intensity that

$$\lambda_d(t) = \mu_d + \sum_{t_k < t}\alpha_{dd_j}e^{-\beta(t - t_k)}, \tag{29}$$

We can also get the interpretation of the deep outputs l_j^d with the view of multi-Hawkes that

$$l_j^d = w_{dd_j}t_j + \log\left(\sum_{k=1}^{j}\alpha_{dd_j}e^{-\beta(t_j - t_k)}\right). \tag{30}$$

As the product of the model, we can also estimate infectivity matrix which users may be interested in by Eq. 30 with linear regression. While Fig. 3(b) shows the estimation of infectivity matrix for 2D-Hawkes process while training INTPP, which can also get the convergence to real coefficients.

Predicition Uncertainty Modeling based on Intensity Measure

Our proposed INTPP has another orthogonal advantage in terms of uncertainty modeling for prediction. This is in contrast to the majority of existing TPP works either only providing point level deterministic prediction (for those RNN based models) or can only provide unstable estimation via numerical integration (for traditional parametric models).

The intensity measure in Eq. 3 satisfies the TCT theorem.

Time Change Theorem (TCT) (Brown et al. 2002)

Theorem 1. *For a temporal point process $\{t_1, t_2, ..., t_j, ...\}$ with conditional intensity $\lambda^*(t)$, the integrated conditional intensity in $(t_j, t_{j+1}]$, i.e. intensity measure, has the form:*

$$\Lambda(t_j, t_{j+1}) = \int_{t_j}^{t_{j+1}}\lambda^*(t)dt.$$

Then $\Lambda(t_j, t_{j+1})$ obeys the exponential distribution with parameter 1.

The theorem above reveals the essence of numerical value for intensity function and intensity measure. It is often used in point process based analysis such as QQ-plot for testing, Least Square (LS) loss for inference etc. Here we use the property to obtain the predictions of next time.

Prediction with Sampling

According to Theorem 1, the intensity measure $\Lambda(t_j, t_{j+1})$ is exponentially distributed with the parameter 1. Then given a current event time t_j and a random number $u \sim \text{uniform}(0, 1)$, we can easily derive the following equation:

$$\Lambda(t_j, t) = -\log(1 - u), \tag{31}$$

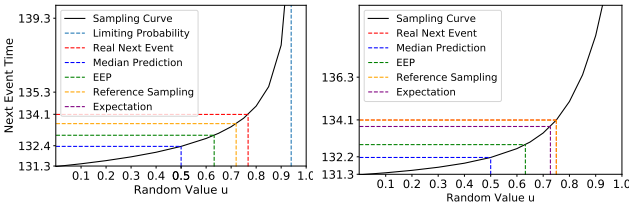
where $-\log(1-u)$ is the sampling of exponential distribution and Λ_j^{-1} is the pseudo inverse of $\Lambda_j(t)$. The above formula paves the road to predicting time. For example, with the intensity of RMTPP in Eq. 5, we can get a simple solution with sampling methods (recall $u \sim \text{uniform}(0, 1)$):

$$\hat{t}_{j+1}^u = t_j + \frac{\log\left(\exp(l_j) - w\log(1 - u)\right) - l_j}{w}. \tag{32}$$

It can be seen that when $w \to 0^+$, the sampling $\hat{t}_{j+1}^u = \frac{-\log(1-u)}{\exp(l_j)}$, which is exactly the simulation of Poisson process with the intensity e^{l_j}.

Besides, Eq. 32 has an implicit condition that $\exp(l_j) - w\log(1 - u) > 0$, which is equal to the inequality that $u < 1 - \exp\left(-\frac{1}{w}e^{l_j}\right)$. In other words, when $u > 1 - \exp\left(-\frac{1}{w}e^{l_j}\right)$, the next event does not exist according to Eq. 32, which is consistent with the drawback of RMTPP. In other words, we can find the limitation of sampling u_{limit} for RMTPP in contrast to the fact $u \sim \text{uniform}(0, 1)$:

$$u_{\text{limit}} = 1 - \exp\left(-\frac{1}{w}\exp(l_j)\right). \tag{33}$$



(a) RMTPP (b) INTPP

Figure 4: Sampling Curve for RMTPP and INTPP with different prediction methods (zoom in for better view).

Fig. 4 shows this issue (the predicted next event time is earlier than current time). We discuss more in experiments.

