Upper confidence bound multi-armed bandits for partially observed Hawkes processes

1st Wen-Hao Chiang
Amazon Web Services
cwenhao@amazon.com

2nd George Mohler
Boston College
mohlerg@bc.edu

***Abstract*—We consider the problem of estimating and ranking a set of self-excited point processes when an action must be taken to observe the events of each process. This situation arises in a number of real-world applications, for example, when crime goes unreported in some regions, or COVID-19 cases are undetected due to a lack of testing resources. Often times, such self-excited events may bear implicit causality. Therefore, we start with Hawkes Processes to model how one event triggers the other. In the scenario of undersampling, we propose Hawkes Process Multi-armed Bandits for learning such point processes to quickly learn the riskiest point processes, while carefully balancing exploitation of known (observed) point processes and exploration of unknown processes. By considering the cumulative number of events of each process as a reward, we derive an upper confidence bound on the counting process to inform actions in the form of which processes to observe in upcoming MAB rounds, based on the history of the partially observed point processes. We then derive a regret bound that scales logarithmically with the number of rounds of observation. We test our model on simulated datasets, crime report data in Vancouver and Los Angeles, and earthquake event data from Alaska, California, and worldwide. Our model outperforms several state-of-the-art MAB algorithms that can be adapted to non-stationary point process estimation across the datasets and performance metrics.**

***Index Terms*—Hawkes processes, Multi-armed bandit, Upper confidence bound, Reinforcement learning**

I. Introduction

A Multi-armed Bandit problem (MAB) is a sequential decision problem in which a set of resources is allocated between choices by a set of actions. After each action, rewards are observed and collected. Our goal is to maximize the total rewards after a sequence of allocations. Given the nature of maximizing payoffs with limited actions, MAB algorithms are designed to address the fundamental trade-off between exploitation and exploration. An enormous body of literature exists on MABs in a variety of settings [1]. These include, but are not limited to, traditional stochastic bandits, non-stationary bandits [2], adversarial bandits [3], contextual bandits [4], and bandits with knapsacks [5]. MABs are applied to a variety of problems, such as clinical trials [6], adaptive routing [7] and financial portfolio design [8].

In this manuscript, we extend the Multi-armed Bandit problem to a new setting, namely that of ranking and selecting a set of point processes when an action must be taken to observe the events of each process. Such a scenario could arise in a number of applications, for example during a disaster event when emergencies may occur in space and time, but are unobserved by first responders due to a power outage. In this setting, there is a tradeoff between exploring regions where the power is out and exploiting information on emergency calls in regions where events have been observed.

We address this problem by learning Hawkes point processes through a Multi-armed Bandit approach. Hawkes processes are non-stationary models for the intensity of events, and are applied to a wide range of phenomena including social media cascades [9], crime patterns [10], earthquake aftershocks [11], financial contagion [12]. disease transmission [13] and disasters [14]. Neural network based point processes can also be used to model high-dimensional event sequence data on graphs [15], hypergraphs [16], knowledge graphs [17], and text [18] [19]. Given that a such model is motivated through modeling self-excited events, it is often times used for causality discovery [20], [21] and it is a good fit for our inference problem. However, existing Hawkes process estimation algorithms require events to be observed, whereas in this work we require an action to be taken in order for events to be observed in a region of space. The goal of Hawkes Process Multi-armed Bandits is to quickly learn the most risky point processes out of a set, while carefully balancing exploitation of known (observed) point processes and exploration of unknown processes. By considering the cumulative number of events of each process as a reward, we derive an upper confidence bound on the counting process to inform actions in the form of which processes to observe in upcoming MAB rounds, based on the history of the partially observed point processes. We then derive a regret bound that scales logarithmically with the number of rounds of observation. A schematic of our Hawkes MAB algorithm, denoted as `HpUCB`, is shown in Figure 1.

We note that there is related work on non-stationary bandits. One typical problem setting is to assume that the reward distribution is piece-wise stationary. The rewards abruptly change at unknown time points, but stay stationary for a short period [22]. One strategy is to discount the previous reward observations. In the work of [23], the authors approach this problem by detecting the changing point and resetting the reward estimation accordingly. In another scenario where the environment is slowly-varying, Cheung et al.[24] build an upper confidence bound for the reward distribution within a sliding window. Zhao et al. [2] eliminate the distortion from outdated observations and restart the estimation periodically to address this problem. However, most non-stationary multi-

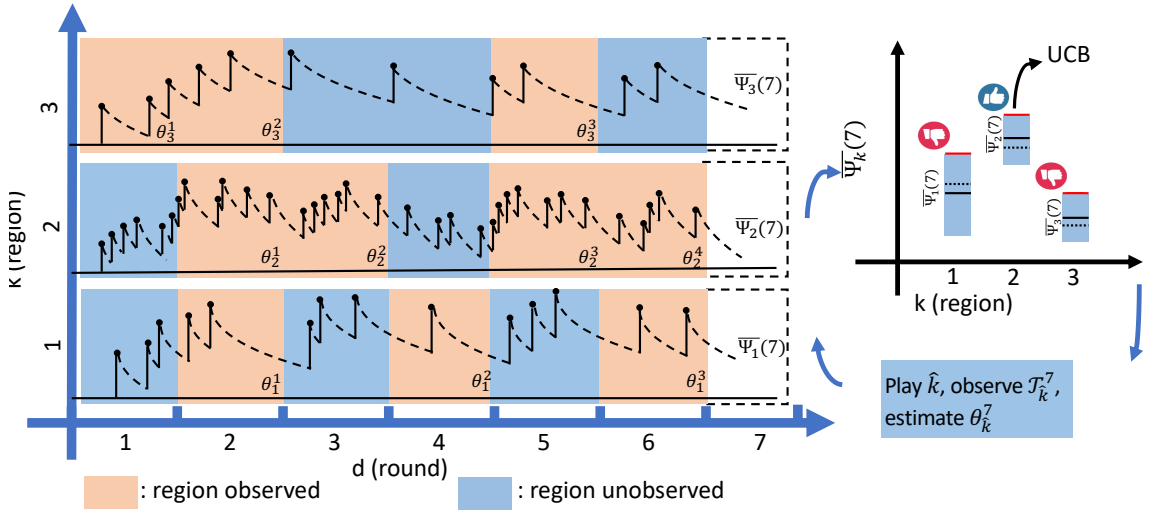

Fig. 1. Schematic diagram of HpUCB where Hawkes process intensities are estimated at each round, in both regions where events were observed through bandit actions and in unobserved regions. An upper confidence bound is then constructed for the expected number of events for each process in the next round, informing the next action.

armed bandit algorithms assume that the reward distribution is bounded. Furthermore, to our knowledge no existing algorithms have been developed to learn non-stationary contagion event processes where the reward is the cumulative number of events observed. We show in both synthetic and real-world experiments that Hawkes MABs achieve state-of-the-art results for learning unobserved point processes compared to existing non-stationary MAB algorithms.

