Uncertainty Quantification for Inferring Hawkes Networks

Haoyun Wang[1], **Liyan Xie**[1], **Alex Cuozzo**[2], **Simon Mak**[2], **Yao Xie**[1]
[1]Georgia Institute of Technology, [2] Duke University

Abstract

Multivariate Hawkes processes are commonly used to model streaming networked event data in a wide variety of applications. However, it remains a challenge to extract reliable inference from complex datasets with uncertainty quantification. Aiming towards this, we develop a statistical inference framework to learn causal relationships between nodes from networked data, where the underlying directed graph implies Granger causality. We provide uncertainty quantification for the maximum likelihood estimate of the network multivariate Hawkes process by providing a non-asymptotic confidence set. The main technique is based on the concentration inequalities of continuous-time martingales. We compare our method to the previously-derived asymptotic Hawkes process confidence interval, and demonstrate the strengths of our method in an application to neuronal connectivity reconstruction.

1 Introduction

Recently, there has been a surge of interest in using Hawkes processes networks to model discrete events data in both the statistics and the machine learning community (see a review in [21]). The popularity of the model can be attributed to its wide range of applications, including seismology, criminology, epidemiology [25], social networks [15], neural activity [23], and so on. The model is capable of capturing a spatio-temporal triggering effect reflected in real-world networks – one event may trigger subsequent events at different locations. Existing works for recovery of the Hawkes network focus on performing point estimators: most of them rely on estimating influence coefficients (representing the magnitude of the influence) and thresholding by a pre-specified value to recover the Hawkes network structure. However, most existing work does not quantify the uncertainty, for instance, in the form of a confidence interval.

One outstanding issue with point estimators is that, without accurate uncertainty quantification, one cannot claim any statistical significance of the results. For instance, it is difficult to assign a probability to the existence of a directed edge between two nodes. This problem is critical for certain problems, such as causal inference. In Hawkes network models, Granger causality has a very simple form: there is a causal relationship between two nodes in the network if and only if there exists an edge between the two nodes, and there is no casual relationship otherwise [8]. Thus, uncertainty quantification (UQ) is critical in scientific studies, because we wish to test whether a causal relationship exists between one node to another node at a given statistical confidence level. The development of easy-to-implement and robust UQ tools is crucial for a variety of scientific problems, from medical data to social networks to neural spike train data and more.

In this paper, we study the uncertainty quantification for maximum likelihood estimates of multivariate Hawkes processes over networks. Each node can represent a location, a region of a brain, or a user in a social network. We are particularly interested in recovering the underlying structure (connections) between different nodes with uncertainty quantification, meaning providing accurate upper and lower confidence intervals for the influence coefficients (which is linked to the causal relationships between

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

these nodes). Since we are particularly interested in the existence (or non-existence) of an underlying edge, one of the most critical aspects of our study is to perform network topology recovery. Motivated by applications where the recovery guarantee is usually required, we focus on quantifying the uncertainty of the maximum likelihood estimate of the unknown parameters by providing confidence intervals (CIs). We proposed a novel non-asymptotic approach to establish a general confidence polyhedral set for hidden network influence parameters (from which the confidence sets can be extracted). The non-asymptotic confidence set is established by constructing a continuous-time martingale using the score function (the gradient of the log-likelihood function) of the network Hawkes process. This enables us to apply a concentration bound for continuous-time martingales. The non-asymptotic confident set is more accurate than the classic asymptotic confidence intervals, since the concentration bound approach captures more than the first and second-order moments (which are essentially what the asymptotic confidence intervals are capturing). We compared the two methods for establishing CIs using synthetic neural activity data, to demonstrate the effectiveness of our approach.

Contributions. Our main contribution can be summarized as follows: (1) We give a non-asymptotic confidence set for the maximum likelihood estimate (MLE) of the Hawkes process over networks, and (2) Our confidence set is more general and can be approximated by a polyhedron; the CIs can be solved efficiently from a linear program. In contrast, the classic CI essentially provides a box in the high-dimensional space for the multi-dimensional parameters.

Related Work. There has been much effort made on network inference for multivariate point processes. Learning algorithms for Granger causality of Hawkes processes has been proposed in [29] using the regularized MLE. [1] proposed a nonparametric way to estimate the mutual inference and causality relationship in multivariate Hawkes processes. [30] considers the spatiotemporal Hawkes process and develop a nonparametric method for network reconstruction. [16] studies the detection of changes in the underlying dynamics. Moreover, recent work has also focused on causal inference for different applications, such as online platforms [14], infectivity matrix estimation [29], etc.

However, there is relatively little literature that provides theoretical guarantees on the significance level of the estimation results. The concentration results for inhomogeneous Poisson processes were studied in [22]. The non-asymptotic tail estimates for the Hawkes process were established in [24]. [7] studies Granger causality for brain networks and characterizes the significance level using numerical methods, while our result gives a theoretical guarantee on the confidence level. The CI for parameter recovery of discrete-time Bernoulli processes is given in [12]. At the same time, this paper focuses on the continuous-time Hawkes process, which is more complicated in uncertainty quantification. In Bayesian statistics literature, works have been done in quantifying the uncertainty of the network parameters by imposing a prior model on the model hyperparameters; the posterior is approximated using Markov chain Monte Carlo [20, 27]. Recently, there has been an effort to establish time-uniform CI based on concentration inequalities [11, 10].

The field of uncertainty quantification itself is very broad, with important applications in computer simulations [26], aerospace engineering [17] and climatology [19]. This literature can be grouped into two categories [28]: inverse UQ (the inference of parameters from a generating model) and forward UQ (the propagation of uncertainty through numerical models). The current work focuses on the inverse problem for the network multivariate Hawkes process and, in particular, in providing the confidence interval for the maximum likelihood estimates of parameters.

2 Background

A temporal point process is a random process whose realization consists of a list of discrete events localized in time. Let $(u_1, t_1), \cdots, (u_n, t_n)$ be a series of events happened during time period $[0, T]$ on a multivariate Hawkes process with D nodes, where t_i denoted the time of the i-th event, and $u_i \in [D]$ is the index of node where the event happens. The intensity function at node i at time t is

$$\lambda_i(t) = \mu_i + \sum_{j:t_j<t} \alpha_{i,u_j} \varphi_{i,u_j}(t - t_j),\ i = 1, \cdots, D,$$

where μ_i is the background rate of events happening at i-th node, $\alpha_{ij} \geq 0$ is a parameter representing the influence of node j to node i, and φ_{ij} is a function supported on $[0, \infty)$. Let $\boldsymbol{N}_t \in \mathbb{N}^D$ be a vector where the i-th entry N_t^i is the number of events happened on node i during $[0, t)$. For any

function f, define the following integral with counting measure

$$\int_0^T f(t)dN_t = \sum_{t\in\mathcal{H}_T} f(t),$$

where $\mathcal{H}_t = \{t_1, \ldots, t_n : t_n < t\}$ denotes the list of times of history events up to but not including time t. Therefore, we can rewrite the intensity function as

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{D} \int_0^t \alpha_{ij}\varphi_{ij}(t-\tau)dN_\tau^i,\ i = 1, \cdots, D. \tag{1}$$

We may consider different types of influence function φ, including: (i) the *gamma function* $\varphi(\Delta t) = (\Delta t)^{k-1}e^{-\Delta t/\beta}/(\Gamma(k)\beta^k)$, $\Delta t \geq 0$. Note that when $k = 1$, it becomes the commonly-used *exponential function*, $\varphi(\Delta t) = \beta e^{-\beta\Delta t}$, $\Delta t \geq 0$, which shows that the influence of events on future intensity is exponentially decaying. The decay starts immediately following the onset (thus, there is no delay); (ii) the *Gaussian function*: $\varphi(\Delta t) = \exp\{-\beta(\Delta t - \tau)^2/\sigma\}/\sqrt{2\pi\sigma}$, $\Delta t \geq 0$, where $\tau \geq 0$ is the unknown delay which means that the influence attains its maximum value τ time after the event happens.

2.1 Decoupled log-likelihood function

Let $A = (\alpha_{ij})_{i,j\in[D]}$, $\alpha_{ij} \geq 0$, be a matrix that contains all the influence parameters between nodes and our parameter-of-interest to be estimated. Given the events on $[0, T]$, the likelihood function is (detailed derivation can be found in, e.g., [21])

$$L(A) = \exp\left(-\sum_{i=1}^{D}\int_0^T \lambda_i(t)dt\right)\prod_{j=1}^{n}\lambda_{u_j}(t_j),$$

where $\lambda_i(t)$ is the intensity as defined in (1). Note that $\lambda_i(t)$ depends on A, but we omit the term for simplicity.

The log-likelihood function can be written in the form of integral with counting measure N_t^i at i-th node,

$$\ell(A) = \log L(A) = \sum_{i=1}^{D}\left(-\int_0^T \lambda_i(t)dt + \int_0^T \log\lambda_i(t)dN_t^i\right).$$

We note that the log-likelihood function $\ell(A)$ can be decoupled into summation of D terms, each for a specific node,

$$\ell(A) = \sum_{i=1}^{D}\ell_i(\boldsymbol{\alpha}_i),$$

where $\boldsymbol{\alpha}_i := [\alpha_{i1}, \cdots, \alpha_{iD}]^\intercal \in \mathbb{R}^D$ is a column vector denoting influence of other nodes to node i, and

$$\ell_i(\boldsymbol{\alpha}_i) = -\int_0^T \lambda_i(t)dt + \int_0^T \log(\lambda_i(t))dN_t^i. \tag{2}$$

Since $\ell_i(\boldsymbol{\alpha}_i)$ only depends on the parameter $\boldsymbol{\alpha}_i$, the statistical inference for each node (therefore each $\boldsymbol{\alpha}_i$) can be decoupled, which enables us to perform the computation in parallel and simplify our analysis. For the rest of this paper, we focus on the inference of a single $\boldsymbol{\alpha}_i$.

