The Neural Hawkes Process

A Neurally Self-Modulating Multivariate Point Process

Hongyuan Mei and Jason Eisner
Center for Language and Speech Processing, Department of Computer Science, Johns Hopkins University

Overview

Events happen at random times $0 < t_1 < t_2 \dots$

At time t_i, there occurs an event of type $k_i \in \{1, 2, \dots, K\}$

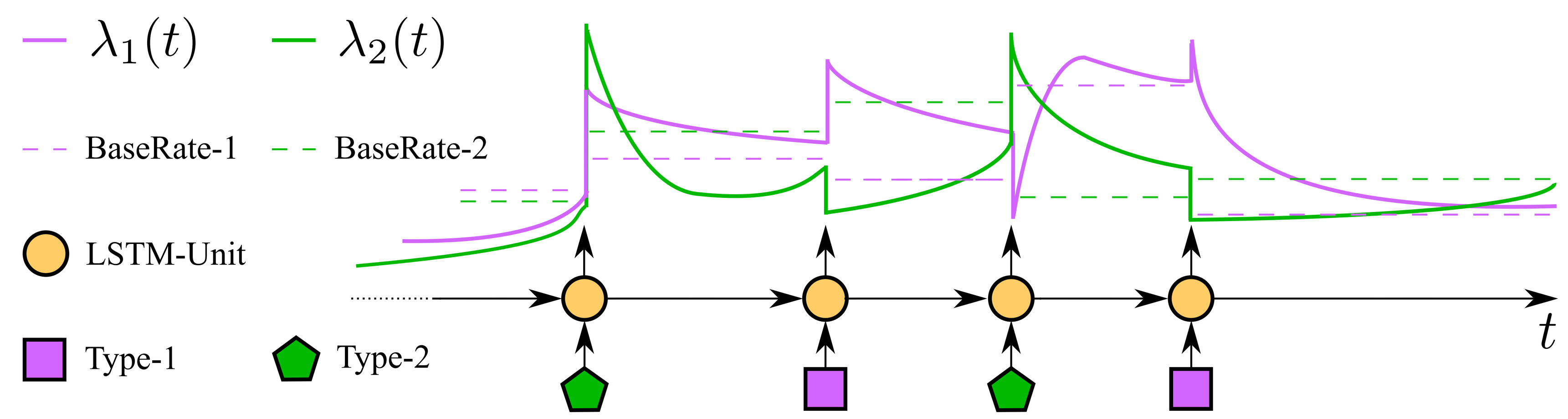
Given past events, what might happen next, and when?

- Generative model $P((k_i, t_i) \mid (k_1, t_1), \dots, (k_{i-1}, t_{i-1}))$
- Medical: patient's visits, tests and diagnoses
- Online shopping: purchasing and feedback
- Social media: posts, shares, comments
- Other: quantified self, news, dialogue, music, etc

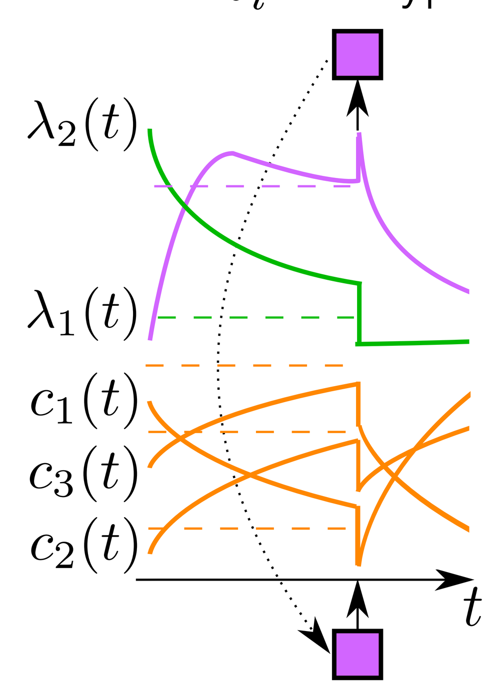
Traditional model is a Hawkes process [1]