The manuscript is organized as follows. In Section II, we introduce the Hawkes MAB methodology and we analyze its regret bound in Section III. We then present our experiments on synthetic data and several real-world datasets, including crime report data in Vancouver and Los Angeles, and earthquake event data from Alaska, California, and worldwide in Section IV.

II. Methodology

In this section, we introduce the Hawkes process based UCB algorithm, denoted as HpUCB. First, we define the problem setting and the rewards. We then derive the expected number of events given the Hawkes process parameters within a specific time period and analyze the regret bound of the algorithm.

A. Problem Formulation

Given a spatial domain, we first divide it into a set of subregions, where self-exciting events take place. We then divide the time span T into several time intervals of length δt. At the round d, the MAB algorithm recommends a shortlist of subregions to visit and the events that happen in the visited regions are observed. We then consider the number of observed events as the reward. We assume that the events in each subregion are generated through an independent Hawkes process. Each Hawkes process k is defined through its conditional intensity:

$$\lambda_k(t) = \nu_k + \sum_{\substack{t>t_i \\ t_i \in \mathcal{T}_k}} \alpha_k e^{-\beta_k (t-t_i)}, \tag{1}$$

where $\mathcal{T}_k$ is the set of timestamps for events in arm k, ν_k represents the background (Poisson) rate of spontaneous events, $\alpha_k = r_k \beta_k$, where r_k is the expected number of offspring generated by one single event and β_k determines the time scale between parent and offspring events in the branching process. We utilize the parametric Hawkes processes because the model needs to be learned from only a few events, given that data is unobserved until actions are taken (preventing deep Hawkes processes from being utilized). In our implementation of HpUCB, the Hawkes process parameters are inferred through the maximum likelihood estimation a quasi-Newton method [1].

We denote these Hawkes process parameters as $\theta_k = (\nu_k, \alpha_k, \beta_k)$. The reward of action k is then defined as the number of events in round d, that is $|\mathcal{T}_k^d|$. Each round lasts for δt. Our goal is to maximize the total reward over rounds, i.e., the total number of observed events in the visited subregions after a total of D rounds.

B. Theoretical Expected Number of Events

During each visit of the MAB Hawkes process, we observe a set of self-exciting events. Based on the observed events, we can estimate a set of Hawkes process parameters. For example, in the round d, we visit the subregion k and observe events $\mathcal{T}_k^d$. We then estimate the parameters based on $\mathcal{T}_k^d$, denoted as $\theta_k^d = (\nu_k^d, \alpha_k^d, \beta_k^d)$. Given the estimated parameters, we can infer the expected number of events for the future and build an upper confidence bound. The key ingredient of such a strategy is the derivation of the theoretical closed-form expression for the expected number of events for Hawkes processes.

In the work of [25], an elementary approach is introduced to obtain moments of Hawkes processes and the intensity of a number of marked Hawkes processes. Following the framework in [25], we can obtain the expected number of

[1] https://github.com/omitakahiro/Hawkes

events for a Hawkes process accumulated up to any time t given the parameter θ as the following:

$$E[N(t)|\theta] = \begin{cases} \frac{-\beta\nu}{\alpha-\beta}t + \frac{\alpha\nu}{(\alpha-\beta)^2}\Big[\exp\big\{(\alpha-\beta)t\big\} - 1\Big] & \alpha \neq \beta \\ \nu t + \frac{1}{2}\nu\alpha t^2 & \alpha = \beta \end{cases}. \quad (2)$$

For simplicity, we use θ to represent the parameters estimated at round d from region k (i.e., θ_k^d). In the MAB setting, we are more interested in the expected number of events during the time period/rounds when we visit the sub-region instead of an accumulation up to any given time. Given the estimated parameter θ, the expected number of events during round d, denoted as $\Psi(d)|\theta$, can be calculated as:

$$\Psi(d)|\theta = E\big[N\big(\delta t \times d\big) - N\big(\delta t \times (d-1)\big)|\theta\big] = E\big[N\big(\delta t \times d\big)\big] - E\big[N\big(\delta t \times (d-1)\big)|\theta\big] = \begin{cases} \frac{-\beta\nu}{\alpha-\beta}\delta t + \frac{\alpha\nu}{(\alpha-\beta)^2}\exp\big\{(\alpha-\beta)\delta t \times d\big\}\big[& \alpha \neq \beta \\ \qquad \exp\big\{(\alpha-\beta)\delta t\big\} - 1\big] & \\ \nu\delta + \frac{1}{2}\nu\alpha(\delta t)^2(2d+1) & \alpha = \beta \end{cases}. \quad (3)$$

C. Construction of Upper Confidence Bound on Future Expectation Number of Events

Suppose that after several rounds we have visited region k non-consecutively with n_k times and estimated n_k sets of parameters, i.e., $\Theta_k = \{\theta_k^j | j = 1, 2, \cdots n_k\}$. Using Equation 3, for any given round d in the future, we can calculate the expected number of events for each θ_k^j in Θ_k. We denote the estimate of the expected number of events as $\Psi_k^j(d) = \Psi(d)|\theta_k^j$, where $j = 1, \cdots, n_k$. We can then construct an upper confidence bound based on the estimation.

Given that the Poisson distribution can be approximated with a Normal distribution when λ is large, we approximate $\Psi_k^j(d)$ with a normal distribution assuming that the number of events in round d are sufficiently large.

