Distribution-Free Conformal Joint Prediction Regions for Neural Marked Temporal Point Processes

Souhaib Ben Taieb
University of Mons, Belgium
April 5, 2024

UMONS Université de Mons

Faculté des Sciences

eunice EUROPEAN UNIVERSITY

Collaborators

Victor Dheur
(PhD student, UMONS)

Tanguy Bosser
(PhD student, UMONS)

Rafaël Izbicki
(Ass. Prof, UFSCar)

Outline

Temporal point processes

Distribution-free uncertainty quantification

Conformal neural temporal point processes

Experiments

UMONS
Université de Mons

Plan

Temporal point processes

Distribution-free uncertainty quantification

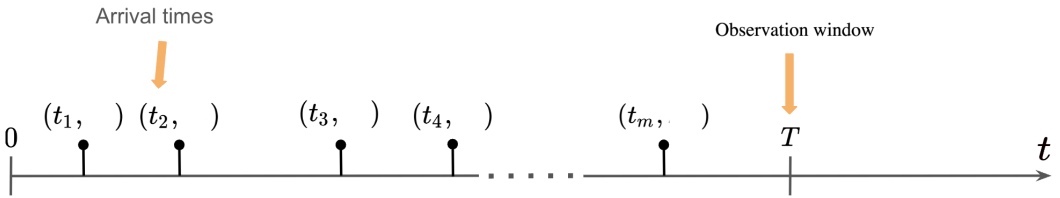
Conformal neural temporal point processes

Experiments

Labeled event sequences

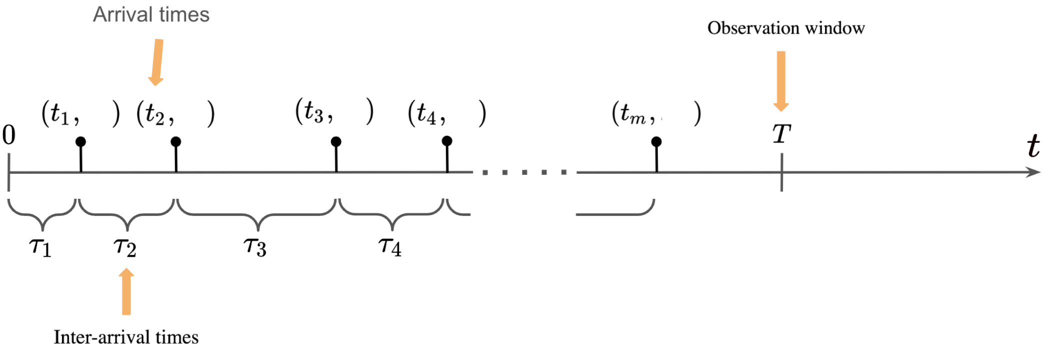
UMONS
Université de Mons

Labeled event sequences



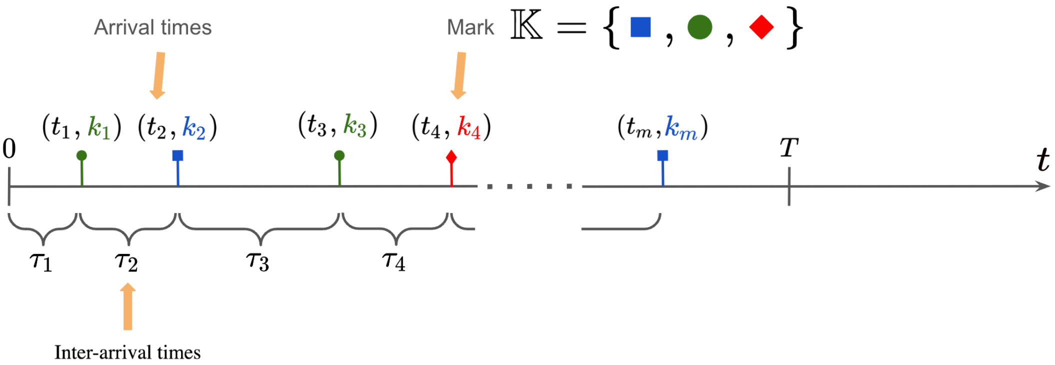
UMONS
Université de Mons

Labeled event sequences

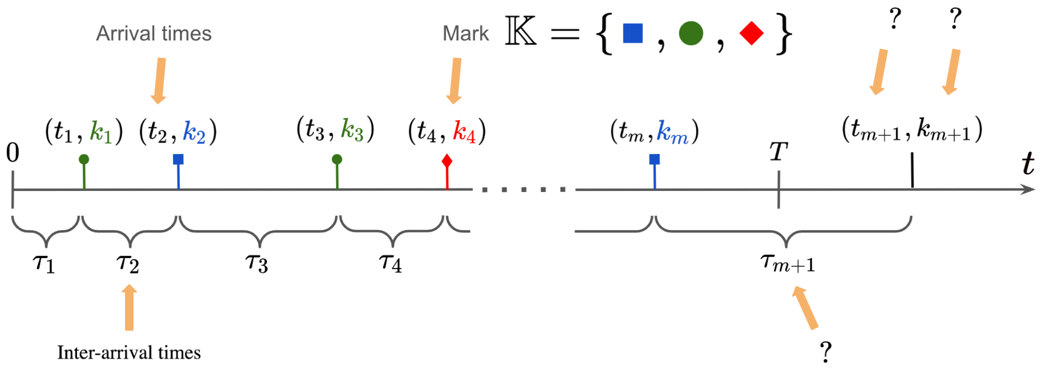


UMONS
Université de Mons

Labeled event sequences



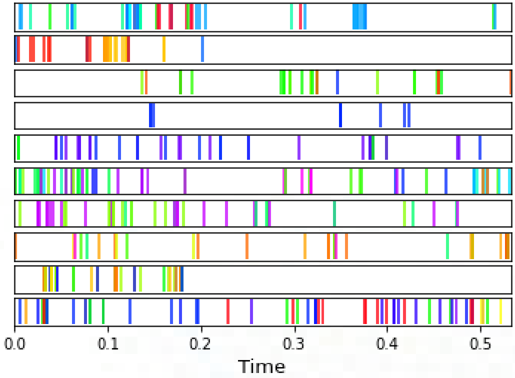
Labeled event sequences



UMONS Université de Mons

Examples of applications

- Social media activity [Far+15]
- Online shopping activity [Cai+18]
- Medical Records [Eng+20]
- Finance [BMM15]
- Earthquakes [Das+23]



UMONS
Université de Mons

Marked Temporal Point Processes

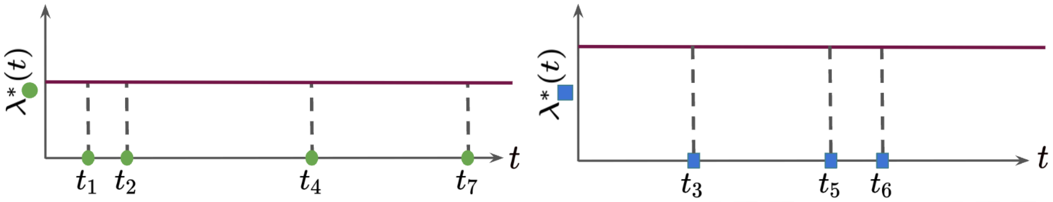
Marked temporal point processes (MTPPs) [D J03] define a probability distribution over label event sequences in **continuous time**.

Marked Temporal Point Processes

Marked temporal point processes (MTPPs) [D J03] define a probability distribution over label event sequences in **continuous time**.

An MTPP can be characterized by its ***marked intensity functions***, defining the expected occurrence **rate** of mark-k events per unit of time, conditional on the history.

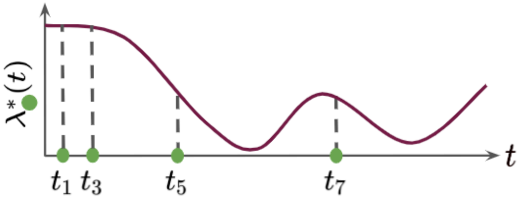
Homogeneous Poisson process: $\lambda_k^*(t) = \lambda_k$

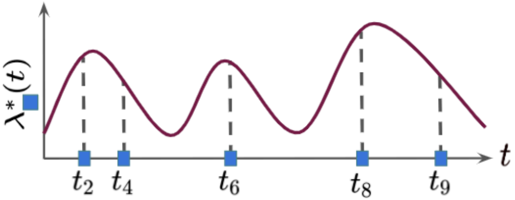


UMONS Université de Mons

Marked Temporal Point Processes

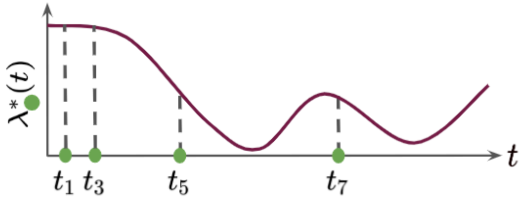
Inhomogeneous Poisson process: $\lambda_k^*(t) = \lambda_k(t)$

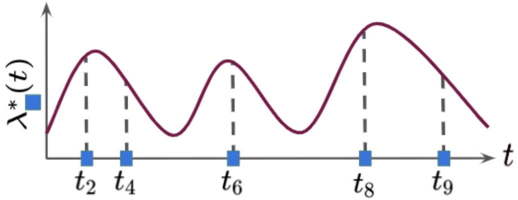




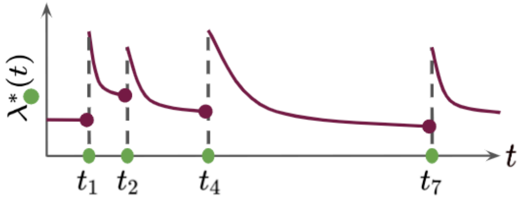
Marked Temporal Point Processes

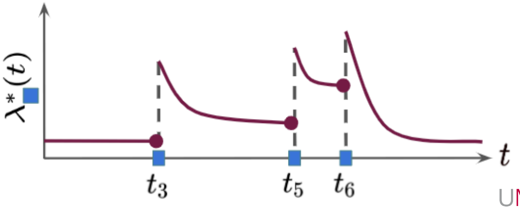
Inhomogeneous Poisson process: $\lambda_k^*(t) = \lambda_k(t)$





