Recent Advance in Temporal Point Process: from Machine Learning Perspective

Junchi Yan
Shanghai Jiao Tong University
yanjunchi@sjtu.edu.cn

Abstract

Temporal point process (TPP) has served as a versatile framework for modeling event sequences in continuous time space. It spans a wide range of applications as event data is prevalent and becoming increasingly available such as online purchase, device failure. Tailored TPP learning algorithms are devised to different special processes, complemented by recent neural network based approaches. In general, traditional statistical TPP models are more interpretable and less data ravenous, which lay their success on appropriate selection of the intensity function via domain knowledge. In contrast, emerging network based models have higher capacity to digest massive event data with less reliance on model selection. However their physical meaning becomes less comprehensible. From machine learning perspective, this survey presents a literature review on these two threads of research. We walk through several working examples to provide a concrete disclosure of representative techniques.

1 Introduction

Many real-world scenarios produce data can be modeled by the temporal point process (TPP). Examples include device disorders with their error codes, earthquake with magnitude and location whereby various types of asynchronous events interact with each other and exhibit complex dynamic patterns in the continuous time domain. Investigating this dynamic process and the underlying causal relationship will lay the foundation for further applications such as micro and macro level event prediction, root cause diagnose. Specifically it has facilitated the tackling of many domain-specific applications, with the increasing availability to large-scale event data e.g. from social media and online commerce.

Despite the abounding literature on time series based sequence models such as Markov chain, hidden Markov model and vector auto-regressive model, learning methods for explicitly addressing problems with asynchronously generated event data only start to emerge over the last decade.

1.1 Preliminaries on temporal point process

Temporal point process (TPP) [Daley and David, 2007] is a classic mathematical tool for modeling stochastic point process in continuous time space which often refers to event sequence as a concrete embodiment[1]. A temporal point process is a random process whose realization involves a sequence of (labeled) events in continuous time. TPP provides a principled treatment by directly absorbing the raw timestamp such that the time information is accurately kept. Compared with the time series representation that converts event sequence by aggregation based on predefined time interval, TPP dismisses the unwanted discretization error which formally refers to the so-called Modifiable Areal Unit Problem [Fotheringham and Wong, 1991] i.e. the learning can be sensitive to the choice of the interval length for aggregation. Moreover, TPP has a remarkably well-established theoretical foundation, and it can mathematically incorporate the whole history without specifying the order as required by Markovian models.

Temporal point process is equivalent to a counting process, denoted by $N(t)$ which counts the number of events before time t. The keystone of TPP is its (conditional) intensity function i.e. the stochastic model for the next event conditioned on the history events. Formally, for an infinitesimal time window $[t, t+dt)$, let $\lambda^*(t)$ be the occurrence rate for the future event conditioned on history $\mathcal{H}_t = \{z_i, t_i | t_i < t\}$ up to but not including time t, we have the following definition:

$$\lambda^*(t)dt = \mathbb{P}(\text{event in } [t+dt] | \mathcal{H}_t) = \mathbb{E}(dN(t)|\mathcal{H}_t)$$

where $\mathbb{E}(dN(t)|\mathcal{H}_t)$ is the expected number of events happened in the interval $(t, t+dt]$ given the historical observations $\mathcal{H}_t$. Note we assume a regular point process [Rubin, 1972] (as most literature do on point process) i.e. two event coincide with likelihood 0 i.e. $dN(t) \in \{0, 1\}$. The $*$ notation reminds us that the function depends on history and we omit $\mathcal{H}_t$ for conciseness. The conditional intensity function has played a central role and many processes vary on how it is parameterized. Readers are referred to the textbook [Daley and David, 2007] for a more detailed and rigorous treatment.

1.2 Motivation of the survey

TPP provides a solid mathematical framework for modeling event sequences [Daley and David, 2007]. However until recently is the machine learning community starting to widely

[1] We use event sequence as the concrete form of stochastic point process for discussion in this paper without loss of generality.

undertake this tool for practical problems. The bond with modern machine learning techniques in turn, has also significantly advanced its theory and methods, e.g. with alternating direction method of multipliers [Zhou *et al.*, 2013a] and adversarial learning [Xiao *et al.*, 2018]. To have a principled picture of recent advances and help readers better understand the technical details to an appropriate extent, a survey from machine learning perspective is welcomed.

There are a few relevant surveys: [González *et al.*, 2016] reviews specific spatio-temporal point process and the latter focuses on self-exciting process. For its theoretical importance and practical dominance in applications, Hawkes process and its applications in finance are reviewed in [Hawkes, 2018] authored by Hawkes himself. More examples and discussion for Hawkes process in finance can be found in another excellent survey [Bacry *et al.*, 2015] which is also tailored to the finance setting. Many relevant works in the machine learning community are missing in these articles, especially for neural network based ones, which is covered in this paper.

The purpose of this survey is to identify recent advances in TPP from the machine learning perspective whereby learning objectives and algorithms are the main focus of this article. The discussion navigates from traditional statistical models to neural network methods, and the latter show its promising capability for learning with massive data and less reliance on prior knowledge. Due to its prevalence in both theoretical study and real-world applications, the self-exciting Hawkes process and its variants are frequently discussed.