- Each event type has an intensity $\lambda_k(t)$
- Each event token occurs with probability $\lambda_k(t)dt$
- Past events temporarily excite future events
- $\lambda_k(t) = \mu_k + \sum_{h:t_h<t} \alpha_{k_h,k} \exp(-\delta_{k_h,k}(t - t_h))$

Model



An event stream from a neural Hawkes process
Time t_i and type k_i depend on details of past history and on each other



Neural Hawkes process: $\lambda_k(t) = f_k(\mathbf{w}_k^\top \mathbf{h}(t))$

- Continuous-time LSTM
- Hidden state $\mathbf{h}(t) = \mathbf{o}_i \odot (2\sigma(2\mathbf{c}(t)) - 1)$
- Cell memory $\mathbf{c}(t) = \bar{\mathbf{c}}_{i+1} + \boldsymbol{\nu}_{i+1}(t)$
- $\boldsymbol{\nu}_{i+1}(t) = (\mathbf{c}_{i+1} - \bar{\mathbf{c}}_{i+1}) \exp(-\boldsymbol{\delta}_{i+1}(t - t_i))$
- Extra gates to compute $\bar{\mathbf{c}}_{i+1}$ and $\boldsymbol{\delta}_{i+1}$

cell $\mathbf{c}(t) \rightarrow$ hidden $\mathbf{h}(t) \rightarrow$ intensity $\boldsymbol{\lambda}(t) \rightarrow$ event? $\rightarrow$ updated $\mathbf{c}(t + dt)$

Algorithms

Train the model by maximizing log-likelihood

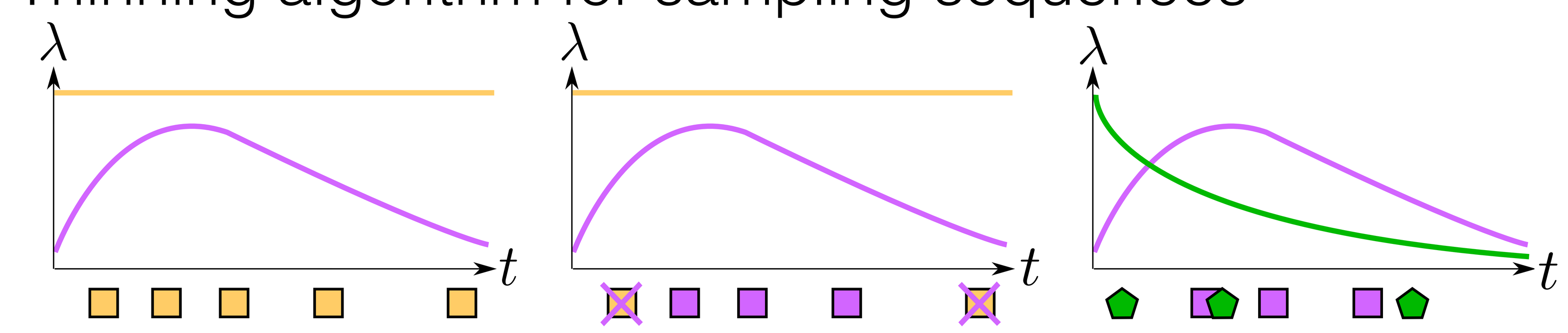
$$\ell = \sum_{i:t_i \le T} \log \lambda_{k_i}(t_i) - \int_{t=0}^{T} \lambda(t)dt$$

- Total intensity $\lambda(t) = \sum_{k=1}^{K} \lambda_k(t)$
- Integral estimation by Monte Carlo simulation

Minimum Bayes Risk prediction

- Density for t_i is $p_i(t) = \lambda(t) \exp(-\int_{s=t_{i-1}}^{t} \lambda(s)ds)$
- Time prediction $\hat{t}_i = \int_{t=t_{i-1}}^{\infty} t p_i(t) dt$
- Type prediction $\hat{k}_i = \arg\max_k \int_{t=t_{i-1}}^{\infty} p_i(t)\lambda_k(t)/\lambda(t)dt$

