Temporal Point Processes

A. De, U. Upadhyay, M. Gomez-Rodriguez
Notes for Human-Centered ML, Saarland University, Winter 2018-19

January 10, 2019

1 Definition

A temporal point process $\mathcal{T}$ is a random process whose realization consists of a list of discrete events localized in time, $\mathcal{H} = \{t_i\}$ with $t_i \in \mathbb{R}^+$ and $i \in \mathbb{Z}^+$. A temporal point process is often represented using a counting process, $N(t)$, which counts the number of events until time t, *i.e.*,

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

where $u(t) = 1$ if $t \geq 0$ and $u(t) = 0$ otherwise[1]. Additionally, we define $\mathcal{H}(t)$ as the history of all events up to, but not including, time t, *i.e.*, $\mathcal{H}(t) = \{t_i \,|\, t_i < t\}$.

Next, it will be useful to define the differential $dN(t) = N(t + dt) - N(t) \in \{0, 1\}$ of a counting process, where dt is an arbitrarily small time interval so that 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:

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

where $\delta(t)$ is the Dirac delta function, which can be informally understood as a function which is zero everywhere except at $t = 0$, where it is infinite, and which is also constrained to satisfy the identity

$$u(t) = \int_{-\infty}^{t} \delta(\tau) d\tau. \tag{3}$$

The delta function can be rigorously defined either as a distribution or as a measure [OWN96], however, an informal understanding of the delta function will be enough for us here. Finally, using Eq. 2 and Eq. 3, it follows that:

$$\begin{aligned} N(t) &= \sum_{t_i \in \mathcal{H}} u(t - t_i) = \sum_{t_i \in \mathcal{H}} \int_{-\infty}^{t - t_i} \delta(\tau) d\tau = \sum_{t_i \in \mathcal{H}} \int_{-\infty}^{t} \delta(s - t_i) ds \\ &= \int_{-\infty}^{t} \sum_{t_i \in \mathcal{H}} \delta(s - t_i) ds = \int_{0}^{t} dN(s), \end{aligned} \tag{4}$$

Figure 1 illustrates the above concepts using the example of a user's timeline in social media.

[1] The function $u(t)$ is called the Heaviside step function [OWN96].

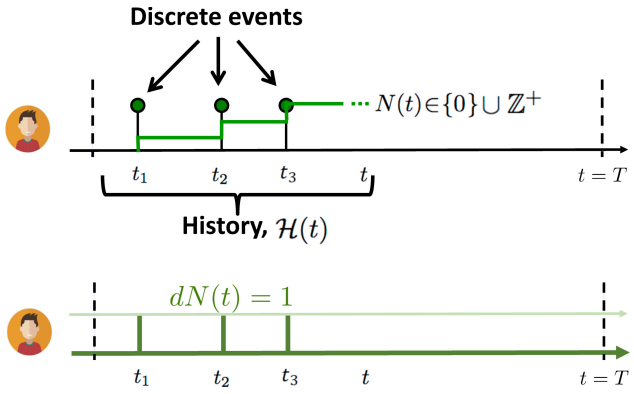

Figure 1: The temporal point process associated with a user's timeline in a social media. The user is posting messages as discrete events during continuous times t_1, t_2, t_3. The upper row indicates that $N(t)$ is the counting process that counts the number of events. The bottom row indicates that the plot of $dN(t)$ is a trail of impulses.

2 Intensity function

By definition, each event time t in a temporal point process is a random variable. Therefore, given $\mathcal{H}(t) = \{t_1, \ldots, t_{i-1}\}$, one could think of characterizing the time t of the next event, the i-th event, using the following three functions, illustrated in Figure 2:

I. A conditional probability density function $f^*(t) = f(t \,|\, \mathcal{H}(t))$, which is the probability that the next event will occur during the interval $[t, t + dt)$ conditioned on the history $\mathcal{H}(t)$.

