Counterfactual Temporal Point Processes

June 14th to June 20th

Background & motivation

- Many ML methods focus on **predicting future events given past sequences;** Recent lines of work have used **reinforcement learning and control theory** to automate interventions
- **real-world high-stakes interventions** (like public health policy) are made by humans, not automatically => a need for tools to **support decision-making**
- assist decision makers at implementing interventions in high-stakes applications (e.g. finance, social networks, epidemiology) => facilitating counterfactual thinking in human decision making

Problem & challenges & Idea

- Real-world data comes from **observed events only** — there's no record of the **rejected events** from the underlying stochastic process.
- This makes it **difficult to simulate counterfactuals**, which requires knowing how outcomes would have changed under a different event-generating mechanism.
- **Transforming the modeling perspective** to reinterpret the observed sequence as if it were the **accepted subset** of a latent, richer process — namely, a **homogeneous Poisson process** with rate λ_max.

Related works & Limitations

- Most existing work on causal inference in TPPs:
  - Focused on **measuring causal influence** (e.g., Granger causality).
  - Predicted outcomes from **interventional distributions** (not counterfactuals).
  - Modeled **only survival settings** or **single-event point processes** (less applicable to multi-event, continuous-time settings).
- Few exceptions:
  - Some works (e.g., Schulam and Saria) study **counterfactual distributions of marks** in **marked TPPs**, not about **intensity-based thinning**.
  - Others explore **counterfactual reasoning in Markov decision processes (MDPs)** using Gumbel-Max — but **not for temporal point processes**. (Oberst and Sontag and Tsirtsis et al.)

Proposed solution & Contribution

- **Augment Lewis' thinning algorithm** using the **Gumbel-Max structural causal model (SCM)** to:
  - Interpret thinning as a causal assignment (similar as **structural function/rule** f in SCM)
  - Apply monotonicity to allow simulation of counterfactual acceptances
- Use **superposition theorem** to:
  - Reconstruct *plausible* sequences of **rejected events**, which are not observed in real data (**statistical approximations**, not deterministic truths)
  - Combine with observed events to simulate **counterfactuals under alternative intensities** (taken as treatment)
- Achievements:
  - simulate counterfactual TPP trajectories under different intensities.
  - Model applicable to **inhomogeneous Poisson processes** and extends to **linear Hawkes processes**.
  - **meaningful insights** when applied to **real-world epidemiological data**

Preliminaries recap

- **TPP**
  - **stochastic process** describing the timing of events over continuous time
  - The realization is a **sequence of discrete events** $\mathcal{H} = \{t_i \in \mathbb{R}^+ \mid i \in \mathbb{N}, t_i < t_{i+1}\}$.
  - characterized by its **intensity function** λ(t), governs the **prob of an event happening in a small interval;** depend on past events

$$\lambda(t)dt = \mathbb{P}[dN(t) = 1] = \mathbb{E}[dN(t)]$$

- **SCM**
  - data-generating process using **structural assignments:** X_i:=f_i (PA_i,U_i)
  - An intervention (e.g., do(Xi=x)), interventional model C^I;
  - **counterfactual distributions:** conditioning on observed values X=x, and using the updated noise distribution P(U | X=x)
    - Challenges: **posterior distribution over noise** may not be identifiable without further assumptions (**multiple functions f_i** and **noise distributions** can be observationally equivalent)

Preliminaries recap

- **Monotonicity Assumption** (to help non-identifiability)
  - higher treatment cannot decrease the chance of observing Y=y
  - It doesn't **ensure uniqueness** of f and P(U)
  - a **partial identification strategy**, not a complete solution

$$P^{\mathcal{C};\mathrm{do}(T=t)}(Y=y) > P^{\mathcal{C};\mathrm{do}(T=t')}(Y=y) \Rightarrow P^{\mathcal{C}|Y=y,T=t';\mathrm{do}(T=t)}(Y=y')=0 \text{ where } y' \neq y$$

