Deep Fourier Kernel for Self-Attentive Point Processes

Shixiang Zhu **Minghe Zhang** **Ruyi Ding** **Yao Xie**
Georgia Institute of Technology

Abstract

We present a novel attention-based model for discrete event data to capture complex nonlinear temporal dependence structures. We borrow the idea from the attention mechanism and incorporate it into the point processes' conditional intensity function. We further introduce a novel score function using Fourier kernel embedding, whose spectrum is represented using neural networks, which drastically differs from the traditional dot-product kernel and can capture a more complex similarity structure. We establish our approach's theoretical properties and demonstrate our approach's competitive performance compared to the state-of-the-art for synthetic and real data.

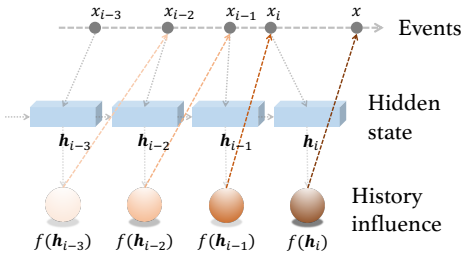
1 Introduction

Discrete event data are ubiquitous in modern applications, ranging from traffic incidents, police incidents, user behaviors in social networks, and earthquake catalogs. Such data consist of a sequence of events that indicate when and where each event occurred and any additional descriptive information about the event (such as category, marks, and free-text). The distribution of events is of scientific and practical interest, both for prediction purposes and for inferring these events' underlying generative mechanism.

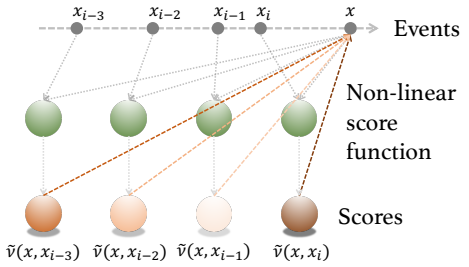
A popular framework for modeling discrete events is point processes, which can be continuous over time and space. Multi-dimensional point processes can be used to model discrete events over networks. An important aspect of this model is to capture the triggering or inhibiting effect of the event on subsequent events in the future. Since the distribution of point processes is completely specified by the conditional intensity function (the occurrence rate of events conditioning on their history), such triggering effect can be conveniently modeled by assuming parametric forms. In the classical statistical framework, the conditional intensity function usually consists of a deterministic background rate plus a stochastic term that includes the influence of the historical events, characterized by a triggering kernel function. For example, the seminar work (Ogata, 1998) proposed the epidemic-type aftershock sequence (ETAS), which suggests an exponentially decaying function over the temporal and spatial distance between events. However, with the increasing complexity and quantity of modern data, we need more expressive models. Recently, there has been much effort in developing neural network-based point processes, leveraging the rich representation power of neural networks (Omi et al., 2019; Zhu et al., 2019). In particular, because of the sequential nature of event data, existing methods rely heavily on Recurrent Neural Networks (RNNs) (Du et al., 2016; Li et al., 2018; Mei and Eisner, 2017; Upadhyay et al., 2018; Xiao et al., 2017a,b; Zhu et al., 2020).

However, there is a notable limitation of existing neural network-based models. The popular RNN models such as Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) are not enough capable of capturing long-range dependencies and still implicitly assumes that the influence of the current event decays monotonically over time (due to their recursive structure). Many real-world applications may not be good candidates to apply these assumptions. For instance, in modeling economic time-series, major economic or historical events (such as economic crisis or shift of policy) will have a much longer impact; their influence may be carried over to current time and should not be "forgotten" by the event model. In modeling traffic events, when a major car accident occurs on the highway, it takes hours to clear the scene. The congestion will not ease during that period – the influence of major traffic incident events may not decay monotonically over time. These motivate us to tackle the long-term and non-homogeneous influence function and capture the influence of past events in a more flexible manner.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.



(a) RNN-based Point Process



(b) Deep Attention Point Process

Figure 1: Comparison between RNN-based models and our DAPP. The color depth of the red balls represent their "importance" in the model. The history influence in (a) are exponentially decaying over the time. The score is a non-linear function with respect to the distance between events and is non-homogeneous over the time.

In the domain of natural language processing (NLP) and computer vision, the self-attention mechanism has been widely adopted as an algorithmic component to tackle the effect of non-linear and long-range dependence (Vaswani et al., 2017). This motivates us to adapt the attention mechanism for the point processes models, leveraging their capabilities to capture long-range and complex dependency in the sequence. However, since the attention mechanism has rarely been used outside of the domains mentioned above, we still need to develop a principled probabilistic (stochastic process) model framework to incorporate the attention mechanism into continuous point processes properly. In particular, unlike the NLP problem (Mikolov et al., 2013), where the similarity between words can be adequately characterized by *dot-product* score (Vaswani et al., 2017) in the conventional attention mechanism, discrete events usually exhibit heterogeneous triggering effects regarding their spatio-temporal distances. Take earthquake catalog data as an example. The dynamics between seismic events are related to the geologic structure of faults. For instance, most aftershocks either occur along the fault plane or other faults within the volume affected by the strain associated with the mainshock.