II. A cumulative distribution function $F^*(t) = F(t \,|\, \mathcal{H}(t)) = \int_{t_{i-1}}^{t} f^*(\tau) d\tau$, which is the probability that the next event will occur before time t conditioned on the history $\mathcal{H}(t)$. Here, t_{i-1} is the last event in $\mathcal{H}(t)$, *i.e.*, the last event before time t.

III. A complementary cumulative distribution function $S^*(t) = S(t \,|\, \mathcal{H}(t)) = 1 - F^*(t)$, also called survival function, which is the probability that the next event will not occur before time t conditioned on the history $\mathcal{H}(t)$.

However, in using the above functions, we would face two main difficulties:

— *Model design:* it would be difficult to build an intuition about the choice of functional form for $f^*(t)$. To complicate things further, we would need to constrain the functional form of $f^*(t)$ to satisfy that $\int_{t_{i-1}}^{\infty} f^*(\tau) d\tau = 1$ in order for it to be a valid probability density function.

— *Model reusability:* it would be difficult to combine several temporal point processes models. For example, assume we are able to accurately characterize the time of the next event in two independent temporal point processes with histories $\mathcal{H}_1(t)$ and $\mathcal{H}_2(t)$ using $f_1^*(t)$ and $f_2^*(t)$, respectively. Then, it is highly nontrivial to characterize the time of the next event in the joint temporal point process with history $\mathcal{H}(t) = \mathcal{H}_1(t) \cup \mathcal{H}_2(t)$ using $f^*(t)$ on the basis of $f_1^*(t)$ and $f_2^*(t)$. In particular, $f^*(t) \neq f_1^*(t) + f_2^*(t)$ and $f^*(t) \neq f_1^*(t) \star f_2^*(t)$, where $\star$ denotes convolution.

To overcome these difficulties, we will characterize the event times of a temporal point process using the conditional intensity function $\lambda^*(t) = \lambda(t \,|\, \mathcal{H}(t))$, which is the conditional probability

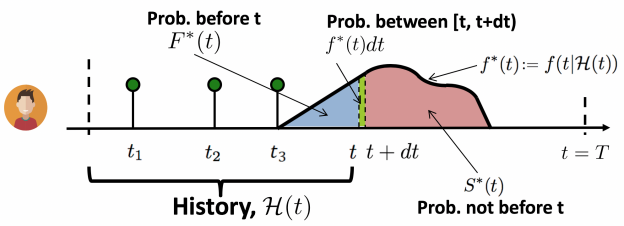

Figure 2: Conditional probability density function $f^*(t)$, cumulative probability distribution function $F^*(t)$, and survival function $S^*(t)$.

that, given the history $\mathcal{H}(t) = \{t_1, \ldots, t_{i-1}\}$ and that it has not happened before t, the next event will happen during $[t, t + dt)$, *i.e.*,

$$\lambda^*(t)dt = \frac{f^*(t)dt}{S^*(t)}. \tag{5}$$

Since, by definition, the differential $dN(t) \in \{0, 1\}$ can only increase by one event at each dt, it readily follows that $\mathbb{P}(dN(t) = 1 \,|\, \mathcal{H}(t)) = \lambda^*(t)dt$ and

$$\begin{aligned}\mathbb{E}[dN(t) \,|\, \mathcal{H}(t)] &= 1 \times \mathbb{P}(dN(t) = 1 \,|\, \mathcal{H}(t)) + 0 \times \mathbb{P}(dN(t) = 0 \,|\, \mathcal{H}(t)) \\ &= \lambda^*(t)dt \end{aligned} \tag{6}$$

Hence, we can also think of the conditional intensity function $\lambda^*(t)$ as an instantaneous rate of events per time of unit, *e.g.*, $\lambda^*(t) = 10$ tweets/minute. Characterizing the evolution of a temporal point process using the intensity function has several advantages:

— *Model design:* using the interpretation of the intensity $\lambda^*(t)$ as a rate, it is easy 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. This is contrast with the conditional probability density function, which needs to integrate to one.