Modeling: From Observed Events to Counterfactual Simulation

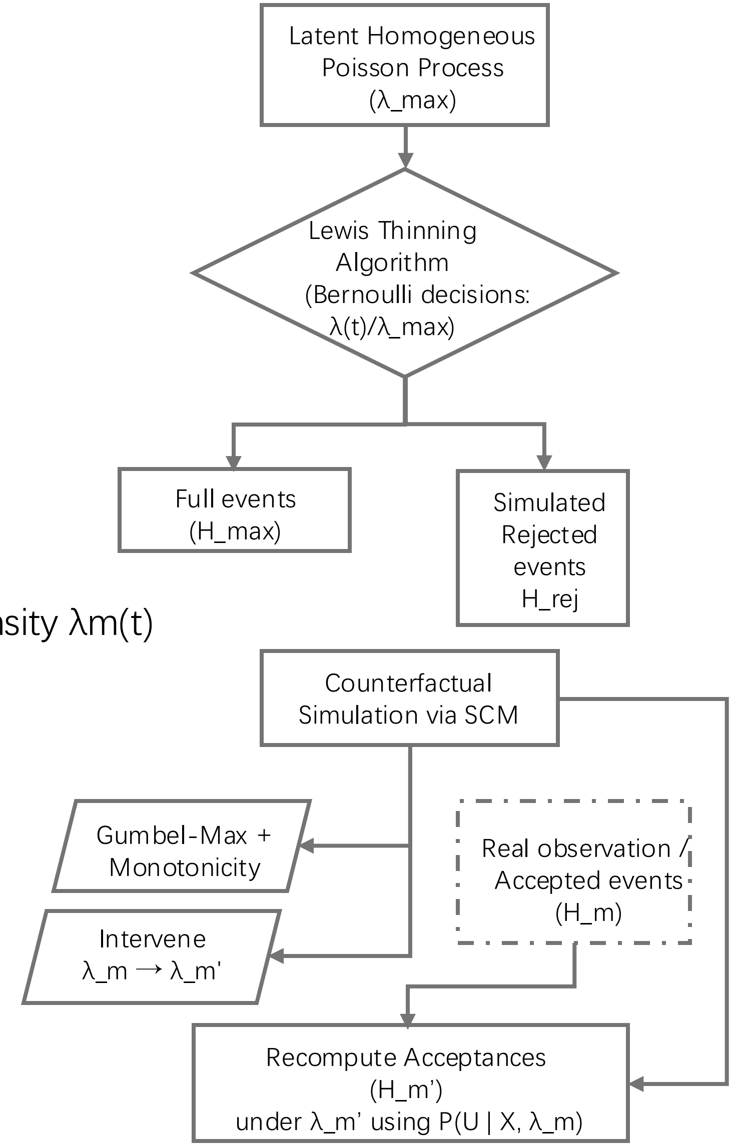
- **Observation setting: TPP**
  - Real-world observation: a set of time-stamped events $\mathcal{H}_m = \{t_1, t_2, ..., t_n\}$
  - assumed to have been generated from an **inhomogeneous Poisson process** with intensity λm(t)
  - process **did not involve thinning**: it came from natural phenomena
- **Lewis' Thinning Algorithm (transformative perspective)**
  - Recast the observed process into the procedure:
    - A **homogeneous Poisson process** with intensity λ_max,
    - events were **accepted or rejected** using Bernoulli trials with probability:

$$p(t_i) = \lambda_m(t_i)/\lambda_{\max}$$

    - Following components:
      - H_max: full proposed points
      - H_m: accepted (observed)
      - H_rej=H_max \ H_m: rejected (not observed, simulated; with feeding lambda_rej inside lewis thinning algorithm)

$$\lambda_{\max} = \lambda_{\text{accepted}}(t) + \lambda_{\text{rejected}}(t). \qquad \mathcal{H}_{\max} = \mathcal{H}_m \cup \text{Simulated } \mathcal{H}_{\text{rej}} \sim \lambda_{\max} - \lambda_m(t)$$

