Meta Learning with Relational Information for Short Sequences

Yujia Xie, Haoming Jiang, Feng Liu, Tuo Zhao, Hongyuan Zha

Georgia Institute of Technology

Introduction

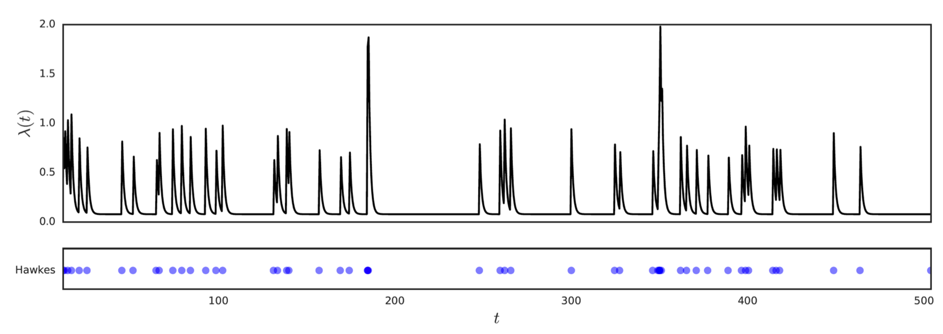
- **Event sequences** – important and informative
  - The timestamps of tweets of a twitter user
  - The job hopping history of a person
- **Short sequences** – widely appeared
  - The event sequences are short in nature
    - Job hopping histories
  - The observation window is narrow
    - The criminal incidents after a regulation is published
- **Short sequences** – hard to infer
  - MLE for each sequence
    - Their lengths are insufficient for reliable inference.
  - Treat the collection of short sequences as i.i.d.
    - Highly biased against certain individuals.

Background - Hawkes Process

- A Hawkes processes is a doubly stochastic temporal point process $\mathcal{H}(\theta)$ with conditional intensity function $\lambda = \lambda(t; \theta, \boldsymbol{\tau})$ defined as

$$\lambda(t; \theta, \boldsymbol{\tau}) = \mu + \sum_{\tau^{(m)} < t} \delta \omega e^{-\omega(t-\tau^{(m)})},$$

where $\theta = \{\mu, \delta, \omega\}$,
μ is the base intensity,
$\boldsymbol{\tau} = \{\tau^{(1)}, \tau^{(2)}, \cdots, \tau^{(M)}\}$ are the timestamps of the events occurring in a time interval $[0, t_{\text{end}}]$.



- Self-exciting
  - The past events always increase the chance of arrivals of new events
- Widely used in many areas
  - behavior analysis
  - financial analysis
  - social network analysis

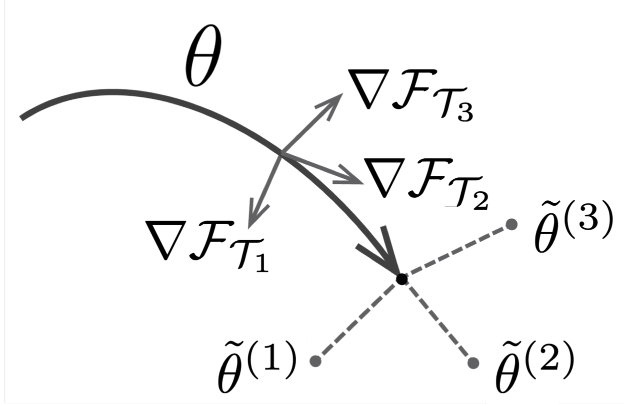
Background - Meta Learning

- **Meta Learning**
  - Given a set of tasks $\Gamma = \{\mathcal{T}_1, \mathcal{T}_2, \cdots, \mathcal{T}_N\}$
  - Each task contains a very small amount of data
- **Model-Agnostic Meta Learning** (MAML)
  - Train a common model for all tasks,

$$\min_{\theta} \sum_{\mathcal{T}_i \in \Gamma} \mathcal{F}_{\mathcal{T}_i}(\theta - \eta \nabla_\theta \mathcal{F}_{\mathcal{T}_i}(\theta))$$

where $\mathcal{F}_{\mathcal{T}_i}$ is the loss function of task $\mathcal{T}_i$,
θ is the parameter of the common model,
η is the step size.