Many UCB algorithms require the reward distribution to be bounded for the guaranteed regret bound [26]. However, the UCB-normal algorithm [27], only requires the reward distribution to be a normal distribution. Since the number of events in the Hawkes process cannot be bounded, UCB-normal is a good choice for us to analyze the regret bound. Under the assumption of $\Psi_k^j(d)$ approximately following a normal distribution, we can build the upper confidence bound for $\Psi_k^j(d)$ at round d following a similar form to UCB-normal, $\overline{\Psi_k}(d) + \sqrt{16\eta\frac{\sum_{j=1}^{n_k}\Big[\Psi_k^j(d)\Big]^2 - n_k\Big[\overline{\Psi_k}(d)\Big]^2 + \epsilon}{n_k - 1}\frac{\ln(d-1)}{n_k}}$ where $\overline{\Psi_k}(d) = \frac{1}{n_k}\sum_{j=1}^{n_k}\Psi_k^j(d)$. Hyper-parameter η is used to adjust the trade-off between the exploration and exploitation. A small number of ϵ can prevent the case when the first several rounds have no events and the model stops exploring. The overall Hawkes process MAB is given in Algorithm 1.

Algorithm 1 Hawkes Processes Multi-armed Bandit Algorithm (HpUCB)

Input K, η, ϵ
Output Hawkes process parameters $\theta_{\hat{k}}^j$

for $d = 1, 2, \cdots, D$ **do**
  If there exists an arm that is played less then $\lceil 8\ln(d)\rceil$, play the arm.
  Otherwise play $\hat{k} = \arg\max_k \overline{\Psi_k}(d) + \sqrt{\eta * 16\frac{\sum_{j=1}^{n_k}\Big[\Psi_k^j(d)\Big]^2 - n_k\Big[\overline{\Psi_k}(d)\Big]^2 + \epsilon}{n_k - 1}\frac{\ln(d-1)}{n_k}}$.
  $n_{\hat{k}} = n_{\hat{k}} + 1$
  Estimate Hawkes process parameter $\theta_{\hat{k}}^j$ based on observed events $\mathcal{T}_{\hat{k}}^d$.
  Update $\overline{\Psi_{\hat{k}}}(d) = \frac{1}{n_{\hat{k}}}\sum_{j=1}^{n_{\hat{k}}}\Psi_{\hat{k}}^j(d)$. $n_{\hat{k}} = n_{\hat{k}} + 1$
  Estimate Hawkes process parameter $\theta_{\hat{k}}^j$ based on observed events$\mathcal{T}_{\hat{k}}^d$.
  Update $\overline{\Psi_{\hat{k}}}(d) = \frac{1}{n_{\hat{k}}}\sum_{j=1}^{n_{\hat{k}}}\Psi_{\hat{k}}^j(d)$.
end for

III. Regret Analysis of HpUCB

In this section, we derive a regret bound for the HpUCB model. To make the regret analysis tractable, we consider the case where η is set to 1 and ϵ is set to 0. Following the work of UCB-Normal in [27], we analyze the regret analysis of HpUCB. One key difference is that the reward distribution is assumed to be constant in UCB-Normal. However, in a real-world application, the reward distribution is usually non-stationary. Our main contribution is to show the regret analysis of HpUCB when the rewards are generated through Hawkes processes stochastically.

Firstly, the regret bound of HpUCB is based on two conjectures which are verified numerically.

Conjecture 1: Given a Student random variable X with s degrees of freedom, for all $0 \leq a \leq \sqrt{2(s+1)}$,

$$P(X \geq a) \leq \mathrm{e}^{-\frac{a^2}{4}}. \quad (4)$$

Conjecture 2: Given a random variable X following a Chi-squared distribution with s degrees of freedom,

$$P(X \geq 4s) \leq \mathrm{e}^{-\frac{(s+1)}{2}}. \quad (5)$$

Such conjectures were proposed in [27], where the authors stated that it can be only verified numerically:

> Here, we show that an index-based policy called UCB1-NORMAL ... However, our proof is based on certain bounds on the tails of the χ^2 and the Student distribution that we could only verify numerically.

We carefully examined the verification and included the results in Appendix A.

We then define $Q_k(d) = \sum_{j=1}^{n_k}\Big[\Psi_k^j(d)\Big]^2$ and $U_k(d) = \sqrt{16\frac{Q_k(d) - n_k\Big[\overline{\Psi_k}(d)\Big]^2}{n_k - 1}\frac{\ln(d-1)}{n_k}}$. The expected regret can be calculated as the following:

$$E\big[R(D)\big] = \sum_{k=1}^{K}\Big\{\sum_{d=1}^{D}P(A_d = k)\Delta_k^d\Big\}, \quad (6)$$

where A_d is the random variable that represents the action at round d; Δ_k^d is the difference between the optimal action and action k, that is, $\Delta_k^d = q^d(k^*) - q^d(k)$. Note that $q^d(k)$ is a function that returns the expected reward at region k at round d and k^* represents the optimal action. To bound the regret, we take the maximal value of Δ_k^d across all the rounds and denote it as $\tilde{\Delta}_k = \max_d \Delta_k^d$, where $0 \leq d \leq D$. The regret can then be bounded as follows:

$$E\big[R(D)\big] \leq \sum_{k=1}^{K} E\big[T_k(D)\big]\tilde{\Delta}_k, \tag{7}$$

where $T_k(D)$ is the number of times arm a_k is played in the first D trials. Our goal is now to bound $E\big[T_k(D)\big]$. We start with the initialization process in which we play all the arms one time and the bound can be further loosed by assuming that the arm a_k has been played l times:

$$\begin{aligned} T_k(D) \leq 1 + \sum_{d=K+1}^{D} \mathbf{1}(A_d = k) \leq \\ l + \sum_{d=K+1}^{D} \mathbf{1}\{A_d = k,\ T_k(d-1) \geq l\}. \end{aligned} \tag{8}$$

After the initialization, we select the arm with a higher upper confidence bound. In specific, the event $A_d = k$ means that at the previous time step, the upper confidence bound of action k was greater than that of the optimal one k^*:

$$\begin{aligned} T_k(D) \leq l + \sum_{d=K+1}^{D} \mathbf{1}\Big\{ \overline{\Psi_{k^*}}\big(d-1\big) + U_{k^*}(d-1) \leq \\ \overline{\Psi_k}\big(d-1\big) + U_k(d-1), T_k(d-1) \geq l \Big\}. \end{aligned} \tag{9}$$

where n_{k^*} in $U_{k^*}(d-1)$ is $T_{k^*}(d-1)$ and n_k in $U_k(d-1)$ is $T_k(d-1)$. Following the Equation 9, the regret bound can be loosed by the fact that the minimum of the upper bounds of the optimal arm for $0 < s < d$ must be less than the maximum of the upper bounds on the arm k for $l \leq b < d$.