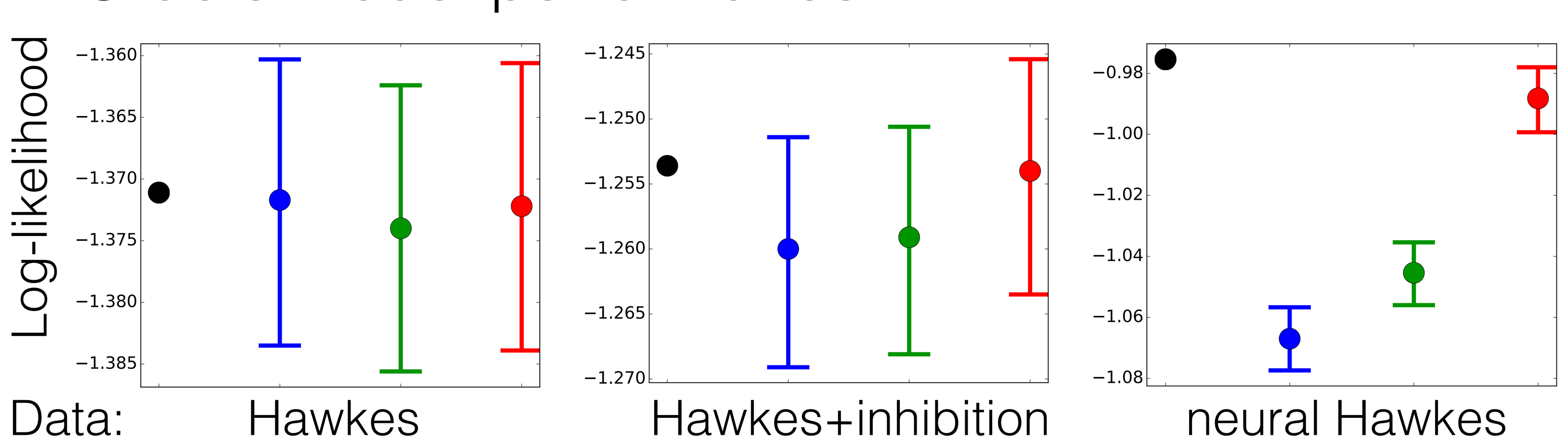
Thinning algorithm for sampling sequences

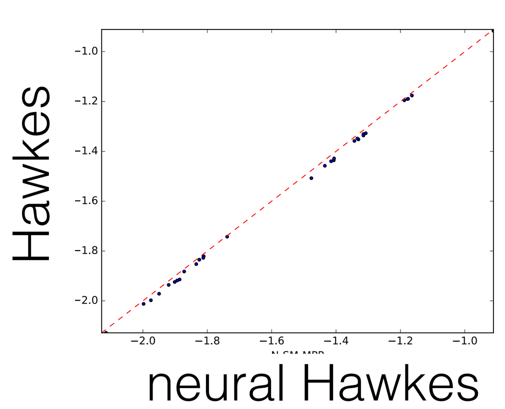


Experiments (many more in paper)

