

Machine Learning for Graphs and Sequential Data

Sequential Data – Temporal Point Processes

lecturer: Prof. Dr. Stephan Günnemann
www.daml.in.tum.de

Summer Term 2020

Data Analytics and
Machine Learning 

Roadmap

- Chapter: Temporal Data / Sequential Data
 1. Autoregressive Models
 2. Markov Chains
 3. Hidden Markov Models
 4. Neural Network Approaches
 - 5. Temporal Point Processes**
 - a) Introduction**
 - b) Selected TPP Models

Event Data

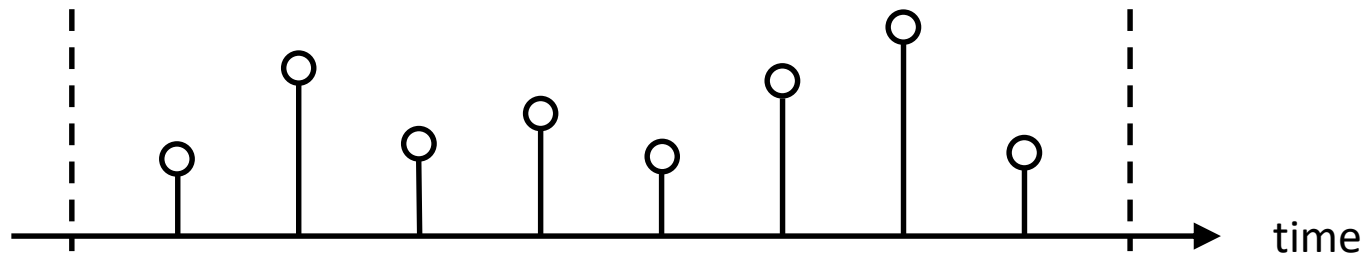
- Our data consists of **discrete** events in **continuous** time, such as
 - Transaction times in finance
 - Messages on social media
 - Visits to hospitals in electronic health records



- Prediction tasks
 - When will the next event happen?
 - How many events will happen in the next hour/day/week?

Difference to Time Series

- Time series
 - Measurements (signal) collected at regular intervals



- (Asynchronous) event data
 - Irregular intervals
 - We care about the time of the occurrence



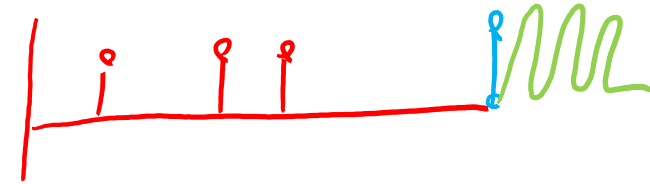
Temporal Point Processes

- Temporal Point Processes (TPP) are a class of probabilistic models that describe the distribution of **discrete** event sequences in **continuous** time

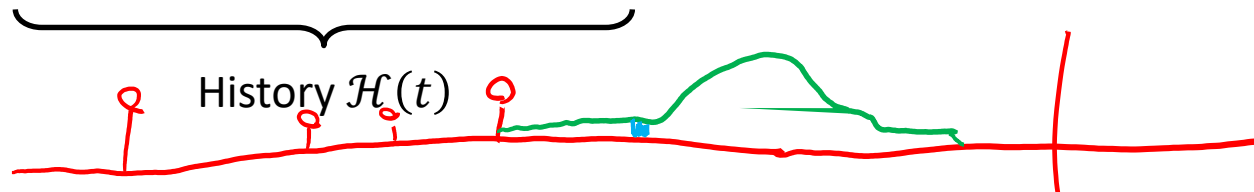
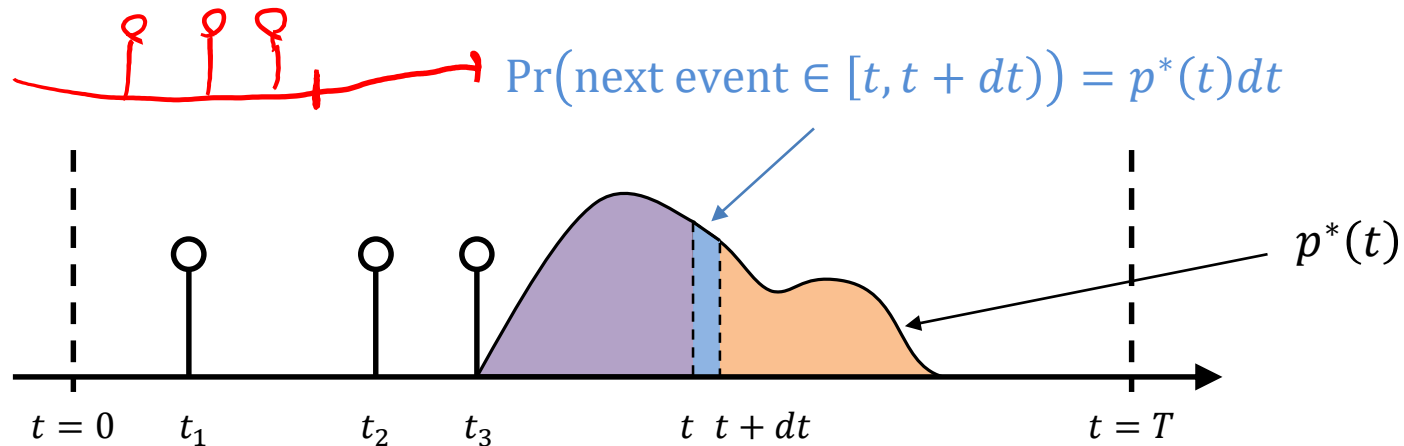


