APMLA: Lecture 5

Caterina De Bacco and Isabel Valera

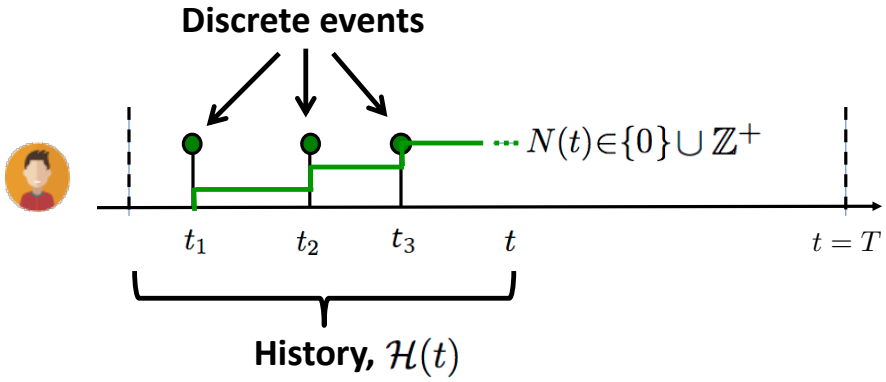

Figure 1: Example of a temporal point process used to represent the activity of a user in online social media, providing a visualization of the relationship between the events $H(t) = \{t_i | t_i < t\}$ generated by the process and its counting process representation $N(t)$.

1 Temporal Point Processes (TPPs)

A *temporal point process* (TPP) is a stochastic process whose realization consists of a sequence of *discrete events localized in continuous time*, i.e., $H(t) = \{t_i | t_i < t\}$ with $t_i \in \mathbb{R}_+$ and $i \in \mathbb{Z}^+$. $H(t)$ is often referred as the *history* of all events until time t. A temporal point process is often represented using a counting process, $N(t)$, which counts the number of events up but not including time t, i.e.,

$$N(t) := \sum_{t_i \in H(t)} u(t - t_i), \tag{1}$$

where $u(t)$ is a step function that takes value one if $t \geq 0$, i.e., $u(t) = 1$ if $t \geq 0$, and zero, otherwise.

Next, it will be useful to define the differential of the counting process $dN(t) = N(t+dt) - N(t) \in \{0, 1\}$ of a counting process, where dt is an arbitrarily small time interval so that either none or only one event can occur in $[t, t + dt)$. Moreover, using the counting process definition given by Eq. 1, we can also write the differential of a counting process as:

$$dN(t) = \sum_{t_i \in H(t)} du(t - t_i) = \sum_{t_i \in H(t)} \delta(t - t_i) dt, \tag{2}$$

where $\delta(\tau)$ is the Dirac delta function which by definition satisfies that $\int_{-\infty}^{\infty} \delta(\tau) d\tau = 1$. Thus $dN(t)$

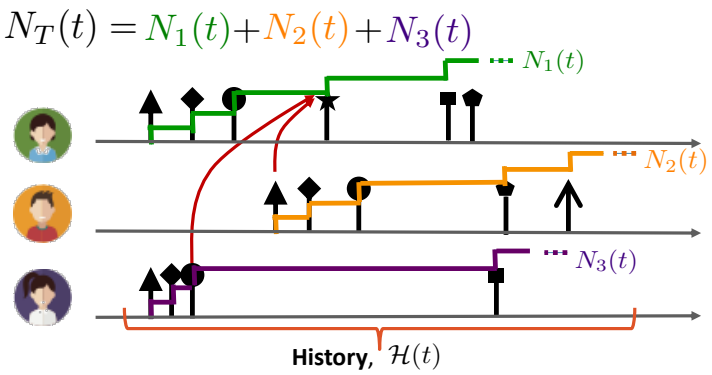

Figure 2: Example of a temporal point process used to represent the activity of a user in online social media, providing a visualization of the relationship between the events $H(t) = \{t_i\}_{t_i \leq t}$ generated by the process and its counting process representation $N(t)$.

results in a train of impulses. Note also that (as expected):

$$\begin{aligned} N(t) &= \sum_{t_i \in H(t)} u(t - t_i) = \sum_{t_i \in H(t)} \int_{-\infty}^{t-t_i} \delta(\tau) d\tau \\ &= \sum_{t_i \in H(t)} \int_{-\infty}^{t} \delta(\tau - t_i) d\tau = \int_{-\infty}^{t} \sum_{t_i \in H(t)} \delta(\tau - t_i) d\tau \\ &= \int_{0}^{t} dN(t). \end{aligned} \tag{3}$$

Figure 1 shows and example of a temporal point process used to represent the activity of a user in online social media.