  - Knowing the plausible latent history: how can we plausibly simulate those **missing rejected events**? (superposition theorem):
    - the recovered points **match the statistical distribution** of the original rejected events
    - statistical surrogate / approximation: **not the exact points**, because that randomness is unobservable (unit level cf not achievable)



Modeling: From Observed Events to Counterfactual Simulation

- **Gumbel-Max SCM: linking to Causal Inference**
  - Classic SCM**:** Y=f(X,U)
  - Model setting**:** H_m =Thinning (H_max, λ_m(t))
  - Transformation of idea:
    - **H_max** plays the role of the **exogenous U** (latent randomness);
    - The **accepted/rejected split** is the realization of the **structural rule**;
    - The **intensity function λ(t)** acts like the **causal mechanism**
  - Reinterpret thinning through an SCM:

random indicator or auxiliary pointer:
turns a *discrete probability distribution* into a *random deterministic choice*

$$\boxed{X_i} = \arg\max_{x\in\{0,1\}} g(x, \Lambda_i, U_{i,x})$$

$$\Lambda_i = \lambda_m(t_i)$$

$$U_{i,x} \sim \text{Gumbel}(0,1)$$

$$g(x, \Lambda, U) = \boxed{\log p(X = x \mid \Lambda) + U}$$

    - Gumbel-Max trick decompose the full random process:
      - Embeds **explicit noise** into the assignment: deterministic function + independent noise.
  - Allows **causal interventions** by changing Λ_i → λ_m′(t_i)

$$P^{\mathcal{C}\,;\,\underline{\mathrm{do}(\Lambda_i=\lambda(t_i))}}(X_i = 1) = p(\lambda(t_i)) = \frac{\lambda(t_i)}{\lambda_{\max}}.$$

- **SCM Interventions → Counterfactual Query**
  - Fixing posterior dist of noise:
    - Conditioning observed known value X_i=x implies ordering constraint of noise term
    - inequality defines a **region** of the joint space (U_{i,0},U_{i,1}) that is **consistent with the observed outcome** X_i=x.

eg: $\log p(x) + U_{i,x} > \log p(1-x) + U_{i,1-x}$ $\quad U_{i,x} - U_{i,1-x} > \log p(1-x) - \log p(x)$

- **SCM Interventions → Counterfactual Query**
  - Counterfactual assignment under new intensity: (same preserved latent noise U_i, **intervened intensity** λm′(ti))

$$P^{\mathcal{C}\,\mid\,X_i=x_i,\Lambda_i=\lambda_m(t_i)\,;\,\underline{\mathrm{do}(\Lambda_i=\lambda_{m'}(t_i))}}(X_i = x) = \mathbb{E}_{U_i\mid X_i=x_i,\Lambda_i=\lambda_m(t_i)}[\mathbf{1}[x = \operatorname*{argmax}_{x'\in\{0,1\}} g(x', \lambda_{m'}(t_i), \boldsymbol{U}_i)]],$$

Modeling: From Observed Events to Counterfactual Simulation

- **Monotonicity and Identifiability:**

$$\lambda_m(t) \le \lambda_{m'}(t) \Rightarrow P(X_i = 1 \mid \mathrm{do}(\lambda_{m'})) \ge P(X_i = 1 \mid \mathrm{do}(\lambda_m))$$

- **Review of Central question:** Given a realized sequence of events H_m under a known intensity function λ_m (t), what would the sequence of events H_m' have been if the event-generating intensity had been λm'(t) instead?