In this paper, we propose a deep attention point process (DAPP) model, with a flexible non-linear score function based on Fourier kernels in the attention mechanism, as shown in Figure 1. We go beyond the recurrent structure of RNN that the historical information can only be passed through the hidden state. Instead, we leverage the attention mechanism to develop a flexible framework that "focuses" on past events with high "importance" scores, regardless of how far away they are. We also present a novel score function via Fourier kernels with spectrum represented using deep neural networks, whose parameters are learned from data. In contrast to the commonly used dot-product score, which essentially performs linear key embedding, our score function performs non-linear kernel-induced feature embedding, capturing more complex similarity structures in events. This can help achieve higher flexibility in retaining the most "significant" historical events relative to the current event. Moreover, to achieve constant memory in the face of streaming data, we develop an online version of DAPP, which is more suitable to process streaming data. We establish the Fourier kernel's theoretical properties and demonstrate the competitive performance of our proposed method relative to the state-of-the-art on a wide range of real and synthetic data sets.

Our contributions include (1) introducing a general probabilistic attention-based point process model for discrete event data; (2) introducing a novel similarity kernel based on Fourier kernel embedding and neural-network represented spectrum (in contrast to the standard dot-product kernel).

Related work. Existing works for point processes modeling, such as Gomez Rodriguez et al. (2010); Yuan et al. (2019); Zhu and Xie (2019), often assume parametric forms of the intensity functions. Such methods enjoy good interpretability and are efficient to perform. However, parametric models are not expressive enough to capture the events' dynamics in some applications. As previously mentioned, recent interest has focused on improving the expressive power of point process models using RNNs (Du et al., 2016; Li et al., 2018; Mei and Eisner, 2017; Upadhyay et al., 2018; Zhu et al., 2020). However, the events' dependence in the conditional intensity is specified as a parametric form. For instance, Du et al. (2016) expresses the influence of two consecutive events in a form of $\exp\{w(t_{i+1} - t_i)\}$, which is an exponential function with respect to the length of the time interval.

There also have been some works that model stochastic processes using the attention mechanism (Zhang et al., 2019; Kim et al., 2019). Kim et al. (2019) uses self-attention to model a class of neural latent variable models, called Neural Processes (Garnelo et al., 2018), which is not for sequential data specifically. In retrospect, we realize a concurrent work (Zhang et al.,

2019; Zuo et al., 2020) which also use the attentive mechanism to model point processes, but the framework is different. An important distinction of their approach from ours is that they rely on a dot-product between features (embedded in a Gaussian argument) in the attention mechanism; we use a more flexible and general Fourier kernel for similarity function, and we also specifically address the design and learning of the kernel by representing the spectrum of the Fourier kernel using neural networks.

2 Background

Marked temporal point processes (MTPPs) (Reinhart, 2017) consist of an ordered sequence of events localized in time, location, and mark spaces. Let $\{x_1, x_2, \ldots, x_{N_T}\}$ represent a sequence of points sampled from a MTPP. We denote N_T as the number of the points generated in the time horizon $[0, T)$. Each point x_i is a marked spatio-temporal tuple $x_i = (t_i, m_i)$, where $t_i \in [0, T)$ is the time of occurrence of the i-th event, and $m_i \in \mathcal{M}$ is the corresponding mark, which may contain location, event type, or other rich description information (such as image or free-text). Here we treat discrete location marks, while sometimes the continuous location is treated separately in spatio-temporal point processes.

The events' distribution in MTPPs are characterized via a conditional intensity function $\lambda(t, m|\mathcal{H}_t)$, which is the probability of observing an event in the marked temporal space $[0, T) \times \mathcal{M}$ given the events' history $\mathcal{H}_t = \{(t_i, m_i)|t_i < t\}$, i.e., $\lambda(t, m|\mathcal{H}_t)|B(m, dm)|dt = \mathbb{E}\left[N([t, t + dt) \times B(m, dm))|\mathcal{H}_t\right]$, where $N(A)$ is the counting measure of events over the set $A \subseteq \mathcal{X}$ and $|B(m, dm)|$ is the Lebesgue measure of the ball $B(m, dm)$ with radius dm. The log-likelihood of observing a sequence with n events denoted as $\boldsymbol{x} = \{(t_i, m_i)\}_{i=1}^{N_T}$ can be obtained by

$$\ell(\boldsymbol{x}) = \sum_{i=1}^{N_T} \log \lambda(t_i, m_i|\mathcal{H}_{t_i}) - \int_{m\in\mathcal{M}} \int_0^T \lambda(t, m|\mathcal{H}_t) dt dm. \tag{1}$$

As self- and mutual-exciting point processes, Hawkes processes (Hawkes, 1971) have been widely used to capture the mutual excitation dynamics among temporal events. The model assumes that influences from past events are linearly additive towards the current event. The conditional intensity function of a Hawkes process is defined as

$$\lambda(t, m|\mathcal{H}_t) = \mu + \sum_{t_i<t} g(t - t_i, m - m_i), \tag{2}$$

where $\mu \geq 0$ is the background intensity of events, $g(\cdot) \geq 0$ is the triggering function that captures spatio-temporal and marked dependencies of the past events. The triggering function can be chosen in advance, e.g., in one-dimensional cases, $g(t, t_i) = \alpha \exp\{-\beta(t - t_i)\}$, where β controls the decay rate and $\alpha > 0$ controls the magnitude of the influence.

3 Proposed Method

This section presents a novel attention-based point process model using the deep Fourier kernel as its score function, which is capable of remembering long-term memory and capturing non-homogeneous triggering effects.

3.1 Self-attention in point processes

Deep Attention Point Processes (DAPP) aims to model the current event's nonlinear dependencies from past events using the attention mechanism. Specifically, we model the conditional intensity function of MTPPs using the attention output. DAPP also adopts the " multi-heads" mechanism, which offers multiple "representation subspaces" for events in the sequence. We describe the DAPP framework for point processes as follows.