Hawkes process: $\lambda_k^*(t) = \lambda_k(t) + \sum_{k'=1}^{K} \sum_{\{(t_j,k_j):t_j<t,\ k_j=k'\}} \alpha_{k'k}\beta_{k'k}e^{-\beta_{k'k}(t-t_j)}$





Marked Temporal Point Processes

The event history until time t:

$$\mathcal{H}_t = \{(t_j, k_j) \mid t_j < t\}.$$

The k-th counting process $(k \in \mathbb{K})$:

$$N_k(t) = \sum_{j=1}^{m} \mathbb{1}(t_j \leq t \ \cap \ k_j = k).$$

The marked intensity functions $(k \in \mathbb{K})$:

$$\lambda_k^*(t) = \lambda_k(t|\mathcal{H}_t) = \lim_{\Delta t \to 0} \frac{\mathbb{E}[N_k(t + \Delta t) - N_k(t)|\mathcal{H}_t]}{\Delta t}.$$

MTPP model training

Dataset composed of m events $\boldsymbol{e}_j = (t_j, k_j)$ where $t_j \in [0, T]$ and $k_j \in \mathbb{K}$:

$$\mathcal{S} = \{(t_1, k_1), (t_2, k_2), \ldots, (t_m, k_m)\} \text{ or } \mathcal{S} = \{(\tau_1, k_1), (\tau_2, k_2), \ldots, (\tau_m, k_m)\}.$$

Negative log-likelihood with $\lambda_k^*(t; \boldsymbol{\theta})$ for $k \in \mathbb{K}$:

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{S}) = -\sum_{j=1}^{m} \log \lambda_{k_j}^*(t_j; \boldsymbol{\theta}) - \int_0^T \sum_{k=1}^{K} \lambda_k^*(t; \boldsymbol{\theta}) dt.$$

MTPP model training

Dataset composed of m events $\boldsymbol{e}_j = (t_j, k_j)$ where $t_j \in [0, T]$ and $k_j \in \mathbb{K}$:

$$\mathcal{S} = \{(t_1, k_1), (t_2, k_2), \ldots, (t_m, k_m)\} \text{ or } \mathcal{S} = \{(\tau_1, k_1), (\tau_2, k_2), \ldots, (\tau_m, k_m)\}.$$

Negative log-likelihood with $\lambda_k^*(t; \boldsymbol{\theta})$ for $k \in \mathbb{K}$:

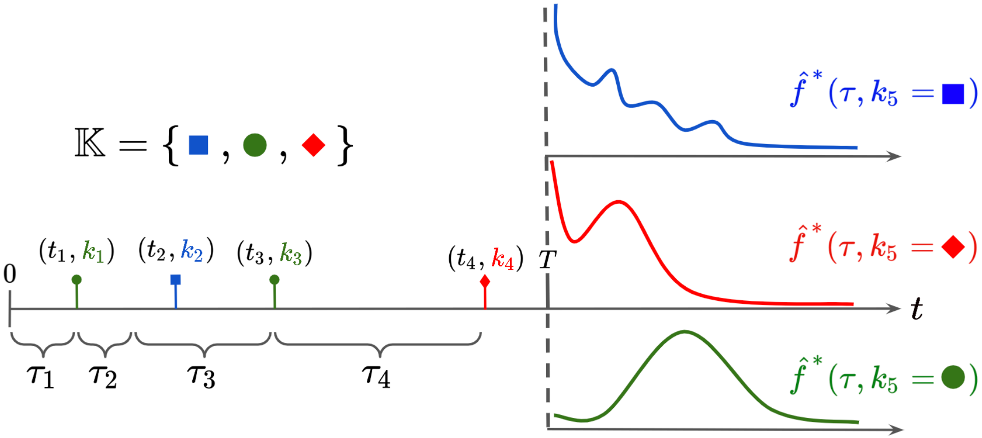
$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{S}) = -\sum_{j=1}^{m} \log \lambda_{k_j}^*(t_j; \boldsymbol{\theta}) - \int_0^T \sum_{k=1}^{K} \lambda_k^*(t; \boldsymbol{\theta}) dt.$$

Negative log-likelihood with $f^*(\tau, k; \boldsymbol{\theta}) = f^*(\tau; \boldsymbol{\theta}) p^*(k|\tau; \boldsymbol{\theta})$:

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{S}) = -\sum_{j=1}^{m} \left[\underbrace{\log f^*(\tau_j; \boldsymbol{\theta})} + \underbrace{\log p^*(k_j|\tau_j; \boldsymbol{\theta})} \right] + \underbrace{\log \left(1 - F^*(T - t_m; \boldsymbol{\theta})\right)},$$

where $F^*(\tau) = \int_0^\tau \sum_{k=1}^K f^*(s, k) ds$.

Joint predictive density with $|\mathbb{K}| = 3$



UMONS
Université de Mons

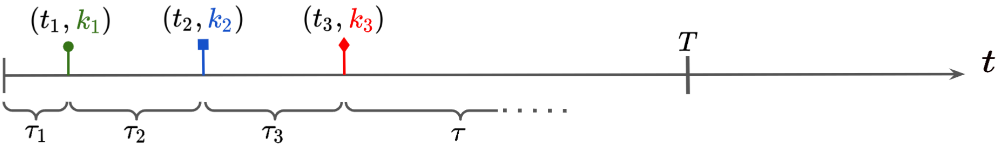
Neural Marked Temporal Point Processes

Classical TPPs lack flexibility to capture **complex dependencies** between past and future events [ME16].

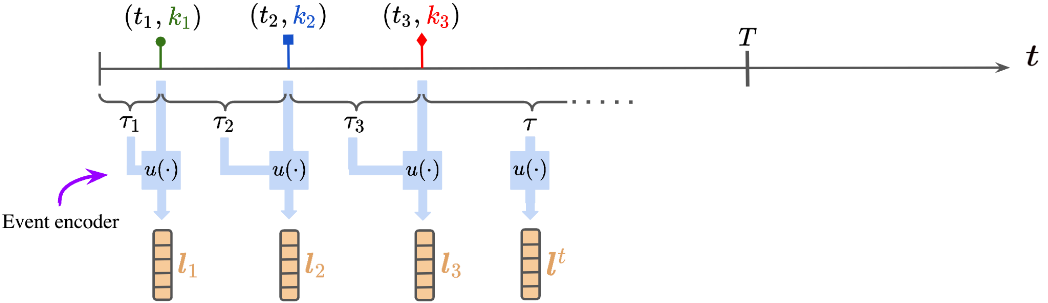
Neural TPPs leverage neural network flexibility to enhance **representation learning** and build **highly flexible** and fully end-to-end trainable models [Shc+21].

- **Neural network architectures**: recurrent architectures [Du+16], attention mechanisms [Zuo+20; Zha+19; Eng+20], non-recurrent architectures [Shc+20].
- **Model parametrizations**: CIF [OUA19], PDF [SBG20], QF [Tai22]
- **Training objectives**: least-squares [Yic+16; Xu+17], adverserial learning [Xia+18], noise constrative estimation [MWE20; GLL18], variational objectives [Boy+20], reinforcement learning [UDG18]

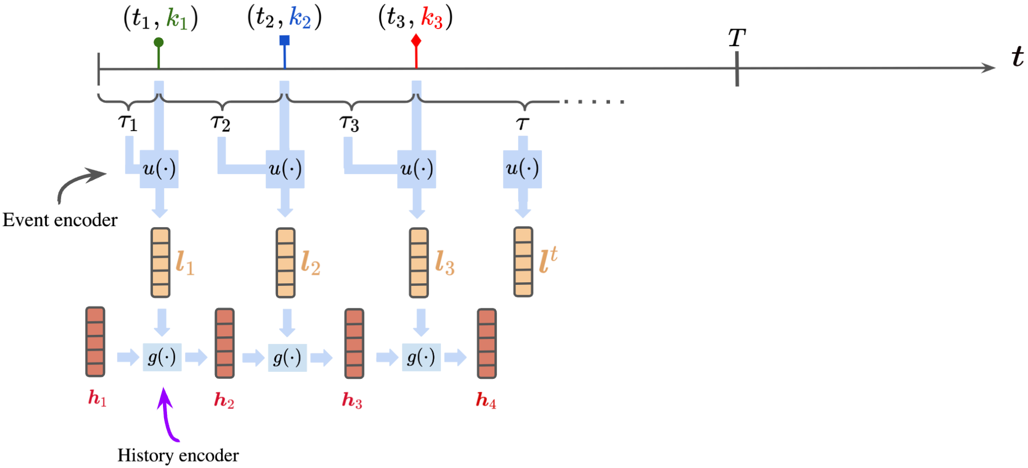
Neural Marked Temporal Point Processes



Neural Marked Temporal Point Processes

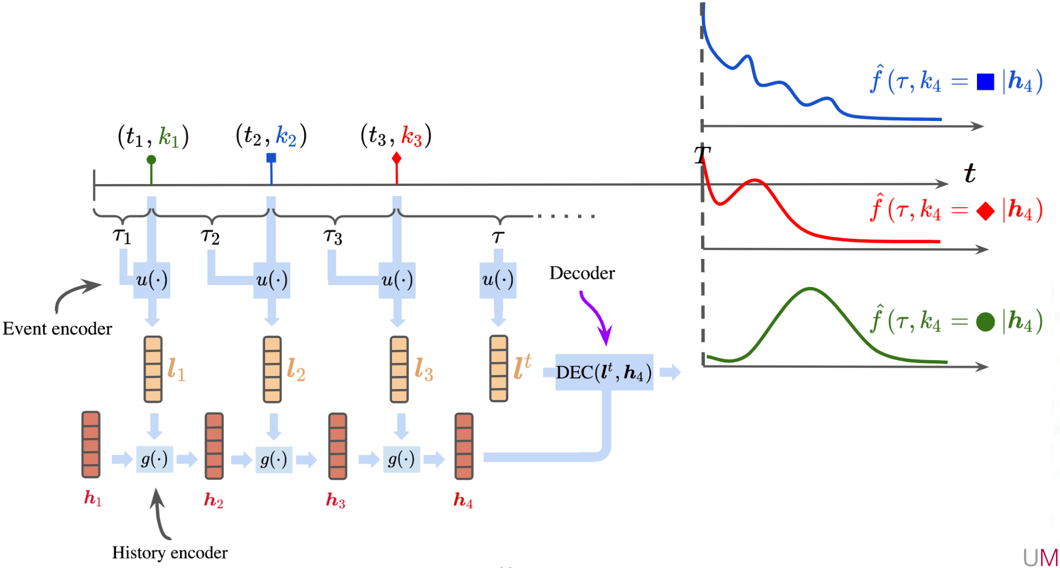


Neural Marked Temporal Point Processes



UMONS
Université de Mons

Neural Marked Temporal Point Processes



The conditional LogNormMix decoder [BB23]

$$\boxed{\hat{f}\left(\tau, k \middle| \boldsymbol{h}\right) = \hat{f}\left(\tau \middle| \boldsymbol{h}\right)\hat{p}\left(k \middle| \tau, \boldsymbol{h}\right)}$$