- TPP defines a generative model for variable-length sequences $\mathbf{t} = \{t_1, \dots, t_N\}$ on the interval $[0, T]$
 - Both locations of the events t_i and their number N are random
- TPPs also provide a likelihood function $p(\{t_1, \dots, t_N\})$

Modeling the Time of the Next Event

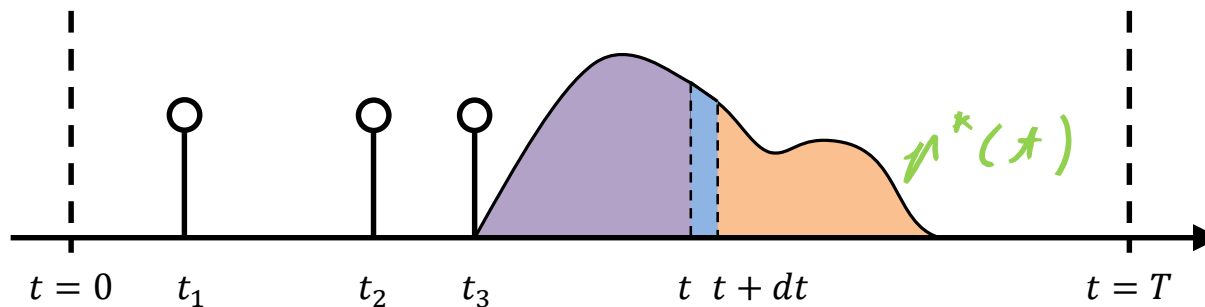


- We can model the distribution $p(\{t_1, \dots, t_N\})$ autoregressively
 - Predict the time of the next event t_i given the history $\mathcal{H}(t) = \{t_j < t\}$
 - Important: $\mathcal{H}(t)$ depends on the specific sample $\{t_1, \dots, t_N\}$!
 - We denote the conditional density as $p^*(t) := p(t|\mathcal{H}(t))$



next event $\in [t, t + dt) \Leftrightarrow$ event in $[t, t + dt)$ & no event in $[t_3, t)$

Alternative Ways to Model the Inter-event Time

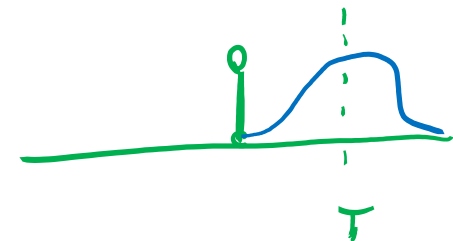


■ Cumulative distribution function (CDF)

- $F^*(t) = \int_{t_{i-1}}^t p^*(u) du$ = Probability that the next event happens in $[t_{i-1}, t]$
- t_{i-1} is the last event that happened before t

■ Survival function

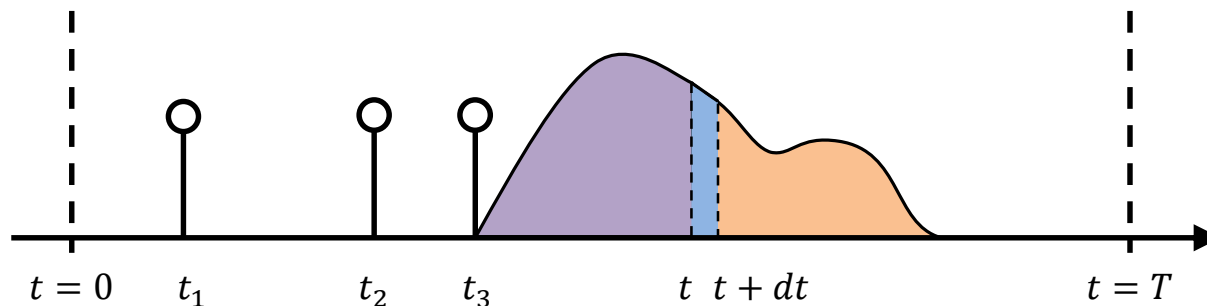
- $S^*(t) = 1 - F^*(t) = \int_t^\infty p^*(u) du$
- Probability that the next event doesn't happen before t
- Probability that the next event happens after t