Experiments on artificial datasets

- Models try to fit data generated by each other
- Oracle model performance — ●



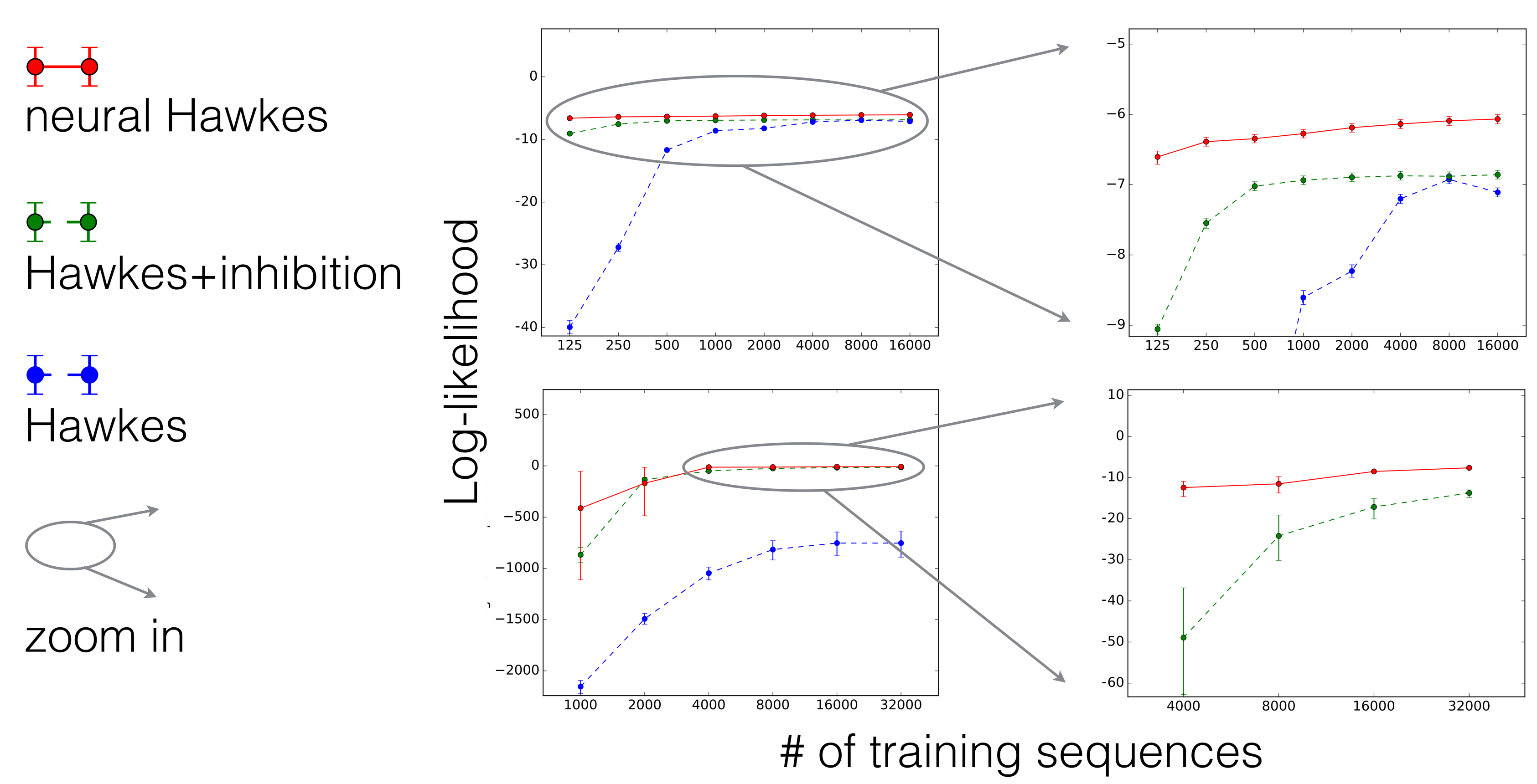


Missing data experiments

★ Censor all events of some types
★ neural Hawkes > Hawkes process
★ Consistent over all combos