Superposition property. An interesting property of the counting process representation of a TTP is the fact that, if we have several (either independent or dependent) TPPs modeling, e.g., the online activity of two users, the overall activity of the two users can be represented as a counting process $N_T(t)$ that is the sum of the two individual counting processes $N_1(t)$ and $N_2(t)$, i.e., $N_T(t) = N_1(t) + N_2(t)$. This is therefore a useful property in practice, since it allows us to easily represent the overall activity in, e.g., an online social network of a set of individuals potentially interacting as the sum of individual actions. See Figure 2 for a visual example. Similarly, the history of events performed in the social network is just the union of the individual histories, i.e., $H_T(t) = H_1(t) \cup H_2(t)$.

Event time as a random variable. It is important to notice that a TPP treats the time t of each new event generated by a temporal point process as a continuous random variable whose probability density function (pdf) $f^*(t) = f(t|H(t))$ (here * indicates that we are conditioning on the history) may depend on the previous events $H(t) = \{t_i\}_{t_i < t}$. Note that here $f^*(t)dt$ can be interpreted as the probability of an event occurring during the interval $[t, t + dt)$ conditioned on the history $H(t)$, which tends to zero as $dt \to 0$. Unfortunately, it is difficult to build an intuition about the choice of the functional form for $f^*(t)$ satisfying that the cumulative function $F^*(t) = \int_{t_n}^{t} f^*(t)dt$, where t_n is the last event time in $H(t)$, evaluated in $t = \infty$ is 1 – which is necessary for $f^*(t)$ to be a valid pdf. To make things more complicated, the pdf $f^*(t)$ does not satisfy the superposition property. That is, even if the if we can write that $N_T(t) = N_1(t) + N_2(t)$ (and $H_T(t) = H_1(t) \cup H_2(t)$), unfortunately $f_T^*(t) \neq f_1(t) + f_2^*(t)$.

Question: How do we design (i.e., parameterize), sample (i.e., generate events from) and learn (the parameters of) a TPP?

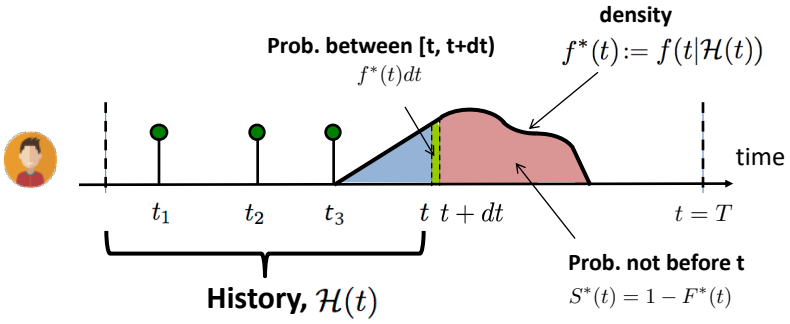

Figure 3: Example of a temporal point process used to represent the activity of a user in online social media, providing a visualization of the relationship between the events $H(t) = \{t_i\}_{t_i \leq t}$ generated by the process and its counting process representation $N(t)$.

2 Intensity function

To overcome the aforementioned difficulties, we characterize the event times of a TPP using the conditional *intensity function* $\lambda^*(t)$, which may depend on the history $H(t)$, and we define as the probability (conditioned on the history $H(t)$) of a new event happening in $[t, t+dt)$ given that no event occur between the last generated event t_n and t, i.e,

$$\lambda^*(t) := \frac{f^*(t)}{1 - F^*(t)} = \frac{f^*(t)}{S^*(t)}, \tag{4}$$

where $S^*(t)$ is often referred to as survival function and accounts for the probability that the next event will not occur before t. Refer to Figure 3 for a visualization of the relationship between $f^*(t)$, $F^*(t)$ and $S^*(t)$.