— *Model reusability:* it is easy to combine several temporal point processes models. For example, assume we are able to accurately characterize two independent temporal point processes, with histories $\mathcal{H}_1(t)$ and $\mathcal{H}_2(t)$, using intensities $\lambda_1^*(t)$ and $\lambda_2^*(t)$, respectively.[2] Then, we can characterize the joint history $\mathcal{H}(t) = \mathcal{H}_1(t) \cup \mathcal{H}_2(t)$ by means of

$$\begin{aligned}\lambda^*(t)dt &= \mathbb{E}[dN(t) \,|\, \mathcal{H}(t)] = \mathbb{E}[dN_1(t) + dN_2(t) \,|\, \mathcal{H}(t)] \\ &= \mathbb{E}[dN_1(t) \,|\, \mathcal{H}_1(t)] + \mathbb{E}[dN_2(t) \,|\, \mathcal{H}_2(t)] = \lambda_1^*(t) + \lambda_2^*(t).\end{aligned}$$

Finally, the following proposition relates $\lambda^*(t)$ to $f^*(t)$ and $S^*(t)$.

[2] In reality, if the conditional intensities $\lambda_1^*(t)$ and $\lambda_2^*(t)$ are conditioned on the joint history $\mathcal{H}(t) = \mathcal{H}_1(t) \cup \mathcal{H}_2(t)$, it is only necessary that both processes are conditionally independent given $\mathcal{H}(t)$. However, for ease of exposition, we require independence.

Proposition 1 *Given a counting process $N(t)$ with $\lambda^*(t)$, $f^*(t)$ and $S^*(t)$, then it holds that:*

$$S^*(t) = \exp\left(-\int_{t_{i-1}}^{t} \lambda^*(\tau)d\tau\right) \tag{7}$$

$$f^*(t) = \lambda^*(t)\exp\left(-\int_{t_{i-1}}^{t} \lambda^*(\tau)d\tau\right) \tag{8}$$

where t_{i-1} is the last event in $\mathcal{H}(t)$, i.e.*, the last event before time t.*

Proof By definition, we have that $S^*(t) = 1 - \int_{t_{i-1}}^{t} f^*(x)dx \implies dS^*(t) = -f^*(t)dt$. Together with Eq. 5, this implies that

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

Then, if we integrate the left and right hand sides in the above equation, we obtain Eq. (7). By definition, we have that $dS^*(t) = -f^*(t)dt$. Together with Eq. (7), this implies that

$$f^*(t) = -\frac{d\exp\left(-\int_{t_{i-1}}^{t} \lambda^*(\tau)d\tau\right)}{dt} = \lambda^*(t)\exp\left(-\int_{t_{i-1}}^{t} \lambda^*(\tau)d\tau\right)$$

■

Using the equations given by the above proposition and a conditional intensity function $\lambda^*_\theta(t)$, with a set of model parameters θ, we can compute the (log-)likelihood that $\lambda^*_\theta(t)$ generates a specific history of events $\mathcal{H}(T) = \{t_1, t_2, \ldots, t_n\}$ as:

$$\begin{aligned}\mathfrak{L}(\mathcal{H}(T);\theta) &= \left(\sum_{i=1}^{n} \log \lambda^*_\theta(t_i) - \int_{t_{i-1}}^{t_i} \lambda^*_\theta(\tau)\, d\tau\right) - \int_{t_n}^{T} \lambda^*_\theta(\tau), \\ &= \sum_{i=1}^{n} \log \lambda^*_\theta(t_i) - \int_{0}^{T} \lambda^*_\theta(\tau)\, d\tau, \quad T \geq t_n. \end{aligned} \tag{9}$$

where, by convention, $t_0 = 0$ and note that, in addition to the likelihood of each event t_i, we also account for the fact that, from t_n to T, no event was generated by means of a survival function. The above expression will later enable us to use maximum likelihood to find the model parameters θ under which an intensity function $\lambda_\theta(t)$ is more likely to generate $\mathcal{H}(T)$, *i.e.*,