Notation. We use $\boldsymbol{\alpha}_i^*$ to denote the true parameter which is unknown, $\widehat{\boldsymbol{\alpha}}_i$ to denote the estimated parameter for i-th node, $\hat{\lambda}_i(t)$ to denote the intensity computed using the estimator $\widehat{\boldsymbol{\alpha}}_i$, and $\lambda_i^*(t)$ to denote the intensity under the true parameter $\boldsymbol{\alpha}_i^*$.

2.2 Score function and Fisher information

The statistical properties of the multivariate Hawkes process are closely related to its score function and the Fisher Information. The score function for i-th node given data, is defined as as

$$S_i(\boldsymbol{\alpha}_i) = \frac{\partial\, \ell_i(\boldsymbol{\alpha}_i)}{\partial \boldsymbol{\alpha}_i} = -\int_0^T \frac{\partial \lambda_i(t)}{\partial \boldsymbol{\alpha}_i} dt + \int_0^T \lambda_i^{-1}(t) \frac{\partial \lambda_i(t)}{\partial \boldsymbol{\alpha}_i} dN_t^i$$
$$= \int_0^T \lambda_i^{-1}(t) \frac{\partial \lambda_i(t)}{\partial \boldsymbol{\alpha}_i} (dN_t^i - \lambda_i(t) dt), \tag{3}$$

where $\partial \lambda_i(t)/\partial \boldsymbol{\alpha}_i$ is a vector independent of the choice of $\boldsymbol{\alpha}_i$. For simplicity, we denote this gradient by $\boldsymbol{\eta}_i(t) \in \mathbb{R}^D$, with j-th entry being

$$\eta_{ij}(t) = \frac{\partial \lambda_i(t)}{\partial \alpha_{ij}} = \int_0^t \varphi_{ij}(t-\tau) dN_\tau^j. \tag{4}$$

Note that η includes information about the influence function between two nodes; this holds for any general influence function φ_{ij}.

The D-by-D Hessian matrix of the log-likelihood function, given observations dN_t^i, is then

$$H_i(\boldsymbol{\alpha}_i) = \frac{\partial^2 l_i(\boldsymbol{\alpha}_i)}{\partial \boldsymbol{\alpha}_i \partial \boldsymbol{\alpha}_i^\intercal} = -\int_0^T \lambda_i^{-2}(t) \boldsymbol{\eta}_i(t) \boldsymbol{\eta}_i^\intercal(t) dN_t^i. \tag{5}$$

The Fisher Information I_i^* is defined as the expected variance of the score function $S_i(\boldsymbol{\alpha}_i)$, and also as the negative expected Hessian of the log-likelihood function over unit time, assuming the process is stationary under true parameter $\{\boldsymbol{\alpha}_i^*\}_{i=1}^D$,

$$I_i^* = \mathbb{E}\left[\int_0^1 \lambda_i^{*-2}(t) \boldsymbol{\eta}_i(t) \boldsymbol{\eta}_i^\intercal(t) dN_t^i\right] = \mathbb{E}\left[\int_0^1 \lambda_i^{*-1}(t) \boldsymbol{\eta}_i(t) \boldsymbol{\eta}_i^\intercal(t) dt\right] = \mathbb{E}\left[\lambda_i^{*-1} \boldsymbol{\eta}_i \boldsymbol{\eta}_i^\intercal\right].$$

An example of the exponentially decaying kernel for the above quantities is given in Appendix A.

Remark 1 (Multiple sequences). In practice, we may not be able to collect data long enough to achieve good confidence bound. Instead, we may have observations of many trials of independent Hawkes processes. We can write down the corresponding log-likelihood function and perform a similar analysis.

3 Main Results

We assume that the influence functions φ_{ij} and background intensities μ_i are *given*, and our goal is to estimate the *unknown* parameters $\boldsymbol{\alpha}_i$. We estimate the unknown parameters $\boldsymbol{\alpha}_i$ by maximizing the log-likelihood function, i.e,

$$\widehat{\boldsymbol{\alpha}}_i := \operatorname*{arg\,max}_{\boldsymbol{x} \in \mathbb{R}^D} \ell_i(\boldsymbol{x}). \tag{6}$$

For each node i, by (1), λ_i is linear in $\boldsymbol{\alpha}_i$, and by (2), we see that the log-likelihood function is concave in λ_i. Therefore, the MLE can be computed efficiently using convex optimization.

Remark 2. In the general case, the Hessian matrix $H_i(\cdot)$ is negative definite everywhere when at least D events happened on node i, and the MLE is unique. Indeed, when at least D events happened in the time interval $[0, T)$, the vectors $\{\boldsymbol{\eta}_i(t) : dN_t^i = 1\}$ will have (in the general case) a linearly independent component of size D, and H_i as the weighted sum of $\{-\boldsymbol{\eta}_i(t)\boldsymbol{\eta}_i^\intercal(t) : dN_t^i = 1, t < T\}$, is negative definite.

In this section, we present two different ways of uncertainty quantification for the MLE $\widehat{\boldsymbol{\alpha}}_i$. The first one is the classic asymptotic CI (for each entry α_{ij}), which uses the fact that the MLE of such Hawkes processes is consistent and asymptotically normal. The second approach is our proposed method, which entails building a general confidence set for $\boldsymbol{\alpha}_i^*$ based on the more precise concentration-bound.

3.1 Asymptotic Confidence Intervals

It is known that the MLE of a temporal point process is asymptotically normal, and the empirical Fisher Information converges to the true Fisher Information with probability 1 [18]. Under our context, this translates into the following theorem.

Theorem 3.1 (Asymptotic CI [18]). *Denote the empirical Fisher Information as*

$$\hat{I}_i(\widehat{\boldsymbol{\alpha}}_i) = -\frac{1}{T}\frac{\partial l_i(\widehat{\boldsymbol{\alpha}}_i)}{\partial \boldsymbol{\alpha}_i \partial \boldsymbol{\alpha}_i^\intercal} = \frac{1}{T}\int_0^T \hat{\lambda}_i^{-2}(t)\boldsymbol{\eta}_i(t)\boldsymbol{\eta}_i^\intercal(t)dN_t^i,$$

then as $T \to \infty$,

$$\sqrt{T}(\widehat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i^*) \to \mathcal{N}(0, I_i^{*-1}),$$
$$\hat{I}_i(\widehat{\boldsymbol{\alpha}}_i) \to I_i^*.$$

Since our focus is on the CI for each α_{ij}, Theorem 3.1 implies that as $T \to \infty$,

$$\sqrt{T}(\widehat{\alpha}_{ij} - \alpha_{ij}^*) \to \mathcal{N}(0, \sigma_{ij}^{*2}),$$

and

$$\widehat{\sigma}_{ij}^2(\widehat{\boldsymbol{\alpha}}_i) \to \sigma_{ij}^{*2},$$

where σ_{ij}^{*2} is the j-th diagonal entry of I_i^{*-1}, $\widehat{\sigma}_{ij}^2(\widehat{\boldsymbol{\alpha}}_i)$ is the j-th diagonal entry of $\hat{I}_i^{-1}(\widehat{\boldsymbol{\alpha}}_i)$. An asymptotic CI on each entry α_{ij} is given by

$$\widehat{\alpha}_{ij} \pm Z_{\varepsilon/2D}\sqrt{\widehat{\sigma}_{ij}^2(\widehat{\boldsymbol{\alpha}}_i)/T},$$

where $Z_{\varepsilon/2D}$ is the corresponding percentage point for standard normal distribution, i,e., $\Phi(Z_{\varepsilon/2D}) = 1 - \varepsilon/2D$, and $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution.

3.2 Generalized confidence sets based on concentration bound

The classic asymptotic CI has a nice form and is easy to compute, but its confidence level has no guarantee when T is not large enough. To be specific, there are three types of convergence involved in the asymptotic behavior of the MLE and the classic CI:

- the score function $S_i(\boldsymbol{\alpha}_i^*)$ is asymptotically normal,
- the empirical Hessian at $\boldsymbol{\alpha}_i^*$ converges to the Fisher Information,
- $\widehat{\boldsymbol{\alpha}}_i \to \boldsymbol{\alpha}_i^*$, and the statistical properties of $\widehat{\boldsymbol{\alpha}}_i$ converge to those of $\boldsymbol{\alpha}_i^*$.

In this section, we propose a general non-asymptotic confidence set for MLE $\widehat{\boldsymbol{\alpha}}_i$ by providing a concentration bound on the first type of convergence. In other words, we give the concentration bound on $S_i(\boldsymbol{\alpha}_i^*)$, which in turn provides the confidence set for $\boldsymbol{\alpha}_i^*$. This concentration result does not seem dependent on the convergence rate of the second term, and we further propose to use the third term's asymptotic behavior to approximate this confidence set and facilitate computation.