  - Find the common model that is expected to produce maximally effective behavior on that task after performing update $\theta - \eta \mathcal{D}(\mathcal{F}_{\mathcal{T}_i}, \theta)$.



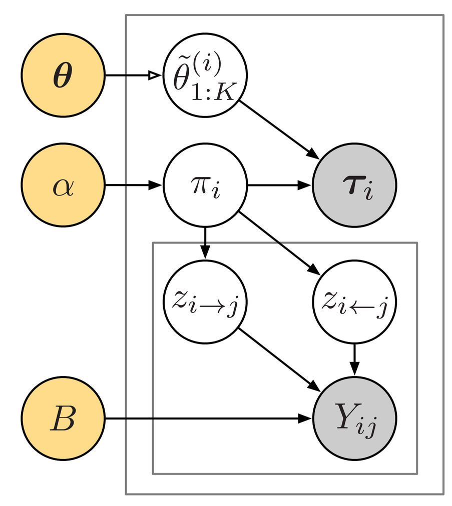
- Variants to alleviate the computational burden:
  - First Order MAML (FOMAML)
  - Reptile

HARMLESS – HAwkes Relational Meta LEarning for Short Sequence

- Given: a collection of sequences $\boldsymbol{T} = \{\boldsymbol{\tau}_1, \boldsymbol{\tau}_2 \cdots, \boldsymbol{\tau}_N\}$
  extra relational information, described as a graph with adjacency matrix as $\boldsymbol{Y}$.
- Key idea: identify and incorporate the relational information between tasks
  - Social graphs often exhibit community patterns
  - Each subject may belong to multiple communities and thus have multiple identities
  - →Assign each subject i a sum-to-one **identity proportion vector** $\pi_i \in [0,1]^K$, where K is the number of communities
- **Identity determines the user's event sequence** – Mixture of Hawkes process model
  For the k-th identity of subject i, we adopt Hawkes process $\mathcal{H}(\widetilde{\theta}_k^{(i)})$ to model the timestamps of the associated events. The likelihood for the i-th sequence $\boldsymbol{\tau}_i$ is

$$p(\boldsymbol{\tau}_i) = \sum_{k=1}^{K} \pi_{i,k} \mathcal{L}_i(\widetilde{\theta}_k^{(i)}). \quad (1)$$

- **Identity determines the user's connection to other users** – Mixed Membership stochastic Blockmodel (MMB)
  - $z_{i \to j}$: the identity of subject i when subject i *approaches* subject j
  - $z_{i \leftarrow j}$: the identity of subject j when j *is approached* by i
  - $z_{i \to j}^T \boldsymbol{B} z_{i \leftarrow j}$: the probability of whether subject i and j have a connection



- **Generative process**:
  - For each node i,
    - Draw a K dimensional identity proportion vector $\pi_i \sim$ Dirichlet(α).
    - Sample the i-th sequence $\boldsymbol{\tau}_i$ from the mixture of Hawkes processes in (1).
  - For each pair of nodes i and j,
    - Draw identity indicator for the initiator $z_{i \to j} \sim$ Categorical(π_i)
    - Draw identity indicator for the receiver $z_{i \leftarrow j} \sim$ Categorical(π_j)
    - Sample whether there is an edge between i and j, $Y_{ij} \sim$ Bernoulli($z_{i\to j}^T \boldsymbol{B} z_{i\leftarrow j}$).