Prediction for INTPP

Different from the intensity measure of RMTPP, we can not get the analytical solution about t_{j+1} in Eq. 31 for INTPP. With the intensity measure in Eq. 14, we set $\Delta = t - t_j$ and let $K(\Delta) = \Lambda(t_j, t)$ which can be written by:

$$K(\Delta) = \frac{\exp(l_j + w\Delta) - \exp(l_j)}{w} + c\Delta + \log(1-u) \quad (34)$$

Obviously we can not get the inverse of K directly. However, given a certainty u, we can find that $K(0) < 0$, $K(\frac{-\log(1-u)}{c}) > 0$ and $K'(\Delta) = \exp(l_j + w\Delta) + c > 0$. Hence Δ (or equally t_{j+1}) has the unique solution, and Newton's method can be used to calculate the unique root. For $K''(\Delta)$ exist and is continuous in $(0, \frac{-\log(1-u)}{c})$, so there is quadratic convergence for Newton's iterative method by:

$$\Delta^{k+1} = \Delta^k - K(\Delta^k)/K'(\Delta^k) \quad (35)$$

Then we can get $\hat{t}^u_{j+1} = t_j + \Delta^*$ where Δ^* is convergence of $\{\Delta^k\}$. We exploit Newton's iterations in Eq. 35 with quadratic convergence and set $u = 0.5$ to predict the next event time as median prediction.

Median prediction, EEP and Reference Sampling

Note Eq. 31 can get the prediction with sampling, which simulates uncertainty. In fact prediction uncertainty is often welcomed such as the expectation in Eq. 8.

For uncertainty based prediction, median prediction is a simple method with a certainty sampling value $u_{\text{MP}} = 0.5$ in Eq. 31. For example, with median sampling, a stable and analytical prediction result can be obtained in Eq. 32, which need no numerical integration otherwise can be time-consuming.

Besides, Exponential Expectation prediction (EEP) is another certainty prediction by setting $-\log(1-u) = 1$:

$$u_{\text{EEP}} = 1 - \exp(-1) \quad (36)$$

The sampling with u_{EEP} is equal to choosing the expectation of next time intensity measure and it is easy to obtain the result $\hat{t}^u_{j+1}$, as another deterministic prediction.

At last, both median prediction and EEP are computed based on different assumptions. However, it cannot guarantee the best way of sampling. Based on training data, we can

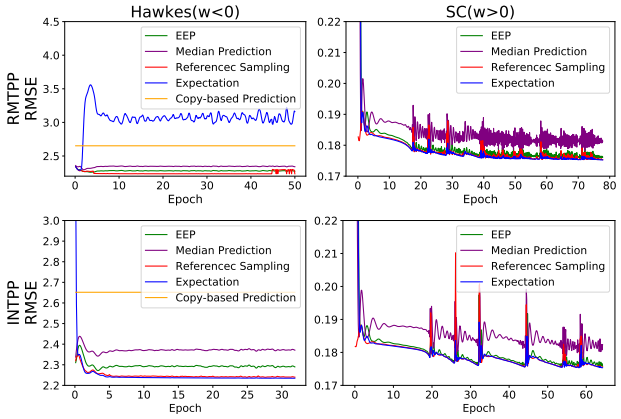


Figure 5: Prediction of RMTPP and INTPP using data generated by HP and SC. It shows RMTPP with expectation is not suitable for prediction when $w < 0$ but INTPP can.

Table 1: Running time for predicting per around 10 thousand events (in seconds).

Predicting Methods	RMTPP			INTPP		
	Hawkes	SC	ACD	Hawkes	SC	ACD
Median prediction	0.125	0.137	0.133	0.680	0.712	0.5902
EEP	0.121	0.148	0.136	0.855	0.608	0.6179
Reference Sampling	2.180	2.178	2.195	14.03	11.21	11.13
Expectation	72.05	102.7	144.6	75.61	130.3	189.6

evaluate the Reference Sampling u_{RS} for testing defined by:

$$u_{\text{RS}} = \arg\min_{u \in (0,1)} \sum_{j=1}^{n} \left(t_{j+1} - \hat{t}_{j+1}(u)\right)^2 \quad (37)$$

where $\hat{t}_{j+1}(u) = \hat{t}^u_{j+1}$ is defined in Eq. 31. It can be proved that the above objective is convex as Λ_j is convex for both RMTPP and INTPP and thus we can get the optimal solution in simple ways such as Bisection method. Using u_{RS} to sample for test data, we can obtain more accurate results.