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \mathfrak{L}(\mathcal{H}(T);\theta). \tag{10}$$

3 Basic intensity functions: inference and sampling

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 perform two common operations:

Algorithm 1: HomogenousPoisson($\mu;\ t_{i-1}$)

Input: μ (parameter), t_{i-1} (time of the last event)
Output: t (time of the next event)
$u \sim \text{Unif}[0, 1]$
$t \leftarrow -\frac{\log(1-u)}{\mu} + t_{i-1}$
return t

(i) Estimate the model parameters that best fits a specific history of events $\mathcal{H}(T)$, *i.e.*, the model parameters under which the intensity function is more likely to generate that specific history of events.

(ii) Sample new events from the intensity function.

Together, the above operations will allow us to (i) train a model from past events and (ii) make predictions about future events.

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.*,

$$\lambda^*_\mu(t) = \mu \geq 0. \tag{11}$$

By definition, the intensity is independent of the history $\mathcal{H}(t)$, 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 mean $\frac{1}{\mu}$.

To estimate the parameter μ that best fits a history of events $\mathcal{H}(t)$, we can just use Eq. 10 with $\lambda^*_\theta(t) = \lambda^*_\mu(t)$ and obtain:

$$\hat{\mu} = \operatorname*{argmax}_\mu \left(n \log \mu - \mu T\right) = \frac{n}{T}, \tag{12}$$

where the last equality follows from setting the derivative to zero to find the maximum.

To sample new events from a Poisson process, we use Algorithm 1. This algorithm uses inversion sampling[3], which goes as follows. First, it samples a uniform random variable $u \sim \text{Unif}[0, 1]$. Then, it computes the time of the next event as $t = (F^*)^{-1}(u)$, where $(F^*)^{-1}(u)$ is the inverse of the cumulative density function $F^*(t)$. In the case of a Poisson process, we can easily compute $(F^*_\mu)^{-1}(u)$, where the subindex μ just denote that the (inverse of the) cumulative density function depends on μ:

$$F^*_\mu(t) = u \implies 1 - \exp(-\mu(t - t_{i-1})) = u \implies t = -\frac{\log(1-u)}{\mu} + t_{i-1}$$

[3]There exist other more efficient methods to sample from (homogeneous) Poisson processes, however, here we utilize inversion sampling since it will generalize to other forms of intensities.

where t_{i-1} is the time of the last generated event (or zero if we are about to generate the first event). It is straight forward to prove that inverse sampling provides us with a valid sample. In particular, we can show that the cumulative density function of $(F^*_\mu)^{-1}(u)$ is $F^*_\mu(t)$ as follows:

$$\mathbb{P}((F^*_\mu)^{-1}(u) \leq t) = \mathbb{P}(u \leq F^*_\mu(t)) = F^*_\mu(t),$$

where the first equality follows by applying F_μ (a monotonous function) to both sides and the next equality follows from the fact that the cumulative density function for a uniform random variable is $\mathbb{P}(u \leq x) = x$.

3.2 Inhomogeneous Poisson Process

An inhomogeneous Poisson process is defined by a time-varying function $g_\theta(t)$, with model parameters θ, *i.e.*,

$$\lambda^*_\theta(t) = g_\theta(t) \geq 0. \tag{13}$$

By definition, the intensity is independent of the history $\mathcal{H}(t)$.

To estimate the parameter θ that best fits a history of events $\mathcal{H}(t)$, we can again use Eq. 10 with $\lambda^*_\theta(t) = g_\theta(t)$ and obtain:

$$\hat{\theta} = \operatorname*{argmax}_\theta \sum_{i=1}^{n} \log g_\theta(t_i) - \int_0^T g_\theta(t)\, dt, \tag{14}$$

The solution to the above optimization problem does depend on the specific parametrization (or functional form) of $g_\theta(t)$. Typically, one looks for parametrizations under which the above optimization problem reduces to a convex program and thus it can be efficiently solved, *e.g.*, using CVX [cvx]. For example, piece-constant intensity functions, *i.e.*,

$$g_\theta(t) = \sum_{j=1}^{N} \theta_j \mathbb{I}(\tau_{j-1} \leq t \leq \tau_j) \tag{15}$$

or mixture of RBF kernels, *i.e.*,

$$g_\theta(t) = \sum_{j=1}^{N} \theta_j \exp(-\beta(t - \tau_j)^2) \tag{16}$$

where $\{\tau_j\}$ and β are given constants, which may be chosen using cross-validation.