1.3 Traditional TPP vs. neural TPP

Temporal point process models are often used for either future prediction or quasi-causality discovery. These two targets are closely related to the central problems in machine learning: model capacity (for prediction accuracy) and model interpretability [Choi *et al.*, 2016]. The history of TPP also evolves from traditional parametric models whereby the conditional intensity function's form is manually prespecified, to more recent neural network based models – we call neural point process in line with the term used in [Mei and Eisner, 2017; Du *et al.*, 2016], which frees the need for explicit parametric intensify form selection. In fact it is generally observed [Mei and Eisner, 2017; Du *et al.*, 2016; Xiao *et al.*, 2017b] that the traditional TPP models with explicit parametric intensity function excel at clear interpretation on the problem for learning, while the neural point process models shows high model capacity for learning arbitrary and unknown distributions. In the following, we will navigate through the works along these two threads. In particular, as the neural point process is a more emerging area, we will describe in more details involving specific formulas used in representative works. As traditional TPP models have been published in many literature in statistics, we call them statistical TPP to differentiate the neural TPP.

2 Traditional Statistical Point Processes

2.1 Likelihood function

We first derive the joint likelihood function based on the conditional intensity function $\lambda^*(t)$. Note that the following derivation is based on the definition of so-called regular point process [Rubin, 1972]. For notational conciseness for a compact survey, in the following we ignore the event mark which will not make significant change of the equations.

The joint density function can be written by:

$$f(t_1, t_2, \ldots, t_n) = \prod_j f^*(t_j|\mathcal{H}_j) \tag{1}$$

where $\mathcal{H}_j = (t_1, t_2, \ldots, t_{j-1})$ is the history for event at t_j starting from t_0 up to time t_j but not including t_j. Recall that the conditional density function $f^*(t_{j+1})$ (again we omit $\mathcal{H}_j$ for f^*) has an elegant relation with the conditional intensity function: $\lambda^*(t)S^*(t) = f^*(t)$ where $S^*(t) = \exp\left(-\int_{t_j}^{t_{j+1}} \lambda(\tau)d\tau\right)$ is the probability that no new event occurs up to time t since t_j. Hence for each t_{j+1} we have

$$f^*(t_{j+1}) = \lambda^*(t_{j+1}) \exp\left(-\int_{t_j}^{t_{j+1}} \lambda(\tau)d\tau\right) \tag{2}$$

Accordingly the log-likelihood function can be written by:

$$\log f(t_1, t_2, \ldots, t_n) = \sum_{j=1}^{n} \log \lambda^*(t_j) - \int_{t_0}^{t_n} \lambda^*(\tau)d\tau \tag{3}$$

Note that for conciseness, the presented equations in this paper are mainly for unmarked point process. Under the same framework, readers are referred to [Liniger, 2009] for the details for multi-dimensional case i.e. marked point process.

2.2 Popular intensity function forms

Traditional works are mostly developed around the innovation of the intensity function such as:

i) **Poisson process**: the basic form is history independent $\lambda(t) = \lambda_0$ which can be dated back to the 1900's. Relaxing the constant constraint to let $\lambda(t)$ a function of time leads to the on-homogeneous Poisson process and further extension to stochastic process leads to the famous doubly stochastic Poisson process, also called Cox Process first appeared in [Cox, 1955]; ii) **Reinforced poisson processes** [Pemantle, 2007]: the model captures the 'rich-get-richer' mechanism by $\lambda(t) = \lambda_0 f(t) i(t)$ where $f(t)$ mimics the aging effect while $i(t)$ is the accumulation of history events; iii) **Self-exciting process** [Hawkes, 1971]: also called Hawkes process, it provides an additive model to capture the self-exciting effect from history events $\lambda(t) = \lambda_0 + \sum_{t_i<t} g_{exc}(t - t_i)$. This model also has an alternative representation by the Poisson branching process [Hawkes and Oakes, 1974]; iv) **Reactive point process** [Ertekin *et al.*, 2015]: it can be regarded as a generalization for the Hawkes process by adding a self-inhibiting term to account for the inhibiting effects from history events $\lambda(t) = \lambda_0 + \sum_{t_i<t} g_{exc}(t-t_i) - \sum_{t_i<t} g_{inh}(t-t_i)$; v) **Self-correcting process** [Isham and Westcott, 1979]: its background part increases steadily, while it is decreased by a constant $e^{-\alpha} < 1$ every time a new event appears.

2.3 Maximum likelihood based learning

Parametric models TPP are mostly learned by optimizing log-likelihood or its lower bound. For instance, efforts have been devoted to learning parametric Hawkes process whereby

the background term and triggering term have explicit forms e.g. a constant and an exponential kernel respectively.

Note Eq. 3 involves an accumulation over previous points rendering analytical optimization intractable. In [Ozaki, 1979], gradients and Hessian of the log-likelihood function are explicitly computed while the convergence can be slow. In the seminal work [Veen and Schoenberg, 2008], an expectation-maximization (EM) framework is devised to construct a bound of the objective and at each iteration the parameters are decoupled such that they can be solved independently. For multi-dimensional process with typed events, techniques e.g. sparse low-rank regularization are used to mitigate the curse of high-dimensionality [Zhou *et al.*, 2013a]. Departure from EM based approaches, other techniques e.g. sampling based methods are adopted in other forms of intensity functions e.g. the reactive point process (RPP) [Ertekin *et al.*, 2015]. In fact the Hawkes process can be regarded as a branching process whereby the background rate and triggering effect can be declustered via EM akin to the estimation for Gaussian mixture model. Such properties do not hold for other processes e.g. RPP.