$$\begin{aligned} T_k(D) \leq l + \sum_{d=K+1}^{D} \mathbf{1}\Big\{ \min_{0<s<d} \overline{\Psi_{k^*}}\big(s\big) + U_s(d-1) \leq \\ \max_{l \leq b < d} \overline{\Psi_k}\big(b\big) + U_b(d-1) \Big\}, \end{aligned} \tag{10}$$

where n_s in $U_s(d-1)$ is $T_s(d-1)$ and n_b in $U_b(d-1)$ is $T_b(d-1)$. Since the particular indices for which they occur are unknown, we consider all possible pairs of indices to loose this bound further.

$$\begin{aligned} E[T_k(D)] \leq l + \sum_{d=1}^{\infty}\sum_{s=1}^{d-1}\sum_{b=l}^{d-1} P(\Big\{ \overline{\Psi_{k^*}}\big(s\big) + \\ U_s(d) \leq \overline{\Psi_k}\big(b\big) + U_b(d) \Big\}). \end{aligned} \tag{11}$$

To make sure the event $\overline{\Psi_{k^*}}\big(s\big) + U_s(d) \leq \overline{\Psi_k}\big(b\big) + U_b(d)$ happens, one of the following three must also occur: (1.) $\overline{\Psi_{k^*}}\big(s\big) + U_s(d) \leq q^s(k^*)$; (2.) $\overline{\Psi_k}\big(b\big) \geq q^b(k) + U_b(d)$; and (3.) $q^s(k^*) < q^b(k) + 2U_b(d)$.

We can loosen the bound through a union of three events.

$$\begin{aligned} E[T_k(D)] \leq l + \sum_{d=1}^{\infty}\sum_{s=1}^{d-1}\sum_{b=l}^{d-1} P(\Big\{ \overline{\Psi_k}\big(b\big) \geq q^b(k) + U_b(d) \Big\} + \\ \Big\{ \overline{\Psi_{k^*}}\big(s\big) + U_s(d) \leq q^s(k^*) \Big\} + \\ \Big\{ q^s(k^*) < q^b(k) + 2U_b(d) \Big\}). \end{aligned} \tag{12}$$

It is shown in [27] that the random variable $\frac{\overline{\Psi_k}\big(b\big) - q^b(k)}{\sqrt{\frac{Q_k(b) - n_b\left[\overline{\Phi_k}(b)\right]^2}{n_b(n_b-1)}}}$ has a Student distribution with $n_k - 1$ degrees of freedom. Given conjecture 1, we bound the probability by letting $s = n_k - 1$ and $a = 4\sqrt{\ln(d)}$:

$$\begin{aligned} P(\Big\{ \overline{\Psi_k}\big(b\big) \geq q^b(k) + U_b(d) \Big\}) = \\ P(\Big\{ \frac{\overline{\Psi_k}\big(b\big) - q^b(k)}{\sqrt{\frac{Q_k(b) - n_b\left[\overline{\Phi_k}(b)\right]^2}{n_b(n_b-1)}}} \geq 4\sqrt{\ln(d)} \Big\}) \leq d^{-4}, \end{aligned} \tag{13}$$

for $n_s \geq 8\ln(d)$. We can then bound $\Big\{ \overline{\Psi_{k^*}}\big(s\big) + U_s(d) \leq q^s(k^*) \Big\}$ through a similar fashion. It is also shown in [27] that the random variable $\frac{Q_k(b) - n_s\left[\overline{\Phi_k}(b)\right]^2}{[\sigma_k^b]^2}$ follows a chi-squared distribution with $n_k - 1$ degree of freedom. Note that here σ_k^b is the variance of the estimation at round b. By letting $s = n_s - 1$ and $a = 4s$ in Conjecture 2 with $s = n_s - 1$ and $a = 4s$, we have

$$\begin{aligned} P(\Big\{ q^s(k^*) < q^b(k) + 2U_b(d) \Big\}) = \\ P(\Big\{ \frac{Q_k(b) - n_b\left[\overline{\Phi_k}(b)\right]^2}{[\sigma_k^b]^2} > (n_b - 1)\frac{[q^s(k^*) - q^b(k)]^2}{[\sigma_k^b]^2}\frac{n_b}{64\ln(d)} \Big\}) \\ \leq P(\Big\{ \frac{Q_k(b) - n_b\left[\overline{\Phi_k}(b)\right]^2}{[\sigma_k^b]^2} > 4(n_b - 1) \Big\}) \leq e^{-\frac{n_b}{2}} \leq d^{-4}, \end{aligned} \tag{14}$$

as long as $b \geq \max\Big\{256\frac{[\sigma_k^b]^2}{[q^s(k^*) - q^b(k)]^2}, 8\Big\}\ln(d)$. This means that as long as we play for no less than $l = \lceil \max\Big\{256\frac{[\sigma_k^b]^2}{[q^s(k^*) - q^b(k)]^2}, 8\Big\}\ln(d) \rceil$ times, we can guarantee that $E[T_k(D)]$ in Equation 12 can be bounded as the following:

$$\begin{aligned} E[T_k(D)] \leq l + \sum_{d=1}^{\infty}\sum_{s=1}^{d}\sum_{b=l}^{d} 3d^{-4} = l + \sum_{d=1}^{\infty} 3d^{-2} \\ \leq 1 + 256\frac{[\sigma_k^b]^2}{[q^s(k^*) - q^b(k)]^2}\ln(D) + 8\ln(D) + \frac{\pi^2}{2} \leq \\ 256\ln(D)\frac{\hat{\sigma}_k^2}{\hat{\Delta}_k^2} + \Big(1 + \frac{\pi^2}{2} + 8\ln(D)\Big), \end{aligned} \tag{15}$$

where $\hat{\sigma}_k = \max_b[\sigma_k^b], 1 < b < D$ and $\hat{\Delta}_k = \min_{s,b}[q^s(k^*) - q^b(k)]$, $1 < s, b < D$. Given the Equation