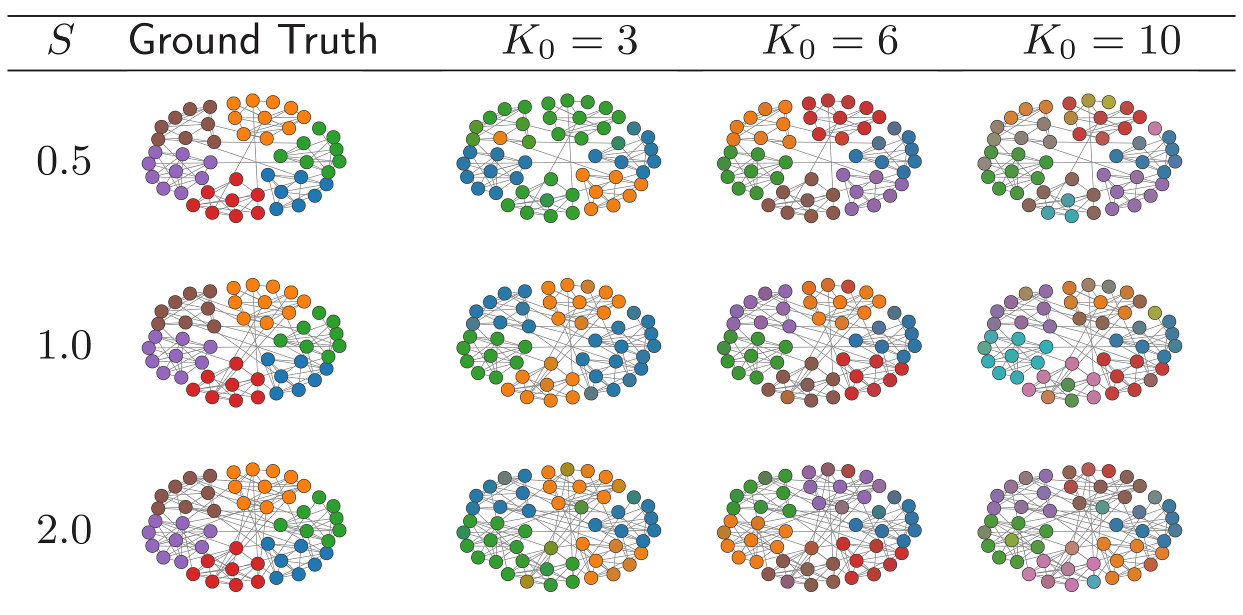
  Here, the observed variables are $\boldsymbol{\tau}_i$ and Y_{ij}. The parameters are α, $\widetilde{\theta}_k^{(i)}$, and $\boldsymbol{B}$. The latent variables are π_i, z_i, $z_{i\to j}$ and $z_{i \leftarrow j}$.

- **Meta inference for $\boldsymbol{\theta}$ and $\widetilde{\theta}$.** Instead of specifying that $\widetilde{\theta}_k^{(i)}$ is sampled from a prior distribution, we adapt the k-th common model $\mathcal{H}(\theta_k)$ to sequence i using MAML-type updates, $\widetilde{\theta}_k^{(i)} = \theta_k - \eta \mathcal{D}(\log \mathcal{L}_i, \theta_k)$.
  The gradient descent step on the log-likelihood of $\boldsymbol{\theta}$ can then be written as