Nonparametric models To improve the model capacity, nonparametric methods are devised whereby the terms of the intensity may not be explicitly parameterized, and implicit regularization are often added. There are also rich literature on nonparametric Hawkes process learning. Due to the aforementioned branching nature, EM based methods are widely used: following the seminal method termed independent stochastic declustering (MISD) [Marsan and Lengline, 2008], an improvement with additional regularizers called maximum penalized likelihood estimation (MPLE) [Lewis and Mohler, 2011] is devised to handle nonparametric form of the triggering kernels, whereby the ordinary differential equation (ODE) is adopted. These works inspire the extension to multi-dimensional Hawkes process [Zhou *et al.*, 2013b].

2.4 Working examples for Hawkes process

Hawkes process so far has been a dominant point process in both statistical and machine learning literature. We show how to learn a parametric self-exciting Hawks process embodiment and its nonparametric variant based on maximum likelihood learning [Lewis and Mohler, 2011].

Parametric example We start from the parametric form. Let the intensity be a classic additive form:

$$\lambda^*(t) = \underbrace{\mu}_{\text{background rate}} + \underbrace{\alpha w \sum_{t_i<t} e^{-w(t-t_i)}}_{\text{exponential exciting } g(t-t_i)} \tag{4}$$

Here w^{-1} inteprets the average waiting time for a new event. Let $t_0 = 0$ and $t_{n+1} = T$, then Eq. 3 can be written by:

$$\begin{aligned}
&\sum_{i=1}^{n} \log \lambda^*(t_i) - \int_0^T \left(\mu + \alpha \sum_{t_j<t} g(t-t_j)dt\right) \\
=&\sum_{i=1}^{n} \log \lambda^*(t_i) - \left(\mu T + \sum_{i=0}^{n} \int_{t_i}^{t_{i+1}} \alpha \sum_{t_j<t} g(t-t_j)dt\right) \\
=&\sum_{i=1}^{n} \log \lambda^*(t_i) - \left(\mu T + \sum_{j=1}^{n} \alpha \left(G(T-t_j) - G(0)\right)\right)
\end{aligned}$$

where $G(t)$ is the integral of exciting term $g(\tau)$ starting from $\tau = 0$. As the log function is concave and $G(0)$=0, we can derive a lower bound by adding auxiliary variable p_{ij} and p_{ii}:

$$\begin{aligned}
&\sum_{i=1}^{n} \log \left(\mu + \sum_{j=1}^{i-1} g(t_i - t_j)\right) - \left(\mu T + \sum_{j=1}^{n} \alpha G(T-t_j)\right) \\
\geq &\sum_{i=1}^{n} \left(p_{ii} \log \frac{\mu}{p_{ii}} + \sum_{j=1}^{i-1} p_{ij} \log \frac{g(t_i - t_j)}{p_{ij}}\right) - \mu T - \sum_{j=1}^{n} \alpha G(T-t_j)
\end{aligned} \tag{5}$$

As shown in [Lewis and Mohler, 2011] which is originated from [Veen and Schoenberg, 2008], an EM method can be derived. The E-step we estimate p_{ij} and p_{ii}:

$$\begin{aligned}
p_{ij}^{k+1} =& \frac{\alpha^k g(t_i - t_j)}{\mu^k + \sum_{j=1}^{i-1} \alpha g(t_i - t_j)}, \quad j = 1, ..., i-1 \\
p_{ii}^{k+1} =& \frac{\mu^k}{\mu^k + \sum_{j=1}^{i-1} \alpha g(t_i - t_j)}
\end{aligned} \tag{6}$$

In the M-step, zeroing partial derivatives $\frac{\partial L}{\partial \mu}$, $\frac{\partial L}{\partial \alpha}$ we have:

$$\mu^{k+1} = \frac{1}{N} \sum_{i=1}^{n} p_{ii}^k, \quad \alpha^{k+1} = \frac{\sum_{i>j} p_{ij}^k}{\sum_{j=1}^{n} G(T-t_j)} \tag{7}$$

For w, it can not be solved in analytical form and one can use the approximation: $e^{-w(T-t_i)} \approx 0$. Then the scale parameter w and α can be simplified by:

$$w^{k+1} = \frac{\sum_{i>j} p_{ij}^k}{\sum_{i>j} (t_i - t_j) p_{ij}^k}, \quad \alpha^{k+1} = \frac{1}{n} \sum_{i>j} p_{ij}^k \tag{8}$$

Nonparametric example We turn the above model to a nonparametric one, with no explicit specification on $\mu(t)$ and $g(t)$. Meanwhile two regularization terms for the background rate and exciting effect are added. Now Eq. 3 becomes:

$$\sum_{j=1}^{n} \log \lambda^*(t_j) - \int_0^T \lambda^*(\tau)d\tau - \alpha_1 R(\mu) - \alpha_2 R(g) \tag{9}$$

Though the above problem can be maximized directly by Euler-Lagrange equation for nonparametric $\mu(t)$ and $g(t)$, an EM treatment can significantly simplify the computing.