7 and Equation 15, the regret bound of HpUCB takes the following form:

$$E\big[R(D)\big] \leq \sum_{k=1}^{K} E\big[T_k(D)\big]\tilde{\Delta}_k \leq \sum_{k=1}^{K} \Big[256\ln(D)\frac{\hat{\sigma}_k^2}{\hat{\Delta}_k^2}\tilde{\Delta}_k + \Big(1+\frac{\pi^2}{2}+8\ln(D)\Big)\tilde{\Delta}_k\Big], \tag{16}$$

where $\hat{\sigma}_k$ can be seen as the estimation variance for arm k (i.e., $\hat{\sigma}_k$) and $\hat{\Delta}_k$ is the maximal difference between the optimal arm and the arm k in all rounds. In UCB-Normal, the mean and standard deviation of the reward distribution remain constant across the rounds. In the setting of Hawkes Processes MAB problem, these properties are dynamic. In specific, the expectation and standard deviation vary at each round. In our regret analysis, we show that the regret can be bounded by $O(K\ln(d))$ under the assumption that the estimation of the expected number of events at round d follows a normal distribution. In our regret analysis, we connect the regret with the precision of the Hawkes process estimation instead of the Hawkes process parameters. In specific, the regret depends on the variance of the estimation (i.e., $\hat{\sigma}_k$) and the difference between the optimal arm and the arm k (i.e., $\hat{\Delta}_k$). The larger the variance is, or the less precise the estimation is, the larger the regret bound grows. This also suggests that as long as we can precisely infer the Hawkes processes, we can minimize the regret by exploiting the best choice. Another factor to govern the regret is $\hat{\Delta}_k$, the difference between the optimal action and action k. The smaller $\hat{\Delta}_k$ is, which means it is harder to differentiate the optimal arm from the others, the larger the regret bound will become.

IV. Experiments

A. Baseline Methods

We compare our model HpUCB with several Multi-armed Bandit baselines that are tailored to handle non-stationary reward distributions. The source code and dataset are included in a anonymous repository [2] for reproducibility. We start with algorithms that update the reward observation with exponential weights, exp3, and its variation exp3S [28]. The name exp3 stands for "Exponential-weight algorithm for Exploration and Exploitation." Essentially, this type of algorithm tracks a weighted list of estimators. The arms are then randomly selected as a function of their weight. After the rewards are received, the weights are updated with respect to the observed value of the payoff. The standard UCB algorithm UCB1 [27] has also been extended to non-stationary settings. In the work of [29], discounted UCB dUCB and sliding-window UCB slideUCB are proposed to deal with changes in the distribution over time. In the dUCB algorithm, the contribution of each observation to the mean and upper bound estimates decays exponentially in time. In the related slideUCB algorithm, a local empirical average of the observed rewards in used to select arms. In more recent work, the monitored-UCB mUCB algorithm was proposed in [23], which introduces a change-point detection component into a classic Upper Confidence Bound (UCB) algorithm. Once the change-point of the reward distribution is detected, mUCB resets and learns the new optimal arm.

B. Datasets

1) Temporal Synthetic Data ($\mathcal{D}_{SIM}$):: We first test our methodology on a synthetic Hawkes process dataset generated through thinning [30]. We denote this dataset as $\mathcal{D}_{\text{SIM}}$. In total, we simulate 40 nodes of independent Hawkes Processes with similar properties (e.g. similar background rates and triggering kernels). We sample the background rate μ uniformly from the range 0.023 to 0.025 across the 40 nodes, the branching parameter r from the range 0.80 to 0.85, and the decay rate β from the range 0.005 to 0.009. In the experiments, we only observe events that happen in the timeframe when we play the arm and only the observed events are considered as rewards. Otherwise, the events are unobserved. The idea of simulating Hawkes processes with similar parameters is to make sure one or several arms do not dominate the others and make estimation trivial. Given the similar parameters, each process has a comparable expected number of events against the others. Such a characteristic makes it difficult for the optimal arm to be identified by standard, stationary MAB algorithms. However, events cluster in different arms at different points in time in this example, which non-stationary MABs can take advantage. We use this example to test our model's ability to track the fluctuation of event numbers and switch between arms to get optimal performance.

2) Records of Earthquakes ($\mathcal{D}_{EARW}$):: In the next example, we use a dataset consisting of the spatial and temporal records of worldwide earthquakes from the United States Geological Survey [31] from 1990 to present which has magnitude above zero. We first categorize these events into subregions and then we consider each subregion as an arm. One way to categorize each earthquake record into different subregions is clustering based on its geolocation. We apply a k-nearest neighbors algorithm to the earthquake dataset and categorize them into 200 subregions, and we retain 40 subregions with the most earthquake events. We denote the dataset as $\mathcal{D}_{\text{EARW}}$.

3) Records of Earthquakes in Alaska and California ($\mathcal{D}_{EACA}$):: To validate our model on a dataset that has a more fine-grained scale, we also test the model on the subset of records from Alaska and California, states that are situated between the Pacific and North American Plates and are the most earthquake-prone areas in the U.S. We map each earthquake record to the closest county or county equivalent. Since most of the earthquakes happen in the coastal areas, we only keep the top-18 subregions with the most earthquake records. The pruned dataset is then denoted as $\mathcal{D}_{\text{EACA}}$.

4) Records of Crime Incidents ($\mathcal{D}_{VAN}$):: We apply the Hawkes MAB to crime incident reports, which also exhibit space-time clustering. We collect crime reports from the Vancouver Open Data Catalogue [3]. These crime types include

[2] https://anonymous.4open.science/r/HawkesMAB-43CB/

[3] http://data.vancouver.ca/datacatalogue/crime-data.htm

theft, break and enter, assault, robbery, arson, etc. In this dataset, the events are separated into 24 subregions/beats by the local police department. We then remove the top-2 subregions with the most crimes, which have substantially more reports compared to other subregions by a large margin. We remove these two subregions since they can be easily learned by MAB algorithms. Therefore, the algorithms may result in similar performances. The dataset is then denoted as $\mathcal{D}_{\text{VAN}}$.

5) Records of Crime Activities in Los Angeles ($\mathcal{D}_{LA}$):: We also test the algorithms on crime report data from Los Angeles [4], containing incidents of crime in the City of Los Angeles dating back to 2010. The events are categorized into 21 divisions. We also removed the top-2 subregions with the most crimes. We then denote this dataset as $\mathcal{D}_{\text{LA}}$.