UMONS
Université de Mons

The conditional LogNormMix decoder [BB23]

$$\mathrm{Softmax}(\mathbf{W}_p \boldsymbol{h} + \boldsymbol{b}_p)_c$$

$$\sum_{c=1}^{C} p(c|\boldsymbol{h}) \frac{1}{\tau \boldsymbol{\sigma}_c \sqrt{2\pi}} \exp\left(-\frac{(\log \tau - \boldsymbol{\mu}_c)^2}{2\boldsymbol{\sigma}_c^2}\right)$$

$$\boxed{\hat{f}(\tau, k|\boldsymbol{h}) = \hat{f}(\tau|\boldsymbol{h})\hat{p}(k|\tau, \boldsymbol{h})}$$

The conditional LogNormMix decoder [BB23]

$$\mathrm{Softmax}(\mathbf{W}_p \boldsymbol{h} + \boldsymbol{b}_p)_c \qquad \exp(\mathbf{W}_\mu \boldsymbol{h} + \boldsymbol{b}_\mu)_c$$

$$\sum_{c=1}^{C} p(c|\boldsymbol{h}) \frac{1}{\tau \boldsymbol{\sigma}_c \sqrt{2\pi}} \exp\left(-\frac{(\log \tau - \boldsymbol{\mu}_c)^2}{2\boldsymbol{\sigma}_c^2}\right)$$

$$\exp(\mathbf{W}_\sigma \boldsymbol{h} + \boldsymbol{b}_\sigma)_c$$

$$\boxed{\hat{f}(\tau, k|\boldsymbol{h}) = \hat{f}(\tau|\boldsymbol{h})\hat{p}(k|\tau, \boldsymbol{h})}$$

The conditional LogNormMix decoder [BB23]

$$\mathrm{Softmax}(\mathbf{W}_p \boldsymbol{h} + \boldsymbol{b}_p)_c$$

$$\exp(\mathbf{W}_\mu \boldsymbol{h} + \boldsymbol{b}_\mu)_c$$

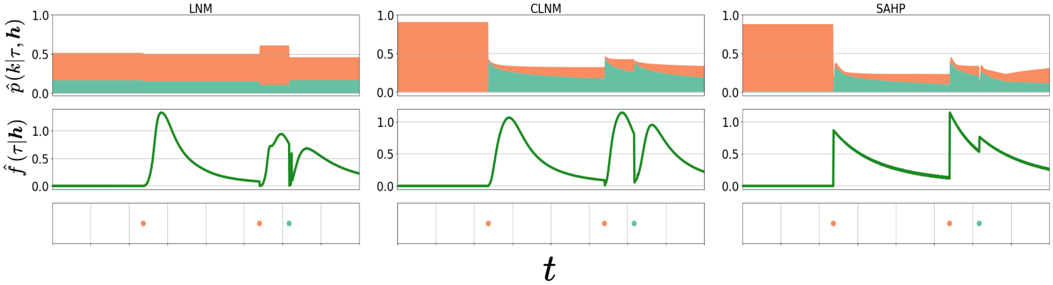
$$\sum_{c=1}^{C} p(c|\boldsymbol{h}) \frac{1}{\tau \boldsymbol{\sigma}_c \sqrt{2\pi}} \exp\left(-\frac{(\log\tau - \boldsymbol{\mu}_c)^2}{2\boldsymbol{\sigma}_c^2}\right)$$

$$\exp(\mathbf{W}_\sigma \boldsymbol{h} + \boldsymbol{b}_\sigma)_c$$

$$\boxed{\hat{f}(\tau, k|\boldsymbol{h}) = \hat{f}(\tau|\boldsymbol{h})\hat{p}(k|\tau, \boldsymbol{h})}$$

$$\mathrm{Softmax}\left(\mathbf{W}_2 \mathrm{ReLU}\left(\mathbf{W}_1 [\boldsymbol{h}||\boldsymbol{l}^t] + \boldsymbol{b}_1\right) + \boldsymbol{b}_2\right)_k$$

Examples of mark and time predictive distributions



UMONS
Université de Mons

Plan

Temporal point processes

Distribution-free uncertainty quantification

Conformal neural temporal point processes

Experiments

More reliable uncertainty quantification with conformal prediction

- Black-box (neural) models provide a "**heuristic**" notion of uncertainty without (finite-sample) prediction **guarantees**.
  - A **reliable** uncertainty quantification is essential for optimal **decision-making** and safe **deployment**
- Many **sources of uncertainty**: model misspecification, noisy and missing data, etc. See "*Sources of Uncertainty in Machine Learning – A Statisticians' View*" [Gru+23]
- With conformal prediction [VGS05], we can generate **distribution-free** *prediction regions* with **finite-sample calibration** guarantees from any model.
- We want our prediction sets to be sufficiently **sharp** to obtain **informative** predictions.

Split conformal prediction algorithm

$\mathcal{D} = \{\, (\boldsymbol{h}_i, \boldsymbol{y}_i)\, \}_{i=1}^{n}$: a dataset consisting of n **exchangeable** pairs.

$\hat{g}$: a model that provides a **heuristic measure of uncertainty** for $\boldsymbol{y}$ given $\boldsymbol{h}$.

The *split conformal algorithm* transforms any $\hat{g}$ into a **rigorous** one [AB21].

UMONS
Université de Mons

Split conformal prediction algorithm

$\mathcal{D} = \{ (\boldsymbol{h}_i, \boldsymbol{y}_i) \}_{i=1}^{n}$: a dataset consisting of n **exchangeable** pairs.

$\hat{g}$: a model that provides a **heuristic measure of uncertainty** for $\boldsymbol{y}$ given $\boldsymbol{h}$.

The *split conformal algorithm* transforms any $\hat{g}$ into a **rigorous** one [AB21].

1. **Split** $\mathcal{D}$ into two *non-overlapping* sets, $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{cal}}$ with $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{cal}} = \mathcal{D}$.
2. Train the **model** with the observations in $\mathcal{D}_{\text{train}}$, to obtain $\hat{g}$.
3. Use $\hat{g}$ to define a **non-conformity score** function $s(\boldsymbol{h}, \boldsymbol{y}) \in \mathbb{R}$
   - It assigns larger value to worse agreement between $\boldsymbol{h}$ and $\boldsymbol{y}$.
4. Compute the **calibration scores** using the observations in $\mathcal{D}_{\text{cal}}$:

$$\{ s_i \}_{i=1}^{|\mathcal{D}_{\text{cal}}|} := \{ s(\boldsymbol{h}, \boldsymbol{y}) : (\boldsymbol{h}, \boldsymbol{y}) \in \mathcal{D}_{\text{cal}} \}$$

Split conformal prediction algorithm

6. Compute the $1-\alpha$ **empirical quantile** of these calibration scores:

$$\hat{q} = \text{Quantile}\left(s_1, \ldots, s_{|\mathcal{D}_{\text{cal}}|} \cup \{\,\infty\,\}\,; \frac{\lceil(|\mathcal{D}_{\text{cal}}|+1)(1-\alpha)\rceil}{|\mathcal{D}_{\text{cal}}|}\right).$$

7. For $\boldsymbol{h}_{n+1}$, use $\hat{q}$ to construct a **prediction region** for $\boldsymbol{y}_{n+1}$ with $1-\alpha$ coverage:

$$\hat{R}_{\boldsymbol{y}}(\boldsymbol{h}_{n+1}) = \{\,\boldsymbol{y} \in \mathcal{Y} : s(\boldsymbol{h}_{n+1}, \boldsymbol{y}) \leq \hat{q}\,\}\,.$$

Split conformal prediction algorithm

6. Compute the $1-\alpha$ **empirical quantile** of these calibration scores:

$$\hat{q} = \text{Quantile}\left(s_1, \ldots, s_{|\mathcal{D}_{\text{cal}}|} \cup \{\,\infty\,\}\,;\, \frac{\lceil (|\mathcal{D}_{\text{cal}}|+1)(1-\alpha)\rceil}{|\mathcal{D}_{\text{cal}}|}\right).$$

7. For $\boldsymbol{h}_{n+1}$, use $\hat{q}$ to construct a **prediction region** for $\boldsymbol{y}_{n+1}$ with $1-\alpha$ coverage:

$$\hat{R}_{\boldsymbol{y}}(\boldsymbol{h}_{n+1}) = \{\,\boldsymbol{y} \in \mathcal{Y} : s(\boldsymbol{h}_{n+1}, \boldsymbol{y}) \le \hat{q}\,\}\,.$$

$$\mathbb{P}\left(\boldsymbol{y}_{n+1} \in \hat{R}_{\boldsymbol{y}}(\boldsymbol{h}_{n+1})\right) = \mathbb{P}(s(\boldsymbol{h}_{n+1}, \tau_{n+1}) \le \hat{q}) \quad \underset{\ge}{\overset{\text{quantile lemma}}{\uparrow}} \quad 1-\alpha$$

Quantile Lemma. If $S_1, \ldots, S_n, S_{n+1}$ are **exchangeable** variables, then

$$\mathbb{P}\left\{S_{n+1} \le \text{Quantile}\left(1-\alpha; \{S_i\}_{i=1}^{n} \cup \{\infty\}\right)\right\} \ge 1-\alpha, \quad \forall \alpha \in (0,1).$$

If ties between $S_1, \ldots, S_n, S_{n+1}$ occur with probability zero, then the rhs is $1-\alpha+\frac{1}{n+1}$.