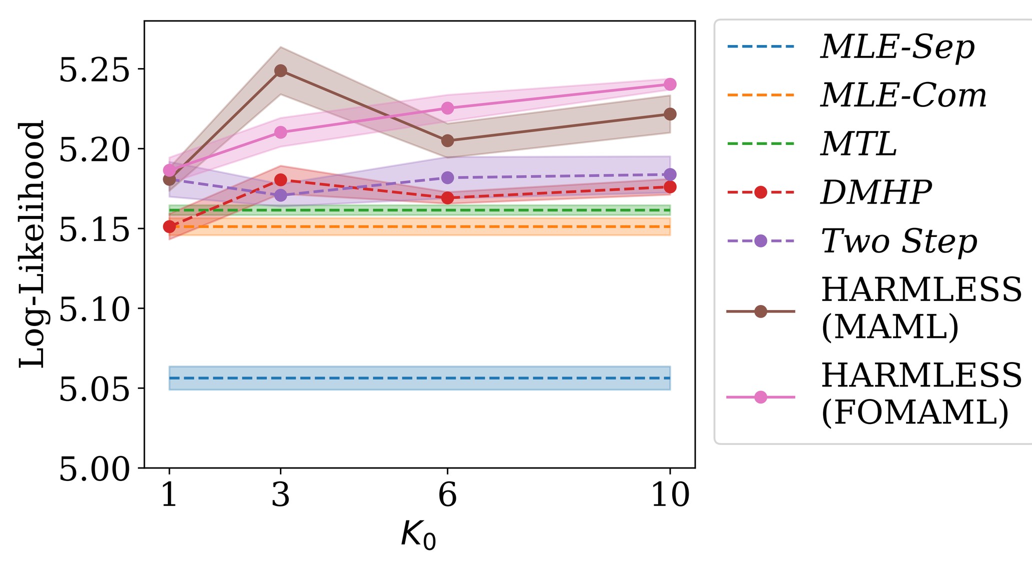
$$\theta_k \leftarrow \theta_k + \eta_{\boldsymbol{\theta}} \nabla_{\theta_k} \left(\sum_{i=1}^{N} \gamma_{i,k} \log \mathcal{L}_i(\theta_k - \eta \mathcal{D}(\log \mathcal{L}_i, \theta_k)) \right).$$

Experiment – Synthetic Graphs

- Data generation: 50 Nodes, 6 Communities, S: Sparsity of the Graph, K_0: Number of Specified Communities
- Experiment: Community Assignment



- Experiment: Likelihood



Experiment – Real Graphs

Dataset	911-Calls	LinkedIn	MathOverflow	StackOverflow
MLE-Sep	4.0030 ± 0.3763	0.8419 ± 0.0251	0.5043 ± 0.0657	0.2862 ± 0.0177
MLE-Com	4.5111 ± 0.3192	0.8768 ± 0.0028	1.7805 ± 0.0345	1.5594 ± 0.0134
DMHP	4.4812 ± 0.3434	0.8348 ± 0.0030	1.5394 ± 0.0347	$N \backslash A$
MTL	4.4621 ± 0.3173	0.9270 ± 0.0027	1.7225 ± 0.0336	1.4910 ± 0.0089
HARMLESS (MAML)	4.5208 ± 0.3256	$\mathbf{1.4070} \pm 0.0105$	1.8563 ± 0.0345	1.3886 ± 0.0082
HARMLESS (FOMAML)	$\mathbf{4.6362} \pm 0.3241$	1.0129 ± 0.004	1.8344 ± 0.0348	1.5988 ± 0.0083
HARMLESS (Reptile)	4.4929 ± 0.3503	0.9540 ± 0.0082	$\mathbf{1.8663} \pm 0.0342$	$\mathbf{1.6017} \pm 0.0097$

Experiment – Ablation Study

Method	Log-Likelihood
HARMLESS (MAML)	$\mathbf{1.4070} \pm 0.0105$
HARMLESS (FOMAML)	1.0129 ± 0.0042
HARMLESS (Reptile)	0.9540 ± 0.0082
Remove inner heterogeneity ($K = 3$)	0.9405 ± 0.0032
Remove inner heterogeneity ($K = 5$)	0.9392 ± 0.0032
Remove grouping (MAML)	0.9432 ± 0.0031
Remove grouping (FOMAML)	0.9376 ± 0.0031
Remove grouping (Reptile)	0.9455 ± 0.0041
Remove graph (MAML)	0.9507 ± 0.0032
Remove graph (FOMAML)	0.9446 ± 0.0032
Remove graph (Reptile)	0.9489 ± 0.0072

Reference

- HAWKES, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*fg, 58 83?90.
- FINN, C., ABBEEL, P. and LEVINE, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org.
- AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. Journal of machine learning research, 9 1981?2014.