C. Reward Performances

In the simulation dataset, we separate the time horizon into 2000 rounds and 350 of them are used to tune the hyperparameters for all the models. In the real-world datasets, we separate the time horizon into 500 rounds, and 150 of them are used in searching the hyper-parameters. For each model, we repeat the multi-armed bandit process for 50 times with different regions to start, and we report the average performance. Experimental results on the earthquake event data from Alaska and California and the crime report data in Los Angeles are presented in the supplementary material due to the page limit of the manuscript.

In table I, table II and table III, we present the performances of all models in various datasets. We evaluate performance through accumulated rewards. Each round, we visit $1, 3$, and 5 regions and record the rewards, respectively. We then denote the metrics as $\overline{\texttt{rewd}}$ @1, $\overline{\texttt{rewd}}$ @3, and $\overline{\texttt{rewd}}$ @5. Overall, the proposed method HpUCB consistently outperforms all the other baseline methods in all $\overline{\texttt{rewd}}$ @1, $\overline{\texttt{rewd}}$ @3, and $\overline{\texttt{rewd}}$ @5. The only exceptions are $\overline{\texttt{rewd}}$ @1 in the dataset. In this exception, our model HpUCB maintains a competitive performance compared to the best model UCB1. It could be because in this dataset, the best region has a lot more events of interest than the others, and the event occurrence remains at a constant rate throughout time. Therefore, the simple stationary model UCB1 can easily find the best choice. However, our model HpUCB can also learn stationary arms via the background rate component in the Hawkes process. When the Hawkes MAB is used in situations where the top arm changes dynamically, by having more events (i.e., $\overline{\texttt{rewd}}$ @3, and $\overline{\texttt{rewd}}$ @5), the advantage of HpUCB starts to show. That is, the improvements in performance against the baseline methods start to increase as we visit more regions each time. This trend can be observed across all the datasets. In general, the experiments show that HpUCB performs well on both the simulated and the real-world datasets. Given the nature of self-exciting properties in criminal activities and earthquakes, the proposed model HpUCB can leverage the temporal information between the events and make good choices for future actions.

[4]https://www.kaggle.com/cityofLA/crime-in-los-angeles

In Figure 2 and 3, we show the cumulative rewards for all models from $\mathcal{D}_{\text{EARW}}$ and $\mathcal{D}_{\text{EACA}}$. In Figure 2 and 3, our model HpUCB learns the optimal arm in the first 200 rounds and consistently chooses the optimal ones. Selection strategies, such as exp3 and exp3S do not perform well in general. One possible explanation is that the probability of sampling the arms grows exponentially based on the rewards. As the rewards grow larger, the probability for multiple competitive arms to be sampled may be saturated. Thus, the algorithms may fail to distinguish those arms. Overall, mUCB and UCB1 are competitive algorithms when the number of the pulled arms is small. Our model, HpUCB, has better performance when multiple competing arms with different event spikes are present.

V. Conclusion

In this paper, we extended Multi-armed Bandits (MAB) to the problem of learning and ranking a set of Hawkes processes when the self-excited events are unobserved until an action is taken. We used the theoretical closed-form expression for the expected number of events of a Hawkes process to estimate future rewards in the MAB problem setting. Following that, we constructed an upper confidence bound as a criterion to pull the arms optimistically. By fusing the self-exciting Hawkes process and Multi-armed bandit framework together, we created a novel Hawkes MAB algorithm that optimizes the tradeoff between exploring regions where Hawkes process events are unobserved and exploiting regions where the Hawkes processes are already learned.

We investigate the factors that affect regret through a regret bound analysis, under an assumption of normality. The regret is bounded by $O(K\ln(d))$ and it is proportional to the Hawkes counting process variance. The model shows promising results for a simulated dataset in which all arms have similar Hawks process parameters but different temporal clustering patterns. This shows that HpUCB can switch between the arms effectively. The performance on crime report and earthquake data demonstrates how our proposed model can be applied to real data, especially when events are self-exciting.

References

[1] A. Slivkins, "Introduction to multi-armed bandits," *arXiv preprint arXiv:1904.07272*, 2019.

[2] P. Zhao, L. Zhang, Y. Jiang, and Z.-H. Zhou, "A simple approach for non-stationary linear bandits," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 746–755.

[3] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem," in *Proceedings of IEEE 36th annual foundations of computer science*. IEEE, 1995, pp. 322–331.

[4] W. Chu, L. Li, L. Reyzin, and R. Schapire, "Contextual bandits with linear payoff functions," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 208–214.

[5] A. Badanidiyuru, R. Kleinberg, and A. Slivkins, "Bandits with knapsacks," in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE, 2013, pp. 207–216.

[6] M. Aziz, E. Kaufmann, and M.-K. Riviere, "On multi-armed bandit designs for dose-finding clinical trials," *Journal of Machine Learning Research*, vol. 22, no. 14, pp. 1–38, 2021.

TABLE I
OVERALL PERFORMANCE ON THE SYNTHETIC DATASET AND THE EARTHQUAKE DATASET

Model	$\mathcal{D}_{\text{SIM}}$			$\mathcal{D}_{\text{EARW}}$			$\mathcal{D}_{\text{EACA}}$		
	$\overline{\texttt{rewd}}$ @1	$\overline{\texttt{rewd}}$ @3	$\overline{\texttt{rewd}}$ @5	$\overline{\texttt{rewd}}$ @1	$\overline{\texttt{rewd}}$ @3	$\overline{\texttt{rewd}}$ @5	$\overline{\texttt{rewd}}$ @1	$\overline{\texttt{rewd}}$ @3	$\overline{\texttt{rewd}}$ @5
exp3	22082.0	66204.8	110369.5	35036.98	107201.8	182437.3	27348.74	84980.46	142215.24
exp3S	22077.6	66216.9	110815.0	38214.14	105540.7	176299.5	28185.16	87948.66	137464.80
mUCB	26265.6	69635.7	113784.5	76027.06	154578.0	218291.0	44527.22	109739.32	161888.84
dUCB	22081.00	66682.1	116890.5	35199.30	184907.3	285493.2	27373.84	114044.32	186503.26
UCB1	26127.4	70433.1	114255.7	74553.12	135262.3	222302.0	**47902.72**	107992.10	166784.96
slideUCB	23723.0	74205.0	119480.2	60541.08	224853.9	279058.9	34088.48	125104.70	190080.82
HpUCB	**26852.7**	**79581.5**	**129435.2**	**82733.1**	**226576.8**	**332122.6**	47304.70	**140628.14**	**221007.26**