Since, by definition, the differential $dN(t) \in \{0, 1\}$ can only increase by one event in infinitesimal amount of time dt, it readily follows that

$$E[dN(t)|H(t)] = 1 \times P(dN(t) = 1|H(t)) + 0 \times P(dN(t) = 0|H(t)) = \lambda^*(t)dt, \tag{5}$$

or equivalently, in a more intuitive form:

$$\lambda^*(t) = \lim_{dt \to 0} \frac{P(N(t+dt) = N(t) + 1)}{dt} = \lim_{dt \to 0} \frac{P(dN(t) = 1)}{dt},$$

which thus can be interpreted as the instantaneous rate of events per time unit, e.g., $\lambda^*(t) = 2$ tweets/minute. Characterizing the evolution of a temporal point process using the intensity function has several advantages:

– *Model design*: The interpretation of the intensity $\lambda^*(t)$ as a rate allows us to easily to build an intuition about the choice of its functional form. For example, one can think that social media users post at a higher rate during the day (waking hours) than during the night (sleeping hours). Moreover, we only need to guarantee that the functional form of $\lambda^*(t)$ is nonnegative, i.e., $\lambda^*(t) \geq 0$ for all t.

– *Superposition property*: The intensity function satisfies the superposition property, making it is easy to combine several temporal point processes models. For example, we can obtain the intensity function $\lambda_T^*(t)$ resulting of the superposition of two TPPs as $N_T(t) = N_1(t) + N_2(t)$ (each of them characterized respectively by the intensity functions $\lambda_1^*(t)$ and $\lambda_2^*(t)$) as:

$$\begin{aligned} \lambda_T^*(t)dt &= E[dN_T(t)|H(t)] = E[dN_1(t) + dN_2(t)|H(t)] = E[dN_1(t)|H(t)] + E[dN_2(t)|H(t)] \\ &= (\lambda_1^*(t) + \lambda_2^*(t))dt. \end{aligned} \tag{6}$$

Finally, we just need to connect the intensity function $\lambda^*(t)$ and $f^*(t)$, $F^*(t)$ and $S^*(t)$ as follows:

Relationship between the intensity and the pdf of a TTP. Given a counting process $N(t)$ with $\lambda^*(t), f^*(t)$ and $S^*(t)$, then it holds that the survival function is given by:

$$S^*(t) = \exp\left(-\int_{t_n}^{t} \lambda^*(\tau) d\tau\right), \tag{7}$$

and the pdf takes the form

$$f^*(t) = \lambda^*(t) \exp\left(-\int_{t_n}^{t} \lambda^*(\tau) d\tau\right), \tag{8}$$

where t_n is the last event in $H(t)$, i.e., the last event before time t.

Proof. By definition, we have that $S^*(t) = 1 - \int_{t_n}^{t} f^*(x)dx \Longrightarrow dS^*(t) = -f^*(t)dt$. Together with Eq. 4, this implies that

$$\lambda^*(t) = \frac{f^*(t)}{S^*(t)} = -\frac{1}{S^*(t)} \frac{dS^*(t)}{dt} = -\frac{d}{dt} \log S^*(t).$$

Then, if we integrate the left and right hand sides in the above equation, we obtain Eq. 7. Also by recalling again on $f^*(t) = -dS^*(t)/dt$ and Eq. 7, we obtain that

$$f^*(t) = -\frac{d \exp\left(-\int_{t_n}^{t} \lambda^*(\tau) d\tau\right)}{dt} = \lambda^*(t) \exp\left(-\int_{t_n}^{t} \lambda^*(\tau) d\tau\right).$$

3 Basic Intensity Functions

In this section, we introduce several basic intensity functions, which are often the building blocks of more complex models. For each of these intensity functions, we will learn how to get samples (i.e., generate data) of the corresponding TPPs.

3.1 Homogeneous Poisson process

A (homogeneous) Poisson process is the simplest temporal point process, where the intensity, or the rate of events, is given by a constant parameter μ, i.e.,

$$\lambda_\mu^*(t) = \mu \geq 0. \tag{9}$$

By definition, the intensity is independent of the history $H(t)$, the occurrence of events happens uniformly at random and the inter-event time, i.e., $t_i - t_{i-1}$ for any i, is exponentially distributed with rate parameter μ.

Generative process. To generate samples $\{t_i\}_{t_i \leq T}$ from a homogenous Poisson process with intensity (or rate) equal to μ in an interval of time T:

- Initialize $t = t_0 = 0$ and $i = 1$;
- While t<T:
  1. Set $t = -\frac{log(1-u)}{\mu} + t_{i-1}$, where $u \sim Uniform[0, 1]$. (Inverse transform sampling)
  2. If $t_i < T$, then $t_i = t$ and $i = i + 1$.

3.2 Inhomogeneous Poisson process

An inhomogeneous Poisson process is defined by a time-varying function

$$\lambda^*(t) = g(t) \geq 0, \tag{10}$$

which, by definition, the intensity is independent of the history $H(t)$. The inhomogeneous Poisson process is often used to model patterns in the temporal dynamics. For example, if we use this process to model the online activity of a user, one may expect the user to be more active during the day-light hours than during sleep hours.