Specifically, one introduce the auxiliary random variable $\mathcal{X}_{ij}$ ($\mathcal{X}_{ij} = 1$ if event i is caused by event j otherwise 0) and $\mathcal{X}_{ii}$ ($\mathcal{X}_{ii} = 1$ if event i is caused by background otherwise 0). Then the objective becomes as follows ($R(\mu)$, $R(g)$ are omitted) as used in [Marsan and Lengline, 2008]:

$$\begin{aligned}
&\sum_{i=2}^{n} \left(\sum_{j=1}^{i-1} \mathcal{X}_{ij} \log(g(t_i - t_j)) - \int_{t_j}^{T} g(\tau - t_j)d\tau\right) \\
&+ \sum_{i=1}^{n} \mathcal{X}_{ii} \log(\mu(t_i)) - \int_0^T \mu(\tau)d\tau
\end{aligned} \tag{10}$$

In [Lewis and Mohler, 2011], the EM solver deals with the following subproblem for μ in each iteration (similar for g):

$$\sum_{i=1}^{n} \mathcal{X}_{ii} \log(\mu(t_i)) - \int_0^T \mu(\tau)d\tau + \alpha_1 \|\mu^{\frac{1}{2}}\|_2^2, \mu > 0, \mu \in L^1(\mathbb{R}) \tag{11}$$

By letting $u = \sqrt{\mu}$ one can obtain an Euler-Lagrange equation which can be solved using off-the-shelf numerical solver:

$$-\alpha u''(t) + Cu(t) = \frac{D}{u(t)} \tag{12}$$

where C, D are fixed coefficients for solving $u(t)$. Similar techniques are used in [Zhou *et al.*, 2013b] for extending to multi-dimensional marked Hawkes process.

Remark Compared with their parametric counterparts, nonparametric models have higher capacity to fit complex data, especially given little knowledge for selecting the appropriate model. However nonparametric models call for more complex algorithms preventing them from adoption by practitioners. A notable trend is resorting to the neural network based approach for modeling TPP, whereby end-to-end using learning can be easily performed by off-the-shelf solvers e.g. stochastic gradient descent and tools e.g. Caffe and Tensorflow. We will discuss these works in Sec. 3. While parametric models can excel when there is a little training data.

3 Neural Point Process

As discussed above, traditional models either suffer from model mis-specification if the chosen parametric intensity function does not fit with the real behavior of the event data, or the learning algorithm can be mathematically very complex for nonparametric models. With the fast development of deep learning theory and techniques, especially for recurrent neural network models, there is a trend for adapting networks to temporal point process learning.

3.1 RNNs for point process

Recurrent neural network (RNN) and its variants e.g. long-short term memory (LSTM) has been the building block for learning neural point process. Taking a sequence $\{\mathbf{x}\}_{t=1}^{T}$ as input, the RNN generates the hidden states $\{\mathbf{h}\}_{t=1}^{T}$ encoding the history, and output a sequence of estimated distribution for event mark and timestamp. One key advantage is that the model can often be learned end-to-end with no need for devising tailored algorithms for tailored models as done in traditional TPP models as described in Sec. 2.

3.2 Maximum-likelihood learning

Akin to the traditional models in Sec. 2, maximum likelihood has been adopted as the learning objective in many network based models [Du *et al.*, 2016; Mei and Eisner, 2017]. Given the sequence set $\{S^i\}$ for $S^i = (t_j^i, y_j^i)_{j=1}^{n_i}$, by assuming the independence of the event mark and timestamp, the objective can be simplified by the factorized model:

$$\max \sum_i \sum_j \Big(\underbrace{\log P(y_{j+1}^i | \mathbf{h}_j)}_{\text{Mark likelihood loss}} + \underbrace{\log f(t_{j+1}^i | \mathbf{h}_j)}_{\text{Time likelihood loss}} \Big) \tag{13}$$

There are many choices for modeling the above two losses. In [Du *et al.*, 2016], $P(y_{j+1}^i|\mathbf{h}_j)$ is specified by the cross-entropy loss and the conditional density function can be directly applied for $f(t_{j+1}^i|\mathbf{h}_j) = f^*(t_{j+1})$. One major difference between [Du *et al.*, 2016] and [Mei and Eisner, 2017] is that the former uses one intensity function for all types while the latter allocates respective intensity functions to each event type. There are technical variants under the above framework. We describe two representative works as follows.

The authors [Xiao *et al.*, 2017b] propose to use two separate RNNs for TPP modeling. The time series RNN can carry such fast-changing dense information e.g. body temperature, heartbeat while the event RNN can capture more abrupt dynamics like clinical trials which happen with a longer interval.

To improve interpretability for neural point process which excels at prediction accuracy while lacks of interpretability, attention based RNN has been developed [Wang *et al.*, 2017a]. The hope is that the impact (either negative or positive) of the preceding events to the current one can be captured such that the hidden network consisted by event type as node and mutual impact as weighted edges can be uncovered. In [Wang *et al.*, 2017a], a coverage strategy is introduced to mitigate the misallocation of attention due to the memoryless of traditional attention mechanism. Using a two-level attention model for visits and clinical variables, the REverse Time AttentIoN model (RETAIN) [Choi *et al.*, 2016] is devised to improve clinical interpretation.