Algorithms

Algorithm 4: Lewis' thinning algorithm

Input: $\lambda(t)$, $\lambda_{\max}$, T.
Initialize: $s = 0$, $\mathcal{H} = \emptyset$.

function $\textsc{Lewis}(\lambda(t), \lambda_{max}, T)$**:**
  while true **do**
    $u_1 \sim \mathrm{Uniform}(0,1)$
    $w \leftarrow -\ln u_1/\lambda_{\max}$
    $s \leftarrow s + w$
    if $s > T$ **then**
      break
    end
    $\mathcal{H}_{\max} \leftarrow \mathcal{H}_{\max} \cup \{s\}$
    $u_2 \sim \mathrm{Uniform}(0,1)$
    if $u_2 \le \lambda(s)/\lambda_{max}$ **then**
      $\mathcal{H} \leftarrow \mathcal{H} \cup \{s\}$
    end
  end
  Return $\mathcal{H}$, $\mathcal{H}_{\max} \backslash \mathcal{H}$
end

sample inter-arrival times

u2 is used later for the **thinning step**, i.e., deciding whether to accept the time point or not

Algorithm 1: It samples a counterfactual sequence of accepted events given a sequence of accepted and rejected events provided by Lewis' thinning algorithm

Input: $\lambda_m(t)$, $\lambda_{m'}(t)$, $\mathcal{H}_m$, $\mathcal{H}_{\max}$, $\lambda_{\max}$.
Initialize: $\mathcal{H}_{m'} = \emptyset$.

function $\textsc{Acc}(\lambda_m(t), \lambda_{m'}(t), \mathcal{H}_m, \mathcal{H}_{\max}, \lambda_{\max})$**:**
  $\mathcal{H}_{m'} \leftarrow \emptyset$
  for $t_i \in \mathcal{H}_{\max}$ **do**
    $x_i \leftarrow \mathbf{1}[t_i \in \mathcal{H}_m]$
    $x'_i \sim P^{\mathcal{C} \mid X_i = x_i, \Lambda_i = \lambda_m(t_i)\,;\, \mathrm{do}(\Lambda_i = \lambda_{m'}(t_i))}(X)$
    if $x'_i = 1$ **then**
      $\mathcal{H}_{m'} \leftarrow \mathcal{H}_{m'} \cup \{t_i\}$
    end
  end
  Return $\mathcal{H}_{m'}$
end

Loop through every proposed time point at full events;
captures the **factual realization** of the thinning

counterfactual resampling step: replaying the **same noise**, but under the **new acceptance threshold**

Algorithm 2: It samples a counterfactual sequence of events given a sequence of observed events from an inhomogeneous Poisson process.

Input: $\lambda_m(t)$, $\lambda_{m'}(t)$, $\mathcal{H}_m$, $\lambda_{\max}$, T.
Initialize: $\mathcal{H}_{m'} = \emptyset$.

function $\textsc{Cf}(\lambda_m(t), \lambda_{m'}(t), \mathcal{H}_m, \lambda_{\max}, T)$**:**
  $\mathcal{H}_{\max}, \_ \leftarrow \textsc{Lewis}(\lambda_{\max} - \lambda_m(t), \lambda_{\max}, T)$
  $\mathcal{H}_{\max} \leftarrow \mathcal{H}_{\max} \cup \mathcal{H}_m$
  $\mathcal{H}_{m'} \leftarrow \textsc{Acc}(\lambda_m(t), \lambda_{m'}(t), \mathcal{H}_m, \mathcal{H}_{\max}, \lambda_{\max})$
  Return $\mathcal{H}_{m'}$
end

gives **rejected candidate events** that would have been proposed by Lewis' algorithm **but not accepted** under λ_m

Limitation & future work

- Only stochastic event generation via intensity: model augmentation
  - Mark, feature: covariates or confounders, consider the case when intensity $\lambda(t)$ could depend on these marks X
- Try different forms of the structural assignments fi(PAi,Ui), replacing Gumbel max trick

Question

- The changes of question/ causal structure/ frameworks, estimation, target of causal inference under TPP? no natural scalar "effect”, then how we interpret it
- Potential of answering questions/queries using TPP