Conditional Intensity Function

$$P(E | N_0) = \frac{P(E, N_0)}{P(N_0)}$$

- There exists another way to describe the conditional distribution



- Conditional intensity

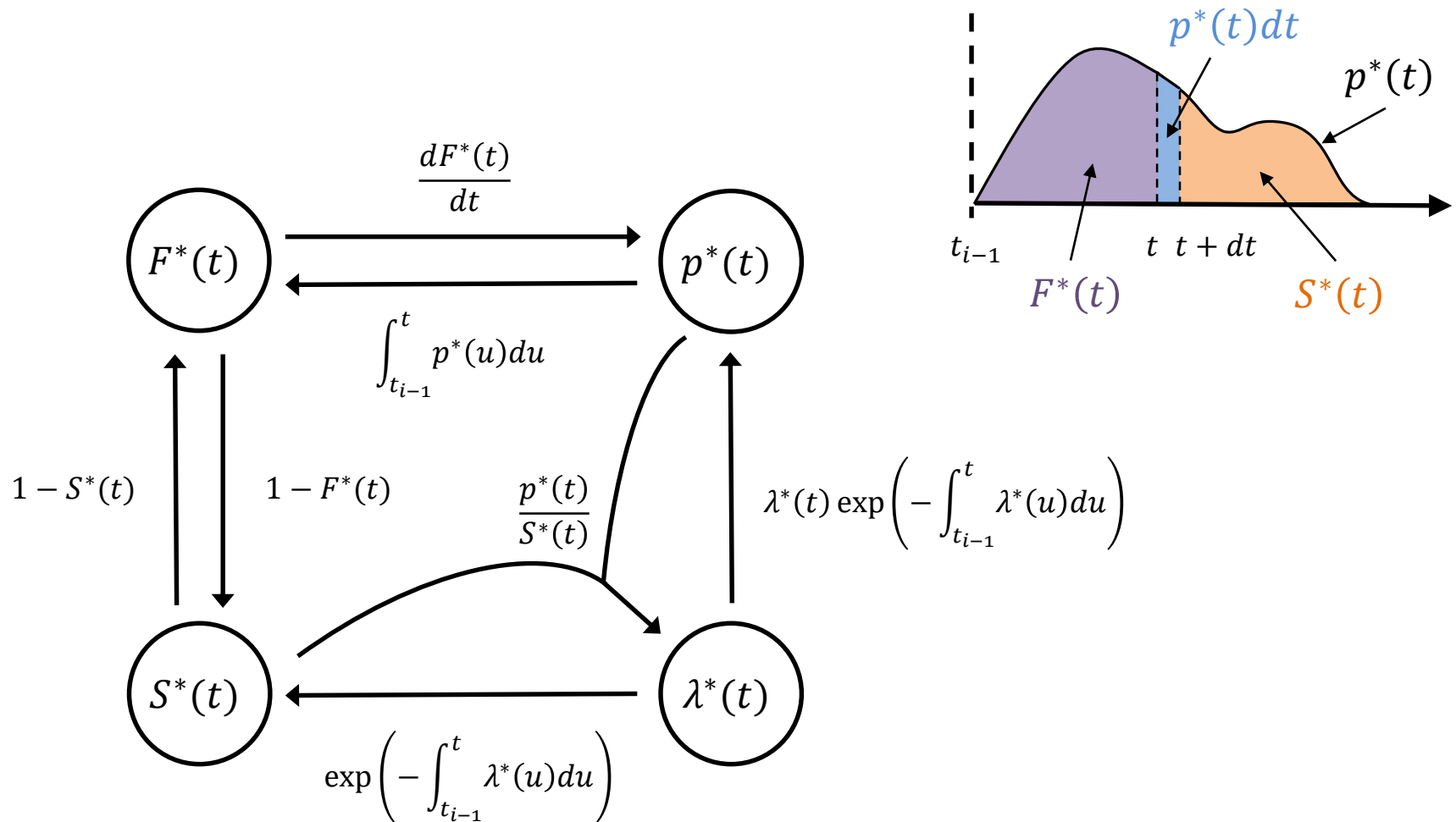
– $\lambda^*(t)dt$ = probability of event in $[t, t + dt)$ **given** no event in $[t_{i-1}, t)$

$$\lambda^*(t)dt = \frac{\Pr(\text{event in } [t, t + dt) \text{ \& no event in } [t_{i-1}, t))}{\Pr(\text{no event in } [t_{i-1}, t))} = \frac{p^*(t)dt}{S^*(t)}$$