A very active area of research

Conformal prediction: Vovk, Gammerman, and Shafer [VGS05]

Conformal regression: Lei, G'Sell, Rinaldo, Tibshirani, and Wasserman [Lei+18], Romano, Patterson, and Candes [RPC19], and Sesia and Romano [SR21]

Conformal classification: Romano, Sesia, and Candès [RSC20]

Conformal density estimation: Izbicki, Shimizu, and Stern [ISS22]

Conditional coverage: Foygel Barber, Candès, Ramdas, and Tibshirani [Foy+20] and Gibbs, Cherian, and Candès [GCC23]

Conformal time series forecasting: Stankeviciute, M Alaa, and Schaar [SMS21], Lin, Trivedi, and Sun [LTS22], and Angelopoulos, Candes, and Tibshirani [ACT23]

Conformal spatial prediction: Mao, Martin, and Reich [MMR20]

Beyond exchangeability: [Bar+22; Tib+19]

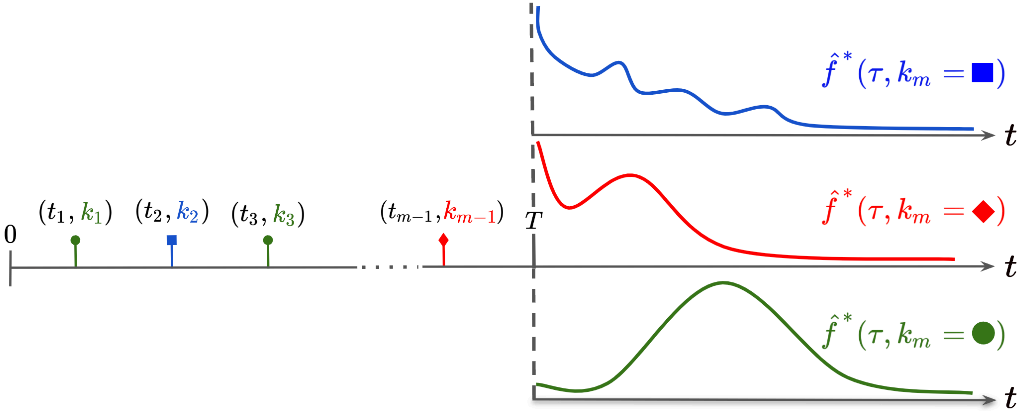
Multi-response: [FBR23; LRW13]

UMONS
Université de Mons

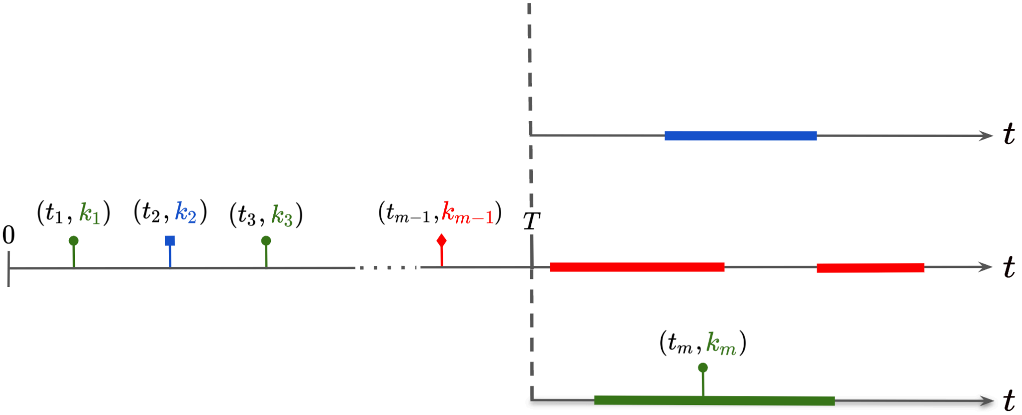
Plan

UMONS
Université de Mons

Conformal neural TPPs



Conformal neural TPPs



UMONS
Université de Mons

Conformal neural TPPs

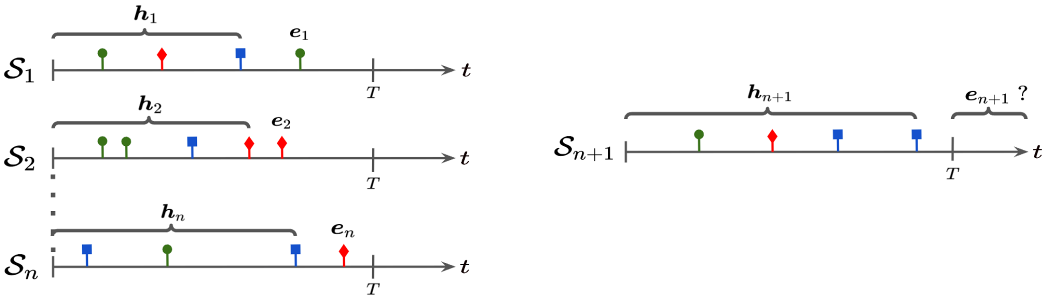
Challenge I. The **events** in a sequence are not exchangeable (temporal dependence).

- In the neural TPP literature, we often assume the **sequences** are exchangeable.
- A similar setting considered in **conformal time series forecasting** [SMS21]

Conformal neural TPPs

Challenge I. The **events** in a sequence are <u>not</u> exchangeable (temporal dependence).

- In the neural TPP literature, we often assume the **sequences** are exchangeable.
- A similar setting considered in **conformal time series forecasting** [SMS21]



UMONS
Université de Mons

Conformal neural TPPs

Challenge II. We need to generate a **joint** prediction region for a **bivariate response**, accommodating both a **continuous** and a **categorical** response.

Given $\boldsymbol{h}_{n+1}$ and $\alpha \in (0, 1)$, our aim is to construct an **informative**, **distribution-free** joint prediction region $\hat{R}_{\tau,k}(\boldsymbol{h}_{n+1}) \in \mathbb{R}^+ \times \mathbb{K}$ for the pair (τ_{n+1}, k_{n+1}) with **finite-sample** marginal coverage, i.e.

$$\mathbb{P}((\tau_{n+1}, k_{n+1}) \in \hat{R}_{\tau,k}(\boldsymbol{h}_{n+1})) \geq 1 - \alpha.$$

UMONS
Université de Mons

Conformal neural TPPs

Challenge II. We need to generate a **joint** prediction region for a **bivariate response**, accommodating both a **continuous** and a **categorical** response.

Given $\boldsymbol{h}_{n+1}$ and $\alpha \in (0,1)$, our aim is to construct an **informative**, **distribution-free** joint prediction region $\hat{R}_{\tau,k}(\boldsymbol{h}_{n+1}) \in \mathbb{R}^+ \times \mathbb{K}$ for the pair (τ_{n+1}, k_{n+1}) with **finite-sample** marginal coverage, i.e.

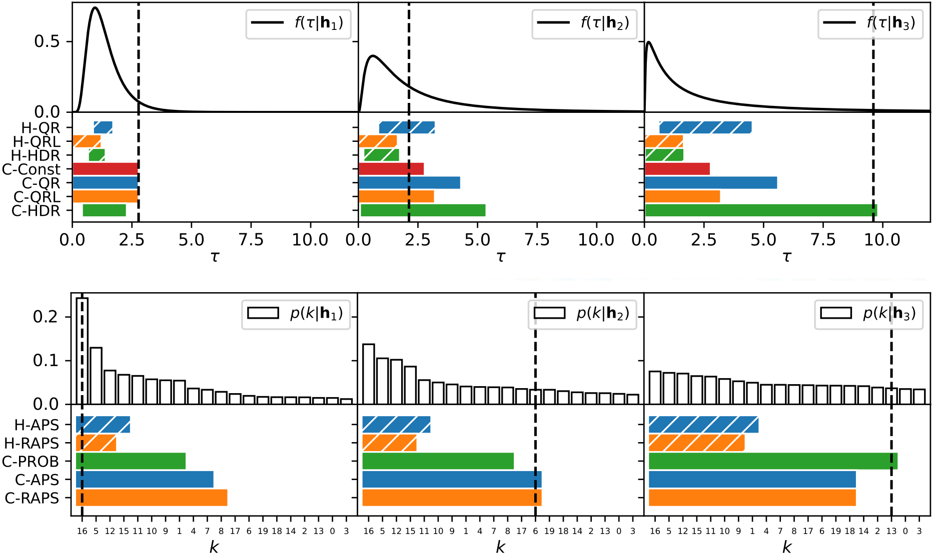
$$\mathbb{P}((\tau_{n+1}, k_{n+1}) \in \hat{R}_{\tau,k}(\boldsymbol{h}_{n+1})) \geq 1 - \alpha.$$

We will explore two approaches:

1. A naive yet valid method combining **individual** prediction sets for τ_{n+1} and k_{n+1}.
2. An approach based on the **highest density regions** (HDRs) of the **joint predictive density** of (τ_{n+1}, k_{n+1}).