To sample new events from an inhomogeneous Poisson process, we use Algorithm 2. This algorithm uses inversion sampling and thinning.[4] First, it samples a time t from a homogeneous Poisson process with rate $g_{max} = \max_\tau g(\tau)$, starting from the time of the last generated event t_{i-1} (or zero if we are about to generate the first event). Then, it accepts this time t as the time of the next event with probability $g(t)/g_{max}$ and, otherwise, it samples once again another time from the same homogeneous Poisson process but starting from the previously sampled time t. The procedure continues until a time is accepted.

[4]Similarly as in the case of homogeneous Poisson, there exist other more efficient methods to sample from inhomogeneous Poisson, however, here we utilize thinning and inversion sampling since it will generalize to other forms of intensities.

Algorithm 2: InhomoPoissonOnline($g_\theta(t);\ t_{i-1}$)

Input: $g_\theta(t)$ (intensity function), t_{i-1} (time of the last event)
Output: t (time of the next event)
$g_{max} \leftarrow \max_\tau g_\theta(\tau)$
$t \leftarrow t_{i-1}$
repeat
  $t \leftarrow \text{HomogenousPoisson}(g_{max}; t)$
  $u \sim \text{Unif}[0, 1]$
until $u \leq \frac{g(t)}{g_{max}}$
return t

One can prove that the above algorithm does provide us with a valid sample, however, it is not as trivial as in the case of homogeneous Poisson.[5] Informally, think that, by rejecting a sample t, we are in reality scaling the constant intensity g_{max} by $g(t)/g_{max}$ and thus the intensity at time t will be $g_{max} \times g(t)/g_{max} = g(t)$, as desired.

3.3 Hawkes Process

An Hawkes process is defined by a history dependent intensity $\lambda^*_{\mu,\alpha}(t)$ defined as follows:

$$\lambda^*_{\mu,\alpha}(t) = \mu + \alpha\kappa_\omega(t) \star dN(t) = \mu + \alpha \sum\nolimits_{t_i \in \mathcal{H}(t)} \kappa_\omega(t - t_i), \tag{17}$$

where,

$$\kappa_\omega(t) := \exp(-\omega t)\, \mathbb{I}[t \geq 0] \tag{18}$$

is an exponential triggering kernel and $\mu \geq 0$ is a baseline intensity independent of the history. Note that the occurrence of each event t_i increases the intensity by a certain amount, determined by the kernel and the parameter $\alpha \geq 0$, making the intensity history dependent and a stochastic process by itself. Perhaps surprisingly, such mutual self-excitation between events naturally fits a very diverse range of real-world scenarios, *e.g.*, information propagation, gang violence, or earth quake prediction.

To estimate the parameters μ and α that best fits a history of events $\mathcal{H}(t)$, we can again use Eq. 10 with $\lambda^*_\theta(t) = \lambda^*_{\mu,\alpha}(t)$ and obtain:

$$\hat{\mu}, \hat{\alpha} = \operatorname*{argmax}_{\mu,\alpha} \sum_{t_i \in \mathcal{H}(T)} \left[\log \left(\mu + \alpha \sum\nolimits_{t_j \in \mathcal{H}(t_i)} \kappa_\omega(t_i - t_j) \right) - \alpha \int_0^T \kappa_\omega(t - t_i) dt \right] - \mu T, \tag{19}$$

Fortunately, the above optimization problem reduces to a convex program and, similarly as in the case of inhomogeneous Poisson processes, can be efficiently solved, *e.g.*, using CVX [cvx]. Finally, note that the triggering kernel parameter ω is typically selected using cross-validation since, otherwise, it would make the above optimization problem nonconvex.