Our proof idea is as follows. We start with a similar step to proving the asymptotic result but follow with a tighter bound for the score function (the gradient of the log-likelihood function), leveraging a concentration bound for the continuous-time martingale.

First, we present the following general result. Since $\lambda_i^*(t)$ denotes the intensity function under true value $\boldsymbol{\alpha}_i^*$, we have the conditional expectation, given observations before time t, of $dN_t^i - \lambda_i^*(t)dt$ is 0, and

$$S_{i,t}(\boldsymbol{\alpha}_i^*) = \int_0^t \lambda_i^{*-1}(\tau)\boldsymbol{\eta}_i(\tau)(dN_\tau^i - \lambda_i^*(\tau)d\tau)$$

can be shown to be a *continuous-time martingale*.

The difficulty for a concentration bound here is that the variance of this process changes over time and cannot be bounded from above. Therefore, a standard Hoeffding or Bernstein type of concentration bound does not apply. Here we derive a concentration bound using the *intrinsic variance*, which depends on the data. Similar to [9], our intrinsic variance is a random process. Then we bound $S_i(\boldsymbol{\alpha}_i^*)$ in K different directions and convert these concentration bounds into a confidence set for $\boldsymbol{\alpha}_i$.

Theorem 3.2 (Confidence set for $\boldsymbol{\alpha}_i$). *Given data, for each* $\boldsymbol{\alpha}_i$, *let*

$$V_i(\boldsymbol{z}, \boldsymbol{\alpha}_i) = \int_0^T \left(\lambda_i(t)\exp(\lambda_i^{-1}(t)\boldsymbol{z}^\intercal\boldsymbol{\eta}_i(t)) - \boldsymbol{z}^\intercal\boldsymbol{\eta}_i(t) - \lambda_i(t)\right)dt. \tag{7}$$

For any given $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K\}$, *a confidence set for* $\boldsymbol{\alpha}_i$ *at level* $1-\varepsilon$ *is given by a polyhedron*

$$\mathcal{C}_{i,\varepsilon} = \left\{\boldsymbol{\alpha}_i \in \mathbb{R}^D : \forall k \in [K], \boldsymbol{z}_k^\intercal S_i(\boldsymbol{\alpha}_i) - V_i(\boldsymbol{z}_k, \boldsymbol{\alpha}_i) \le \ln(K/\varepsilon)\right\}. \tag{8}$$

UQ for each α_{ij}. Based on the confidence set for vector $\boldsymbol{\alpha}_i$, we can construct the CI for each entry. So, naturally, we would like our confidence set $\mathcal{C}_{i,\varepsilon}$ to resemble an orthotope parallel to the axes. Based on the mean value theorem, for any $\boldsymbol{\alpha}_i$ in the confidence set, there exists $\tilde{\boldsymbol{\alpha}}_i$ between $\boldsymbol{\alpha}_i$ and $\widehat{\boldsymbol{\alpha}}_i$ such that

$$S_i(\boldsymbol{\alpha}_i) - S_i(\hat{\boldsymbol{\alpha}_i}) = H_i(\tilde{\boldsymbol{\alpha}}_i)(\boldsymbol{\alpha}_i - \widehat{\boldsymbol{\alpha}}_i).$$

Since $\widehat{\boldsymbol{\alpha}}_i$ is the maximum likelihood estimate, we have $S_i(\widehat{\boldsymbol{\alpha}}_i) = 0$. The confidence set is supposed to be a small neighborhood of $\boldsymbol{\alpha}_i^*$, and when T is large, we have

$$S_i(\boldsymbol{\alpha}_i) = H_i(\tilde{\boldsymbol{\alpha}}_i)(\boldsymbol{\alpha}_i - \widehat{\boldsymbol{\alpha}}_i) \approx TI_i^*(\boldsymbol{\alpha}_i - \widehat{\boldsymbol{\alpha}}_i).$$

Intuitively, we can let $K = 2D$, $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{2D}$ be in the same direction with the columns of $\pm I_i^{*-1}$, each is supposed to correspond to the upper/lower bound on the entries of $\boldsymbol{\alpha}_i$, and the confidence set $\mathcal{C}_{i,\varepsilon}$ will approximately be a box around the MLE. Formally speaking, we have the following lemma.

Lemma 3.1. *Under the assumption that the moment generating function of $\boldsymbol{\eta}_i$ exists, there exists a neighborhood U of $\mathbf{0}$, such that*

$$\|S_i(\boldsymbol{\alpha}_i) - TI_i^*(\boldsymbol{\alpha}_i - \widehat{\boldsymbol{\alpha}}_i)\| \le O(T)\|\boldsymbol{\alpha}_i - \widehat{\boldsymbol{\alpha}}_i\|^2 + o(T)\|\boldsymbol{\alpha}_i - \widehat{\boldsymbol{\alpha}}_i\|, \tag{9}$$

and

$$\left|V_i(\boldsymbol{z}, \boldsymbol{\alpha}_i) - \frac{T}{2}\boldsymbol{z}^\intercal I_i^* \boldsymbol{z}\right| \le o(T)\|\boldsymbol{z}\|^2 + O(T)\|\boldsymbol{\alpha}_i - \widehat{\boldsymbol{\alpha}}_i\|\|\boldsymbol{z}\|^2 + O(T)\|\boldsymbol{z}\|^3, \tag{10}$$

uniformly for any $\boldsymbol{\alpha}_i \ge 0$, $\boldsymbol{z} \in U$ with high probability. Above, $\|\cdot\|$ denotes ℓ_2 norm.

To make the confidence set at level $1-\varepsilon$ as small as possible, we can choose $\boldsymbol{z}$ as the following:

Proposition 1. *Let $K = 2D$, and $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{2D}$ be*

$$\pm\sqrt{\frac{2\ln(2D/\varepsilon)}{T\sigma_{ij}^{*2}}} I_i^{*-1}\boldsymbol{e}_j,\ j = 1, \cdots, D, \tag{11}$$

respectively. As $T \to \infty$, for any $j \in [D]$, the width of $\mathcal{C}_{i,\varepsilon}$ in the direction of α_{ij} is $2\sqrt{2\ln(2D/\varepsilon)\sigma_{ij}^{*2}/T}(1+o(1))$ with probability 1.

Note that the width of $\mathcal{C}_{i,\varepsilon}$ is asymptotically a constant times the classic asymptotic confidence intervals.

3.3 Concentration confidence bound with adapted $\boldsymbol{z}$

In reality, we don't have the true parameter $\boldsymbol{\alpha}_i^*$ or the Fisher Information I_i^*. So to make the confidence set as small as possible, we will have to estimate the Fisher Information. The challenge is that we cannot estimate the Fisher Information by simulation, because we don't know the true parameter, and simulation using the MLE will make our choice of $\boldsymbol{z}_k$ depend on the data. What we can do, though, is use data that comes earlier to estimate the Fisher Information I_i^* and the proper choice of $\boldsymbol{z}$ for future data. This leads to our concentration bound with adapted $\boldsymbol{z}$:

Theorem 3.3 (Martingale concentration for score function with adapted $\boldsymbol{z}$)**.** *For any measurable process $(\boldsymbol{z}(t) \in \mathbb{R}^D)_{t=0}^T$ adapted to $(\mathcal{H}_{t^-})_{t=0}^T$, where $(\mathcal{H}_t)_{t=0}^T$ is the filtration of the Hawkes process, any $\varepsilon \in (0,1)$, we have*

$$\Pr\left(\int_0^T \boldsymbol{z}^\intercal(t) dS_{i,t}(\boldsymbol{\alpha}_i^*) - V_i(\boldsymbol{z}, \boldsymbol{\alpha}_i^*) \ge \ln(1/\varepsilon)\right) \le \varepsilon, \tag{12}$$

where

$$dS_{i,t}(\boldsymbol{\alpha}_i) = \lambda_i(t)^{-1}\boldsymbol{\eta}_i(t)(dN_t^i - \lambda_i(t)dt),$$

and

$$\begin{aligned} V_i(\boldsymbol{z}, \boldsymbol{\alpha}_i) &= \int_0^T \log\left(\mathbb{E}\left(\exp\left(\boldsymbol{z}^\intercal(t) dS_{i,t}(\boldsymbol{\alpha}_i)\right)\middle|\mathcal{H}_{t^-}\right)\right) \\ &= \int_0^T \left(\lambda_i(t)\exp\left(\lambda_i^{-1}(t)\boldsymbol{z}^\intercal(t)\boldsymbol{\eta}_i(t)\right) - \boldsymbol{z}^\intercal(t)\boldsymbol{\eta}_i(t) - \lambda_i(t)\right) dt. \end{aligned}$$

Corollary 1 (UQ for each α_{ij})**.** *For any $\boldsymbol{\alpha}_i$, $t \in [0,T]$, let $\hat{I}_i(\boldsymbol{\alpha}_i, t)$ be some estimator for the Fisher Information given data up to time t^-. Let $\boldsymbol{z}_1(t,\boldsymbol{\alpha}_i), \cdots, \boldsymbol{z}_{2D}(t,\boldsymbol{\alpha}_i)$ be*

$$\pm\sqrt{\frac{2\ln(2D/\varepsilon)}{T\boldsymbol{e}_j^\intercal \hat{I}_i^{-1}(\boldsymbol{\alpha}_i,t)\boldsymbol{e}_j}}\hat{I}_i^{-1}(\boldsymbol{\alpha}_i,t)\boldsymbol{e}_j,\ j=1,\cdots,D,$$

respectively. Then

$$\mathcal{C}_{i,\varepsilon} = \left\{\boldsymbol{\alpha}_i \in \mathbb{R}^D : \int_0^T \boldsymbol{z}_k^\intercal(t,\boldsymbol{\alpha}_i)dS_{i,t}(\boldsymbol{\alpha}_i) - V_i(\boldsymbol{z}_k,\boldsymbol{\alpha}_i) \le \ln(2D/\varepsilon), k=1,\cdots,2D\right\}$$

is a confidence set for $\boldsymbol{\alpha}_i$ at level $1-\varepsilon$.