For notational simplicity, we denote the d-dimensional marked temporal sapce as $\mathcal{X} \coloneqq [0, T) \times \mathcal{M} \subset \mathbb{R}^d$. Let data tuple of the current event be $x \coloneqq (t, m) \in \mathcal{X}$, and the data tuple of an arbitrary past event be $x' \coloneqq (t', m') \in \mathcal{X}$ for any $t' < t$. For the k-th *attention head*, we first score the current event against its past event using score function $\nu^{(k)} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. For the event x, the score $\nu^{(k)}(x, x')$ determines how much *attention* to place on the past event x' as we encode the history information. More details about the score formulation will be presented in Section 3.2. The normalized score $\widetilde{\nu}^{(k)}(x, x') \in [0, 1]$ for the event x and x' is obtained by employing the softmax function over the score, which is defined as

$$\widetilde{\nu}^{(k)}(x, x') = \frac{\nu^{(k)}(x, x')}{\sum_{t_i<t} \nu^{(k)}(x, x_i)}, \ k = 1, \ldots, K, \tag{3}$$

Then we map past events to the *value* embedding space via $\varphi^{(k)} : \mathcal{X} \to \mathbb{R}^p$, where p is the dimension of the value embedding. Here the value embedding is a linear transformation of the event's data tuple, i.e., $\varphi^{(k)}(x) = W_v^{(k)} x \in \mathbb{R}^p$, where $W_v^{(k)} \in \mathbb{R}^{p\times d}$ is the weight matrix. Therefore, the k-th attention head $\boldsymbol{h}^{(k)}(x) \in \mathbb{R}^p$ for the event x can be obtained by multiplying each value embedding by the score and adding them up, which is formally defined as

$$h^{(k)}(x) = \sum_{t_i<t} \widetilde{\nu}^{(k)}(x, x_i)\varphi^{(k)}(x_i), \ k = 1, \ldots, K, \tag{4}$$

Note that events x, x_i are analogous to the *query* and the i-th *key*, the embedding of the i-th event $\varphi^{(k)}(x_i)$ is analogous to the *value* in the attention mechanism. The

multi-head attention $\boldsymbol{h}(x) \in \mathbb{R}^{Kp}$ is the concatenation of K single attention heads:

$$\boldsymbol{h}(x) = \text{concat}\left(h^{(1)}(x), \ldots, h^{(K)}(x)\right).$$

We highlight that the attention $\boldsymbol{h}(x)$ is able to "emphasize" (or "de-emphasize") events, which are most (or least) influential in their future, by directly assigning them larger (smaller) scores. In comparison, RNN-based models pass the history information sequentially via a hidden state, where the recent memory will override the long-term memory. This has led RNNs to "overemphasize" the recent events and fail to capture the events' influences in the distant past.

Follow the similar idea of Mei and Eisner (2017), we consider a non-linear transformation of the multi-head attention $\boldsymbol{h}(x)$ as the historical information before event x, the conditional intensity function λ can be specified as:

$$\lambda(x|\boldsymbol{h}(x)) = \underbrace{\mu(x)}_{\text{base intensity}} + \underbrace{g\left(\boldsymbol{h}(x)^\top W + b\right)}_{\text{triggering effect}}, \quad (5)$$

where $W \in \mathbb{R}^{Kp}, b \in \mathbb{R}$ are the weight matrix and the bias term, where $g : \mathbb{R} \to \mathbb{R}^+$ is a monotonically increasing function, and here we choose the function $g(x) := \text{softplus}(x) = \log(1 + e^x) > 0$ is a smooth approximation of the ReLU function, which ensures the intensity strictly positive at all times when an event could possibly occur and avoid infinitely bad log-likelihood. The $\mu(x) > 0$ is the base intensity, which can be estimated from the data.

3.2 Score function via deep Fourier kernel

As introduced in the previous section, the score function ν directly quantifies how likely one event is triggered by the other in a sequence, which plays a similar role as the triggering function in Hawkes processes defined in (2). Usually, the *dot-product score* has been widely used in most attention models. Specifically, two points $x, x' \in \mathbb{R}^d$ are first projected onto another space via $W_u x$ and $W_u x'$ (the so-called *key embeddings*), where $W_u \in \mathbb{R}^{r \times d}$ is a linear mapping and r is the dimension of key embeddings. Then the score is obtained by computing $x^\top W_u^\top W_u x'$, which essentially is their inner product in the embedding space. However, for some real applications, inner product or Euclidean distance may be limited when the triggering effects between events are non-homogeneous.

We aim to find a more representative score function without specifying any parametric form to capture the complex interactions between events, which goes beyond the linear assumption on the dot-product score. To this end, we propose a novel deep Fourier kernel as the score function in the attention mechanism, where

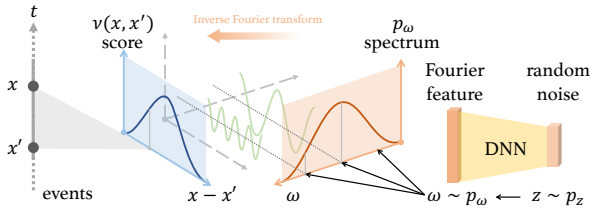


Figure 2: An illustration for the Fourier kernel score. The score is computed via an inverse Fourier transform, where the distribution of Fourier features is represented by a deep neural network.

the key embedding $W_u x$ is substituted with the kernel-induced feature mapping $\Phi(x)$. As shown in Figure 2, these feature mappings are randomly sampled from an optimal high-dimensional power spectrum. The optimal spectrum (the distribution of power) is represented by a deep neural network. The network's inputs are random normal noises, and the outputs are Fourier features sampled from the optimal spectrum.