Interval Prediction and Interval Test

In fact the probability that the event occurs at a single time point is almost 0. With the sampling method, one can get interval prediction with probability p in $[\tau_1, \tau_2]$. For instance, given the first event time τ_1 and the probability p, τ_2 can be calculated $\tau_2 = \hat{t}^{u_2}_{j+1}$ where u_2 can be specified as:

$$u_2 = p - \left(1 - \exp\left(-\Lambda_j(\tau_1)\right)\right) \quad (38)$$

Here $u_1 = 1 - \exp\left(-\Lambda_j(\tau_1)\right)$ corresponds to the first sampling value of t_1 and Λ_j is mentioned in Eq. 31.

So given two of the three prior values τ_1, τ_2, p, we can calculate the rest and give the probability p that the next event happens in the interval $[\tau_1, \tau_2]$ according to the model.

Besides, by interval prediction, Interval Test can be used to test the model's uncertainty modeling capability by separately setting u as $(0.05, 0.95), (0.1, 0.9), \ldots, (0.45, 0.55)$, with corresponding theoretical probability: $0.9, 0.8, \ldots, 0.1$.

Experiments

Compared Methods For predicting time evaluation, we compare INTPP with classical point process methods:

Copy-based Prediction (CBP). One simple and memory based way is set $t_{j+1} = t_j$ as the copy-based prediction.

Homogeneous Poisson Process. When the intensity is a constant value, the point process is well-known as Homogeneous Poisson Process which is independent on history.

Hawkes Process (HP) (Hawkes 1971). It is a classical point process with intensity $\lambda^*(t) = \mu + \alpha \sum_{t_i<t} e^{\beta(t-t_i)}$, as used to fit temporal event data with EM algorithm (Yan et al. 2016; Eichler, Dahlhaus, and Dueck 2017).

Self-Correcting Process (SC) (Isham and Westcott 1979). Self-Correcting process is a regular temporal point process with the intensity $\lambda(t) = \exp \mu t - \sum_{t_i<t} \alpha$, which has two parameters μ and α to be learned.

Autoregressive Conditional Duration (ACD) (Engle and Russell 1998). Its intensity function is $\lambda(t) = \gamma_0 + \sum_{j=0}^{m} \alpha_j d_{i-j}$ for event t_{i+1}, where d_{i-j} is the duration between the j^{th} event and i^{th} event.

For comparing marker predictions or both time and markers with INTPP, we choose the following baselines:

Majority Prediction. It is also well-known as the 0-order Markov Chain (MC-0), where at each time step, which is applied to predict the most popular marker regardless of the history. We use MC-0 as the first baseline.

Markov Chain. MC-0 can also be generalized to MC-n with Markov models with varying orders. We set $n = 1, 2$ to compare with other marked prediction model.

RMTPP (Du et al. 2016). The state-of-the-art TPP model. The model is based on RNN and the size of hidden layers of RMTPP is set to be same as ours.

Synthetic Data To show the robustness and effectiveness of the proposed INTPP and our prediction methods, we simulate data with the following point process in line with the protocol in (Du et al. 2016):

1) Hawkes Process. Given the conditional intensity $\lambda(t) = \mu + \alpha \sum_{t_j<t} e^{-\beta(t-t_j)}$ with $\mu = 0.2, \alpha = 0.8, \beta = 1$.

2) Self-correcting Process. The conditional intensity function of Self-correcting Process is given by $\lambda(t) = \exp(\mu t - \sum_{t_i<t} \alpha)$ with $\mu = 1$ and $\alpha = 0.2$.

3) Auto-regressive Conditional Duration. ACD has the intensity $\lambda(t) = \left(\mu_0 + \sum_{j=1}^{m} \alpha d_{i-j}\right)^{-1}$ for event t_{i+1}, where d_{i-j} is the duration between the j^{th} event and i^{th} event. Here we set $m = 2$, $\mu = 0.2$, $\alpha = 0.25$.

Specifically, 0.1 million events (640 sequences) are simulated in the experiments for each generated dataset and we use around 90% of them for training and the rest for testing.

