An Explicit Characterization of Consistency for Regularized MLEs of Multivariate Hawkes Processes

Xin Guo * Anran Hu † Renyuan Xu ‡ Junzi Zhang §

January 23, 2019

Abstract

Multivariate Hawkes process (MHP) is a counting process. Its popularity comes from its mutual exciting property which has been widely exploited to model the causality in real-world events and to measure how occurrence of one event triggers the arrival of others. In practice, in order to capture desirable patterns and structures such as sparsity and low-rankness in social network and crime patterns, regularized maximum likelihood estimators (MLEs) have been adopted for empirical studies. This paper builds the theoretical foundation by establishing the consistency property of three variants of regularized MLEs for a general class of MHPs. Our approach combines the intrinsic connection between MHPs and branching processes and the proof techniques developed by [Oga78] for the vanilla MLEs of point processes.

1 Introduction

MHPs. Multivariate Hawkes processes (MHPs) are counting/point processes originally introduced in [Haw71a, Haw71b] to model the causality between multiple events. Since then, MHPs have been applied to model a wide range of real-world phenomena, including the arrival patterns of earthquakes and the aftershocks triggered by the earthquakes in seismology ([Oga88, Oga98]), the mutual interactions of different users and posts in social networks [BBH12], the limit order book transactions in algorithm trading and finance [GZZ18], the neural activity and genomics in biology [BM96], as well as the crime events in criminology [YEHK17]. The popularity of the MHP comes from its mutual exciting property: in social...
networks, it models how activities of one user affect activities of other users; in financial markets, it captures the herd behavior of trading activities.

**MLEs for MHPs in practice.** The standard parameter estimation approach for MHPs is the maximum likelihood estimator (MLE). To capture desired patterns and structures such as sparsity and low-rankness and to improve stability, regularized MLEs have been proposed by adding a penalty term to the standard MLEs. In [ZZS13a], $l_1$ and nuclear norm penalties were added to the likelihood function to enforce sparsity and low-rankness in MHPs, with significant improvement using both synthetic and real social network datasets. Similarly, in [XFZ16], an $l_{1,2}$-norm penalty was proposed to obtain group sparsity in MHPs, with similar performance improvements. In [YEHK17], a Tikhonov (quadratic) regularization term was added to improve stability, creating a state-of-the-art online algorithm for computing MLEs, with successful applications to news agencies impact modeling and crime pattern inference. Regularization was also added to nonparametric MLEs in [ZZS13b] with success. Despite the empirical evidence, corresponding theoretical studies on the consistency of regularized MLEs for MHPs are virtually non-existent.

There are further computational and statistical issues when applying regularized MLEs. One issue is that in general regularized MLEs can only be computed numerically with approximation. Another issue is the issue of missing data, where samples are incomplete due to problems in the process of data collection or data transmission errors [TSL18]. Several algorithms have been proposed to address the problem on the empirical side [Le18, SQS18], yet again with no theoretical analysis.

**Our work.** In this paper, we take the first step towards building the theoretical foundation for the regularized MLE, by establishing the consistency property for three variants of the regularized MLEs for a general class of MHPs: regularized MLEs, approximate regularized MLEs, and regularized MLEs for missing data. Our proof techniques are inspired by the deep work of [Oga78], and extensively exploits the intriguing connection between MHPs and the branching processes.

The major contribution of our work is two-fold.

- **General regularization.** We provide the first theoretical analysis for regularized MLEs in the context of MHPs. In particular, we establish a consistency result for both exact and approximate MLEs as well as MLEs for missing data, with an arbitrary continuous regularization. As a by-product, we also provide some novel tools and techniques, including the first higher-order moment bounds on general linear MHPs.

- **Explicit characterization.** We propose a set explicit and verifiable assumptions on the deterministic components that define an MHP, which are completely different from the abstract assumptions proposed in the literature ([Oga78, PT86]), while not restricted to deal with specific types of MHPs (e.g., the Markovian [GS18] and exponential [CY17] cases). We also discuss through concrete examples how the assumptions can be ensured
and verified, which provides practitioners with guidance on when and why an MHP model is learnable through regularized MLEs.