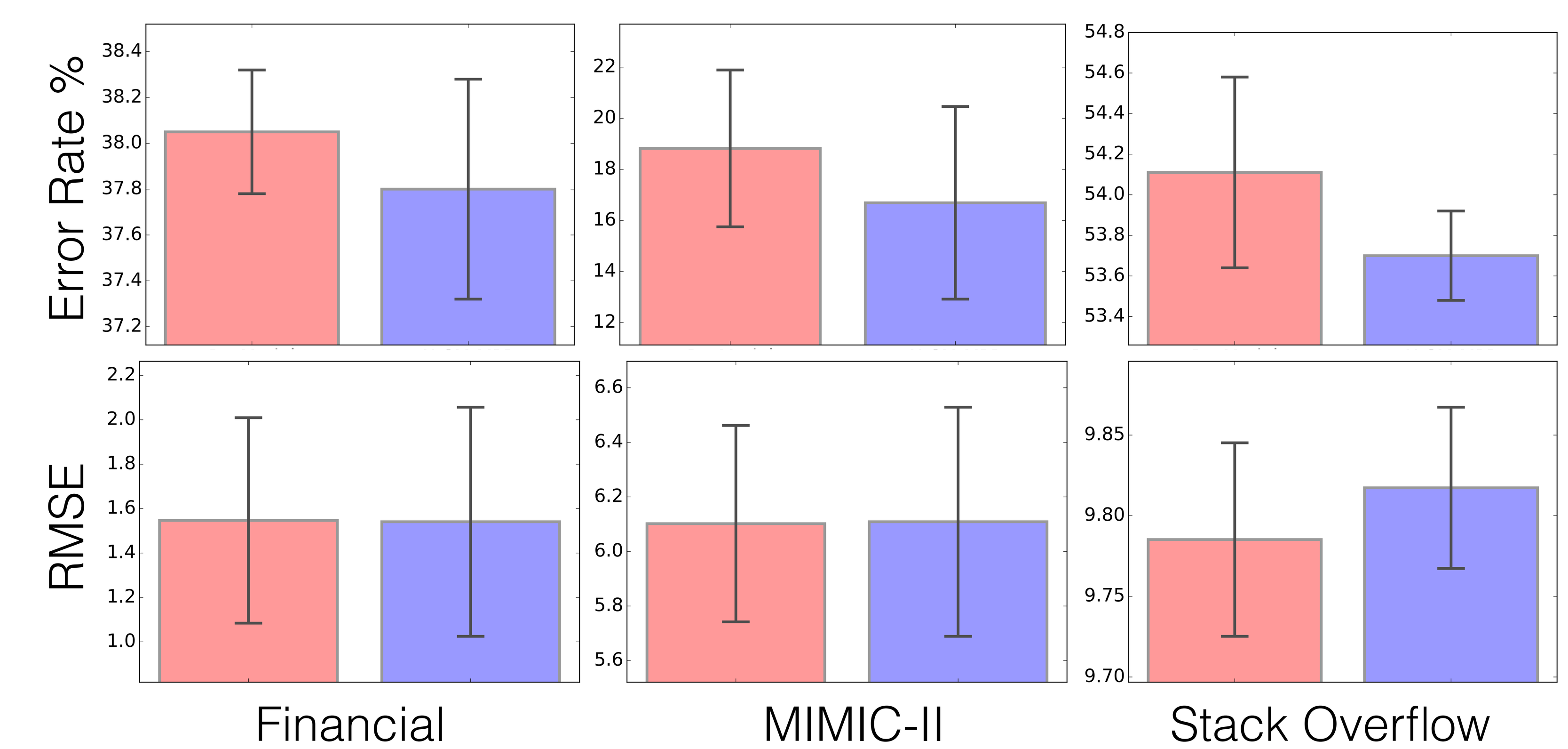
Experiments on real-world social media datasets

- Retweet (top): long sequences with $K = 3$
- MemeTrack (bottom): short sequences with $K = 5000$



Neural Hawkes process (blue) vs. similar work (red) [2]

- Prediction error for type (upper) and time (lower)



Neural Hawkes is winner (4/5, 5/5, and 5/5) on type prediction
No clear winner on time prediction

[1] Hawkes, Alan G. Spectra of some self-exciting and mutually exciting point processes. 1971. [2] Du, Nan, et. al. Recurrent marked temporal point processes: Embedding event history to vector. 2016.