Formally, this score formulation relies on Bochner's Theorem (Rudin, 1962), which states that any bounded, continuous and shift-invariant kernel is a Fourier transform of a bounded non-negative measure:

Theorem 1 (Bochner (Rudin, 1962)). *A continuous kernel of the form $\nu(x, x') = \kappa(x - x')$ defined over a locally compact set $\mathcal{X} \subset \mathbb{R}^d$ is positive definite if and only if g is the Fourier transform of a non-negative measure:*

$$\nu(x, x') = \kappa(x - x') = \int_\Omega p(\omega) e^{jw^\top (x - x')} d\omega, \quad (6)$$

where p is a non-negative measure, Ω is the Fourier feature space, and kernels of the form $\nu(x, x')$ are called shift-invariant kernel.

If a shift-invariant kernel $\kappa(\cdot)$ is properly scaled such that $\kappa(0) = 1$, Bochner's theorem guarantees that its Fourier transform $p(\omega)$ is a proper probability distribution.

Suppose an optimal spectrum that best describes how the "energy" of events' interaction in each attention head is distributed with Fourier features. Here we assume $p_\omega^{(k)}$ is the optimal distribution of Fourier features $\omega \in \Omega \subset \mathbb{R}^r$ in the k-th attention head, where r is dimension of Fourier features. We also substitute $\exp\{jw^\top (x - x')\}$ with a real-valued feature mapping, such that the probability distribution p_ω and the kernel ν are real (Rahimi and Recht, 2008). We, therefore, obtain a score formulation of the k-th attention head in (3) between two events $x, x' \in \mathcal{X} \subset \mathbb{R}^d$ that satisfies these conditions as the following proposition (see proof in Appendix A):

Proposition 1 (Score function via Fourier kernel embedding). *Let the score $\nu^{(k)}$, $k = 1, \ldots, K$ be a con-*

tinuous real-valued shift-invariant kernel and $p_\omega^{(k)}$ *be a probability distribution, we have the following definition:*

$$\nu^{(k)}(x, x') := \mathbb{E}\big[\phi_\omega^{(k)}(x) \cdot \phi_\omega^{(k)}(x')\big], \tag{7}$$

where $\phi_\omega^{(k)}(x) := \sqrt{2}\cos(\omega^\top W_u^{(k)} x + b_u)$, *and* $W_u^{(k)} \in \mathbb{R}^{r\times d}$ *is a linear mapping. These Fourier features* $\omega \in \Omega \subset \mathbb{R}^r$ *are sampled from* $p_\omega^{(k)}$ *and* b_u *is drawn uniformly from* $[0, 2\pi]$.

We can conclude from the proposition that (1) the score function is defined by the optimal spectrum $p_\omega^{(k)}$ and the weight $W_u^{(k)}$. Here $W_u^{(k)} x$ resembles the *key embedding* in the dot-product score, which projects event x to a high-dimensional embedding space; (2) this representation enables us to conveniently estimate the score from samples, i.e.,

$$\nu^{(k)}(x, x') \approx \frac{1}{D}\sum_{j=1}^{D} \phi_{\omega_j}^{(k)}(x)\cdot\phi_{\omega_j}^{(k)}(x') = \Phi^{(k)}(x)^\top \Phi^{(k)}(x'), \tag{8}$$

where $\omega_j, j = 1, \ldots, D$ are D Fourier features sampled from the distribution $p_\omega^{(k)}$. The vector

$$\Phi^{(k)}(x) := [\phi_{\omega_1}^{(k)}(x), \ldots, \phi_{\omega_D}^{(k)}(x)]^\top,$$

can be viewed as the approximation of the kernel-induced feature mapping for the score function.

In the following proposition, we will show this empirical estimation converges uniformly over a compact domain $\mathcal{X}$ as D grows and is a lower variance approximation to (7) (see the proof in Appendix B):

Proposition 2 (Concentration of empirical scores)**.** *Assume* $\sigma_p^2 = \mathbb{E}_{\omega\sim p_\omega^{(k)}}[\omega^\top\omega] < \infty$ *and* $\mathcal{X} \subset \mathbb{R}^d$. *Let* R *denote the radius of the Euclidean ball containing* $\mathcal{X}$, *then for the kernel-induced feature mapping* $\Phi^{(k)}$ *defined in* (8), *we have*

$$\mathbb{P}\left\{\sup_{x,x'\in\mathcal{X}}\left|\Phi^{(k)}(x)^\top\Phi^{(k)}(x') - \nu^{(k)}(x,x')\right| \geq \epsilon\right\} \leq \left(\frac{48R\sigma_p}{\epsilon}\right)^2 \exp\left\{-\frac{D\epsilon^2}{4(d+2)}\right\}. \tag{9}$$

The proposition guarantees that a good estimate of the score function can be found, with high probability, by sampling a finite number of Fourier features. In particular, for an absolute error of at most ϵ, the number of samples needed is on the order of $D = O(d\log(R\sigma_p/\epsilon)/\epsilon^2)$, which grows linearly as data dimension d increases.

3.3 Fourier feature generator

To represent the distribution $p_\omega^{(k)}$ over Fourier feature ω, we define a prior (generator) on an input noise variable $z \sim p_z$, then represent a mapping to feature space as $G : \mathbb{R}^q \to \mathbb{R}^r$ as shown in Figure 2, where G is a differentiable function characterized by a deep neural network with parameters $\theta^{(k)}$ and q is the dimension of the noise, such that roughly speaking the distribution functions are the same $p_\omega^{(k)} \approx G(z)$. Note that the score function's representative power is jointly decided by the generator's parameters and the weight matrix of the key embedding.