UMONS
Université de Mons

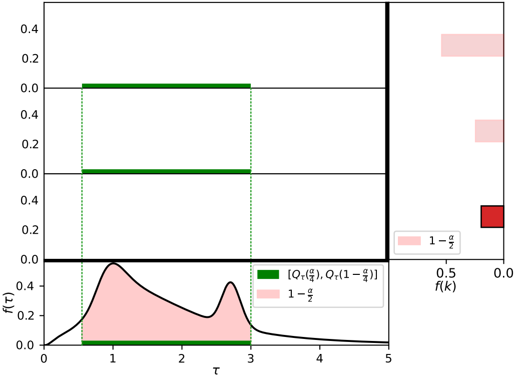
Individual prediction regions



UMONS
Université de Mons

Naive bivariate prediction regions

We construct a $1 - \alpha$ **bivariate** prediction region for (τ_{n+1}, k_{n+1}) by combining **individual** predictions regions $\hat{R}_\tau(\boldsymbol{h}_{n+1})$ and $\hat{R}_k(\boldsymbol{h}_{n+1})$, each with coverage $1 - \alpha/2$.



$$
\begin{aligned}
&\hat{R}_{\tau,k}(\boldsymbol{h}_{n+1}) \\
&= \hat{R}_\tau(\boldsymbol{h}_{n+1}) \times \hat{R}_k(\boldsymbol{h}_{n+1}) \\
&= \{(\tau', k') | \tau' \in \hat{R}_\tau(\boldsymbol{h}_{n+1}), k' \in \hat{R}_k(\boldsymbol{h}_{n+1})\}
\end{aligned}
$$

Naive bivariate prediction regions

By the union bound, we have:

$$
\begin{aligned}
&\mathbb{P}((\tau_{n+1}, k_{n+1}) \in \hat{R}_\tau(\boldsymbol{h}_{n+1}) \times \hat{R}_k(\boldsymbol{h}_{n+1})) \\
&= \mathbb{P}(\tau_{n+1} \in \hat{R}_\tau(\boldsymbol{h}_{n+1}) \cap k_{n+1} \in \hat{R}_k(\boldsymbol{h}_{n+1})) \\
&= 1 - \underbrace{\mathbb{P}(\tau_{n+1} \notin \hat{R}_\tau(\boldsymbol{h}_{n+1}) \cup k_{n+1} \notin \hat{R}_k(\boldsymbol{h}_{n+1}))}_{\leq \alpha/2 + \alpha/2} \\
&\geq 1 - \alpha.
\end{aligned}
$$

However, this method can be overly **conservative**, resulting in large and inflexible prediction regions. Indeed, the joint prediction region generated by this approach yields **the same prediction interval for the arrival time** across all selected marks.

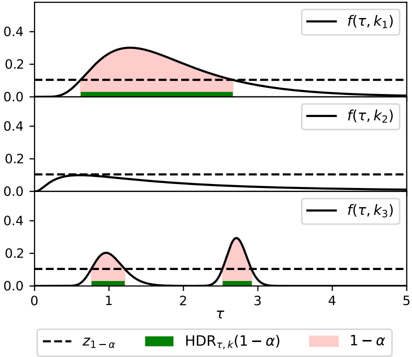
Bivariate HDRs

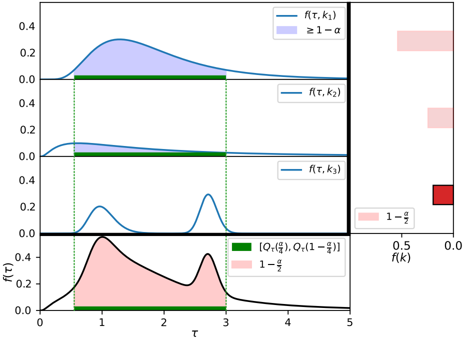
The **HDR** of $\hat{f}(\tau, k|\boldsymbol{h}_{n+1})$ with nominal coverage level $1-\alpha$ is defined as:

$$\text{HDR}(1-\alpha|\boldsymbol{h}_{n+1}) = \left\{ (\tau, k) \;\middle|\; \hat{f}(\tau, k|\boldsymbol{h}_{n+1}) \geq z_{1-\alpha} \right\},$$

where

$$z_{1-\alpha} = \sup \left\{ z' \;\middle|\; \mathbb{P}(\hat{f}(\tau, k|\boldsymbol{h}_{n+1}) \geq z') \geq 1-\alpha \right\}.$$





UMONS
Université de Mons

Bivariate HDRs

- With the **joint predictive density**, we account for the **dependence** between τ and k.
  - We exclude **unlikely combinations** of the two variables while maintaining the pre-specified coverage level.
- With multimodal distributions, an HDR is a **union of intervals** collectively **shorter** than a single interval with the same coverage level.
- The oracle HDR has the useful property of generating the **smallest possible region** that guarantees conditional coverage.

The joint HDR can be expressed as

$$\hat{R}_{\tau,k}(\boldsymbol{h}_{n+1}) = \mathrm{HDR}(1-\alpha|\boldsymbol{h}_{n+1}) = \bigcup_{k' \in \hat{R}_k(\boldsymbol{h}_{n+1})} \{(\tau', k')|\tau' \in \hat{R}_\tau^{(k')}(\boldsymbol{h}_{n+1})\}$$

where

$\hat{R}_k(\boldsymbol{h}_{n+1}) = \{k'|\exists\, \tau \in \mathbb{R}^+ : \hat{f}(\tau, k'|\boldsymbol{h}_{n+1}) \geq z_{1-\alpha}\}$ and $\hat{R}_\tau^{(k)}(\boldsymbol{h}_{n+1}) \quad = \{\tau'|\hat{f}(\tau', k|\boldsymbol{h}_{n+1}) \geq z_{1-\alpha}\}$.

Conformal bivariate HDRs

By definition[1], we have

$$\begin{aligned}\mathrm{HDR}(\hat{q}) &= \left\{ \boldsymbol{y} \mid \hat{f}(\boldsymbol{y}) \geq z_{\hat{q}} \right\}, \text{where } z_{\hat{q}} = \sup\left\{ z' \mid \mathbb{P}(\hat{f}(\boldsymbol{y}) \geq z') \geq \hat{q} \right\} \\ &= \left\{ \boldsymbol{y} \mid F_z(\hat{f}(\boldsymbol{y})) \geq 1 - \hat{q} \right\},\end{aligned}$$

This implies that

$$\boldsymbol{y}_{n+1} \in \mathrm{HDR}(\hat{q}) \iff F_z(\hat{f}(\boldsymbol{y}_{n+1})) \geq 1 - \hat{q} \iff \underbrace{1 - F_z(\hat{f}(\boldsymbol{y}_{n+1}))}_{\mathrm{HPD}(\boldsymbol{y}_{n+1})} \leq \hat{q},$$

where

$$\mathrm{HPD}(\boldsymbol{y}) = 1 - F_z(\hat{f}(\boldsymbol{y})) = \mathbb{P}(z \geq \hat{f}(\boldsymbol{y})) = \int_{\left\{ \boldsymbol{y}' \mid \hat{f}(\boldsymbol{y}') \geq \hat{f}(\boldsymbol{y}) \right\}} \hat{f}(\boldsymbol{y}')d\boldsymbol{y}'.$$

[1] To simplify notations, we remove the dependence on $\boldsymbol{h}$ and write $\boldsymbol{y} = (\tau, k)$.

Conformal bivariate HDRs

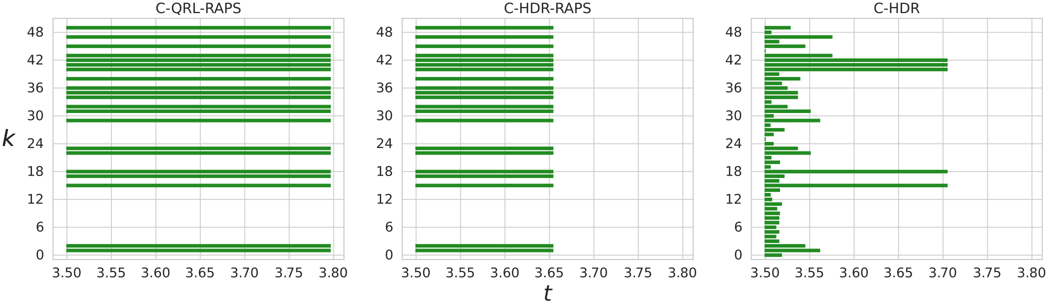
This is a generalization of the univariate HPD-split method [ISS22] for bivariate responses, denoted **C-HDR**.

C-HDR is based on the following **non-conformity score**:

$$s_{\text{C-HDR}}(\boldsymbol{h},(\tau,k)) = \text{HPD}(\tau,k|\boldsymbol{h}) = \sum_{k'\in\mathbb{K}} \int_{\{\,\tau' \,\mid\, \hat{f}(\tau',k'|\boldsymbol{h})\geq \hat{f}(\tau,k|\boldsymbol{h})\,\}} \hat{f}(\tau',k'|\boldsymbol{h})d\tau',$$

where $\text{HPD}(\tau,k|\boldsymbol{h})$ calculates the **probability coverage** of pairs (τ',k') having a higher density than (τ,k).

Examples of joint predictions regions



Plan

Experimental setup

	#Seq.	#Events	Mean Length	Max Length	Min Length	#Marks
LastFM	856	193441	226.0	6396	2	50
MOOC	7047	351160	49.8	416	2	50
Reddit	4278	238734	55.8	941	2	50
Retweets	12000	1309332	109.1	264	50	3
Stack Overflow	7959	569688	71.6	735	40	22

The sequences are **randomly split** into $\mathcal{D}_{\text{train}}/\mathcal{D}_{\text{cal}}/\mathcal{D}_{\text{test}}$ with sizes 75%, 15% and 10%. This procedure is repeated **5 times**. For each dataset, the model is trained using $\mathcal{D}_{\text{train}}$, and the results are averaged over the 5 $\mathcal{D}_{\text{test}}$ splits.

We present the results for the **CLNM** nerual TPP model, for $\alpha = 0.2$. We consider both **heuristic** (H-) and **conformal** (C-) methods.