- Intuitive meaning of $\lambda^*(t)$: Expected # of events / unit of time
 - We will demonstrate this later

$$P(E | N_0)$$

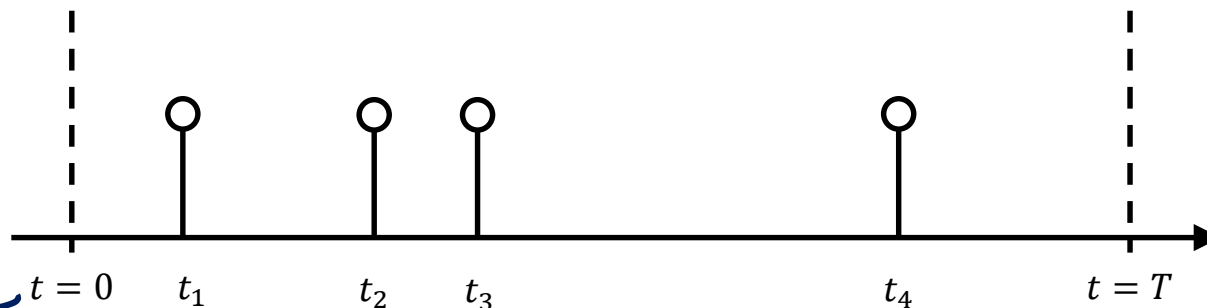
Relation between p^*, F^*, S^*, λ^*



$$n \times n^{(*)}$$

Likelihood of an Entire Sequence

- How can we compute the likelihood of a realization $\{t_1, \dots, t_N\}$?



$$s^*(t_2) \cdot s^*(t_1) \dots$$

$$\uparrow s^*(t_4) \cdot s^*(T)$$

$$p(\{t_1, t_2, t_3, t_4\}) = p^*(t_1) p^*(t_2) p^*(t_3) p^*(t_4) S^*(T)$$

$$= \lambda^*(t_1) \lambda^*(t_2) \lambda^*(t_3) \lambda^*(t_4) \exp\left(-\int_0^T \lambda^*(u) du\right)$$

$$= \frac{\lambda^*(t_1)}{s^*(t_1)} \cdot \frac{\lambda^*(t_2)}{s^*(t_2)} \dots$$

- Remember that the number of events can vary

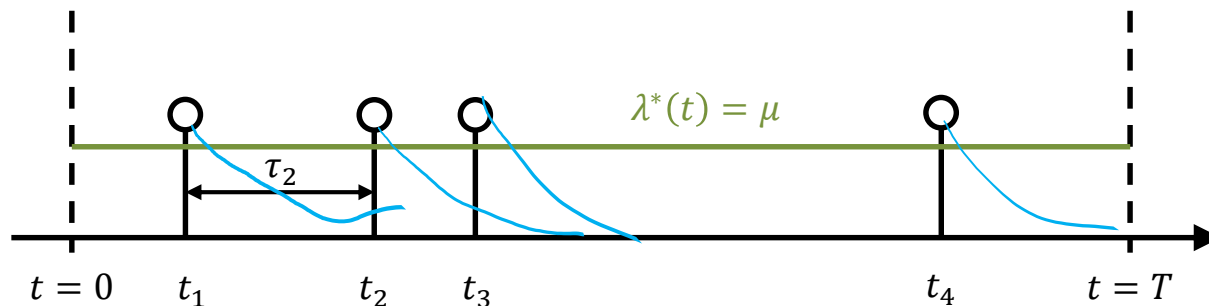
Roadmap

- Chapter: Temporal Data / Sequential Data
 1. Autoregressive Models
 2. Markov Chains
 3. Hidden Markov Models
 4. Neural Network Approaches
 - 5. Temporal Point Processes**
 - a) Introduction
 - b) Selected TPP Models**

Models based on Conditional Intensity

- Defining TPPs in terms of $\lambda^*(t)$ has several advantages
- 1. Easy to define TPPs with pre-defined behavior
 - Global trend, burstiness, repulsiveness
 - Intensity is more interpretable
- 2. Easy to combine different TPPs with different $\lambda^*(t)$'s
- 3. Efficient sampling

Homogeneous Poisson Process (HPP)



- Simplest possible model: constant intensity

$$\lambda^*(t) = \mu$$

- Inter-event times follow exponential distribution

$$p^*(t) = \mu \exp\left(-\int_{t_{i-1}}^t \mu \, du\right) = \mu \exp\left(-\underbrace{\mu(t - t_{i-1})}_{\text{inter-event time } \tau_i}\right)$$

Simulating an HPP

- We can simulate an HPP by generating the inter-event times

```
arrival_times = []  
t = 0  
while t < T:  
    tau ~ Exponential(mu)  
    t += tau  
    if t < T:  
        arrival_times.append(t)
```