Real World Data Different datasets are tested including:

1) NYSE. A book order dataset from NYSE with 0.7 million transactions in total(Du et al. 2016). Each transaction contains the time (in millisecond) and the possible action (buy/sell). We cut sequence with 1929600 events for training and 482400 for testing. The type of actions is treated as a marker and we predict when and which action will occur.

2) ATM. The dataset is mainly for the predictive ATM maintenance problem, which is comprised of the event logs involving error reporting and failure tickets (Xiao et al. 2017b). The type of ATM has 2 main type 'ticket' and 'error' within 7 moths in America. Besides the 'error' is divided into 6 subtypes: printer (PRT), cash dispenser module (CNG), internet data center (IDC), communication part (COMM), printer monitor (LMTP), miscellaneous. We deal them with 7 types to test the model.

3) MIMIC II. The medical dataset (Saeed et al. 2002) collects de-identified clinical visit records of ICU patients for 7 years, which has been shared. We select samples of MIMIC II with more than three records, which exists 487 sequences for training and 57 sequences for testing.

Table 2: RMSE results with different methods.

Methods	Hawkes	SC	ACD	NYSE	ATM	MIMIC
CBP	2.551	0.270	1.564	1.561	1.665	1.025
OPT	0.757	0.175	1.109	-	-	-
Poisson	2.235	0.182	1.204	1.552	1.658	0.944
Hawkes	2.242	0.182	1.170	1.554	1.526	0.932
SC	2.354	0.175	1.204	1.812	1.672	1.068
ACD	2.675	0.181	1.110	1.477	1.672	1.150
RMTPP	2.279	0.176	1.109	1.422	1.513	0.912
INTPP	2.233	0.176	1.109	0.415	1.502	0.908

Table 3: Interval Test.

Probability	Hawkes		SC		ACD	
	RMTPP	INTPP	RMTPP	INTPP	RMTPP	INTPP
0.1	0.093	0.101	0.097	0.104	0.100	0.101
0.2	0.182	0.198	0.198	0.200	0.202	0.201
0.3	0.273	0.297	0.303	0.302	0.303	0.303
0.4	0.370	0.398	0.399	0.405	0.403	0.401
0.5	0.462	0.495	0.491	0.505	0.503	0.501
0.6	0.555	0.595	0.586	0.604	0.602	0.601
0.7	0.653	0.698	0.684	0.697	0.701	0.700
0.8	0.757	0.799	0.779	0.798	0.800	0.800
0.9	0.863	0.897	0.884	0.899	0.901	0.900

Table 4: ACC results (%).

DataSets	MC-0	MC-1	MC-2	MC-3	RMTPP	INTPP
NYSE	50.56	61.94	61.88	61.82	63.2	63.9
ATM	28.09	72.08	95.04	94.47	95.48	95.88
MIMIC	23.67	86.0	59.33	32.67	87.21	87.34

Experiment Results Figure 5 shows the MSE results predicted with different methods for testing data and the large deviation exists for RMTPP with expectation prediction. Both expectation and Reference Sampling can achieve the best results. However, according to Table 1, Reference Sampling only takes about one-tenth of the time cost.

Table 2 shows the results of comparison with different models. OPT is the optimal estimator which predicts with real coefficients. The predictive performance of INTPP is almost consistent with the respective optimal estimator. Note we use Reference Sampling as final results and choose u with mean of u_{RS} in the last epoch.

Table 3 gives the interval test, which tests RMTPP and INTPP from the probabilistic view. Compared with the theoretical probability based on RMTPP (or INTPP) and the exact time of the next event in the particular interval, INTPP performs better. Besides, Table 4 gives the results of accurate of mark predictions and INTPP also outperforms.

Conclusion

We have presented a general and interpretable neural point process, which can be connected with parametric TPP. Experiments show its advantages against existing (neural) TPP models in terms of prediction flexibility and interpretability.

References

Brown, E. N.; Barbieri, R.; Ventura, V.; Kass, R. E.; and Frank, L. M. 2002. The time-rescaling theorem and its application to neural spike train data analysis. *Neural computation* 14(2):325–346.

Du, N.; Dai, H.; Trivedi, R.; Upadhyay, U.; Gomez-Rodriguez, M.; and Song, L. 2016. Recurrent marked temporal point processes: Embedding event history to vector. In *KDD*.

Eichler, M.; Dahlhaus, R.; and Dueck, J. 2017. Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*.

Elman, J. L. 1990. Finding structure in time. *Cognitive science* 14(2):179–211.