**Related works.** In addition to the well-known work of [Oga78] on the consistency of MLEs of general point processes, [PT86] and [GS18] also studied the consistency of MLEs for some classes of point processes. Moreover, [CY17] established the convergence of moments and hence asymptotic normality and consistency of MLEs for exponential MHPs. On the other hand, [YEHK17] established bounds on the likelihood regret (instead of the convergence of the MLEs). A detailed comparison with these works is summarized in Table 1.

<table>
<thead>
<tr>
<th>Literature</th>
<th>Main assumptions</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Oga78]</td>
<td>moment bounds on stochastic intensity processes $\lambda$, $\hat{\lambda}$ and their derivatives, no regularization</td>
<td>consistency and asymptotic normality and efficiency for general point processes</td>
</tr>
<tr>
<td>[PT86]</td>
<td>similar to [Oga78], and a.s. bounded $\lambda$, no regularization</td>
<td>consistency and asymptotic normality for general point processes</td>
</tr>
<tr>
<td>[YEHK17]</td>
<td>a.s. bounded $\lambda$, linear basis combination (RKHC) triggering functions, quadratic regularization</td>
<td>finite sample bound on likelihood regret for an online variant of general (linear) MHP MLE</td>
</tr>
<tr>
<td>[CY17]</td>
<td>exponential triggering functions with some mild regularity conditions, no regularization</td>
<td>convergence of moments ($\Rightarrow$ consistency + asymptotic normality)</td>
</tr>
<tr>
<td>[GS18]</td>
<td>Markovian $\lambda$, no regularization</td>
<td>asymptotic normality ($\Rightarrow$ consistency) of a filtered variant of MLE</td>
</tr>
<tr>
<td>This work</td>
<td>all assumptions specified on the deterministic components $\mu_i$ and $g_{ij}$, general regularization</td>
<td>consistency for general (linear) MHP MLE</td>
</tr>
</tbody>
</table>

**Table 1:** Comparison of existing literature on MLEs for MHPs

The connection between MHPs and branching process first appeared in [HO74] and was ingeniously adapted by [JHR15] to study the higher order cumulants of MHPs.

Apart from the MLE, there is also the least-squares estimator (LSE) (see for example [BGM15]) and moment-matching estimators [WBNL16, ABG17]. Besides the parametric approach, the nonparametric approach has been reported in [Hal12, Ras13, BM14a, LA14, DFA15, CH16, Kir16, EDD17]. Incidentally, [BGM15] reports the empirical superiority of MLE over LSE for MHPs, which also serves as a motivation for seeking an explicit characterization of the MLE consistency of MHPs in our paper.

**Outline of the paper.** After introducing three different version of regularized MLEs (Section 2), appropriate model assumptions and examples of MHPs are given before the main results on their consistency properties (Section 3), which are then followed by the illustration of proof ideas and key technical lemmas, as well as the outline of the proof for the main
results (Section 4). Detailed proofs for all the lemmas and the main theorems are given in the appendix due to space limit.

2 Problem settings

MHPs and intensity processes. A \(K\)-dimensional MHP is a \(K\)-dimensional point process \(\mathbf{N} = (N_1, \cdots, N_K)\). Here each \(N_i(a, b]\) is a one-dimensional point process counting the number of arrivals of the \(i\)-th type within the interval \((a, b]\). For notational simplicity, we also use \(N_i(t) := N_i(0, t]\) in some cases.

A point process such as the MHP is characterized by its intensity process which represents the instantaneous likelihood of event arrivals given the history. Here the intensity process \(\lambda_i(t)\) for \(N_i(t)\), is defined as

\[
\lambda_i(t) = \lim_{\delta \to 0} \frac{1}{\delta} \mathbb{P}(N_i(t, t + \delta] > 0 | \mathcal{F}_t),
\]

where the filtration \(\mathcal{F}_t\), the history up to time \(t\), is defined to be the \(\sigma\)-field generated by \(\{\mathbf{N}(-\infty, s] | s \leq t\}\).