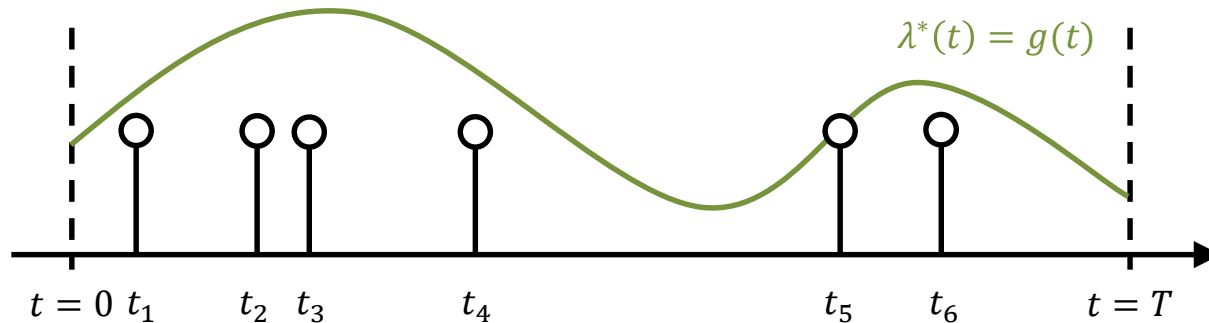
- How to sample from the exponential distribution? – Inverse CDF transform

$$u = F(\tau) = 1 - \exp(-\mu\tau) \Rightarrow \tau = F^{-1}(u) = -\frac{1}{\mu} \log(1 - u)$$

where $u \sim \text{Uniform}(0, 1)$ and F is the CDF of the exponential distribution

Inhomogeneous Poisson Process (IPP)

$$g(x) = a \cdot x^2 + b \cdot x + c$$



- The intensity changes over time

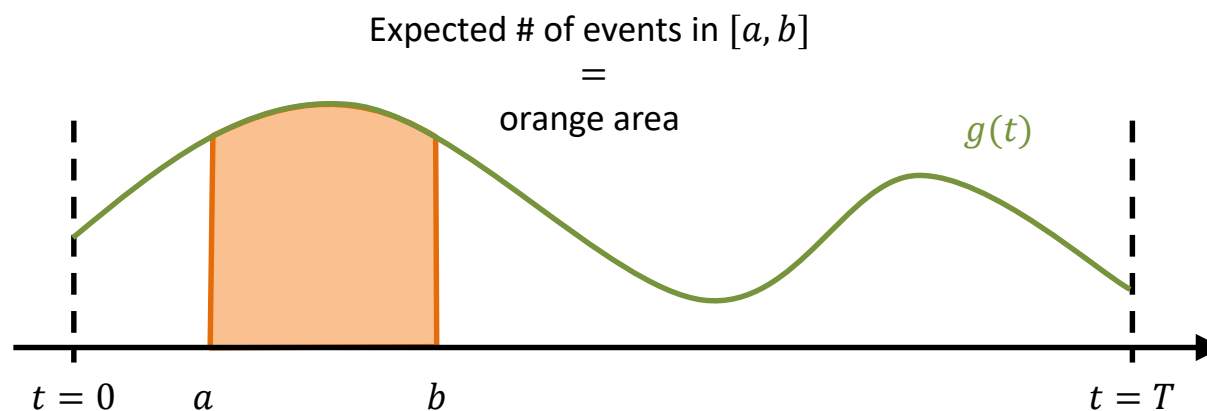
$$\lambda^*(t) = g(t) \geq 0$$
- Intensity is independent of the history
- Captures global trend
 - More events happen in the regions with higher intensity

Expected Number of Events

- Number of events in an interval $[a, b] \subseteq [0, T]$ follows Poisson distribution

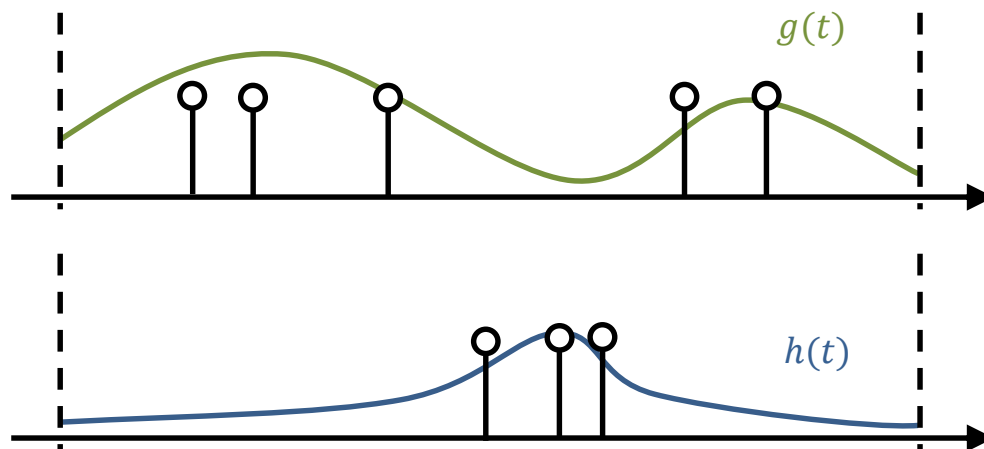
$$N([a, b]) \sim \text{Poisson} \left(\int_a^b g(t) dt \right)$$