3.3 Likelihood-free learning

Likelihood maximization is asymptotically equivalent to minimizing the Kullback-Leibler (KL) divergence requiring strict matching between two probability distributions, which is sensitive to noise and outliers especially given multi-modal distributions. Generative adversarial networks (GAN) have proven to be a promising alternative with extensive theoretical foundation and empirical verification. Recent improvement of Wasserstein GAN (WGAN) [Arjovsky *et al.*, 2017] replaces the Jensen-Shannon (JS) distance adopted in the original GAN [Goodfellow *et al.*, 2014] by the earth moving (EM) i.e. Wasserstein distance. Compared with the KL distance, the W-distance is more sensitive to the underlying geometry structure of samples and robust to issues like mode dropping in case of multi-modal distribution. We review two state-of-the-art methods adapting Wasserstein GAN to TPP that cover both unconditional generative point process models [Xiao *et al.*, 2017a] and conditional ones [Xiao *et al.*, 2018].

Figure 1 illustrates the overview of the architecture for conditional and random based GAN model for TPP.

3.4 Working examples for Wasserstein TPP

Two working examples are also presented in the following.

Random input based Wasserstein learning

We start with the Wassestein distance between two distributions $\mathbb{P}_r$ and $\mathbb{P}_\theta$ which is defined by:

$$\begin{aligned} W(\mathbb{P}_r, \mathbb{P}_\theta) &= \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_\theta)} \mathbb{E}_{(\xi,\eta)\sim\gamma}[c(\xi,\eta)] \\ &= \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_\theta)} \int_{\mathcal{X}\times\mathcal{X}} c(\xi,\eta) d\gamma(\xi,\eta) \end{aligned} \tag{14}$$

where $\Pi(\mathbb{P}_r, \mathbb{P}_\theta)$ denotes the set of all joint distributions $\gamma(\xi,\eta)$ whose marginals are respectively $\mathbb{P}_r$ and $\mathbb{P}_\theta$, and $c(\xi,\eta)$ is the cost function c: $\mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+$.

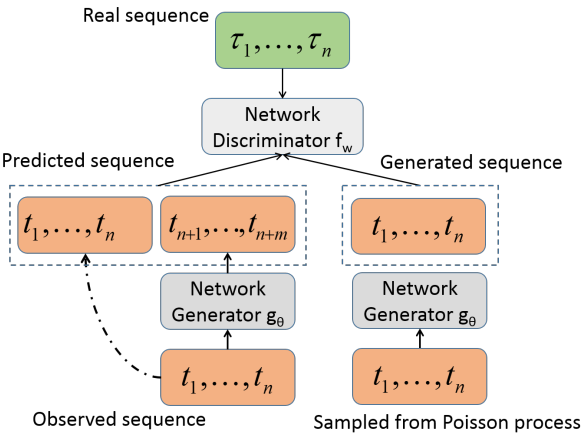

Figure 1: Input and output for adversarial TPP learning. Left branch: conditional GAN model; right: random input based model.

Instead of directly solving Eq. 14, [Arjovsky *et al.*, 2017] turns to its Kantorovich-Rubinstein duality written by:

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{||f_w||_L \leq 1} \mathbb{E}_{\xi \sim \mathbb{P}_r}[f_w(\xi)] - \mathbb{E}_{\eta \sim \mathbb{P}_\theta}[f_w(\eta)] \quad (15)$$

for all the 1-Lipschitz functions f mapping event sequence to a real number: $\mathcal{X} \mapsto \mathbb{R}$. Here we parameterize f_w with parameter w to approximate its search space.

In the context of WGAN, one aims to find a parameterized generator g_θ whose generated distribution is close to the real data ξ w.r.t. Wasserstein distance and the mapping function f_w is called the discriminator or critic. Hence we let $\eta = g_\theta(\zeta_m)$. Moreover to enforce the Lipschitz constraints, meanwhile avoiding the computation of the gradient which can be costly, regularization can be added with weight ν:

$$\min_\theta \max_w \underbrace{\frac{1}{L}\sum_{l=1}^{L} f_w(\xi_l) - \sum_{l=1}^{L} f_w(g_\theta(\zeta_l))}_{\text{Wasserstein distance (random input)}} - \underbrace{\nu \sum_{l,m=1}^{L} \left| \frac{|f_w(\xi_l) - f_w(g_\theta(\zeta_m))|}{|\xi_l - g_\theta(\zeta_m)|_\star} - 1 \right|}_{\text{Gradient-free Lipschitz regularizer}} \quad (16)$$

We analyze the terms in the formula as follows:

i) **Unconditional generative model.** Here g_θ is the sequence generator and ζ_l is the sequence sampled from a Poisson process, akin to the role of the uniform distribution $[0, 1]$ used in GAN for vector like data generation.

ii) **Gradient-free Lipschitz regularizer.** The key for this regularizer is how to define the Wasserstein distance between two sequence (i.e. the denominator $|\xi_l - g_\theta(\zeta_m)|_\star$). In fact it has been shown that this is equivalent to the optimal transport problem which involves the doubly stochastic matrix for mapping sequence points in the two parts. Moreover, by Birkhoff's theorem, the extreme points of the set of doubly stochastic matrices is a permutation. Accordingly the Wasserstein distance can be specified as follows which is proved indeed a valid norm [Xiao *et al.*, 2017a]:

$$\min_\sigma \sum_{i=1}^{n} \|x_i - y_{\sigma(i)}\| + \sum_{i=n+1}^{m} \|s - y_{\sigma(i)}\| \quad (17)$$

where s is a given limiting point at the border of the compact space S and the minimum covers all permutations $1 \ldots m$ such that the second term penalizes unmatched points.