The intensity process \(\lambda_i(t)\) of the MHP has two components: the baseline intensity and the triggering functions. The former is similar to the intensity for a simple Poisson process, and the latter carry the mutual exciting property and measure how much the occurrence of one event “triggers” the arrival of others.

There are two standard definitions of intensity processes for MHPs in the literature [Lin09].

The first definition is from the original work of [Haw71a, Haw71b], where the intensity process \(\mathbf{\lambda} = (\lambda_1, \cdots, \lambda_K)\) is given as the following. For each \(i = 1, \cdots, K\),

\[
\lambda_i(t; \theta) = \mu_i + \sum_{j=1}^{K} \int_{-\infty}^{t} g_{ij}(t - s; \eta_j) N_j(ds).
\] (2.1)

Here \(\mu_i\) is the baseline intensity of \(N_i\), and \(g_{ij} : \mathbb{R}^+ \times \mathbb{R}^D \to \mathbb{R}\) is the triggering function that captures the mutual excitation or the causality between \(N_i\) and \(N_j\) for \(i \neq j\) and the self excitation for \(i = j\). When the arrivals time of \(N_i(t)\) is specified by \(\{t_i^n\}\), the above definition can be rewritten as

\[
\lambda_i(t; \theta) = \mu_i + \sum_{j=1}^{K} \sum_{t_j^n < t} g_{ij}(t - t_j^n; \eta_j).
\] (2.2)

This expression explicitly characterizes the impact of one event on the future event: the arrival of event type \(j\) at time \(t_j^n\) increases the intensity by the amount of \(g_{ij}(t - t_j^n; \eta_j)\). Generally, \(\theta = (\mu, \eta) = (\mu_1, \cdots, \mu_K, \eta_1, \cdots, \eta_D) \in \Theta \subseteq \mathbb{R}^{K+D}\) is referred to as the underlying parameter for the MHP.
The second definition is used in most of the engineering literature, where the intensity process \( \hat{\lambda} = (\hat{\lambda}_1, \cdots, \hat{\lambda}_K) \) takes the following form

\[
\hat{\lambda}_i(t; \theta) = \mu_i + \sum_{j=1}^{K} \int_{0}^{t} g_{ij}(t - s; \eta) N_j(ds).
\] (2.3)

Note that the first definition (2.1) is infeasible for parameter estimations, as it needs samples from an infinite time interval \((-\infty, t)\). However, this intensity process has a nice stationary property under mild conditions (to be specified in Section 3) on \( g_{ij} \) and \( \mu_i \).

The second definition involves only samples over the interval \([0, t)\), and is thus computationally tractable for parameter estimations in practice. However, this intensity process is nonstationary. This stationarity property and (lack of it) is critical for parameter estimation of MHP (Section 3).

**Maximum likelihood estimators (MLEs).** The log-likelihood functions for the intensity processes (2.1) and (2.3) on the interval \([0, T]\) take the following respective forms,

\[
L_T(\theta) := \sum_{i=1}^{K} \left[ -\int_{0}^{T} \lambda_i(t; \theta) dt + \int_{0}^{T} \log(\lambda_i(t; \theta)) N_i(dt) \right],
\]

\[
\hat{L}_T(\theta) := \sum_{i=1}^{K} \left[ -\int_{0}^{T} \hat{\lambda}_i(t; \theta) dt + \int_{0}^{T} \log(\hat{\lambda}_i(t; \theta)) N_i(dt) \right].
\]

Their respective MLEs are

\[
\theta_T := \arg\max_{\theta \in \Theta} L_T(\theta), \quad \hat{\theta}_T := \arg\max_{\theta \in \Theta} \hat{L}_T(\theta).
\] (2.4)

As mentioned earlier, the integrand \( \lambda_i(t; \theta) \) is infeasible for parameter estimation. We will thus focus on the consistency property of the regularized MLE for \( \hat{L}_T(\theta) \). Despite the nonstationarity of \( \hat{\lambda} \), the consistency property for \( \hat{\theta}_T \) can be established by exploiting the closeness between \( \lambda_i(t; \theta) \) and \( \hat{\lambda}_i(t; \theta) \), and the stationarity property of \( \lambda_i(t; \theta) \).