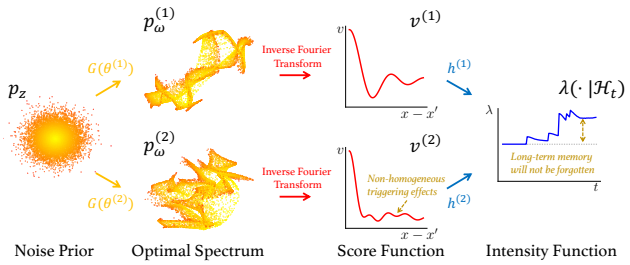
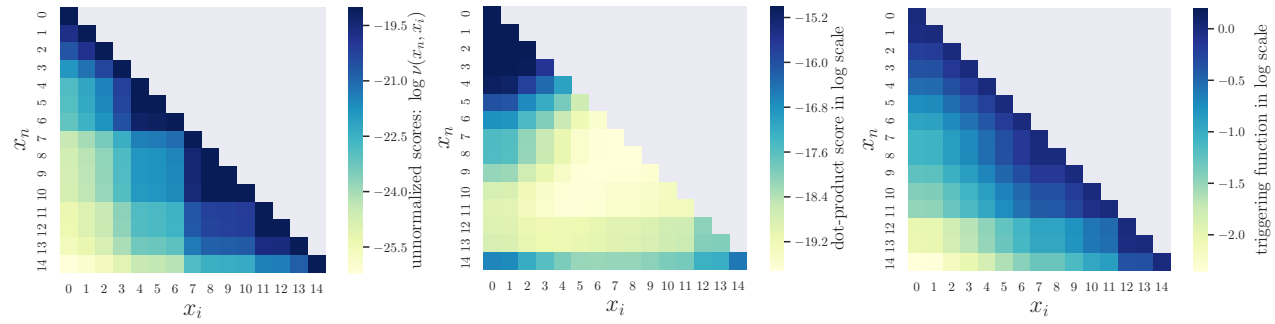


Figure 3: A real example of optimal spectrums, score function, and corresponding intensity function learned from DAPP. Here, the DAPP is trained using real 911 calls-for-service data with recorded time only in 2017 provided by Atlanta Police. There are 10,000 Fourier features sampled from the optimal spectrums being used to reconstruct score functions. The right-most sub-figure represents the intensity of a 911 call sequence reported in a single day at beat 702. We can see that Fourier kernel score is able to capture non-homogeneous triggering effects of events and long-term memory will also not be forgotten in this case.

Figure 3 gives an intuitive example of representing the intensity of events using our DAPP with two attention heads ($K = 2$). Here, we choose $q = r = 2$ to visualize the noise prior and the optimal spectrums in a 2D space for ease of presentation. The optimal spectrum learned from data in each attention head uniquely specifies a score function, which is capable of capturing various types of non-linear triggering effects. Unlike Hawkes processes, underlying long-term influences of some events, in this case, can be preserved in the intensity function. As shown in Figure 4, we present two examples of pairwise scores calculated by the proposed Fourier score and dot-product score under the same architecture, respectively, which enables a visual comparison. To make these two methods comparable, we trained two models using the same synthetic data set, and its exact triggering function is also provided as the "ground truth". This ablation study confirms that our Fourier score in a single attention head is expressive enough to accurately capture the triggering effects.

3.4 Online attention for streaming data

The attention calculation may be computationally intractable for streaming data since the number of past events would overgrow as time goes on. Here, we propose an adaptive online attention algorithm to address



(a) Fourier kernel (b) dot-product (c) truth

Figure 4: Pairwise scores between events learned from DAPP using a synthetic Hawkes process data set. The x_n denotes the current event and x_i denotes past events, where $t_i < t_n$. The color of the entry at n-th row and i-th column in these figures indicates: (a) Fourier kernel scores; (b) dot-product score (using the same architecture by only substituting Fourier kernel score with dot-product score); (c) true triggering effects evaluated by triggering function $g(t, t_i) = \alpha \exp\{-\beta(t - t_i)\}$ (does not exactly correspond to the score, but reveals some key facts on the correlation of events, e.g., exponential decaying over time).

this issue. Only a fixed number of "important" historical events with high average scores will be remembered for the attention calculation in each attention head. The procedure for collecting "important" events in each attention head is demonstrated as follows.

When the i-th event occurs, for a past event $x_j, t_j < t_i$ in k-th attention, we denote the set of its score against future events as $\mathscr{S}_j^{(k)} := \{\widetilde{\nu}^{(k)}(x_i, x_j)\}_{i:t_j \le t_i}$. Then the average score of the event x_j can be computed by

$$\bar{\nu}_j^{(k)} = (\sum\nolimits_{s \in \mathscr{S}_j^{(k)}} s)/|\mathscr{S}_j^{(k)}|,$$

where $|A|$ denotes the number of elements in set A. Hence, a recursive definition of the set of active events $\mathscr{A}_i^{(k)}$ in the k-th attention head up until the occurrence of the event x_i is written as:

$$\mathscr{A}_i^{(k)} = \mathcal{H}_{t_{i+1}}, \ \forall i \le \eta,$$
$$\mathscr{A}_i^{(k)} = \mathscr{A}_{i-1}^{(k)} \cup \{x_i\} \setminus \arg\min_{j:t_j < t_i} \left\{\bar{\nu}_j^{(k)}\right\}, \ \forall i > \eta,$$

where η is the maximum number of events we want to remember. The exact event selection is carried out by Algorithm 3 shown in Appendix C. To perform the online attention, we substitute $\mathcal{H}_{t_n}$ in (3) and (4) with $\mathscr{A}_n^{(k)}$ for all attention heads.