[5] It is out of the scope for the course. For a formal proof, you can read Section 4.2 in [thi].

Algorithm 3: Hawkes($\lambda^*_{\mu,\alpha}(t);\ t_{i-1}$)

Input: $\lambda^*_{\mu,\alpha}(t)$ (intensity function), t_{i-1} (time of the last event)
Output: t (time of the next event)
$\lambda_{\max} \leftarrow \lambda^*_{\mu,\alpha}(t_{i-1})$
$t \leftarrow t_{i-1}$
repeat
    $t \leftarrow \text{HomogenousPoisson}(\lambda_{\max}; t)$
    $u \sim \text{Unif}[0, 1]$
until $u \leq \lambda^*_{\mu,\alpha}(t)/\lambda_{\max}$
return t

To sample new events from a Hawkes process, we use Algorithm 3. This algorithm uses inversion sampling and thinning and it is very similar to Algorithm 2, which we used for inhomogeneous Poisson process. However, in this case, the maximum value λ_{max} changes every time an event happens.

3.4 Terminating point process

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

$$\lambda^*_\theta(t) = g^*_\theta(t)(1 - N(t)), \tag{20}$$

where $N(t)$ is the corresponding counting process, $g^*_\theta(t)$ is a nonnegative intensity function, with parameters θ, and the intensity $\lambda^*_\theta(t)$ becomes zero if an events happens. Parameter estimation and sampling can be done similarly as before and we leave it as an exercise for the reader.

4 Marked temporal point processes

A marked temporal point process $\mathcal{T}$ is a random process whose realization consists of discrete *marked* events localized in time, $\{(t_i, x_i)\}$ with $t_i \in \mathbb{R}^+$, $x_i \in \mathcal{X}$, and $i \in \mathbb{Z}^+$ and the domain of the marks $\mathcal{X}$ is application dependent. In this case, the history of all events $\mathcal{H}(t)$ up to, but not including, time t contains both times and marks, *i.e.*, $\mathcal{H}(t) = \{(t_i, x_i)|t_i < t\}$.

Similarly as in the case of temporal point processes without marks, the event times are represented using a counting process $N(t)$ and its corresponding intensity function $\lambda^*(t)$, which may depend however on past mark values. However, one needs to also represent the marks distribution. Here, one may consider one of the following choices:

— *Independent identically distributed marks:* the marks x_i are *i.i.d.* random variables, *i.e.*,

$$x_i \sim p(x), \tag{21}$$

and, thus, are independent of the history. Here, one can resort to known methods for fitting and sampling the mark distribution.

— *SDEs with jumps:* the marks x_i are history dependent random variables, which are defined using stochastic differential equation (SDE) with jumps, *i.e.*,

$$x_i = x(t_i) \tag{22}$$

$$dx(t) = f(x(t), t)dt + h(x(t), t)dN(t) \tag{23}$$

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. For fitting and sampling the function f and h, one can resort to least squares minimization, as it will become clear in later lectures.

— *SDEs with jumps + noise:* the marks x_i are history dependent random variables, which are defined using a probability distribution whose parameters are defined using stochastic differential equation with jumps, *i.e.*,

$$x_i \sim p(x; \theta(t)) \tag{24}$$

$$d\theta(t) = f(\theta(t), t)dt + h(\theta(t), t)dN(t) \tag{25}$$

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. For fitting and sampling the functions f and h, one can resort to maximum likelihood estimation, as it will become clear in later lectures.

Acknowledgements

We would like to thank Sophie Henning and Hui-Syuan Yeh for helpful comments and typo fixes.

References

[cvx] https://cvxopt.org.

[OWN96] Alan V. Oppenheim, Alan S. Willsky, and S. Hamid Nawab. *Signals &Amp; Systems (2Nd Ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1996.

[thi] https://www.math.fsu.edu/~ychen/research/Thinning%20algorithm.pdf.