Remark 3. An example of estimator for the Fisher Information is

$$\hat{I}_i(\boldsymbol{\alpha}_i,t) = -\frac{1}{t}\int_0^t \lambda_i^{-2}(\tau)\boldsymbol{\eta}_i(\tau)\boldsymbol{\eta}_i^\intercal(\tau)dN_\tau^i,$$

and if the estimator is rank deficient, we simply take it to be the identity matrix.

For simplicity, we use $g_k(\boldsymbol{\alpha}_i)$ to denote

$$\int_0^T \boldsymbol{z}_k^\intercal(t,\boldsymbol{\alpha}_i)dS_{i,t}(\boldsymbol{\alpha}_i) - V_i(\boldsymbol{z}_k,\boldsymbol{\alpha}_i),\ k=1,\cdots,2D.$$

The CI of entry α_{ij} is then $\{\alpha_{ij} : g_k(\boldsymbol{\alpha}_i) \le \ln(2D/\varepsilon), k=1,\cdots,2D\}$, This CI can be computed by numerically inverting the functions $g_k, k=1,\cdots,2D$, but it may be time-consuming. Since $\widehat{\boldsymbol{\alpha}}_i \to \boldsymbol{\alpha}_i^*$ with probability one when $T\to\infty$, we can approximate $g_k(\boldsymbol{\alpha}_i^*)$ using first order Taylor expansion at $\widehat{\boldsymbol{\alpha}}_i$. Let

$$\tilde{g}_k(\boldsymbol{\alpha}_i) = g_k(\widehat{\boldsymbol{\alpha}}_i) + (\boldsymbol{\alpha}_i - \widehat{\boldsymbol{\alpha}}_i)^\intercal \frac{\partial g_k(\widehat{\boldsymbol{\alpha}}_i)}{\partial \boldsymbol{\alpha}_i},$$

an approximated confidence set for $\boldsymbol{\alpha}_i$ is

$$\mathcal{C}_{i,\varepsilon}^p = \left\{\boldsymbol{\alpha}_i \in \mathbb{R}^D : \tilde{g}_k(\boldsymbol{\alpha}_i) \le \ln(2D/\varepsilon), k=1,\cdots,2D\right\},$$

which is a polyhedron.

With the polyhedron $\mathcal{C}_{i,\varepsilon}^p$, we can easily get CI on each entry α_{ij}

$$\left[\min\{\alpha_{ij} : \boldsymbol{\alpha}_i \in \mathcal{C}_{i,\varepsilon}^p\}, \max\{\alpha_{ij} : \boldsymbol{\alpha}_i \in \mathcal{C}_{i,\varepsilon}^p\}\right].$$

using linear optimization.

Algorithm 1 summarizes how to find the concentration-bound based confidence set.

Algorithm 1: Polyhedral Confidence Set for $\boldsymbol{\alpha}_i$

Input: confidence level $1-\varepsilon$**, data** $\{(t_i,u_i)\}$**, estimator** $\hat{I}_i^{-1}(\cdot,\cdot)$**;**
Compute the MLE $\widehat{\boldsymbol{\alpha}}_i$ **by convex optimization** (6)**;**
for $k=1,\cdots,2D$ **do**

$$g_k(\widehat{\boldsymbol{\alpha}}_i) = \int_0^T dS_{i,t}(\widehat{\boldsymbol{\alpha}}_i)\boldsymbol{z}_k(t,\widehat{\boldsymbol{\alpha}}_i) - V_i(\boldsymbol{z}_k,\widehat{\boldsymbol{\alpha}}_i),$$
$$g_k'(\widehat{\boldsymbol{\alpha}}_i) := \frac{\partial g_k(\widehat{\boldsymbol{\alpha}}_i)}{\partial \boldsymbol{\alpha}_i}.$$

end
Output: $\mathcal{C}_{i,\varepsilon}^p := \left\{\boldsymbol{\alpha}_i \in \mathbb{R}^D : g_k(\widehat{\boldsymbol{\alpha}}_i) + (\boldsymbol{\alpha}_i - \widehat{\boldsymbol{\alpha}}_i)^\intercal g_k'(\widehat{\boldsymbol{\alpha}}_i) \le \ln(2D/\varepsilon), k=1,\cdots,2D\right\}$

4 Numerical Experiment

In this section, we present a numerical example based on synthetic data to demonstrate the performance of the proposed confidence intervals. We compare the coverage ratio of the confidence intervals: the percentage of confidence intervals that contain the true parameters, for the same nominal confidence level $(1-\varepsilon)$.

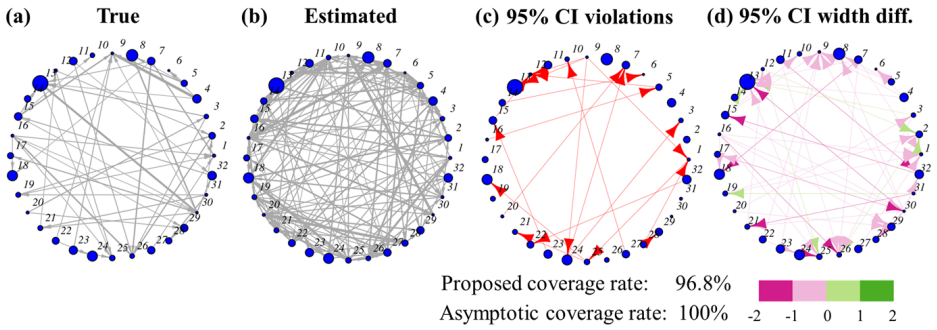

Figure 1: (a) and (b): Visualizing the "true" and estimated influence matrix A as edges and background rates $\boldsymbol{\mu}$ as nodes. Wider edges indicate greater influence, and larger nodes indicate greater background rates. (c): Edges whose 95% CIs do not cover the true influence parameter for the proposed CI method. The coverage rate for the proposed and asymptotic CIs are 96.8% and 100%, respectively. (d): Visualizing the difference in 95% CI widths between the proposed and asymptotic CIs. Purple edges and green edges indicate narrower and wider widths for the proposed CI, respectively.

We study uncertainty quantification for reconstructing neuronal networks. Recent developments in neural engineering have allowed researchers to simultaneously record precise spike train data from large numbers of biological neurons [13]. A key challenge is harnessing this data to learn the connectivity of biological neural networks, which provides insight on the functions of such networks. We show next how the proposed method can quantify the uncertainty of the reconstructed neuronal connectivity from spiking data. This uncertainty is crucial for neuronal reconstruction: it provides a principled statistical framework for testing different neurological theories and hypotheses.

The experimental set-up is as follows. The neural spike train data is simulated via the PyNN Python package [6] with the NEURON simulator [5], which was chosen over in vivo recordings for straightforward data collection. The neuronal network consists of excitatory and inhibitory networks in a ratio of 4 to 1, which are connected sparsely and at random. The neurons are modeled as exponential integrate-and-fire neurons with default parameters, which have been shown to accurately capture biological neural dynamics [3]. Following [4], each excitatory neuron receives a stochastic Poisson process-distributed excitation from an external source, reflecting the external inputs from biological networks either from the environment or from neurons which are not being recorded.

Using the above network structure with $D = 32$ neurons, we simulate a *long* sequence (2000 seconds) of spiking data, and fit a Hawkes network using an exponential influence function with a decay rate of 1 millisecond. This fitted model (with estimates of the influence matrix A and background rate vector $\boldsymbol{\mu}$) can be viewed as the Hawkes network "closest" to the complex neuroscience model which generated the data. The fitted parameters for A and $\boldsymbol{\mu}$ (see Figure 1 (a)) are then set as the "true" parameters for evaluating CI coverage. We then simulate a *shorter* sequence (400 seconds) of spiking data for constructing the proposed (non-asymptotic) and asymptotic CIs on A. Figure 1 (b) shows the MLE of A, estimated using this shorter sequence. Note that, while the connectivity for the "true" topology is quite sparse, the estimated connectivity is noticeably more dense, perhaps due to the limited data in the shorter sequence. In this limited data setting, there is an increasing need for uncertainty quantification to validate neuronal connectivity.