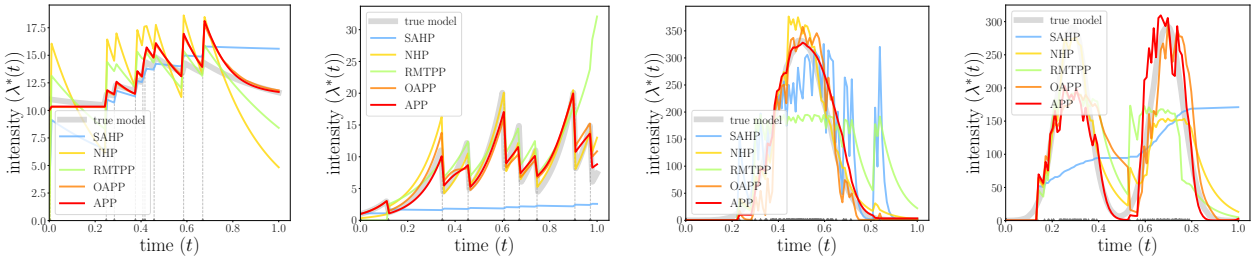
3.5 Learning and simulation

The proposed model is jointly parameterized by $\boldsymbol{\theta} = \{W, b, \{\theta^{(k)}, W_u^{(k)}, W_v^{(k)}\}_{k=1,\dots,K}\}$, which can be learned via *maximum likelihood estimation* using the stochastic gradient descent. The log-likelihood function of the model can be obtained by substituting (5) into (1) defined in Section 2. The exact learning algorithm is carried out by Algorithm 1 shown in Appendix C.

A default way to generate events from a point process is to use the thinning algorithm (Daley and Vere-Jones, 2008; Gabriel et al., 2013). However, the vanilla thinning algorithm suffers from low sampling efficiency as it needs to sample in the space $\mathcal{X}$ uniformly with the upper limit of the conditional intensity $\bar{\lambda}$ and only very few candidate points will be retained in the end. To improve sampling efficiency, we use an efficient thinning algorithm summarized in Algorithm 2, Appendix C. The "proposal" density is a non-homogeneous MTPP, whose intensity function is defined from the previous iterations. This analogous to the idea of rejection sampling (Ogata, 1981).

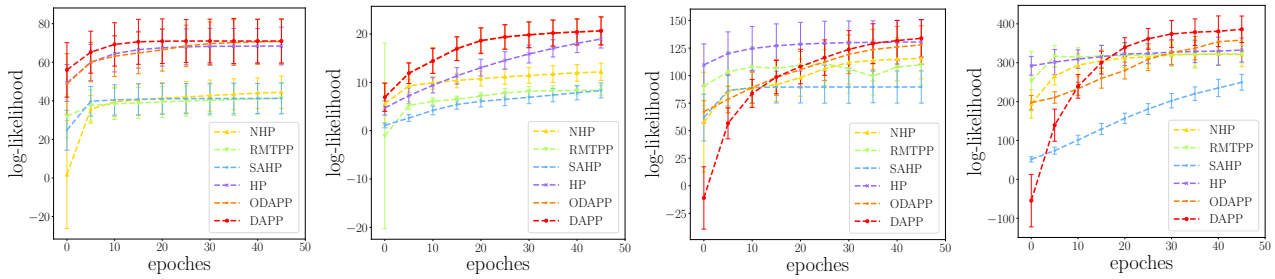
4 Experiments

In this section, we conduct experiments on four synthetic data sets and four large-scale real-world data sets. We compare our DAPP and its online version (ODAPP) with the other four baselines by evaluating the mean square intensity-recovering error and the likelihood value, which have been widely adopted in the related works (Mei and Eisner, 2017; Omi et al., 2019; Zhang et al., 2019). The implementation details of baselines are discussed in Section 4.1. We describe the experiment configurations as follows: we consider two attention heads ($K = 2$) in DAPP and ODAPP, where the Fourier feature generator $\theta^{(k)}$ of the k-th head is characterized by a fully-connected neural network with three hidden layers, where the widths of each layer are 128, 256, and 128, respectively. To learn DAPP and its associated optimal spectrums more efficiently, we adopt the stochastic gradient descent method and only sample a few points of Fourier features ($D = 20$) for each mini-batch. For accurate intensity recovery, a more significant number of Fourier features ($D = 10,000$) will be sampled in a bid to reconstruct a high-resolution optimal spectrum. Besides, there is only a 50% number of events are retained for training ODAPP, i.e., $\eta = 0.5n$, where n is the maximum length of sequences in each data set.



(a) Hawkes (b) self-correction (c) non-homogeneous 1 (d) non-homogeneous 2

Figure 5: Conditional intensity function estimated from synthetic data sets. Triangles at the bottom of each panel represent events. The ground truth of conditional intensities is indicated by the grayline.



(a) stock (b) MIMIC-III (c) meme (d) traffic

Figure 6: The average log-likelihood of real data sets versus training epochs. For each real data set, we evaluate performance of the five methods according to the final log-likelihood averaged per event calculated for the test data.

4.1 Baseline methods

Recurrent Marked Temporal Point Process (RMTPP) assumes the following form for the conditional intensity function λ^* in point processes, denoted as $\lambda^*(t) = \exp\big(\boldsymbol{v}^\top \boldsymbol{h}_j + \omega(t-t_j) + b\big)$, where the j-th hidden state in the RNN $\boldsymbol{h}_j$ is used to represent the history influence up to the nearest happened event j, and $w(t-t_j)$ represents the current influence. The $\boldsymbol{v}, \omega, b$ are trainable parameters (Du et al., 2016).