To address the consistency issue of regularized MLEs, we first define the regularized MLE as follows.

**Definition 1** (regularized MLE). \( \hat{\theta}^{reg}_T \) is called a regularized MLE if

\[
\hat{\theta}^{reg}_T := \arg\max_{\theta \in \Theta} \hat{L}_T(\theta) - P(\theta) = \arg\max_{\theta \in \Theta} \hat{L}^{reg}_T(\theta),
\] (2.5)

where \( P(\theta) \) is a regularization term.

Throughout the paper, \( P(\theta) \) is assumed to be continuous, which is inclusive of all commonly used regularization terms, such as \( l_1, l_2, l_{1,2}, l_1 + \) nuclear norm, and SCAD. Notice that here \( P \) is assumed to remain unchanged as \( T \) increases, which leads to a vanishing penalty with order \( O(1/T) \) compared to the scaled log-likelihood \( \hat{L}_T/T \). This is also consistent with the penalty scaling used in [BGM15] up to logarithmic terms.

To study the consistency issue in approximating the regularized MLE, we define \( \hat{\theta}^{approx}_T \).
Definition 2 (Approximate regularized MLE). \( \hat{\theta}_T^{\text{approx}} \) is called an \( \epsilon \)-approximation of \( \hat{\theta}_T^{\text{reg}} \) if 
\[
\hat{L}_T^{\text{reg}} \left( \hat{\theta}_T^{\text{approx}} \right) > \max_{\theta \in \Theta} L_T^{\text{reg}}(\theta) - \epsilon.
\]

Finally, to analyze the regularized MLEs for missing data, suppose that the recorded arrivals of events form a sequence of disjoint closed intervals \( \{[l_n, u_n]\}_{n=0}^{\infty} \), where \( l_0 = 0 \) and \( l_n \leq u_n < l_{n+1} \leq u_{n+1} \) for all \( n \geq 0 \), and the arrival records are complete within intervals \([l_n, u_n]\), while records are missing in between the intervals (i.e., in \([u_n, l_{n+1}]\) for \( n \geq 0 \)). Then we propose the following definition.

Definition 3 (regularized MLE for missing data). \( \tilde{\theta}_N^{\text{reg}} \) is called a regularized MLE for missing data if 
\[
\tilde{\theta}_N^{\text{reg}} := \arg\max_{\theta \in \Theta} \tilde{L}_N(\theta) - P(\theta) = \arg\max_{\theta \in \Theta} \tilde{L}_N^{\text{reg}}(\theta).
\]

Here the log-likelihood 
\[
\tilde{L}_N(\theta) := \sum_{i=1}^{K} \left[ - \int_{I_n} \tilde{\lambda}_i(t; \theta) dt + \int_{I_n} \log(\tilde{\lambda}_i(t; \theta)) N_i(dt) \right],
\]
with the truncated intensity process 
\[
\tilde{\lambda}_i(t; \theta) := \mu_i + \sum_{j=1}^{K} \int_{l_n}^{t-} g_{ij}(t-s; \eta) N_j(ds), \ t \in I_n,
\]
with \( I_n \subseteq [l_n, u_n] \).