Consider now the coverage performance of the proposed (non-asymptotic) and asymptotic CIs. At a confidence level of 95%, the coverage rate of the proposed method (over all influence parameters in A) is 96.8%, whereas the coverage rate for the asymptotic method is 100%. Hence, the proposed CIs indeed provide similar coverage to the desired confidence level of 95%, whereas the asymptotic CIs are too wide and over-covers the true parameters. Figure 1 shows the edges with influence parameters *not* covered by the proposed method. All of these edges have a true influence of 0, i.e., such edges were not in the true topology, but had positive CIs. Figure 1 visualizes the difference in CI widths between the proposed and asymptotic CIs, for edges with non-zero true influence. Here, purple edges and green edges indicate narrower and wider widths for the proposed CI, respectively. We see that the proposed method yields noticeably narrower CIs compared to the asymptotic approach, which enables more precise inference on the influence matrix. This in turn provides greater certainty on the reconstructed neuronal network, particularly given limited experimental data.

Broader Impact

Our method can be useful for many applications involving Hawkes processes, including seismology, social networks, neuroscience and more. In particular, it is useful for performing causal inference and making statistically significant claims. Recent developments in neuroscience and engineering have allowed researchers to simultaneously record precise spiking data from large numbers of biological neurons. A key challenge is harnessing this experimental data to learn the underlying connectivity of biological neural networks, which is integral for understanding the functions of such networks. We show how the proposed model can be used to both learn this connectivity information and quantify uncertainty from observed neural spike data.

Acknowledgement

This work is partially funded by an NSF CAREER Award CCF-1650913, CMMI-2015787, DMS-1938106, and DMS-1830210.

References

[1] Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering causality from multivariate hawkes integrated cumulants. *The Journal of Machine Learning Research*, 18(1):6998–7025, 2017.

[2] Emmanuel Bacry and Jean-François Muzy. First-and second-order statistics characterization of hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62(4):2184–2202, 2016.

[3] Romain Brette and Wulfram Gerstner. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of neurophysiology*, 94(5):3637–3642, 2005.

[4] Romain Brette, Michelle Rudolph, Ted Carnevale, Michael Hines, David Beeman, James M Bower, Markus Diesmann, Abigail Morrison, Philip H Goodman, Frederick C Harris, et al. Simulation of networks of spiking neurons: a review of tools and strategies. *Journal of computational neuroscience*, 23(3):349–398, 2007.

[5] Nicholas T Carnevale and Michael L Hines. *The NEURON book*. Cambridge University Press, 2006.

[6] Andrew P Davison, Daniel Brüderle, Jochen M Eppler, Jens Kremkow, Eilif Muller, Dejan Pecevski, Laurent Perrinet, and Pierre Yger. Pynn: a common interface for neuronal network simulators. *Frontiers in neuroinformatics*, 2:11, 2009.

[7] Mingzhou Ding, Jue Mo, Charles E Schroeder, and Xiaotong Wen. Analyzing coherent brain networks with granger causality. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5916–5918. IEEE, 2011.

[8] Michael Eichler, Rainer Dahlhaus, and Johannes Dueck. Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242, 2017.

[9] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Exponential line-crossing inequalities. *arXiv preprint arXiv:1808.03204*, 2018.

[10] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*, 2018.

[11] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, Jasjeet Sekhon, et al. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

[12] Anatoli Juditsky, Arkadi Nemirovski, Liyan Xie, and Yao Xie. Convex recovery of marked spatio-temporal point processes. *arXiv preprint arXiv:2003.12935*, 2020.

[13] Ryota Kobayashi, Shuhei Kurita, Anno Kurth, Katsunori Kitano, Kenji Mizuseki, Markus Diesmann, Barry J Richmond, and Shigeru Shinomoto. Reconstructing neuronal circuitry from parallel spike trains. *Nature Communications*, 10(1):1–13, 2019.

[14] Tomasz Kusmierczyk and Manuel Gomez-Rodriguez. On the causal effect of badges. In *Proceedings of the 2018 World Wide Web Conference*, pages 659–668, 2018.

[15] Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[16] Shuang Li, Yao Xie, Mehrdad Farajtabar, Apurv Verma, and Le Song. Detecting changes in dynamic events over networks. *IEEE Transactions on Signal and Information Processing over Networks*, 3(2):346–359, 2017.

[17] Yixing Li, Xingjian Wang, Simon Mak, Chih-Li Sung, C F Jeff Wu, and Vigor Yang. Uncertainty quantification of flame transfer function under a bayesian framework. In *2018 AIAA Aerospace Sciences Meeting*, page 1187, 2018.

[18] Yoshiko Ogata. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30(1):243–261, 1978.

[19] Yun Qian, Charles Jackson, Filippo Giorgi, Ben Booth, Qingyun Duan, Chris Forest, Dave Higdon, Z Jason Hou, and Gabriel Huerta. Uncertainty quantification in climate modeling and projection. *Bulletin of the American Meteorological Society*, 97(5):821–824, 2016.

[20] Jakob Gulddahl Rasmussen. Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, 15(3):623–642, 2013.

[21] Alex Reinhart et al. A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3):299–318, 2018.

[22] Patricia Reynaud-Bouret. Adaptive estimation of the intensity of inhomogeneous poisson processes via concentration inequalities. *Probability Theory and Related Fields*, 126(1):103–153, 2003.

[23] Patricia Reynaud-Bouret, Vincent Rivoirard, and Christine Tuleau-Malot. Inference of functional connectivity in neurosciences via hawkes processes. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 317–320. IEEE, 2013.

[24] Patricia Reynaud-Bouret, Emmanuel Roy, et al. Some non asymptotic tail estimates for hawkes processes. *Bulletin of the Belgian Mathematical Society-Simon Stevin*, 13(5):883–896, 2007.

[25] Marian-Andrei Rizoiu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. Sir-hawkes: linking epidemic models and hawkes processes to model diffusions in finite populations. In *Proceedings of the 2018 World Wide Web Conference*, pages 419–428, 2018.

[26] Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. Design and analysis of computer experiments. *Statistical Science*, pages 409–423, 1989.

[27] Farnood Salehi, William Trouleau, Matthias Grossglauser, and Patrick Thiran. Learning hawkes processes from a handful of events. In *Advances in Neural Information Processing Systems*, pages 12715–12725, 2019.

[28] Ralph C Smith. *Uncertainty Quantification: Theory, Implementation, and Applications*, volume 12. SIAM, 2013.

[29] Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. Learning granger causality for hawkes processes. In *International Conference on Machine Learning*, pages 1717–1726, 2016.

[30] Baichuan Yuan, Hao Li, Andrea L Bertozzi, P Jeffrey Brantingham, and Mason A Porter. Multivariate spatiotemporal hawkes processes and network reconstruction. *SIAM Journal on Mathematics of Data Science*, 1(2):356–382, 2019.

A Example: Exponential decay function

Here we give the analysis for the score function and the Fisher information under exponential decay function $\varphi_{ij}(\Delta t) = \beta e^{-\beta \Delta t}$. The score function is,

$$S_i(\boldsymbol{\alpha}_i^*) = \int_0^T \frac{\int_0^t \beta e^{-\beta(t-\tau)} d\boldsymbol{N}_\tau}{\mu_i + (\boldsymbol{\alpha}_i^*)^\intercal \int_0^t \beta e^{-\beta(t-\tau)} d\boldsymbol{N}_\tau} (dN_t^i - \lambda_i^*(t)dt),$$

where $d\boldsymbol{N}_t = (dN_t^1, \cdots dN_t^D)^\intercal$.

We show that $S_i(\boldsymbol{\alpha}_i^*)$ is small by giving an upper bound of its covariance matrix, which is T times the Fisher information.

Assume the Hawkes process with parameter $\boldsymbol{\alpha}_i, \mu_i, i = 1, \cdots, D, \beta$ is stationary, we have

$$I_i^* = \mathbb{E}\left[\frac{\left(\int_{-\infty}^t \beta e^{-\beta(t-\tau)} d\boldsymbol{N}_\tau\right)\left(\int_{-\infty}^t \beta e^{-\beta(t-\tau)} d\boldsymbol{N}_\tau\right)^\intercal}{\mu_i + (\boldsymbol{\alpha}_i^*)^\intercal \int_{-\infty}^t \beta e^{-\beta(t-\tau)} d\boldsymbol{N}_\tau} \right]$$

Since $\boldsymbol{\alpha}_i^T \int_{-\infty}^t \beta e^{-\beta(t-\tau)} dN_\tau \geq 0$, we have

$$I_i^* \preceq \mu_i^{-1} \underbrace{\mathbb{E}\left[\left(\int_{-\infty}^t \beta e^{-\beta(t-\tau)} d\boldsymbol{N}_\tau\right)\left(\int_{-\infty}^t \beta e^{-\beta(t-\tau)} d\boldsymbol{N}_\tau\right)^\intercal \right]}_{W},$$

where W has a close-form expression for Hawkes processes with exponential influence function, derived from [2] and [16].

Lemma A.1.

$$W = \Lambda\Lambda^\intercal + \frac{\beta}{2}\Sigma + \frac{\beta}{4} A(\mathbb{I} - A)^{-1}\Sigma + \frac{\beta}{4}\Sigma A^\intercal (\mathbb{I} - A^\intercal)^{-1},$$

where $\mathbb{I}$ is the identity matrix, $A = (\boldsymbol{\alpha}_1^*, \cdots, \boldsymbol{\alpha}_D^*)^\intercal$, $\Lambda = (\mathbb{I} - A)^{-1}\boldsymbol{\mu}$ is the expected intensity, and $\Sigma = diag(\Lambda)$.