- This means, the expected number of events inside $[a, b]$ is equal to the total intensity over this region

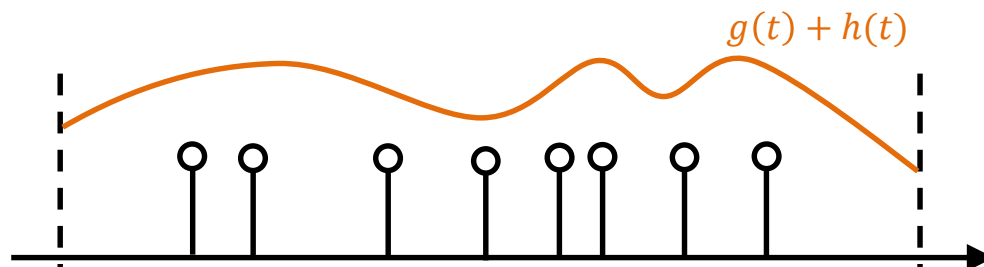


Superposition of IPPs

- Consider two IPPs with intensities $g(t)$ and $h(t)$



- Combination of the two IPPs is again an IPP with intensity $g(t) + h(t)$



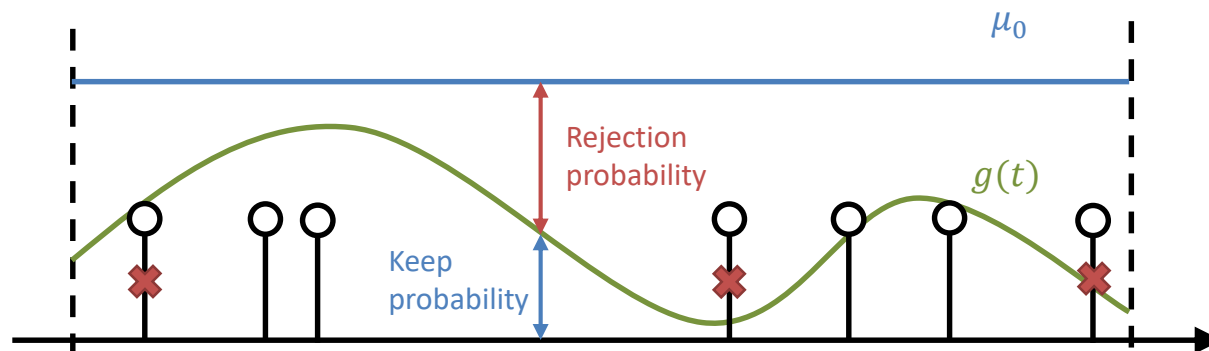
- This also applies to a general $\lambda^*(t)$, but showing this is more involved

Simulating an IPP

$$g'(x) = \mu_0 - g(x)$$

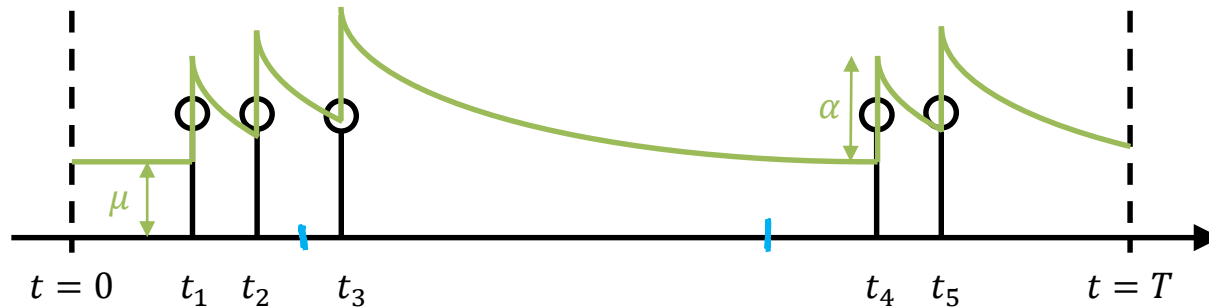
$$g(x) + g'(x) = \mu_0$$

- Simulating the inter-event times is hard (requires integration)
- Better alternative – thinning



1. Find an upper bound $\mu_0 \geq g(t)$ for all t
2. Simulate candidate events $\{t_1, t_2, \dots\}$ from a HPP with rate μ_0
3. Keep each t_i with probability $g(t_i)/\mu_0$