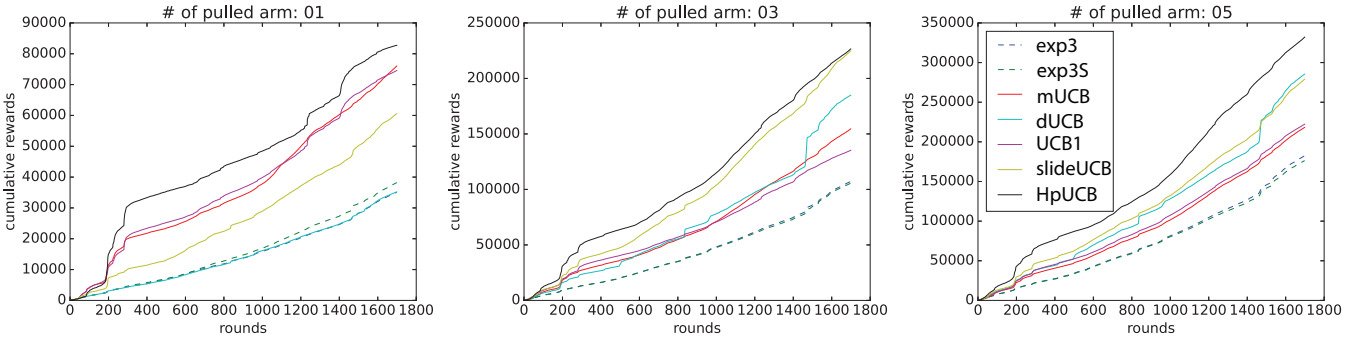

Fig. 2. Cumulative rewards of records of worldwide earthquakes $\mathcal{D}_{\text{EARW}}$

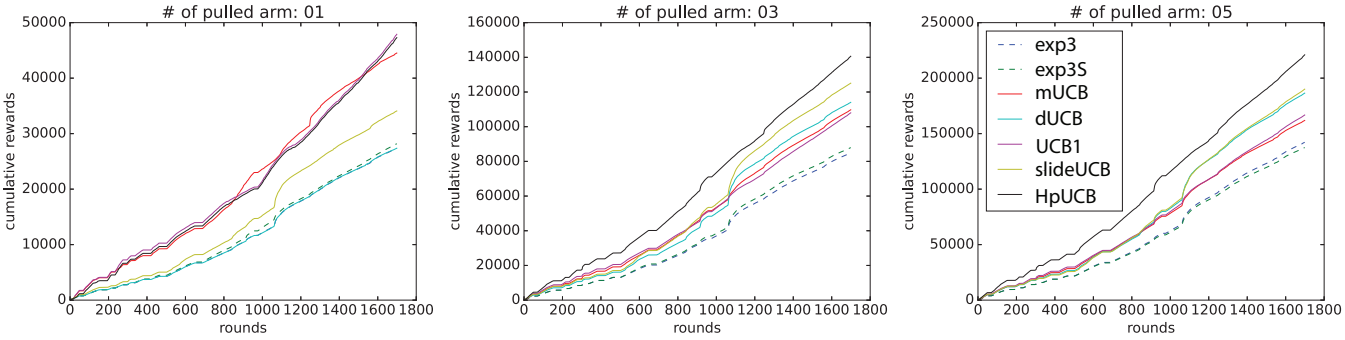

Fig. 3. Cumulative earthquake rewards of records in AK CA ($\mathcal{D}_{\text{EACA}}$)

[7] B. Awerbuch and R. D. Kleinberg, "Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches," in *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, 2004, pp. 45–53.

[8] W. Shen, J. Wang, Y.-G. Jiang, and H. Zha, "Portfolio choices with orthogonal bandit learning," in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[9] M.-A. Rizoiu, Y. Lee, S. Mishra, and L. Xie, "A tutorial on hawkes processes for events in social media," *arXiv preprint arXiv:1708.06401*, 2017.

[10] G. Mohler, J. Carter, and R. Raje, "Improving social harm indices with a modulated hawkes process," *International Journal of Forecasting*, vol. 34, no. 3, pp. 431–439, 2018.

[11] B. Yuan, H. Li, A. L. Bertozzi, P. J. Brantingham, and M. A. Porter, "Multivariate spatiotemporal hawkes processes and network reconstruction," *SIAM Journal on Mathematics of Data Science*, vol. 1, no. 2, pp. 356–382, 2019.

[12] B. M. Lapham, "Hawkes processes and some financial applications," Master's thesis, University of Cape Town, 2014.

[13] W.-H. Chiang, X. Liu, and G. Mohler, "Hawkes process modeling of covid-19 with mobility leading indicators and spatial covariates," *International journal of forecasting*, vol. 38, no. 2, pp. 505–520, 2022.

[14] L. Lesage, M. Deaconu, A. Lejay, J. A. Meira, G. Nichil *et al.*, "Hawkes processes framework with a gamma density as excitation function: application to natural disasters for insurance," *Methodology and Computing in Applied Probability*, pp. 1–29, 2022.

[15] S. Xue, X. Shi, H. Hao, L. Ma, J. Zhang, S. Wang, and S. Wang, "A graph regularized point process model for event propagation sequence," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–7.

[16] Z.-H. Cheng, J.-W. Liu, and Z. Cao, "Hypergraph neural network hawkes process," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–6.

[17] Q. Li, J. Chen, X. Tang, H. Yu, and H. Song, "Modeling time decay effect in temporal knowledge graphs via multivariate hawkes process," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[18] L.-n. Zhang, J.-w. Liu, Z.-y. Song, X. Zuo, W.-m. Li, and Z.-y. Liu, "Universal transformer hawkes process," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–7.