Proof. By Lemma 2 and 3 in [16], we have

$$\mathbb{E}[d\boldsymbol{N}_t] = \Lambda dt,$$

and

$$\mathrm{Cov}[d\boldsymbol{N}_t, d\boldsymbol{N}_{t'}^\intercal] = c(t - t')dtdt',$$

where

$$c(\tau) = \begin{cases} \beta e^{-\beta(\mathbb{I}-A)\tau} A\left(\mathbb{I} + \frac{1}{2}(\mathbb{I} - A)^{-1}A\right)\Sigma, & \tau > 0; \\ \Sigma\delta(\tau), & \tau = 0; \\ c(-\tau)^\intercal, & \tau < 0, \end{cases}$$

where $\delta(\cdot)$ is the Dirac delta function. Then

$$\begin{aligned} W &= \mathbb{E}\left[\left(\int_{-\infty}^t \beta e^{-\beta(t-\tau)} d\boldsymbol{N}_\tau\right)\left(\int_{-\infty}^t \beta e^{-\beta(t-\tau)} d\boldsymbol{N}_\tau\right)^\intercal \right] \\ &= \mathbb{E}\left[\int_{-\infty}^0 \int_{-\infty}^0 \beta^2 e^{\beta(t+t')} d\boldsymbol{N}_t d\boldsymbol{N}_{t'}^\intercal \right] \\ &= \int_{-\infty}^0 \int_{-\infty}^0 \beta^2 e^{\beta(t+t')} \mathbb{E}[d\boldsymbol{N}_t d\boldsymbol{N}_{t'}] \\ &= \Lambda\Lambda^\intercal + \int_{-\infty}^0 \int_{-\infty}^0 \beta^2 e^{\beta(t+t')} \mathrm{Cov}[d\boldsymbol{N}_t, d\boldsymbol{N}_{t'}] \\ &= \Lambda\Lambda^\intercal + \int_{-\infty}^0 \beta^2 e^{2\beta t}\Sigma dt + \iint_{t\leq 0, \tau \in (0, -t]} \beta^2 e^{\beta(2t+\tau)}(c(\tau) + c(-\tau))dtd\tau \end{aligned}$$

$$
\begin{aligned}
&= \Lambda\Lambda^\intercal + \frac{\beta}{2}\Sigma + \int_0^\infty (c(\tau)+c(-\tau))d\tau \int_{-\infty}^{-\tau} \beta^2 e^{\beta(2t+\tau)} dt \\
&= \Lambda\Lambda^\intercal + \frac{\beta}{2}\Sigma + \int_0^\infty (c(\tau)+c(-\tau))\frac{\beta}{2} e^{-\beta\tau} d\tau \\
&= \Lambda\Lambda^\intercal + \frac{\beta}{2}\Sigma + \int_0^\infty \frac{\beta^2}{2}\Bigg[e^{-\beta(2\mathbb{I}-A)\tau} A(\mathbb{I} + \frac{1}{2}(\mathbb{I}-A)^{-1}A)\Sigma \\
&\qquad\qquad + \left(e^{-\beta(2\mathbb{I}-A)\tau} A(\mathbb{I} + \frac{1}{2}(\mathbb{I}-A)^{-1}A)\Sigma \right)^\intercal \Bigg] d\tau \\
&= \Lambda\Lambda^\intercal + \frac{\beta}{2}\Sigma + \frac{\beta}{2}\Bigg[(2\mathbb{I}-A)^{-1} A(\mathbb{I} + \frac{1}{2}(\mathbb{I}-A)^{-1}A)\Sigma \\
&\qquad\qquad + \left((2\mathbb{I}-A)^{-1} A(\mathbb{I} + \frac{1}{2}(\mathbb{I}-A)^{-1}A)\Sigma \right)^\intercal \Bigg].
\end{aligned}
$$

We notice that

$$
(2\mathbb{I}-A)^{-1}A(\mathbb{I} + \frac{1}{2}(\mathbb{I}-A)^{-1}A) = (2\mathbb{I}-A)^{-1}A(\mathbb{I}-A)^{-1}(\mathbb{I}-A/2) = \frac{1}{2}A(\mathbb{I}-A)^{-1}.
$$

Together, we prove the lemma. □

Lemma A.2. *For any vector $\boldsymbol{z}$,*

$$
P\left(\boldsymbol{z}^\intercal S_i(\boldsymbol{\alpha}_i^*) \geq \varepsilon\sqrt{T}\right) \leq \frac{\mu_i^{-1}\boldsymbol{z}^\intercal W \boldsymbol{z}}{\varepsilon^2}.
$$

Proof.

$$
\mathrm{Var}\left[\boldsymbol{z}^\intercal S_i(\boldsymbol{\alpha}_i^*)\right] = T\boldsymbol{z}^\intercal I_i^* \boldsymbol{z} \leq \mu_i^{-1} T \boldsymbol{z}^\intercal W \boldsymbol{z},
$$

by Markov's inequality on the random variable $(\boldsymbol{z}^\intercal S_i(\boldsymbol{\alpha}_i^*))^2$, we proof the lemma.

□

B Proofs

The proof of Theorem 3.2 and Theorem 3.3 is an immediate results of the following two lemmas.

Lemma B.1. *For any measurable random process $(\boldsymbol{z}(t) \in \mathbb{R}^D)_{t\in[0,T]}$ adapted to the same filtration $(\mathcal{H}_t)_{t\in[0,T]}$ with the Hawkes process, let the intrinsic variance of $\int_0^t \boldsymbol{z}^\intercal(\tau) dS_{i,\tau}(\boldsymbol{\alpha}_i^*)$ (denoted by $V_{i,t}(\boldsymbol{z})$) be a random process also adapted to $(\mathcal{H}_t)_{t\in[0,T]}$, such that there exists a supermartingale $(M_t(\boldsymbol{z}))_{t\in[0,T]}$ with respect to $(\mathcal{H}_t)_{t\in[0,T]}$,*

$$
\exp\left(\int_0^t \boldsymbol{z}^\intercal(\tau) dS_{i,\tau}(\boldsymbol{\alpha}_i^*) - V_{i,t}(\boldsymbol{z}) \right) \leq M_t(\boldsymbol{z})
$$

almost surely. Then $\forall \varepsilon \in (0,1)$,

$$
\Pr\left(\int_0^T \boldsymbol{z}^\intercal(t) dS_{i,t}(\boldsymbol{\alpha}_i^*) - V_{i,T}(\boldsymbol{z}) \geq \ln(\mathbb{E}[M_0(\boldsymbol{z})]/\varepsilon) \right) \leq \varepsilon.
$$

Proof. By the property of a supermartingale, we have

$$
\mathbb{E}\left[\exp\left(\int_0^T \boldsymbol{z}^\intercal(t) dS_{i,t}(\boldsymbol{\alpha}_i^*) - V_{i,T}(\boldsymbol{z}) \right) \right] \leq \mathbb{E}[M_T(\boldsymbol{z})] \leq \mathbb{E}[M_0(\boldsymbol{z})],
$$

and by Markov's inequality,

$$\begin{aligned}
&\Pr\left[\int_0^T \boldsymbol{z}^\intercal(t)dS_{i,t}(\boldsymbol{\alpha}_i^*) - V_{i,T}(\boldsymbol{z}) \geq \ln(\mathbb{E}[M_0(\boldsymbol{z})]/\varepsilon)\right] \\
=\ &\Pr\left[\exp\left(\int_0^T \boldsymbol{z}^\intercal(t)dS_{i,t}(\boldsymbol{\alpha}_i^*) - V_{i,T}(\boldsymbol{z})\right) \geq \mathbb{E}[M_0(\boldsymbol{z})]/\varepsilon\right] \\
\leq\ &\frac{\mathbb{E}\left[\exp\left(\int_0^T \boldsymbol{z}^\intercal(t)dS_{i,t}(\boldsymbol{\alpha}_i^*) - V_{i,T}(\boldsymbol{z})\right)\right]}{\mathbb{E}[M_0(\boldsymbol{z})]/\varepsilon} \leq \varepsilon.
\end{aligned}$$

□

Moreover, the intrinsic variance can be characterized explicitly by the following result.

Lemma B.2. *Let*

$$V_{i,t}(\boldsymbol{z}) = \int_0^t \left(\lambda_i^*(\tau)\exp(\lambda_i^{*-1}(\tau)\boldsymbol{z}^\intercal(\tau)\boldsymbol{\eta}_i(\tau)) - \boldsymbol{z}^\intercal(\tau)\boldsymbol{\eta}_i(\tau) - \lambda_i^*(\tau)\right)d\tau. \tag{13}$$

$$M_t(\boldsymbol{z}) = \exp\left(\int_0^t \boldsymbol{z}^\intercal(\tau)dS_{i,\tau}(\boldsymbol{\alpha}_i^*) - V_{i,t}(\boldsymbol{z})\right)$$

is a supermartingale, with $M_0(\boldsymbol{z}) = 1$ almost surely.