Hawkes Process



- Also known as self-exciting process

$$\lambda^*(t) = \mu + \alpha \sum_{t_j \in \mathcal{H}(t)} k_{\omega}(t - t_j)$$

- Triggering kernel $k_{\omega}(t - t_i) = \exp(-\omega(t - t_i))$
 - Parameters $\mu, \alpha, \omega \geq 0$
- Intensity depends on the history
- Clustered (“bursty”) event occurrences

Parameter Estimation in TPPs

- Pick a parametric conditional intensity $\lambda_{\theta}^*(t)$ (e.g. Hawkes, IPP)
- Maximize the log-likelihood of the observed sequences $\mathcal{D}_{\text{train}}$

$$\max_{\theta} \sum_{t=\{t_1, \dots, t_N\} \in \mathcal{D}_{\text{train}}} \log p_{\theta}(\{t_1, \dots, t_N\})$$

- The log-likelihood of a single sequence is

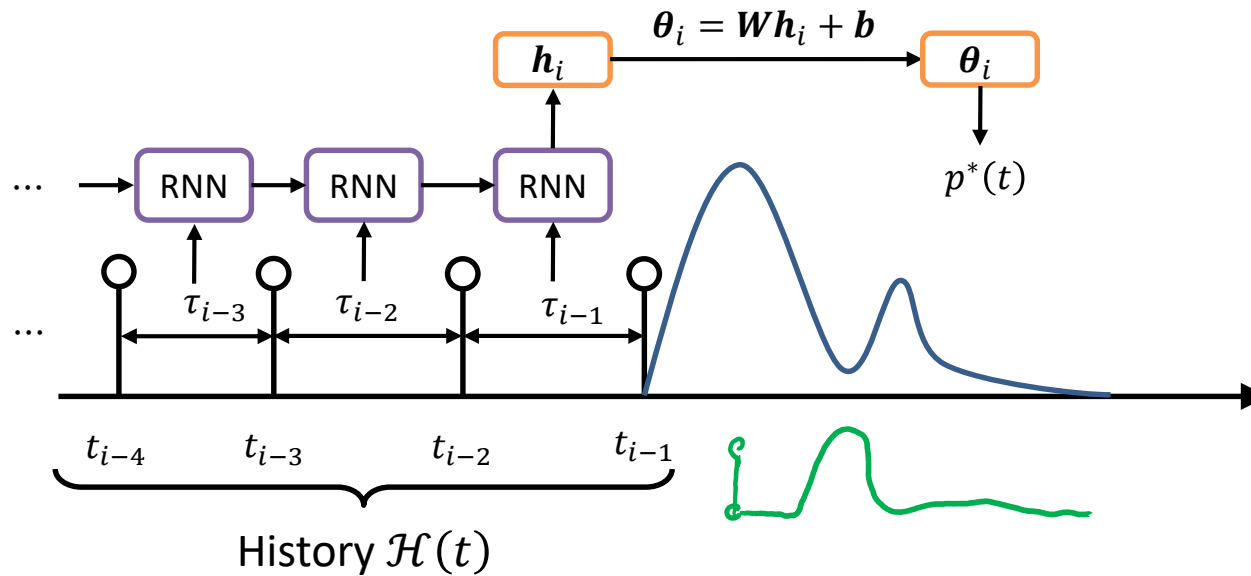
$$\log p_{\theta}(\{t_1, \dots, t_N\}) = \sum_{i=1}^N \log \lambda^*(t_i) - \int_0^T \lambda^*(u) du$$

- Remember, different sequences have different length N
- Lots of different optimization techniques possible
 - Simple models like HPP allow closed-form solutions
 - For Hawkes process we can use convex optimization methods
 - Always possible to use gradient descent (with modifications for constraints)

Conditional Intensity: Summary

- Conditional intensity $\lambda^*(t)$ provides an alternative to the conditional density $p^*(t)$ when constructing TPPs
- Advantages
 - Easy to define models with simple behavior
 - Interpretable
 - Efficient sampling
- Limitations
 - Integration required to compute the log-likelihood – might be intractable
 - Not clear how to define flexible models with arbitrary dynamics
- We will define more flexible TPPs by going back to $p^*(t)$ and using RNNs