It can be further shown that the identity permutation i.e. $\sigma(i) = i$ (σ is permutation mapping) is the minimizer in (17) and it leads to the reduced form for a time window $[0, T)$:

$$\|\xi - \rho\|_\star = \sum_{i=1}^{n} |t_i - \tau_i| + \sum_{i=n+1}^{m} (T - y_i) \quad (18)$$

Conditional Wasserstein learning

However, the above method can only model the overall distribution of the whole training set, which cannot be directly used for individual sequence modeling. The conditional GAN technique is explored by [Xiao *et al.*, 2018] to enable individual level prediction from its history observation. Specifically, the overall objective function becomes:

$$\min_\theta \max_w \underbrace{\sum_{l=1}^{L} f_w(\{\zeta_l, \rho_l\}) - \sum_{l=1}^{L} f_w(\{\zeta_l, g_\theta(\zeta_l)\})}_{\text{Wasserstein distance between two distribution (conditional)}} - \underbrace{\lambda |f'_w(\hat{x}) - 1|}_{\text{1-Lipschitz regularizer}} - \underbrace{\sigma \log(P_\theta(\rho|\zeta))}_{\text{Likelihood loss}} \quad (19)$$

where η_l is the observed sequence, ρ_l is the real sequence for prediction, λ, σ are weights. We discuss this formula:

i) **Conditional generative model.** Here ζ_l is the observed history for each individual sequence l, differing from the unconditional model sampling from the Poisson process. In fact, $g_\theta(\zeta_l)$ is embodied by a sequence-to-sequence recurrent network (seq2seq LSTM) in [Xiao *et al.*, 2018].

ii) **New Lipschitz regularizer.** Inspired by the improved technique for WGAN [Arjovsky *et al.*, 2017], the gradient based regularizer is used. Note $\hat{x}$ is the interpolation of $\{\zeta_l, \rho_l\}$ and $\{\zeta_l, g_\theta(\zeta_l)\}$ which can be randomly sampled.

iii) **Combining with likelihood loss.** Though the likelihood loss or KL divergence only considers the relative probability of two samples instead of their closeness, the advantage is that it is an unbiased estimation of parameters while Wasserstein distance has biased gradients. Hence it is natural to add the likelihood loss to make the best of the two worlds.

4 Applications and Further Discussion

4.1 Scenarios and practical challenge

Prediction Among the popular models, the self-exciting Hawkes process has received extensive attentions which is first used for seismology. Recently it spans a wide range of applications such as finance [Errais *et al.*, 2010], bioinformatics [Reynaud-Bouret *et al.*, 2010], criminology [Mohler *et al.*, 2011], equipment maintenance [Ertekin *et al.*, 2015], terrorist [Porter *et al.*, 2012], and social network [Du *et al.*, 2015], to name a few. These applications typically involve generative models which aim to estimate the model parameters via maximum likelihood estimator for the observed history events.

Given learned TPP models, the application often involves accurate future event prediction. Due to the stochastic nature, related techniques often involve approximations and heuristics corrections e.g. RPP [Gao *et al.*, 2015], SEISMIC [Zhao

et al., 2015]. The classic and general approach is by simulating future events using Shedler-Lewis thinning algorithm or Ogata's modified thinning algorithm [Ogata, 1981] which suffers from the edge effects. Interestingly [Wang *et al.*, 2017b] presents a principled paradigm for linking microscopic event data to macro scope prediction, with a jump stochastic differential equation model.

Clustering Compared with time series, clustering a set of event sequences into different clusters is more challenging due to the representation difficulty of event sequence. A few works tend to learn each sequence's point process model and then perform clustering based on the distance among the learned models either in a parametric [Luo *et al.*, 2015] or nonparametric way [Lian *et al.*, 2015]. The drawback is that the clustering step can be sensitive to the learning results, while the learning involves too many models incurring overfitting. In the seminal work [Xu and Zha, 2017], a method for joint clustering and point process model learning is presented i.e. the clustered event sequences share the same learned model. The method is a Dirichlet mixture model of Hawkes processes and the local identifiability problem is also studied.

Another typical scenario is intra sequence clustering, i.e. given one event sequence, one aims to relabel each event into different subsequences which can be overlapped with each other over time. For instance, in [Du *et al.*, 2015] the authors aim to cluster streaming news overtime whereby each news is labeled with different clusters, and the topic and temporal model of news in the same cluster is learned by the Dirichlet-hawkes processes model. An offline setting is considered in [Yang and Zha, 2013] for clustering news in the diffusion network with clustered temporal and content mixtures. Apparently this setting is more challenging as all events are mixed in a single sequence. Hence, additional information e.g. textual content for news event is infused to help disambiguation.

Causality discovery One fundamental task for multi-dimensional point process is to learn the Granger causality [Granger, 1969] as originally applied to processes in discrete time. Extensions to continuous time marked point process have been made in [Didelez, 2008] which builds a directed Granger causality graph (or local independence graph) over the dimensions of point process. Different from the time series based Granger causality which can be captured by vector auto-regressive (VAR) model, causality learning for MPP is more challenging. For the Hawkes process, the connection between the Granger causality and impact functions has been revealed in [Eichler *et al.*, 2017]: uncovering whether type-u event Granger-causes type-v event or not is equivalent to detecting whether the impact function $\phi_{uv(t)}$ is all-zero or not. Based on this, [Xu *et al.*, 2016] extends the EM method for learning the impact function with a series of basis kernels.