UMONS
Université de Mons

Heuristic and conformal methods for individual regions

Quantile regression (**QR**):

$$\hat{R}_{\tau,\text{H-QR}}(\boldsymbol{h}_{n+1}) = [\hat{Q}_\tau(\alpha|\boldsymbol{h}_{n+1}), \hat{Q}_\tau(1-\alpha|\boldsymbol{h}_{n+1})]$$
$$\hat{R}_{\tau,\text{C-QR}}(\boldsymbol{h}_{n+1}) = [\hat{Q}_\tau(\alpha|\boldsymbol{h}_{n+1}) - \hat{q}, \hat{Q}_\tau(1-\alpha|\boldsymbol{h}_{n+1}) + \hat{q}]$$

Quantile regression left (**QRL**):

$$\hat{R}_{\tau,\text{H-QRL}}(\boldsymbol{h}_{n+1}) = [0, \hat{Q}_\tau(1-\alpha|\boldsymbol{h}_{n+1})]$$
$$\hat{R}_{\tau,\text{C-QRL}}(\boldsymbol{h}_{n+1}) = [0, \hat{Q}_\tau(1-\alpha|\boldsymbol{h}_{n+1}) + \hat{q}]$$

Univariate Highest Density Regions (**HDR-T**):

$$\hat{R}_{\tau,\text{H-HDR-T}}(\boldsymbol{h}_{n+1}) = \{\tau|\hat{f}(\tau|\boldsymbol{h}_{n+1}) \geq z_{1-\alpha}\}, \quad z_{1-\alpha} = \sup\left\{ z' \,\middle|\, \mathbb{P}(\hat{f}(\tau|\boldsymbol{h}_{n+1}) \geq z') \geq 1-\alpha \right\}$$
$$\hat{R}_{\tau,\text{C-HDR-T}}(\boldsymbol{h}_{n+1}) = \{\tau|\hat{f}(\tau|\boldsymbol{h}_{n+1}) \geq z_{\hat{q}}\}, \quad z_{\hat{q}} = \sup\left\{ z' \,\middle|\, \mathbb{P}(\hat{f}(\tau|\boldsymbol{h}_{n+1}) \geq z') \geq \hat{q} \right\}$$

(Regularized) adaptive prediction sets (**(R)APS**):

$$\hat{R}_{k,\text{H-(R)APS}}(\boldsymbol{h}_{n+1}) = \left\{ k' \in \mathbb{K} : s_{\text{(R)APS}}(\boldsymbol{h}_{n+1}, k') \leq 1-\alpha \right\}$$
$$\hat{R}_{k,\text{C-(R)APS}}(\boldsymbol{h}_{n+1}) = \left\{ k' \in \mathbb{K} : s_{\text{(R)APS}}(\boldsymbol{h}_{n+1}, k') \leq \hat{q} \right\}$$

Heuristic and conformal methods for bivariate regions

- The naive method which combines individual prediction regions
    - **C-QRL-RAPS** : C-QRL for $\hat{R}_\tau(\boldsymbol{h}_{n+1})$ and C-RAPS for $\hat{R}_k(\boldsymbol{h}_{n+1})$.
    - **C-HDR-RAPS** : C-HDR-T for $\hat{R}_\tau(\boldsymbol{h}_{n+1})$ and C-RAPS for $\hat{R}_k(\boldsymbol{h}_{n+1})$.
- The conformal highest density regions method (**C-HDR**) based on the joint density of the arrival time and the mark.
- The heuristic counterparts, denoted as **H-QRL-RAPS**, **H-HDR-RAPS** and **H-HDR**.

Evaluation metrics

Marginal coverage:

$$\mathsf{MC} = \hat{\mathbb{P}}_{\mathcal{D}_{\text{test}}}\left(\boldsymbol{y}_i \in \hat{R}_{\boldsymbol{y}}(\boldsymbol{h}_i)\right) = \frac{1}{|\mathcal{D}_{\text{test}}|}\sum_{i=1}^{|\mathcal{D}_{\text{test}}|} \mathbb{1}\left[\boldsymbol{y}_i \in \hat{R}_{\boldsymbol{y}}(\boldsymbol{h}_i)\right].$$

Average length:

$$\mathsf{Length} = \frac{1}{|\mathcal{D}_{\text{test}}|}\sum_{i=1}^{|\mathcal{D}_{\text{test}}|} |\hat{R}_{\boldsymbol{y}}(\boldsymbol{h}_i)|.$$

Geometric average of the lengths[2]:

$$\mathsf{G.\ Length} = \frac{1}{|\mathcal{D}_{\text{test}}|}\sum_{i=1}^{|\mathcal{D}_{\text{test}}|} \log(|\hat{R}_{\boldsymbol{y}}(\boldsymbol{h}_i)| + \epsilon)^3,$$

[2] to decrease the weight of large lengths and increase the weight of small lengths

[3] ϵ is a small value to handle the case where $|\hat{R}_{\boldsymbol{y}}(\boldsymbol{h}_i)| = 0$

Evaluation metrics

Worst slab coverage (**WSC**, [CGD21]), the coverage conditionally to the worst slab $\boldsymbol{v} \in \mathbb{R}^{d_h}$:

$$\mathsf{WSC} = \min_{\boldsymbol{v}_j \in \mathbb{S}^{d-1}} \inf_{a<b} \left\{ \hat{\mathbb{P}}_{\mathcal{D}_{\text{test}}} \left(\boldsymbol{y}_i \in \hat{R}_{\boldsymbol{y}}(\boldsymbol{h}_i) |\ a \le \boldsymbol{v}^\intercal \boldsymbol{h}_i \le b \right) \,\middle|\, \hat{\mathbb{P}}_{\mathcal{D}_{\text{test}}}(a \le \boldsymbol{v}^\intercal \boldsymbol{h}_i \le b) \ge \delta \right\},$$

each containing at least a proportion δ of the total mass, where $0 < \delta \le 1$.

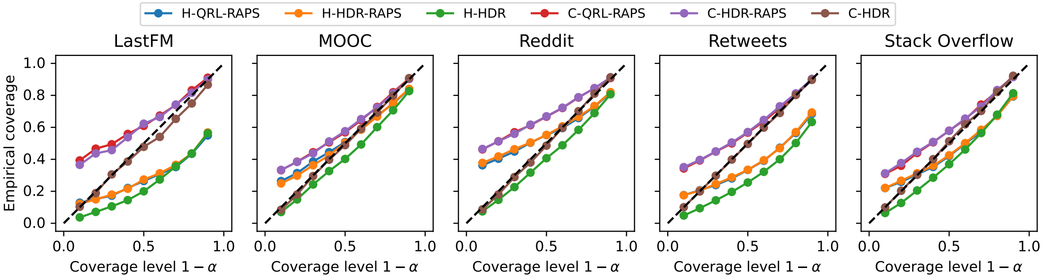
Conditional coverage error (**CCE**), the average coverage error over different clusters $A_1, \ldots, A_J$:

$$\mathsf{CCE} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i=1}^{|\mathcal{D}_{\text{test}}|} \sum_{j=1}^{J} \left(\hat{\mathbb{P}}_{\mathcal{D}_{\text{test}}} \left(\boldsymbol{y}_i \in \hat{R}_{\boldsymbol{y}}(\boldsymbol{h}_i) \,\middle|\, \boldsymbol{h}_i \in A_j \right) - (1-\alpha) \right)^2,$$

where the clusters are determined by the k-means++ algorithm using the 2-Wasserstein distance function to cluster instances whose HPD values Z are similary distributed [ISS22]:

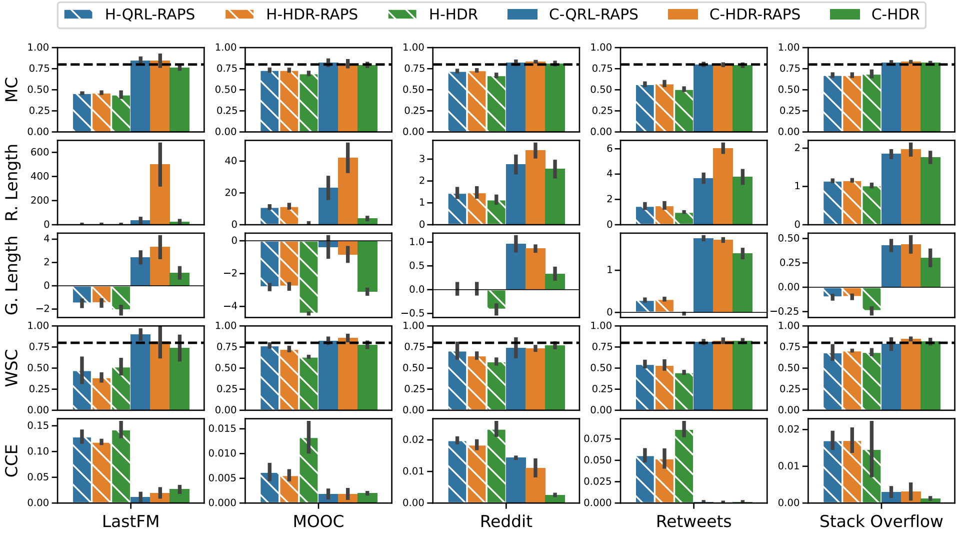
$$d_Z(\boldsymbol{h}_a, \boldsymbol{h}_b) = \left(\int_0^1 \left| F_Z^{-1}(u \mid \boldsymbol{h}_a) - F_Z^{-1}(u \mid \boldsymbol{h}_b) \right|^2 du \right)^{\frac{1}{2}}.$$

Marginal coverage for different coverage levels



- Heuristic methods **undercover** for large coverage levels
- Combining individual regions leads to **overcoverage**
- C-HDR obtains the **right coverage** at all coverage levels

Results for the bivariate prediction regions



UMONS
Université de Mons

Summary

V. Dheur, T. Bosser, S. Ben Taieb. Distribution-Free Conformal Joint Prediction Regions for Neural Marked Temporal Point Processes (2024). arxiv.org/abs/2401.04612

- We want to generate **distribution-free**, **calibrated**, and **informative** bivariate prediction regions for the arrival time and mark from **neural TPP models**.
- The **naive approach** which combines individual prediction sets can be overly **conservative**, resulting in large and inflexible prediction regions.
- We proposed a conformal approach based on **HDRs** which efficiently excludes **unlikely combinations** of the two variables while maintaining the pre-specified coverage level.
- Future work
  - **Relax** the assumption of exchangeable sequences using **block structures**.
  - Other **stronger** (achievable) coverage criteria.
  - Continuous or/and **multivariate** mark, and **spatio-temporal** marked processes

References I

[AB21] Anastasios N Angelopoulos and Stephen Bates. "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification". In: (July 2021). arXiv: 2107.07511 [cs.LG].

[ACT23] Anastasios N Angelopoulos, Emmanuel J Candes, and Ryan J Tibshirani. "Conformal PID Control for Time Series Prediction". In: (July 2023). arXiv: 2307.16895 [cs.LG].