Neural Hawkes Process (NHP) specifies the conditional intensity function in point processes using a continuous-time LSTM, denoted as $\lambda^*(t) = f(\boldsymbol{\nu}^\top \boldsymbol{h}_t)$, where the hidden state of the LSTM up to time t represents the history influence, the $f(\cdot)$ is a softplus function which ensure the positive output given any input (Mei and Eisner, 2017).

Self-Attentive Hawkes Process (SAHP) adopts self-attention mechanism to model the historical information in the conditional intensity function, which is specified as $\lambda^*(t) = \text{softmax}\big(\mu + \alpha \exp\{\omega(t-t_j)\}\big)$, where μ, α, w are computed via three non-linear mappings: $\mu = \text{softplus}(\boldsymbol{h}W_\mu)$, $\alpha = \tanh(\boldsymbol{h}W_\alpha)$, $\omega = \text{softplus}(\boldsymbol{h}W_\omega)$. The $W_\mu, W_\alpha, W_\omega$ are trainable parameters (Zhang et al., 2019).

Hawkes Process (HP) specifies the conditional intensity function as $\lambda^*(t) = \mu + \alpha \sum_{t_j<t} \beta \exp\{-\beta(t-t_j)\}$, where parameters μ, α, β can be estimated via maximizing likelihood (Hawkes, 1971).

4.2 Synthetic data sets

The synthetic data are obtained by the following four generative processes: (1) *Hawkes process*: the conditional intensity function is given by $\lambda^*(t) = \mu + \alpha \sum_{t_j<t} \beta \exp -\beta((t-t_j))$, where $\mu = 10$, $\alpha = 1$, and $\beta = 1$; (2) *self-correction point process*: the conditional intensity function is given by $\lambda^*(t) = \exp\big(\mu t - \sum_{t_i<t} \alpha\big)$, where $\mu = 10$, $\alpha = 1$; (3) *non-homogeneous Poisson* 1: The intensity function is given by $\lambda^*(t) = c \cdot \Phi(t-0.5) \cdot U[0,1]$ where $c = 100$ is the sample size, the $\Phi(\cdot)$ is the PDF of standard normal distribution, and $U[a,b]$ is uniform distribution between a and b; (4) *non-homogeneous Poisson* 2: The intensity function is a composition of two normal functions, where $\lambda^*(t) = c_1 \cdot \Phi(6(t-0.35)) \cdot U[0,1] + c_2 \cdot \Phi(6(t-0.75)) \cdot U[0,1]$, where $c_1 = 50$, $c_2 = 50$. Each synthetic data set contains 5,000 sequences with an average length of 30, where each data point in the sequence only contains the occurrence time of the event.

4.3 Real data sets

Traffic Congestions (traffic): We collect the data of traffic congestions from the Georgia Department of Transportation (GDOT, 2019) over 178 days from 2017 to 2018, including 15,663 congestion events recorded by 86 observation sites. Each event consists of time, location, and congestion level. We partition the data into 178 sequences by day, and each sequence has an average length of 88.

Table 1: the mean square error of recovering the intensity.

Data set	HP	SAHP	NHP	RMTPP	DAPP+dot-prod	ODAPP+dot-prod	DAPP+NN	ODAPP+NN	DAPP	ODAPP
hawkes	0.031	18.3	49.9	35.9	15.3	19.3	1.221	1.893	0.258	**0.166**
self-correction	74.3	130.8	25.8	36.1	117.4	133.3	47.2	51.3	**21.8**	27.3
non-homo 1	1672.3	7165.5	1431.6	6852.3	6201.5	7014.8	972.5	1124.1	**605.7**	1511.8
non-homo 2	3210.3	9858.9	2063.1	3854.8	9812.7	9733.1	1449.5	1722.7	**1351.4**	1527.9

Electrical Medical Records (MIMIC-III): Medical Information Mart for Intensive Care III (MIMIC-III) (Johnson et al., 2016) contains de-identified clinical visit time records from 2001 to 2012 for more than 40,000 patients. We select 2,246 patients with at least three visits. Each patient's visit history will be considered an event sequence, and each clinical visit will be considered an event.

Financial Transactions (stock): We collected data from the NYSE of the high-frequency transactions. It contains 0.7 million transaction records, each of which records the time (in milliseconds) and the possible action (sell or buy). We partition the raw data into 5,756 sequences with an average length of 48 by days.

Memes (meme): MemeTracker (Leskovec et al., 2007) tracks the meme diffusion over public media, which contains more than 172 million news articles or blog posts. The memes are sentences, such as ideas, proverbs, and recorded when it spreads to specific websites. We randomly sample 22,003 sequences of memes with an average length of 24.

4.4 Synthetic data experiment results

In the following experiments with synthetic data, we confirmed that our deep attention point process model could capture synthetic events' spatio-temporal dynamics. We first summarized the mean square error of recovering the true intensity in Table 1, where our methods achieve the minimal error in recovering intensities. We also visualized recovered intensity over time given a randomly-picked sequence from each data set in Figure 5. The solid grey line represents the true intensity of the sequence. The result shows that our methods can accurately recover the temporal intensity. It is noteworthy that our approach can also capture the sequences' dynamics with non-homogeneous temporal intensity, as shown in Figure 5 (a), (b), which is extremely challenging to characterize by the other baselines. Besides, we emphasize that the online version of our approach (ODAPP) also shows competitive performances against other methods, where only 50% of events are used.

Table 2: the average log-likelihood.