Engle, R. F., and Russell, J. R. 1998. Robert f engle and jeffrey r russell. 1998. autoregressive conditional duration: a new model for irregularly spaced transaction data. In *Econometrica*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*.

Hawkes, A. 1971. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)* 438–443.

Huang, H.; Wang, H.; and Mak, B. 2019. Recurrent poisson process unit for speech recognition. In *AAAI*.

Isham, V., and Westcott, M. 1979. A self-correcting point process. *Stochastic Processes and Their Applications* 8(3):335–347.

Li, S.; Xiao, S.; Zhu, S.; Du, N.; Xie, Y.; and Song, L. 2018. Learning temporal point processes via reinforcement learning. In *NIPS*.

Liniger, T. J. 2009. Multivariate hawkes processes. *PhD thesis, Swiss Federal Institute Of Technology, Zurich*.

Mei, H., and Eisner, J. 2018. The neural hawkes process: A neurally self-modulating multivariate point process. In *NIPS*.

Mei, H.; Qin, G.; and Eisner, J. 2019. Imputing missing events in continuous-time event streams. In *ICML*.

Rasmussen, J. G. 2018. Lecture notes: Temporal point processes and the conditional intensity function. *arXiv preprint arXiv:1806.00221*.

Saeed, M.; Lieu, C.; Raber, G.; and Mark, R. G. 2002. Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring. In *Computers in cardiology*.

Saha, A.; Samanta, B.; Ganguly, N.; and De, A. 2018. Crpp: Competing recurrent point process for modeling visibility dynamics in information diffusion. In *CIKM*.

Upadhyay, U.; De, A.; and Rodriguez, M. G. 2018. Deep reinforcement learning of marked temporal point processes. In *NIPS*.

Wang, Y.; Liu, S.; Shen, H.; Gao, J.; and Cheng, X. 2017a. Marked temporal dynamics modeling based on recurrent neural network. In *PAKDD*.

Wang, Y.; Shen, H.; Liu, S.; Gao, J.; and Cheng, X. 2017b. Cascade dynamics modeling with attention-based recurrent neural network. In *IJCAI*.

Wu, Q.; Yang, C.; Zhang, H.; Gao, X.; Weng, P.; and Chen, G. 2018a. Adversarial training model unifying feature driven and point process perspectives for event popularity prediction. In *CIKM*.

Wu, W.; Yan, J.; Yang, X.; and Zha, H. 2018b. Decoupled learning for factorial marked temporal point processes. In *KDD*.

Wu, W.; Yan, J.; Yang, X.; and Zha, H. 2019. Reinforcement learning with policy mixture model for temporal point processes clustering. In *arXiv:1905.12345*.

Xiao, S.; Yan, J.; Li, C.; Jin, B.; Wang, X.; Yang, X.; Chu, S.; and Zha, H. 2016. On modeling and predicting individual paper citation count over time. In *IJCAI*.

Xiao, S.; Farajtabar, M.; Ye, X.; Yan, J.; Song, L.; and Zha, H. 2017a. Wasserstein learning of deep generative point process models. In *NIPS*.

Xiao, S.; Yan, J.; Yang, X.; Zha, H.; and Chu, S. 2017b. Modeling the intensity function of point process via recurrent neural networks. In *AAAI*.

Xiao, S.; Yan, J.; Farajtabar, M.; Song, L.; Yang, X.; and Zha, H. 2019. Learning time series associated event sequences with recurrent point process networks. *IEEE TNNLS*.

Xu, H.; Wu, W.; Nemati, S.; and Zha, H. 2016. Icu patient flow prediction via discriminative learning of mutually-correcting processes. *TKDE*.

Yan, J.; Xiao, S.; Li, C.; Jin, B.; Wang, X.; Ke, B.; Yang, X.; and Zha, H. 2016. Modeling contagious merger and acquisition via point processes with a profile regression prior. In *IJCAI*.

Yan, J.; Liu, X.; Shi, L.; Li, C.; and Zha, H. 2018. Improving maximum likelihood estimation of temporal point process via discriminative and adversarial learning. In *IJCAI*.

Yang, G.; Cai, Y.; and K.Reddy, C. 2018. Recurrent spatio-temporal point process for check-in time prediction. In *CIKM*.

Zhou, K.; Zha, H.; and Song, L. 2013. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *AISTATS*.