Modeling TPPs with RNNs



- Directly model the conditional distribution $p^*(t)$ using an RNN
 1. Every time an event happens, we feed τ_i into the RNN
 2. Use the hidden state $\mathbf{h}_i \in \mathbb{R}^D$ of the RNN as the history embedding
 3. Use \mathbf{h}_i to generate the parameters $\boldsymbol{\theta}_i$ of the distribution $p^*(t)$

$$p^*(t) = p(t|\mathcal{H}(t)) = p(t|\mathbf{h}_i)$$

How to Model $p^*(t)$

$p(t)$

- The sequence of events must be increasing: $t_i > t_{i-1}$ for all i
- We can instead model the distribution of the inter-event times τ_i
 - It's sufficient to ensure that $\tau_i > 0$
- How to define a flexible and tractable $p^*(\tau_i)$
- Simple nonnegative distribution
 - Exponential, Gamma, Weibull, Gompertz, ...
- Mixture distribution
 - Take a convex combination of simple densities
- Normalizing flows
 - Use transformations like $\exp(x)$ or $\log(1 + \exp(x))$ to ensure non-negativity
 - Combine with other transformations (e.g., polynomials, NNs with positive weights) to add flexibility

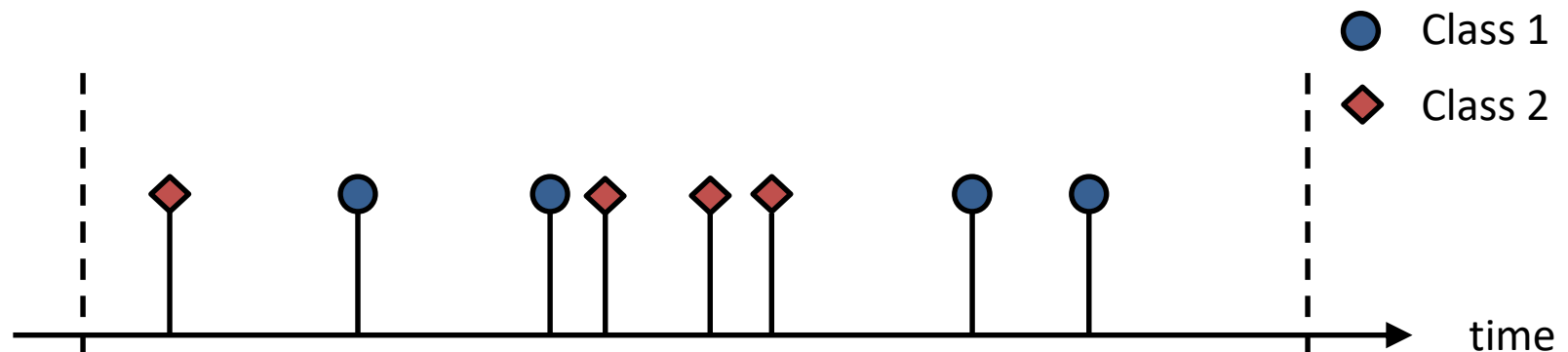
$p(\tau)$

What we haven't covered

- Modeling TPPs with marks
 - <https://www.research-collection.ethz.ch/handle/20.500.11850/151886>
 - <https://www.kdd.org/kdd2016/papers/files/rpp1081-duA.pdf>
- More efficient sampling techniques
 - <https://web.ics.purdue.edu/~pasupath/PAPERS/2011pasB.pdf>
- Spatial and spatio-temporal point processes – modeling events in space
 - <https://arxiv.org/abs/1708.02647>

Marked Temporal Point Processes

- Most common type: categorical marks
 - Each event has an associated class (i.e., category, type)
 - Events of different classes may influence each other
 - E.g., activity of each use is represented by a different mark



- Continuous marks also possible
 - E.g., magnitude of the earthquake, amount of money spent by a customer

Questions – TPP

1. Is it possible to obtain the conditional intensity $\lambda^*(t)$ if you know only the survival function $S^*(t)$ and don't know the conditional PDF $p^*(t)$?
2. Would you use (a) Hawkes process or (b) inhomogeneous Poisson process to model the following event data?
 - Customers visiting a supermarket (event = customer enters the supermarket)
 - Messages sent by a single user on WhatsApp (event = message sent)
 - Taxi rides in a city (event = a trip starts)
3. What can you say about a TPP with the following conditional intensity function? What kind of behavior does it model?

$$\lambda^*(t) = \exp\left(t - \sum_{t_i \in \mathcal{H}(t)} 1\right)$$

Acknowledgments

- These slides are based on the ICML 2018 tutorial by Manuel Gomez Rodriguez & Isabel Valera (<http://learning.mpi-sws.org/tpp-icml18/>)

Recommended Reading

- Lecture notes on TPPs by De, Upadhyay and Gomez-Rodriguez
 - <http://courses.mpi-sws.org/hcml-ws18/lectures/TPP.pdf>
 - Except Section 3.4, 4

- Alternatively, lecture notes by Rasmussen
 - <https://arxiv.org/abs/1806.00221>
 - Except Sections 2.4, 3.2, 4.2, 5, 6

- Modeling TPPs with recurrent neural networks
 - <https://arxiv.org/abs/1909.12127>
 - <https://www.kdd.org/kdd2016/papers/files/rpp1081-duA.pdf>