Generative process. To generate samples $\{t_i\}_{t_i \leq T}$ from a inhomogeneous Poisson process with intensity equal to $\lambda^*(t) = g(t)$ in an interval of time T:

- Initialize $t = t_0 = 0$, $i = 1$
- Set $g_{max} = \max_\tau g(\tau)$;
- While t<T:
  1. Set $t = -\frac{log(1-u_1)}{g_{max}} + t$, where $u_1 \sim Uniform[0,1]$. (Inverse transform sampling)
  2. Sample $u_2 \sim Uniform[0,1]$
  3. If $u_2 \leq \frac{g(t)}{g_{max}}$, then: (Rejection sampling)
     - $t_i = t$;
     - $i = i + 1$.

3.3 Hawkes process

Hawkes processes (HPs) are a particular type of TPPs especially suited to capture the *self-excitatory* nature of many real-world temporal dynamics, where each new event has the ability to excite (increase) the arrival rate of future events for a certain period of time [2]. As an example, in epidemiology, the probability of a new infection increases with the number of already infected subjects, or equivalently, any new infection event may increase the probability of new infections.

The intensity function of an HP can be written, using the superposition property, as:

$$\lambda^*(t) = \mu(t) + \sum_{t_i \in H(t)} \gamma(t - t_i), \tag{11}$$

where we remark again that the (instantaneous) intensity depends on the history of previous events $H(t)$, and can be divided into two terms: the first term that is independent of previous events (i.e., it is history independent), and the other that depends on the history of events up to time t, $H(t)$. The first term accounts for the *exogenous* intensity, a.k.a., baseline intensity, $\mu(t) \geq 0$ and models the events that are exogenous to the process, i.e., events that are not triggered by previous events as a consequence of self-excitement. The second term in (11), accounts for those events that are due to the self-excitatory nature of the HP, and is often referred as *endogenous* intensity. Here, $\gamma(t) \geq 0$ corresponds the excitatory (triggering) kernel with positive finite measure [2].

A very common parametrization of the HP assumes the exogenous intensity to be constant and the triggering kernel to be exponential, such that:

$$\lambda^*(t) = \mu + \alpha \sum_{t_i \in H(t)} k_w(t - t_i), \tag{12}$$

where $\alpha \geq 0$ captures the maximum influence of previous events in the intensity, and $k_w(t) = \exp(-wt)[t \geq 0]$, modeling w the memory of the HP.

Generative process. To generate samples $\{t_i\}_{t_i \leq T}$ from a Hawkes process with intensity equal to $\lambda^*(t)$ and constant base rate μ in an interval of time T:

- Initialize $t = t_0 = 0$, $i = 1$, and $\lambda_{max} = \lambda^*(t_0) = \mu$;
- While t<T:
  1. Set $t = -\frac{log(1-u_1)}{\lambda_{max}} + t$, where $u_1 \sim Uniform[0,1]$. (Inverse transform sampling)
  2. Sample $u_2 \sim Uniform[0,1]$
  3. If $u_2 \leq \frac{\lambda^*(t)}{\lambda_{max}}$, then: (Rejection sampling)
     - $t_i = t$;
     - $\lambda_{max} = \lambda^*(t_i)$;
     - $i = i + 1$.

3.4 Terminating (a.k.a. survival) point process

A terminating point process finishes once an event happens, i.e.,

$$\lambda^*(t) = g^*(t)(1 - N(t)), \tag{13}$$

where $N(t)$ is the corresponding counting process, $g^*(t)$ is a nonnegative intensity function which may depend on the history, and the intensity $\lambda^*(t)$ becomes zero if an event happens. This TPP is often used to model events that may only happen once, e.g., it can be used to estimate the time of failure of a sensor after installation, being $g^*(t)$ history independent; or the time of an individual getting infected by the flu given the infection times of other individuals, being in this case $g^*(t)$ increasing every time another individual gets infected, and thus, history dependent.

Generative process. One may get samples from a terminating by sampling one event from the corresponding TPP with intensity function of the form $g^*(t)$.