[Ang+22] Anastasios Angelopoulos et al. *Uncertainty Sets for Image Classifiers using Conformal Prediction*. 2022. arXiv: 2009.14193 [cs.CV].

[Bar+22] Rina Foygel Barber et al. "Conformal prediction beyond exchangeability". In: (Feb. 2022). arXiv: 2202.13415 [stat.ME].

[BB23] Tanguy Bosser and Souhaib Ben Taieb. *Revisiting the Mark Conditional Independence Assumption in Neural Marked Temporal Point Processes*. ESANN. 2023.

[BMM15] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. *Hawkes processes in finance*. Market Microstructure and Liquidity. 2015.

[Boy+20] Alex Boyd et al. *User-Dependent Neural Sequence Models for Continuous-Time Event Data*. Neurips. 2020.

UMONS Université de Mons

References II

[Cai+18] Renqin Cai et al. *Modeling Sequential Online Interactive Behaviors with Temporal Point Process.* CIKM. 2018.

[CGD21] Maxime Cauchois, Suyash Gupta, and John C Duchi. "Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction". In: *Journal of machine learning research: JMLR* 22.1 (Jan. 2021), pp. 3681–3722.

[D J03] D Vere-Jones D. J. Daley. *An Introduction to the Theory of Point Processes (Volume I: Elementary Theory and Methods)*. Springer-Verlag New York, 2003.

[Das+23] Kelian Dascher-Cousineau et al. "Using deep learning for flexible and scalable earthquake forecasting". en. In: *Geophysical research letters* 50.17 (Sept. 2023).

[Du+16] Nan Du et al. "Recurrent Marked Temporal Point Processes: Embedding Event History to Vector". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Aug. 2016, pp. 1555–1564.

[Eng+20] Joseph Enguehard et al. *Neural Temporal Point Processes For Modelling Electronic Health Records.* Machine Learning for Health (ML4H). 2020.

References III

[Far+15] Mehrdad Farajtabar et al. *COEVOLVE: A Joint Point Process Model for Information Diffusion and Network Co-evolution.* International Conference on Neural Information Processing Systems. 2015.

[FBR23] Shai Feldman, Stephen Bates, and Yaniv Romano. "Calibrated Multiple-Output Quantile Regression with Representation Learning". In: *Journal of machine learning research: JMLR* 24.24 (2023), pp. 1–48.

[Foy+20] Rina Foygel Barber et al. "The limits of distribution-free conditional predictive inference". en. In: *Information and Inference: A Journal of the IMA* 10.2 (Aug. 2020), pp. 455–482.

[GCC23] Isaac Gibbs, John J Cherian, and Emmanuel J Candès. "Conformal Prediction With Conditional Guarantees". In: (May 2023). arXiv: 2305.12616 [stat.ME].

[GLL18] Ruocheng Guo, Jundong Li, and Huan Liu. "INITIATOR: Noise-contrastive estimation for marked temporal point process". In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, July 2018.

References IV

[Gru+23] Cornelia Gruber et al. "Sources of Uncertainty in Machine Learning – A Statisticians' View". In: (May 2023). arXiv: 2305.16703 [stat.ML].

[ISS22] Rafael Izbicki, Gilson Shimizu, and Rafael B Stern. "CD-split and HPD-split: Efficient Conformal Regions in High Dimensions". In: *Journal of machine learning research: JMLR* 23.87 (2022), pp. 1–32.

[Lei+18] Jing Lei et al. "Distribution-Free Predictive Inference for Regression". In: *Journal of the American Statistical Association* 113.523 (Mar. 2018), pp. 1094–1111.

[LRW13] Jing Lei, James Robins, and Larry Wasserman. "Distribution Free Prediction Sets". en. In: *Journal of the American Statistical Association* 108.501 (2013), pp. 278–287.

[LTS22] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. "Conformal Prediction with Temporal Quantile Adjustments". In: (May 2022). arXiv: 2205.09940 [stat.ML].

[ME16] Hongyuan Mei and Jason Eisner. *The Neural Hawkes Process: A Neurally Self-Modulating Multivariate Point Process*. Neurips. 2016.

[MMR20] Huiying Mao, Ryan Martin, and Brian Reich. "Valid model-free spatial prediction". In: (June 2020). arXiv: 2006.15640 [stat.ME].

UMONS Université de Mons

References V

[MWE20] Hongyuan Mei, Tom Wan, and Jason Eisner. "Noise-Contrastive Estimation for Multivariate Point Processes". In: *Advances in Neural Information Processing Systems*. Ed. by H Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 5204–5214.

[OUA19] Takahiro Omi, Naonori Ueda, and Kazuyuki Aihara. "Fully Neural Network based Model for General Temporal Point Processes". In: *Advances in Neural Information Processing Systems 32*. Ed. by H Wallach et al. Curran Associates, Inc., 2019, pp. 2120–2129.

[RPC19] Yaniv Romano, Evan Patterson, and Emmanuel Candes. "Conformalized quantile regression". In: *Advances in neural information processing systems* (2019).

[RSC20] Yaniv Romano, Matteo Sesia, and Emmanuel J Candès. "Classification with valid and adaptive coverage". In: (Mar. 2020). arXiv: 2006.02544 [stat.ME].

[SBG20] Oleksandr Shchur, Marin Biloš, and Stephan Günnemann. "Intensity-Free Learning of Temporal Point Processes". In: *International Conference on Learning Representations (ICLR)*. 2020.

[Shc+20] Oleksandr Shchur et al. "Fast and Flexible Temporal Point Processes with Triangular Maps". In: *Advances in Neural Information Processing Systems (NeurIPS)*. June 2020.

UMONS Université de Mons

References VI

[Shc+21] Oleksandr Shchur et al. "Neural temporal point processes: A review". In: *arXiv preprint arXiv:2104.03528* (2021).

[SMS21] Kamile Stankeviciute, Ahmed M Alaa, and Mihaela van der Schaar. "Conformal Time-series Forecasting". In: *Advances in neural information processing systems* 34 (2021).

[SR21] Sesia and Romano. "Conformal prediction using conditional histograms". In: *Advances in neural information processing systems* (2021).

[Tai22] Souhaib Ben Taieb. *Learning Quantile Functions for Temporal Point Processes with Recurrent Neural Splines*. AISTATS. 2022.

[Tib+19] Ryan J Tibshirani et al. "Conformal Prediction Under Covariate Shift". In: (Apr. 2019). arXiv: 1904.06019 [stat.ME].

[UDG18] Utkarsh Upadhyay, Abir De, and Manuel Gomez-Rodriguez. *Deep Reinforcement Learning of Marked Temporal Point Processes*. Neurips. 2018.

[VGS05] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer International Publishing, 2005.

References VII

[Xia+18] S Xiao et al. “Learning conditional generative models for temporal point processes”. In: *Thirty-Second AAAI* (2018).

[Xu+17] Hongteng Xu et al. “Benefits from Superposed Hawkes Processes”. In: (2017). AISTATS.

[Yic+16] Wang Yichen et al. *Isotonic Hawkes Processes.* International Conference on Machine Learning. 2016.

[Zha+19] Qiang Zhang et al. *Self-Attentive Hawkes Processes.* Machine Learning Research. 2019.

[Zuo+20] Simiao Zuo et al. *Transformer Hawkes Process.* ICML. 2020.

Appendix

TPP model training

From $\lambda_k^*(t)$, one can compute the **joint density** of (inter-)arrival times and marks,

$$f^*(\tau, k) = \lambda_k^*(t_{j-1} + \tau)(1 - F^*(\tau)) = \lambda_k^*(t_{j-1} + \tau) \text{ exp} \left(- \sum_{k=1}^{K} \Lambda_k^*(t) \right),$$

where $F^*(\tau) = \int_0^\tau \sum_{k=1}^K f^*(s,k)ds$ and $\Lambda_k^*(t) = \int_{t_{j-1}}^t \lambda_k^*(s)ds$.

Given a sequence $\mathcal{S}$ of n events observed in $[0, T]$, and $\lambda_k^*(t; \boldsymbol{\theta})$, the parameters $\boldsymbol{\theta}$ can be estimated by MLE, i.e. by minimizing the **negative log-likelihood** (NLL):

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{S}) = \sum_{j=1}^{m} \text{log } \lambda_k^*(t_j; \boldsymbol{\theta}) + \int_0^T \sum_{k=1}^{K} \lambda_k^*(t; \boldsymbol{\theta})dt.$$

If $f^*(\tau, k; \boldsymbol{\theta}) = f^*(\tau; \boldsymbol{\theta})p^*(k|\tau; \boldsymbol{\theta})$, the NLL is

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{S}) = \sum_{j=1}^{m} [\text{log } f^*(\tau_j; \boldsymbol{\theta})) + \text{log } (p^*(k_j|\tau_j; \boldsymbol{\theta}))] + \text{log } (1 - F^*(T - t_m; \boldsymbol{\theta})).$$

The conditional LogNormMix model

The conditional LogNormMix model [BB23] computes

$$\hat{f}(\tau, k|\boldsymbol{h}) = \hat{f}(\tau|\boldsymbol{h})\hat{p}(k|\tau, \boldsymbol{h}),$$

where

$$\hat{f}(\tau|\boldsymbol{h}) = \sum_{c=1}^{C} p(c|\boldsymbol{h}) \frac{1}{\tau\sigma_c\sqrt{2\pi}} \exp\left(- \frac{(\log\ \tau - \mu_c)^2}{2\sigma_c^2} \right),$$

with

$$\begin{aligned} p(c|\boldsymbol{h}) &= \mathrm{Softmax}\big(\mathbf{W}_p \boldsymbol{h} + \boldsymbol{b}_p\big)_c, \\ \mu_c &= (\mathbf{W}_\mu \boldsymbol{h} + \boldsymbol{b}_\mu)_c, \\ \sigma_c &= \exp(\mathbf{W}_\sigma \boldsymbol{h} + \boldsymbol{b}_\sigma)_c, \end{aligned}$$