[19] X. Wang, F. Qiang, L. Ma, P. Zhang, H. Yang, Z. Li, and J. Zhang, "Federated transformer hawkes processes for distributed event sequence

TABLE II
CRIMES IN VANCOUVER $\mathcal{D}_{\text{VAN}}$

Model	$\overline{\texttt{rewd}}$ @1	$\overline{\texttt{rewd}}$ @3	$\overline{\texttt{rewd}}$ @5
exp3	8467.2	25517.7	42729.4
exp3S	8434.3	25269.7	42111.2
mUCB	16619.2	40339.2	57674.3
dUCB	11468.0	39930.8	61504.2
UCB1	**17552.2**	44436.1	61726.7
slideUCB	16945.5	44027.1	61444.0
HpUCB	17311.9	**50971.7**	**80531.4**

TABLE III
CRIMES IN LOS ANGELES $\mathcal{D}_{\text{LA}}$

Model	$\overline{\texttt{rewd}}$ @1	$\overline{\texttt{rewd}}$ @3	$\overline{\texttt{rewd}}$ @5
exp3	41110.18	123348.12	205715.82
exp3S	41102.44	123235.70	205530.36
mUCB	46108.64	131706.22	214805.94
dUCB	43018.02	131227.52	215557.94
UCB1	**46579.32**	133710.34	215999.10
slideUCB	46399.08	133618.60	215751.74
HpUCB	46277.78	**139447.44**	**226297.88**

prediction," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[20] A. Jalaldoust, K. Hlaváčková-Schindler, and C. Plant, "Causal discovery in hawkes processes by minimum description length," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 6, 2022, pp. 6978–6987.

[21] W. Chen, J. Chen, R. Cai, Y. Liu, and Z. Hao, "Learning granger causality for non-stationary hawkes processes," *Neurocomputing*, vol. 468, pp. 22–32, 2022.

[22] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in *International Conference on Algorithmic Learning Theory*. Springer, 2011, pp. 174–188.

[23] Y. Cao, Z. Wen, B. Kveton, and Y. Xie, "Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 418–427.

[24] W. C. Cheung, D. Simchi-Levi, and R. Zhu, "Learning to optimize under non-stationarity," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1079–1087.

[25] L. Cui, A. Hawkes, and H. Yi, "An elementary derivation of moments of hawkes processes," *Advances in Applied Probability*, vol. 52, no. 1, pp. 102–137, 2020.

[26] K. H. Zhao, "Regret-based algorithms for multi-armed bandits," Ph.D. dissertation, 2020.

[27] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multi-armed bandit problem," *Machine learning*, vol. 47, no. 2, pp. 235–256, 2002.

[28] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," *SIAM journal on computing*, vol. 32, no. 1, pp. 48–77, 2002.

[29] L. Kocsis and C. Szepesvári, "Discounted ucb," in *2nd PASCAL Challenges Workshop*, vol. 2, 2006, pp. 51–134.

[30] J. Møller and J. G. Rasmussen, "Perfect simulation of hawkes processes," *Advances in applied probability*, vol. 37, no. 3, pp. 629–646, 2005.

[31] United States Geological Survey, "Usgs real-time earthquake data," 2022. [Online]. Available: `http://earthquake.usgs.gov/earthquakes/map/`

APPENDIX

In Fig. 4 and 5, we show the numerical results to verify the conjectures proposed in [27]. These conjectures provide certain bounds on the tails of the χ^2 and the Student's t-distribution. In Figure 4, we show that given a Student's t-distribution with the degree of freedom s, the tail probability of $P(X > a)$ is bounded by $\mathrm{e}^{-\frac{a^2}{4}}$ as long as $0 \leq a \leq \sqrt{2(s+1)}$. We show that the conjecture holds under different scenarios (i.e., $s = 5, 10, 50, 100$). Given the cumulative distribution function of a Student's t-distribution $F_x(a)$, we plug in all different a and get $P(X \geq a) = 1 - F_x(a)$.

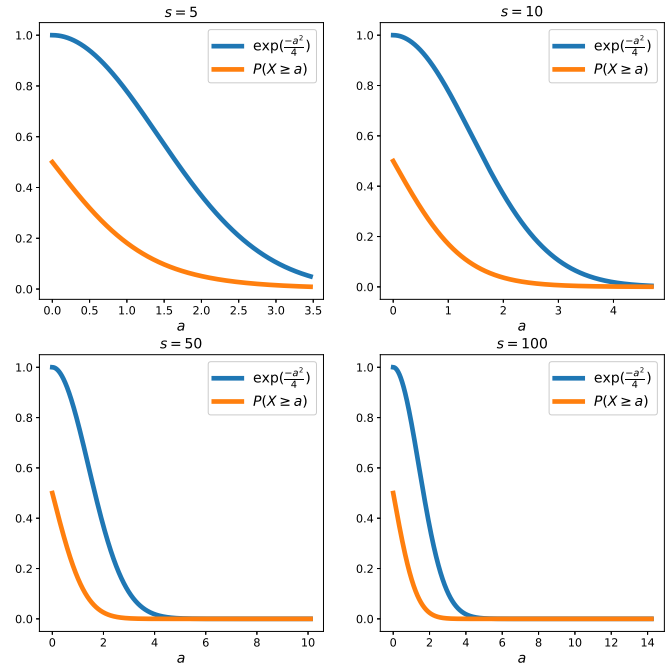

Fig. 4. Numerical verification for conjecture 1. Given a Student random variable X with s degrees of freedom, for all $0 \leq a \leq \sqrt{2(s+1)}$, $P(X \geq a) \leq \mathrm{e}^{-\frac{a^2}{4}}$.

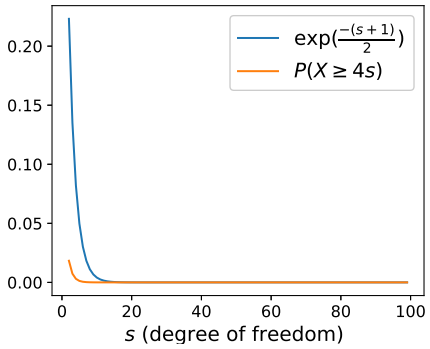

Fig. 5. Numerical verification for conjecture 2. Given a random variable X following a Chi-squared distribution with s degrees of freedom, $P(X \geq 4s) \leq \mathrm{e}^{-\frac{(s+1)}{2}}$.

Comparing it with $\mathrm{e}^{-\frac{a^2}{4}}$, we can find out the $P(X \geq a) \leq \mathrm{e}^{-\frac{a^2}{4}}$. This holds for all s. In Figure 5, we show that the tail probability of a Chi-squared distribution $P(X \geq 4s)$ can be bounded by $\mathrm{e}^{-\frac{(s+1)}{2}}$. This can be simply verified by plugging different s into the cumulative distribution function of the Chi-squared distribution.