4 Maximum Likelihood Estimation.

Let's assume now that we have collected a sequence of events, e.g., the timestamps of the tweet of a user, in a time window $[0, T)$, such that $H(T) = \{t_i\}_{t_i < T}$. Let's also assume that they have been generated according to a TTP with intensity function $\lambda^*_{\boldsymbol{\theta}}(t)$ which is parametrized by the set of parameters $\boldsymbol{\theta}$. Then, we can use the results from the previous section to write the log-likelihood function as:

$$\begin{aligned} \mathscr{L}(\boldsymbol{\theta}|H(T)) = \log \prod_{i=1}^{n} f^*(t_i) &= \sum_{i=1}^{n} \left(\log \lambda^*_{\boldsymbol{\theta}}(t_i) - \int_{t_{i-1}}^{t_i} \lambda^*_{\boldsymbol{\theta}}(\tau) d\tau \right) \\ &= \sum_{i=1}^{n} \log \lambda^*_{\boldsymbol{\theta}}(t_i) - \int_0^T \lambda^*_{\boldsymbol{\theta}}(\tau) d\tau. \end{aligned} \tag{14}$$

Thus, we can use the above expression to find the parameters $\boldsymbol{\theta}$ that maximize the likelihood as:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \quad \mathscr{L}(\boldsymbol{\theta}|H(T)).$$

5 Marked Temporal TPPs

So far, we have assumed that every event in a TPP is only defined by its timestamp. However, in real world applications we are often interested in modeling events that are not only characterized by the

time when they occur but also by "what happens" at this time, i.e., the observation at this time. For example, in online social media every post performed by a user can be seen as an event characterized by the timestamp of the post and the post itself. In such cases, each event can thus be represented as a tuple of the form:

$$e := (t, \mathbf{x}) \tag{15}$$

where here x is a random variable (the post content in our example), often referred as *mark*, characterizing the event. As an example, events may represent posts on a social network, m could encode either post itself, which may be represented, e.g., as a bag of words, or the post sentiment. In addition, the mark may also include geolocation information about the post, leading to spatio-temporal representation of events.

Formally, a *marked* (TPP) is a stochastic process whose realization consists of a sequence of discrete *marked* events localized in continuous time, i.e., $H(t) = \{(t_i, \mathbf{x}_i) | t_i < t\}$ with $t_i \in \mathbb{R}_+$, $\mathbf{x}_i \in \mathscr{X}$ and $i \in \mathbb{Z}^+$. $H(t)$ is still referred as the *history* of all events until time t.

Similarly as in the case of standard TPPs, the event times are represented using a counting process $N(t)$ and its corresponding intensity function $\lambda^*(t)$, which may depend on past events. However, we need now also a mark distribution. Here, we consider the following choices, which have been used in the literature:

– *Independent and identically distributed marks:* The marks are i.i.d. samples from a fixed distribution $p(\mathbf{x})$, i.e.,

$$\mathbf{x}_i \sim p(\mathbf{x}), \tag{16}$$

being therefore independent of the history.

– *Stochastic differential equations (SDEs) with Jumps:* the marks $\mathbf{x}_i$ are history dependent random variables, which are defined using stochastic differential equation (SDE) with jumps, i.e.,

$$\begin{aligned} \mathbf{x}_i &= \mathbf{x}(t_i) \\ d\mathbf{x}(t) &= f(\mathbf{x}(t), t)dt + h(\mathbf{x}(t)), t)dN(t) \end{aligned} \tag{17}$$

where $f(\cdot)$ and $h(\cdot)$ are domain dependent functions and the second term in the SDE with jumps accounts for the influence of previous events. Here, note that the above SDE with jumps defines the mark values for all values of t, however, a mark only gets realized whenever an event happens.

– *SDEs with Jumps & noise:* the marks $\mathbf{x}_i$ are history dependent random variables, which are defined using stochastic differential equation (SDE) with jumps, i.e.,

$$\begin{aligned} \mathbf{x}_i &\sim p(\mathbf{x}|\boldsymbol{\theta}(t)) \\ d\boldsymbol{\theta}(t) &= f(\boldsymbol{\theta}(t), t)dt + h(\boldsymbol{\theta}(t), t)dN(t) \end{aligned} \tag{18}$$

where as before $f(\cdot)$ and $h(\cdot)$ are domain dependent functions and the second term in the SDE with jumps accounts for the influence of previous events. Note that in this case, we are assuming a distribution over the marks, whose parameters $\boldsymbol{\theta}(t)$ evolve over time as indicated by the SDE.

In the next lecture, we will show an example of marked TPPs with i.i.d. marks, where we will perform clustering of event data. However, in case of interest in marks based on SDEs, please refer for example to [1].

References

[1] A. De, I. Valera, N. Ganguly, S. Bhattacharya, and M. Gomez Rodriguez. Learning and forecasting opinion dynamics in social networks. In *Advances in Neural Information Processing Systems 29*, pages 397–405. 2016.

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