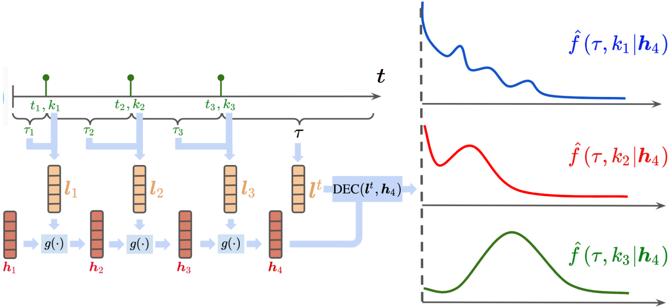
and

$$\hat{p}(k|\tau, \boldsymbol{h}) = \mathrm{Softmax}\left(\mathbf{W}_2 \mathrm{ReLU}\left(\mathbf{W}_1 [\boldsymbol{h}||\boldsymbol{l}^t] + \boldsymbol{b}_1\right) + \boldsymbol{b}_2\right)_k.$$

Neural Marked Temporal Point Processes

A neural MTPP model can be decomposed into three components:

1. **An event encoder:** For each $\boldsymbol{e}_j = (t_j, k_j) \in \mathcal{S}$, generate $\boldsymbol{l}_j \in \mathbb{R}^{d_e}$.
2. **A history encoder:** For each $\boldsymbol{e}_j$, generate $\boldsymbol{h}_j \in \mathbb{R}^{d_h}$ from past event encodings $\{\boldsymbol{l}_{j-1}, \ldots, \boldsymbol{l}_{j-p}\}$ where p is the lag.
3. **A decoder:** For a query time $t > t_j$, parametrize $\hat{\lambda}_k(t|\boldsymbol{h}_j)$ or $\hat{f}(\tau, k|\boldsymbol{h}_j)$ using $\boldsymbol{h}_j$ and $\boldsymbol{l}^t$ for all $k \in \mathbb{K}$.



Individual prediction regions

For both **inter-arrival times** and **marks**, our goal is to construct prediction regions that achieve finite-sample marginal coverage at level $1 - \alpha$.

Given a dataset $\mathcal{D} = \{(\boldsymbol{h}_i, y_i)\}_{i=1}^n$ where $y_i = \tau_i$ or $y_i = k_i$, and a new test input $\boldsymbol{h}_{n+1}$, the objectives are as follows:

1. **Inter-arrival times**: Construct a prediction region $\hat{R}_\tau(\boldsymbol{h}_{n+1}) \subseteq \mathbb{R}^+$ for τ_{n+1}, ensuring
$$\mathbb{P}(\tau_{n+1} \in \hat{R}_\tau(\boldsymbol{h}_{n+1})) \geq 1 - \alpha. \tag{1}$$
2. **Marks**: Generate a prediction set $\hat{R}_k(\boldsymbol{h}_{n+1}) \subseteq \mathbb{K}$ for k_{n+1}, guaranteeing
$$\mathbb{P}(k_{n+1} \in \hat{R}_k(\boldsymbol{h}_{n+1})) \geq 1 - \alpha. \tag{2}$$

UMONS
Université de Mons

Prediction regions for the arrival time

If $\hat{Q}_\tau(\alpha|\boldsymbol{h})$ is the α-**conditional quantile** estimated by quantile regression (QR) on $\mathcal{D}$, we can construct the following equal-tailed prediction interval:

$$\hat{R}_{\tau,\mathsf{QR}}(\boldsymbol{h}_{n+1}) = [\hat{Q}_\tau(\alpha/2|\boldsymbol{h}_{n+1}), \hat{Q}_\tau(1-\alpha/2|\boldsymbol{h}_{n+1})],$$

Conformalized Quantile Regression (CQR) [RPC19] computes an adjusted interval

$$\hat{R}_{\tau,\mathsf{CQR}}(\boldsymbol{h}_{n+1}) = [\hat{Q}_\tau(\alpha/2|\boldsymbol{h}_{n+1}) - \hat{q}, \hat{Q}_\tau(1-\alpha/2|\boldsymbol{h}_{n+1}) + \hat{q}],$$

which satisfies **marginal coverage** at level $1-\alpha$, i.e.

$$\mathbb{P}(\tau_{n+1} \in \hat{R}_{\tau,\mathsf{CQR}}(\boldsymbol{h}_{n+1})) \geq 1-\alpha.$$

Conformalized Quantile Regression

We can write

$$\begin{aligned}
&\tau_{n+1} \in \hat{R}_{\tau,\mathsf{CQR}}(\boldsymbol{h}_{n+1}) \\
\iff &\tau_{n+1} \in [\hat{Q}_\tau(\alpha/2|\boldsymbol{h}_{n+1}) - \hat{q}, \hat{Q}_\tau(1-\alpha/2|\boldsymbol{h}_{n+1}) + \hat{q}] \\
\iff &\hat{Q}_\tau(\alpha/2|\boldsymbol{h}_{n+1}) - \hat{q} \leq \tau_{n+1} \text{ and } \tau_{n+1} \leq \hat{Q}_\tau(1-\alpha/2|\boldsymbol{h}_{n+1}) + \hat{q} \\
\iff &\hat{Q}_\tau(\alpha/2|\boldsymbol{h}_{n+1}) - \tau_{n+1} \leq \hat{q} \text{ and } \tau_{n+1} - \hat{Q}_\tau(1-\alpha/2|\boldsymbol{h}_{n+1}) \leq \hat{q} \\
\iff &\underbrace{\max\left\{ \hat{Q}_\tau(\alpha/2 \,\middle|\, \boldsymbol{h}_{n+1}) - \tau_{n+1}, \tau_{n+1} - \hat{Q}_\tau(1-\alpha/2|\boldsymbol{h}_{n+1}) \right\}}_{s_{\mathsf{CQR}}(\boldsymbol{h}_{n+1},\tau_{n+1})} \leq \hat{q}
\end{aligned}$$

The finite-sample coverage guarantee is obtained using the quantile lemma:

$$\mathbb{P}(\tau_{n+1} \in \hat{R}_\tau(\boldsymbol{h}_{n+1}) = \mathbb{P}(s_{\mathsf{CQR}}(\boldsymbol{h}_{n+1}, \tau_{n+1}) \leq \hat{q}) \geq 1-\alpha$$

Conformalized Quantile Regression for Left intervals

In practice, the arrival times often show a **skewed distribution** with a significant concentration of probability mass close to 0.

By construction, CQR does not encompass these high density regions, potentially leading to large predictions intervals.

We consider **Conformalized Quantile Regression for Left intervals** (CQRL) approach that defines an **asymmetric prediction interval** for τ_{n+1}

$$\hat{R}_{\tau,\mathsf{CQRL}}(\boldsymbol{h}_{n+1}) = [0, \hat{Q}_\tau(1-\alpha|\boldsymbol{h}_{n+1}) + \hat{q}], \tag{3}$$

where the nonconformity score is

$$s_{\mathsf{CQRL}}(\boldsymbol{h}, \tau) = \tau - \hat{Q}_\tau(1-\alpha|\boldsymbol{h}). \tag{4}$$

Prediction sets for the mark

If $\hat{p}(\cdot|\boldsymbol{h})$ is the mark conditional PMF, **reguralized adaptive prediction sets** (RAPS) [Ang+22] defines the following non-conformity score:

$$s_{\text{RAPS}}(\boldsymbol{h}, k) = \sum_{k':\hat{p}(k'|\boldsymbol{h}) \geq \hat{p}(k|\boldsymbol{h})} \hat{p}(k'|\boldsymbol{h}) + u \cdot \hat{p}(k|\boldsymbol{h}) + \gamma \left(o(k) - k_{\text{reg}}\right)^{+},$$

where

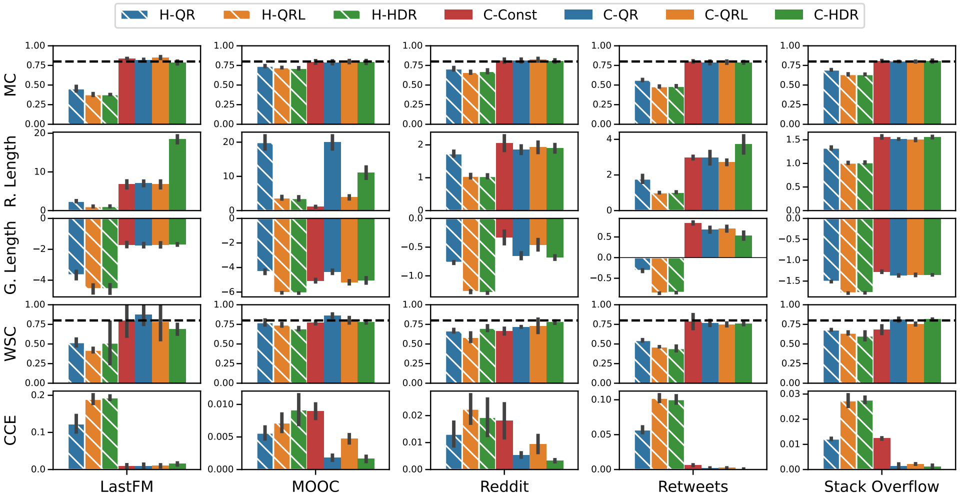
- u is a uniform random variable handling discrete jumps in the cumulative sum of $\hat{p}(k|\boldsymbol{h})$.
- $o(k) = |\{ k' \in \mathbb{K} : \hat{p}(k' \mid \boldsymbol{h}) \geq \hat{p}(k|\boldsymbol{h}) \}|$ is the ranking of the observed mark k among the probabilities in $\hat{p}(\cdot|\boldsymbol{h})$.
- $(x)^{+}$ denotes the positive part of x, and $\gamma, k_{\text{reg}} \geq 0$ are regularization parameters.

We also consider the unreguralized version of the previous method, called **adaptive prediction sets** (APS) [RSC20], i.e. $\gamma = 0$.

We construct the following prediction set for k_{n+1}:

$$\hat{R}_k(\boldsymbol{h}_{n+1}) = \{ k' \in \mathbb{K} : s_{\text{(R)APS}}(\boldsymbol{h}_{n+1}, k') \leq \hat{q} \},$$

Results for the arrival time prediction regions



Results for the mark prediction sets