Proof. For any t, since $V_{i,t}$ is continuous and the right derivative exists,

$$\begin{aligned}
&\lim_{\Delta t\to 0^+} \frac{\log\mathbb{E}\left[M_{t+\Delta t}(\boldsymbol{z})/M_t(\boldsymbol{z})|\mathcal{H}_t\right]}{\Delta t} \\
=\ &\lim_{\Delta t\to 0^+} \frac{\log\mathbb{E}\left[\exp\left(\int_t^{t+\Delta t} \boldsymbol{z}^\intercal(\tau)dS_{i,\tau}(\boldsymbol{\alpha}_i^*) - \Delta V_{i,t}(\boldsymbol{z})\right)|\mathcal{H}_t\right]}{\Delta t} \\
=\ &\lim_{\Delta t\to 0^+} \frac{\log\mathbb{E}\left[\exp\left(\boldsymbol{z}^\intercal(t)\boldsymbol{\eta}_i(t)(\lambda_i^{*-1}\Delta N_t^i - \Delta t\right)|\mathcal{H}_t\right]}{\Delta t} - (\lambda_i^*\exp(\lambda_i^{*-1}\boldsymbol{z}^\intercal(t)\boldsymbol{\eta}_i(t)) - \boldsymbol{z}^\intercal(t)\boldsymbol{\eta}_i(t) - \lambda_i^*(t)) \\
=\ &\lim_{\Delta t\to 0^+} \frac{\log\left(\lambda_i^*(t)\Delta t\exp(\lambda_i^{*-1}\boldsymbol{z}^\intercal(t)\boldsymbol{\eta}_i(t)) + (1-\lambda_i^*(t)\Delta t)\exp(-\boldsymbol{z}^\intercal(t)\boldsymbol{\eta}_i(t)\Delta t)\right)}{\Delta t} \\
&- (\lambda_i^*\exp(\lambda_i^{*-1}\boldsymbol{z}^\intercal(t)\boldsymbol{\eta}_i(t)) - \boldsymbol{z}^\intercal(t)\boldsymbol{\eta}_i(t) - \lambda_i^*(t)) \\
=\ &0.
\end{aligned}$$

From this we can see that M_t is actually a martingale. □

Proof of Theorem 3.2, Theorem 3.3 and Corollary 1. From Lemma B.1 and Lemma B.2, we have immediately

$$\Pr\left[\int_0^T \boldsymbol{z}^\intercal(t)dS_{i,t}(\boldsymbol{\alpha}_i^*) - V_{i,T}(\boldsymbol{z}) \geq \ln(1/\varepsilon)\right] \leq \varepsilon,\ \forall \boldsymbol{z}\in\mathbb{R}^D, \forall\varepsilon\in(0,1), \tag{14}$$

where $V_{i,T}(\boldsymbol{z})$ is chosen as (13). Moreover, we can also choose multiple $\boldsymbol{z}$ to bound $S_i(\boldsymbol{\alpha}_i^*)$ in all directions. By simple union bound, it holds that

$$\Pr\left[\exists k\in[K], \int_0^T \boldsymbol{z}_k^\intercal(t)dS_{i,t}(\boldsymbol{\alpha}_i^*) - V_{i,T}(\boldsymbol{z}_k) \geq \ln(K/\varepsilon)\right] \leq K\varepsilon/K = \varepsilon,\ \forall \boldsymbol{z}_1,\cdots,\boldsymbol{z}_K\in\mathbb{R}^D, \forall\varepsilon\in(0,1).$$

We define continuous process $V_{i,t}$ for any $\boldsymbol{\alpha}_i$ as (7), $\boldsymbol{\alpha}_i^*$ falls into the confidence set $\mathcal{C}_{i,\varepsilon}$ with probability at least $1-\varepsilon$. □

The proof of Lemma 3.1 relies on the following lemma:

Lemma B.3 (Ogata [18], Lemma 2)**.** *If ξ_t is a stationary predictable process, then*

$$\frac{1}{T}\int_0^T \xi_t dt \to \mathbb{E}\left[\xi\right]$$

with probability 1. In addition, if ξ_t has finite second moment, then

$$\frac{1}{T}\int_0^T \xi_t \frac{dN_t}{\lambda(t)} \to \mathbb{E}\left[\xi\right]$$

with probability 1.

Proof of Lemma 3.1. Denote

$$\Delta\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_i - \hat{\boldsymbol{\alpha}}_i.$$

Using Taylor expansion and based on the mean value theorem, there exists $\tilde{\boldsymbol{\alpha}}_i$ between $\boldsymbol{\alpha}_i$ and $\hat{\boldsymbol{\alpha}}_i$, such that

$$\frac{1}{T}S_i(\boldsymbol{\alpha}_i) = \frac{1}{T}H_i(\hat{\boldsymbol{\alpha}}_i)\Delta\boldsymbol{\alpha}_i + \frac{1}{2T}\sum_{k,l\in[D]}\frac{\partial^2 S_i(\tilde{\boldsymbol{\alpha}}_i)}{\partial\alpha_{ik}\partial\alpha_{il}}\Delta\alpha_{ik}\Delta\alpha_{il}.$$

For any $j,k,l\in[D]$, the j-th entry of $\frac{\partial^2 S_i(\tilde{\boldsymbol{\alpha}}_i)}{\partial\alpha_{ik}\partial\alpha_{il}}$ is

$$\int_0^T \frac{2\eta_{ij}(t)\eta_{ik}(t)\eta_{il}(t)}{\tilde{\lambda}_i^3(t)}dN_t^i.$$

Under the assumption that the moment generating function of $\boldsymbol{\eta}_i$ exists, and $\tilde{\lambda}_i \geq \mu_i > 0$, by Lemma B.3, as $T\to\infty$,

$$\frac{1}{T}\int_0^T \frac{2\eta_{ij}(t)\eta_{ik}(t)\eta_{il}(t)}{\tilde{\lambda}_i^3(t)}dN_t^i \leq \frac{1}{T\mu_i^3}\int_0^T 2\eta_{ij}(t)\eta_{ik}(t)\eta_{il}(t)dN_t^i$$

is uniformly bounded for any $\tilde{\boldsymbol{\alpha}}_i \geq 0$ with probability 1. Now we have for any $\boldsymbol{\alpha}_i$,

$$\|S_i(\boldsymbol{\alpha}_i) - H_i(\hat{\boldsymbol{\alpha}}_i)\Delta\boldsymbol{\alpha}_i\| \leq O(T)\|\Delta\boldsymbol{\alpha}_i\|^2.$$

Since

$$\frac{1}{T}H_i(\hat{\boldsymbol{\alpha}}_i) \to I_i^*$$

with probability 1, we have

$$\|S_i(\boldsymbol{\alpha}_i) - TI_i^*\Delta\boldsymbol{\alpha}_i\| \leq O(T)\|\Delta\boldsymbol{\alpha}_i\|^2 + \|(H_i(\hat{\boldsymbol{\alpha}}_i) - TI_i^*)\Delta\boldsymbol{\alpha}_i\| \leq O(T)\|\Delta\boldsymbol{\alpha}_i\|^2 + o(T)\|\Delta\boldsymbol{\alpha}_i\|,$$

which is (9).

For (10), similarly we use the Taylor expansion at $\boldsymbol{z}=0$ and $\hat{\boldsymbol{\alpha}}_i$, by the mean value theorem,

$$\begin{aligned}
V_i(\boldsymbol{z},\boldsymbol{\alpha}_i) &= V_i(\mathbf{0},\boldsymbol{\alpha}_i) + \frac{\partial V_i(\mathbf{0},\boldsymbol{\alpha}_i)}{\partial\boldsymbol{z}^\intercal}\boldsymbol{z} + \frac{1}{2}\boldsymbol{z}^\intercal\frac{\partial^2 V_i(\mathbf{0},\boldsymbol{\alpha}_i)}{\partial\boldsymbol{z}\partial\boldsymbol{z}^\intercal}\boldsymbol{z} + \frac{1}{6}\sum_{j,k,l\in[D]}\frac{\partial^3 V_i(\tilde{\boldsymbol{z}},\boldsymbol{\alpha}_i)}{\partial z_j\partial z_k\partial z_l}z_j z_k z_l \\
&= \frac{1}{2}\boldsymbol{z}^\intercal\frac{\partial^2 V_i(\mathbf{0},\boldsymbol{\alpha}_i)}{\partial\boldsymbol{z}\partial\boldsymbol{z}^\intercal}\boldsymbol{z} + \frac{1}{6}\sum_{j,k,l\in[D]}\frac{\partial^3 V_i(\tilde{\boldsymbol{z}},\boldsymbol{\alpha}_i)}{\partial z_j\partial z_k\partial z_l}z_j z_k z_l \\
&= \frac{1}{2}\boldsymbol{z}^\intercal\frac{\partial^2 V_i(\mathbf{0},\hat{\boldsymbol{\alpha}}_i)}{\partial\boldsymbol{z}\partial\boldsymbol{z}^\intercal}\boldsymbol{z} + \frac{1}{2}\sum_{j,k,l\in[D]}\frac{\partial^3 V_i(\mathbf{0},\tilde{\boldsymbol{\alpha}}_i)}{\partial z_j\partial z_k\partial\alpha_{il}}z_j z_k\Delta\alpha_{il} + \frac{1}{6}\sum_{j,k,l\in[D]}\frac{\partial^3 V_i(\tilde{\boldsymbol{z}},\boldsymbol{\alpha}_i)}{\partial z_j\partial z_k\partial z_l}z_j z_k z_l,
\end{aligned}$$

for some $\tilde{\boldsymbol{z}}$ between $\mathbf{0},\boldsymbol{z}$, some $\tilde{\boldsymbol{\alpha}}_i$ between $\hat{\boldsymbol{\alpha}}_i,\boldsymbol{\alpha}_i$. By the assumption that the moment generating function of $\boldsymbol{\eta}_i$ exists and by Lemma B.3, for the first term