Censored sequence Most existing TPP methods assume the observation window is complete from scratch which in real world can hardly be satisfied. For instance, a patient usually visits more than one hospitals in her life, and one hospital can only record a subsequence of visits. Hence the observation window for one hospital is often censored. This problem has been well studied for survival analysis [Klein and Moeschberger, 2005]. Recent works present global [Streit, 2010] and local [Fan, 2009] maximum likelihood estimators for point process, which is different from the traditional bootstrap method [Cowling *et al.*, 1996]. For Hawkes process, [Xu *et al.*, 2017] proposes a sampling-stitching synthesis method to recombine short censored sequences.

Event attribution Event attribution involves hypothesizing about the unobserved attributes e.g. actors, diffusion paths, types. The feasibility of this problem depends on the nature for how the events are generated by the process. For instance, the missing event labels cannot be recovered for a set of independent homogenous Poisson processes. Fortunately most real-world processes show highly non-homogenous and history-dependent temporal patterns, suggesting the non-trivial correlation (recall the concept Granger causality).

By assuming the model parameters of the Hawkes process are known, the unknown actors in gang network is estimated in [Stomakhin *et al.*, 2011]. While [Hegemann *et al.*, 2012; Li and Zha, 2013; Cho *et al.*, 2014] manage to iteratively estimate the missing actors and point process model parameters and the mean-field variational optimization approach [Li and Zha, 2013] provides a more tight surrogate objective function, while the work [Cho *et al.*, 2014] devises a latent point process model and a variational EM algorithm for both learning and inference whereby both space and time are considered.

4.2 Further discussion

In the above examples, the target sequence is learned by TPP for different tasks. Compared with feature based regression/classification models, TPP enjoys several benefits:

From the input perspective, regression/classification methods ignore the fine-grained temporal dynamics and require laborious and ad-hoc feature engineering to convert timestamps into fixed-length features (e.g. mean, maximum, minimum, variance, etc.). While TPP provides a more principled approach whereby the accurate time information is retained and the whole history can be incorporated to facilitate the tasks of prediction, clustering and causality discovery.

For output, in particular, TPP can generate predictions for arbitrary time window as the objective is often for explaining the joint probability of the history events (at least for maximum likelihood based methods), which has nothing to do with a specific future window for prediction. In contrast, regression/classification based models often involve a fixed-length vector as the prediction target such that the prediction window is predefined in learning and model testing. This limits their flexibility in for prediction with different periods. For instance, for equipment failure prediction [Ertekin *et al.*, 2015], repairman may be more interested in knowing the risk for next week while the maintenance budget planner cares more about the overall risk for the next whole year.

5 Concluding Remarks

We have witnessed the fast adoption and development of TPP by machine learning community, which covers both traditional statistical approaches and emerging network based models. We believe the neural point process methodology opens up the new space for more expressive modeling of point process, less demand for prior knowledge, which pushes the frontier of machine learning towards event data.

References

[Arjovsky *et al.*, 2017] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017.

[Bacry *et al.*, 2015] E. Bacry, I. Mastromatteo, and J. Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.

[Cho *et al.*, 2014] Y. Cho, A. Galstyan, P J. Brantingham, and G. Tita. Latent self-exciting point process model for spatial-temporal networks. *Discrete & Continuous Dynamical Systems-B*, 19(5):1335–1354, 2014.

[Choi *et al.*, 2016] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NIPS*, pages 3504–3512, 2016.

[Cowling *et al.*, 1996] A. Cowling, P. Hall, and M. J Phillips. Bootstrap confidence regions for the intensity of a poisson point process. *Journal of the American Statistical Association*, 91(436):1516–1524, 1996.

[Cox, 1955] David R Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 129–164, 1955.

[Daley and David, 2007] D.J. Daley and Vere-Jones David. *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Springer Science & Business Media, 2007.

[Didelez, 2008] V. Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):245–264, 2008.

[Du *et al.*, 2015] N. Du, M. Farajtabar, A. Ahmed, A. J Smola, and L. Song. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *SIGKDD*, pages 219–228. ACM, 2015.

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

[Eichler *et al.*, 2017] M. Eichler, R. Dahlhaus, and J. Dueck. Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242, 2017.

[Errais *et al.*, 2010] E. Errais, K. Giesecke, and L. R Goldberg. Affine point processes and portfolio credit risk. *SIAM Journal on Financial Mathematics*, 1(1):642–665, 2010.

[Ertekin *et al.*, 2015] S. Ertekin, C. Rudin, and T. H McCormick. Reactive point processes: A new approach to predicting power failures in underground electrical systems. *The Annals of Applied Statistics*, 9(1):122–144, 2015.

[Fan, 2009] Chun-Po Steve Fan. *Local likelihood for interval-censored and aggregated point process data*. PhD thesis, 2009.

[Fotheringham and Wong, 1991] A S. Fotheringham and D. WS Wong. The modifiable areal unit problem in multivariate statistical analysis. *Environment and planning A*, 23(7):1025–1044, 1991.

[Gao *et al.*, 2015] S. Gao, J. Ma, and Z. Chen. Modeling and predicting retweeting dynamics on microblogging platforms. In *WSDM*, pages 107–116. ACM, 2015.