Data set	HP	SAHP	NHP	RMTPP	DAPP	ODAPP
hawkes	22.0	20.8	20.0	19.7	**21.2**	21.1
self-correction	3.9	3.5	5.4	6.9	7.1	**7.1**
non-homo 1	437.8	432.4	445.6	443.1	442.3	**457.0**
non-homo 2	399.4	364.3	410.1	405.1	**428.3**	420.1
mimic-iii	17.1	11.7	14.4	8.7	**21.5**	21.2
stock	66.3	43.1	43.4	44.0	**72.9**	72.9
meme	129.8	84.0	113.4	106.0	**131.0**	128.5
traffic	313.8	326.7	324.4	339.2	**458.5**	387.2

To further investigate the effects of deep Fourier kernel on discrete event modeling, we consider the following two ablation studies: we replace the deep Fourier kernel score function in the DAPP by (a) `DAPP+dot-prod`: the conventional dot-product; (b) `DAPP+NN`: a fully-connected neural network (with the same configuration as the generator in deep Fourier kernel), where the network's input is the concatenation of projections of two events, and the scalar output is the score. As we can see from Table 1, the non-linearity of the score function plays a pivotal role in modeling the spatio-temporal dynamics between these events and drastically reduces the mean square error. In particular, the deep Fourier kernel enjoys a greater expressive power in representing non-linear triggering effects comparing to a simple neural network. The above result confirms that the combination of the attention and our deep Fourier kernel-based score leads to event modeling and prediction success.

4.5 Real data experiment results

This section evaluates the performance of our methods on real-world data sets from a diverse range of domains, including a spatio-temporal data set and three other temporal data sets. Due to a lack of true knowledge of intensity in real data, the recovery error is unavailable. Here, we reported the average log-likelihood of each method over training epochs on the testing data in Figure 6 and summarized the highest average log-likelihood each method could obtain after the convergence in Table 2. As we can see, our DAPP and ODAPP outperform the other alternatives with higher average log-likelihood values on various data sets.

5 Conclusion

We proposed an attention-based point processes model with a deep Fourier kernel, where the spectrum represented by neural networks captures complex non-linear dependence on past events. As demonstrated by our experiments with synthetic and real data, our method achieves competitive performance in achieving higher log-likelihood and smaller recovery error for conditional intensity function of a point process compared to the state-of-the-art.

References

Daley, D. J. and Vere-Jones, D. (2008). *An introduction to the theory of point processes. Vol. II.* Probability and its Applications (New York). Springer, New York, second edition. General theory and structure.

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 *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1555–1564, New York, NY, USA. Association for Computing Machinery.

Gabriel, E., Rowlingson, B., and Diggle, P. (2013). stpp: An r package for plotting, simulating and analyzing spatio-temporal point patterns. *Journal of Statistical Software*, 53:1–29.

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S. M. A., and Teh, Y. W. (2018). Neural processes.

GDOT (2019). Traffic analysis and data application (tada). http://www.dot.ga.gov/DS/Data.

Gomez Rodriguez, M., Leskovec, J., and Krause, A. (2010). Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 1019–1028, New York, NY, USA. Association for Computing Machinery.

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

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, S. M. A., Rosenbaum, D., Vinyals, O., and Teh, Y. W. (2019). Attentive neural processes. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Leskovec, J., Backstrom, L., and Kleinberg, J. (2007). Memetracker. http://www.memetracker.org/data.html.

Li, S., Xiao, S., Zhu, S., Du, N., Xie, Y., and Song, L. (2018). Learning temporal point processes via reinforcement learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS '18, pages 10804–10814, Red Hook, NY, USA. Curran Associates Inc.

Mei, H. and Eisner, J. M. (2017). The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems 30*, pages 6754–6764. Curran Associates, Inc.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Ogata, Y. (1981). On lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402.

Omi, T., ueda, n., and Aihara, K. (2019). Fully neural network based model for general temporal point processes. In *Advances in Neural Information Processing Systems 32*, pages 2120–2129. Curran Associates, Inc.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc.

Reinhart, A. (2017). A review of self-exciting spatio-temporal point processes and their applications.

Rudin, W. (1962). *Fourier analysis on groups*, volume 121967. Wiley Online Library.

Upadhyay, U., De, A., and Gomez Rodriguez, M. (2018). Deep reinforcement learning of marked temporal point processes. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 3168–3178. Curran Associates, Inc.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Xiao, S., Farajtabar, M., Ye, X., Yan, J., Song, L., and Zha, H. (2017a). Wasserstein learning of deep generative point process models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS '17, pages 3250–3259, Red Hook, NY, USA. Curran Associates Inc.

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

of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI '17, pages 1597–1603. AAAI Press.

Yuan, B., Li, H., Bertozzi, A. L., Brantingham, P. J., and Porter, M. A. (2019). Multivariate spatiotemporal hawkes processes and network reconstruction. *SIAM Journal on Mathematics of Data Science*, 1(2):356–382.

Zhang, Q., Lipani, A., Kirnap, O., and Yilmaz, E. (2019). Self-attentive hawkes processes.

Zhu, S., Li, S., and Xie, Y. (2019). Interpretable generative neural spatio-temporal point processes.

Zhu, S. and Xie, Y. (2019). Spatial-temporal-textual point processes with applications in crime linkage detection.

Zhu, S., Yuchi, H. S., and Xie, Y. (2020). Adversarial anomaly detection for marked spatio-temporal streaming data. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8921–8925.

Zuo, S., Jiang, H., Li, Z., Zhao, T., and Zha, H. (2020). Transformer hawkes process. *arXiv preprint arXiv:2002.09291*.