$$\frac{1}{2T}\boldsymbol{z}^\intercal\frac{\partial^2 V_i(\mathbf{0},\hat{\boldsymbol{\alpha}}_i)}{\partial\boldsymbol{z}\partial\boldsymbol{z}^\intercal}\boldsymbol{z} = \frac{1}{2T}\boldsymbol{z}^\intercal\int_0^T\frac{\boldsymbol{\eta}_i(t)\boldsymbol{\eta}_i(t)^\intercal}{\hat{\lambda}_i(t)}dt\boldsymbol{z} \to \frac{\boldsymbol{z}^\intercal I_i^*\boldsymbol{z}}{2}$$

with probability 1. For the second term, for any $j,k,l\in[D]$,

$$\left|\frac{1}{T}\frac{\partial^3 V_i(\mathbf{0},\tilde{\boldsymbol{\alpha}}_i)}{\partial z_j\partial z_k\partial\alpha_{il}}\right| = \left|\frac{1}{T}\int_0^T\frac{\eta_{ij}(t)\eta_{ik}(t)\eta_{il}(t)}{\tilde{\lambda}_i^2(t)}dt\right| \leq \left|\frac{1}{T}\int_0^T\frac{\eta_{ij}(t)\eta_{ik}(t)\eta_{il}(t)}{\mu_i^2}dt\right|.$$

is uniformly bounded for any $\tilde{\boldsymbol{\alpha}}_i \geq 0$ with probability 1. For the third term, for any $i, j, k \in [D]$,

$$\left|\frac{1}{T}\frac{\partial^3 V_i(\tilde{\boldsymbol{z}}, \boldsymbol{\alpha}_i)}{\partial z_j \partial z_k \partial z_l}\right| = \left|\frac{1}{T}\int_0^T \frac{\eta_{ij}(t)\eta_{ik}(t)\eta_{il}(t)}{\lambda_i^2(t)} \exp\left(\lambda_i^{-1}\boldsymbol{\eta}_i^\intercal(t)\tilde{\boldsymbol{z}}\right) dt\right|$$
$$\leq \left|\frac{1}{T}\int_0^T \frac{\eta_{ij}(t)\eta_{ik}(t)\eta_{il}(t)}{\mu_i^2} \min\left\{\exp\left(\mu_i^{-1}\boldsymbol{\eta}_i^\intercal(t)\tilde{\boldsymbol{z}}\right), 1\right\} dt\right|,$$

is convex in $\boldsymbol{z}$. There exists a neighborhood U of $\mathbf{0}$ such that the expectation of the term above for any $\boldsymbol{z} \in U$ is finite, and by its convexity, it is uniformly bounded in U with probability 1.

Together, we have

$$\left|V_i(\boldsymbol{z}, \boldsymbol{\alpha}_i) - \frac{T}{2}\boldsymbol{z}^\intercal I_i^* \boldsymbol{z}\right| \leq o(T)\|\boldsymbol{z}\|^2 + O(T)\|\Delta\boldsymbol{\alpha}_i\|\|\boldsymbol{z}\|^2 + O(T)\|\boldsymbol{z}\|^3.$$

□

Proof of Proposition 1. We prove a slightly weaker version: for any neighborhood U_1 of $\boldsymbol{\alpha}_i^*$, such that the diameter of U_1 is $o(1)$, the width of $\mathcal{C}_{i,\varepsilon} \cap U_1$ in α_{ij} converges to $2\sqrt{2\ln(K/\varepsilon)\sigma_{ij}^2/T}$ with probability 1.

Before proving the proposition, we explain why we choose $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_K$ this way. Let $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{2D}$ be $\pm c_j I_i^{*-1}\boldsymbol{e}_j$, $c_j > 0$, $j = 1, \cdots, D$. By Lemma 3.1, we have

$$(\pm c_j I_i^{*-1}\boldsymbol{e}_j)^\intercal S_i(\boldsymbol{\alpha}_i) = (\pm c_j I_i^{*-1}\boldsymbol{e}_j)^\intercal T I_i^*(\boldsymbol{\alpha}_i - \hat{\boldsymbol{\alpha}}_i) + c_j\left(O(T)\|\boldsymbol{\alpha}_i - \hat{\boldsymbol{\alpha}}_i\|^2 + o(T)\|\boldsymbol{\alpha}_i - \hat{\boldsymbol{\alpha}}_i\|\right),$$

for any $\boldsymbol{\alpha}_i \geq 0$. Note that

$$(\pm c_j I_i^{*-1}\boldsymbol{e}_j)^\intercal T I_i^*(\boldsymbol{\alpha}_i - \hat{\boldsymbol{\alpha}}_i) = \pm c_j T(\alpha_{ij} - \hat{\alpha}_{ij}).$$

By (10), we have

$$V_i(\pm c_j I_i^{*-1}\boldsymbol{e}_j, \boldsymbol{\alpha}_i) = \frac{c^2 T}{2}\boldsymbol{e}_j^\intercal I_i^{*-1}\boldsymbol{e}_j + c^2\left(o(T) + O(T)\|\boldsymbol{\alpha}_i - \hat{\boldsymbol{\alpha}}_i\|\right) + O(T)c^3.$$

The constraints

$$\boldsymbol{z}_k^\intercal S_i(\boldsymbol{\alpha}_i) - V_i(\boldsymbol{z}_k, \boldsymbol{\alpha}_i) \leq \ln(K/\varepsilon), \quad k = 1, \cdots, 2D$$

becomes

$$c_j T|\alpha_{ij} - \hat{\alpha}_{ij}| - \frac{c_j^2 T}{2}\sigma_{ij}^2 + o(T)c_j^2 + O(T)c_j^3 + (O(T)c_j^2 + o(T)c_j)\|\boldsymbol{\alpha}_i - \hat{\boldsymbol{\alpha}}_i\| + O(T)c_j\|\boldsymbol{\alpha}_i - \hat{\boldsymbol{\alpha}}_i\|^2$$
$$\leq \ln(K/\varepsilon), \quad j = 1, \cdots, D.$$

If all the $o(\cdot), O(\cdot)$ terms are negligible when $T \to \infty$, the width of $\mathcal{C}_{i,\varepsilon}$ in α_{ij} is

$$2\left(\frac{\ln(K/\varepsilon)}{c_j T} + \frac{c_j\sigma_{ij}^2}{2}\right),$$

and is minimized when

$$c_j = \sqrt{\frac{2\ln(K/\varepsilon)}{T\sigma_{ij}^2}}.$$

The $o(\cdot), O(\cdot)$ terms are indeed negligible with this choice of c_j, because the constraints now becomes

$$\sqrt{2T\ln(K/\varepsilon)/\sigma_{ij}^2}|\alpha_{ij} - \hat{\alpha}_{ij}| + o(T^{1/2})\|\boldsymbol{\alpha}_i - \hat{\boldsymbol{\alpha}}_i\| + O(T^{1/2})\|\boldsymbol{\alpha}_i - \hat{\boldsymbol{\alpha}}_i\|^2 \leq 2\ln(K/\varepsilon) + o(1), \quad (15)$$

$j = 1, \cdots, D$. Let U_1 be any neighborhood of $\boldsymbol{\alpha}_i^*$ with diameter $o(1)$. For any $\boldsymbol{\alpha}_i \in \mathcal{C}_{i,\varepsilon} \cap U_1$, we choose

$$j' = \arg\max_{j \in [D]} |\alpha_{ij} - \hat{\alpha}_{ij}|/\sigma_{ij}.$$

By the way we choose j', $\|\boldsymbol{\alpha}_i - \hat{\boldsymbol{\alpha}}_i\|$ can be upper bounded by $|\alpha_{ij'} - \hat{\alpha}_{ij'}|$ up to some constant scale, and $|\alpha_{ij'} - \hat{\alpha}_{ij'}| = o(1)$. There is

$$\sqrt{2T\ln(K/\varepsilon)/\sigma_{ij'}^2}|\alpha_{ij'} - \hat{\alpha}_{ij'}| + o(T^{1/2})|\alpha_{ij'} - \hat{\alpha}_{ij'}| \leq 2\ln(K/\varepsilon) + o(1),$$

and

$$\frac{|\alpha_{ij'} - \hat{\alpha}_{ij'}|}{\sigma_{ij'}} \leq \sqrt{\frac{2\ln(K/\varepsilon)}{T}}(1 + o(1)).$$

Again by the way we choose j', this inequality holds for any $j \in [D]$. So the width of $\mathcal{C}_{i,\varepsilon}$ in α_{ij} is upper bounded by $2\sqrt{2\ln(K/\varepsilon)\sigma_{ij}^2/T}(1 + o(1))$ with high probability. It is easy to see from (15) that there exists $\boldsymbol{\alpha}_i \in \mathcal{C}_{i,\varepsilon}$ with $\alpha_{ij} = \hat{\alpha}_{ij} \pm \sqrt{2\ln(K/\varepsilon)\sigma_{ij}^2/T}(1 - o(1))$. Together, we know that the width of $\mathcal{C}_{i,\varepsilon}$ converges to $2\sqrt{2\ln(K/\varepsilon)\sigma_{ij}^2/T}$ with probability 1. □