[González *et al.*, 2016] J. A González, F. J Rodríguez-Cortés, O. Cronie, and J. Mateu. Spatio-temporal point process statistics: a review. *Spatial Statistics*, 18:505–544, 2016.

[Goodfellow *et al.*, 2014] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[Granger, 1969] C. WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[Hawkes and Oakes, 1974] A. G Hawkes and D. Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974.

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

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

[Hegemann *et al.*, 2012] R Hegemann, E Lewis, and A Bertozzi. An "estimate & score algorithm" for simultaneous parameter estimation and reconstruction of missing data on social networks. *Security Informatics*, 2:1–14, 2012.

[Isham and Westcott, 1979] V. Isham and M. Westcott. A self-correcting pint process. *Advances in Applied Probability*, 37:629–646, 1979.

[Klein and Moeschberger, 2005] J. P Klein and M. L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

[Lewis and Mohler, 2011] E. Lewis and E. Mohler. A nonparametric em algorithm for multiscale hawkes processes. *Journal of Nonparametric Statistics*, 2011.

[Li and Zha, 2013] L. Li and H. Zha. Dyadic event attribution in social networks with mixtures of hawkes processes. In *CIKM*, pages 1667–1672. ACM, 2013.

[Lian *et al.*, 2015] W. Lian, R. Henao, V. Rao, J. Lucas, and L. Carin. A multitask point process predictive model. In *ICML*, pages 2030–2038, 2015.

[Liniger, 2009] T. Liniger. *Multivariate hawkes processes*. PhD thesis, ETH Zurich, 2009.

[Luo *et al.*, 2015] D. Luo, H. Xu, Y. Zhen, X. Ning, H. Zha, X. Yang, and W. Zhang. Multi-task multi-dimensional hawkes processes for modeling event sequences. In *IJCAI*, pages 3685–3691, 2015.

[Marsan and Lengline, 2008] D. Marsan and O. Lengline. Extending earthquakes' reach through cascading. *Science*, 319(5866):1076–1079, 2008.

[Mei and Eisner, 2017] H. Mei and J. M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *NIPS*, pages 6757–6767, 2017.

[Mohler *et al.*, 2011] G. O Mohler, M. B Short, P J. Brantingham, F. P. Schoenberg, and G. E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

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

[Ozaki, 1979] T. Ozaki. Maximum likelihood estimation of hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, 1979.

[Pemantle, 2007] R. Pemantle. A survey of random processes with reinforcement. *Probability Survey*, 4(0):1–79, 2007.

[Porter *et al.*, 2012] M. D Porter, G. White, et al. Self-exciting hurdle models for terrorist activity. *The Annals of Applied Statistics*, 6(1):106–124, 2012.

[Reynaud-Bouret *et al.*, 2010] P. Reynaud-Bouret, S. Schbath, et al. Adaptive estimation for hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5):2781–2822, 2010.

[Rubin, 1972] I. Rubin. Regular point processes and their detection. *IEEE Transactions on Information Theory*, 18(5):547–557, 1972.

[Stomakhin *et al.*, 2011] A. Stomakhin, M. B Short, and A. L Bertozzi. Reconstruction of missing data in social networks based on temporal patterns of interactions. *Inverse Problems*, 27(11):115013, 2011.

[Streit, 2010] R. L Streit. *Poisson point processes: imaging, tracking, and sensing*. Springer Science & Business Media, 2010.

[Veen and Schoenberg, 2008] A. Veen and F. P Schoenberg. Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624, 2008.

[Wang *et al.*, 2017a] Y. Wang, H. Shen, S. Liu, J. Gao, and X. Cheng. Cascade dynamics modeling with attention-based recurrent neural network. In *AAAI*, pages 2985–2991, 2017.

[Wang *et al.*, 2017b] Y. Wang, X. Ye, H. Zhou, H. Zha, and L. Song. Linking micro event history to macro prediction in point process models. In *AISTATS*, pages 1375–1384, 2017.

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

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

[Xiao *et al.*, 2018] S. Xiao, H. Xu, J. Yan, M. Farajtabar, X. Yang, L. Song, and H. Zha. Learning conditional generative models for temporal point processes. In *AAAI*, 2018.

[Xu and Zha, 2017] H. Xu and H. Zha. A dirichlet mixture model of hawkes processes for event sequence clustering. In *NIPS*, pages 1354–1363, 2017.

[Xu *et al.*, 2016] H. Xu, M. Farajtabar, and H. Zha. Learning granger causality for hawkes processes. In *ICML*, pages 1717–1726, 2016.

[Xu *et al.*, 2017] H. Xu, D. Luo, and H. Zha. Learning hawkes processes from short doubly-censored event sequences. In *ICML*, pages 3831–3840, 2017.

[Yang and Zha, 2013] S. Yang and H. Zha. Mixture of mutually exciting processes for viral diffusion. In *ICML*, pages 1–9, 2013.

[Zhao *et al.*, 2015] Q. Zhao, M. A Erdogdu, H. Y He, A. Rajaraman, and J. Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *SIGKDD*, pages 1513–1522. ACM, 2015.

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

[Zhou *et al.*, 2013b] K. Zhou, H. Zha, and L. Song. Learning triggering kernels for multi-dimensional hawkes processes. In *ICML*, pages 1301–1309, 2013.