Notice that the definition of \( \tilde{L}_N(\theta) \) only uses \( \tilde{\lambda}_i(t; \theta) \) on \( \bigcup_{n=0}^{\infty} I_n \). This is because the truncated intensity process is only “trustworthy” on the set of \( t \) with relatively sufficient data (denoted as \( I_n \)). The exact choice of \( I_n \) to guarantee consistency of \( \tilde{\theta}_N^{\text{reg}} \) is given in Section 3.

3 Assumptions and consistency results

In this section, we will establish the consistency property for the three regularized MLEs, i.e., their convergence in probability to the (unknown) true parameter, denoted as \( \theta^* = (\mu^*, \eta^*) \).

3.1 General assumptions

We begin by introducing the assumptions that will be used throughout the paper for establishing the consistency of MLEs. For simplicity, we denote \( \Theta_{\eta} := \{ \eta \mid \exists \mu, (\mu, \eta) \in \Theta \} \), and \( \Theta_{\mu} := \{ \mu \mid \exists \eta, (\mu, \eta) \in \Theta \} \). We will also sometimes use the shorthand \( \lambda_i(t) := \lambda_i(t; \theta^*) \) to denote the intensity process corresponding to the underlying true MHP \( N \).
Assumption 1 (Regularity). \( \Theta \) is nonempty and compact. \( \exists \mu > 0 \), s.t. for any \( \mu = (\mu_1, \ldots, \mu_K) \in \Theta_{\mu} \), \( \mu_i \geq \mu \) (\( i = 1, \ldots, K \)). For \( i, j = 1, \ldots, K \), \( g_{ij} \) is bounded, nonnegative, left continuous and integrable over \([0, \infty)\) with respect to \( t \) for any \( \eta \in \Theta_{\eta} \). Moreover, there is an open set \( \tilde{\Theta} \supset \Theta \) such that \( g_{ij}(t; \eta) \) is differentiable with respect to \( \eta \) in \( \tilde{\Theta} \).

This regularity assumption is standard. In particular, the compactness of \( \Theta \) defines the range of the true parameters, and the bounds on \( g_{ij} \) ensure the well-definedness for the consistency property of MLEs.

The next assumption is also standard in literature \cite{Haw71a, Haw71b, HO74, BM14b}. It ensures that the MHP and its underlying intensity process \( \lambda(t; \theta^*) \) are stationary and ergodic (cf., Theorem 7 in \cite{BM96}). In particular, this stationarity assumption implies that \( \bar{\lambda}_i := \mathbb{E}[\lambda_i(t; \theta^*)] \) is independent of \( t \).

Assumption 2 (Stationarity). The spectral radius of matrix \( G := [G_{ij}]_{K \times K} \) is smaller than 1, where \( G_{ij} := \int_0^\infty g_{ij}(t; \eta^*)dt \).

When \( K = 1, i.e., \) in the one dimensional case, Assumption 2 corresponds to \( \int_0^\infty g(t)dt < 1 \). When the triggering function is exponential such that \( g(t) = \alpha e^{-\beta t} \), this assumption holds when \( |\alpha| < \beta \), which means that the decay is faster than clustering so that the Hawkes process avoids exploding. In the multi-dimensional case, e.g., when \( g_{ij}(t) = \alpha_{ij} e^{-\beta t} \) (\( i, j = 1, \ldots, K \)), the assumption holds if \( \rho(\alpha) < \beta \), where \( \alpha := [\alpha_{ij}]_{K \times K} \) and \( \rho(\cdot) \) denotes the spectral radius. More examples of this stationarity condition on different triggering functions and their combinations can be found in \cite{LC17}.

The next assumption is essential for distinguishing the true parameter from the others.

Assumption 3 (Identifiability). For any \( \eta \neq \eta' \in \Theta_{\eta} \), there exists a set of \( t \in \mathbb{R}_+ \) with nonzero measure such that \( G(t; \eta) \neq G(t; \eta') \). Here \( G(t; \eta) := [g_{ij}(t; \eta)]_{K \times K} \).

Finally, the summability condition below requires that the triggering functions decay sufficiently fast.

Assumption 4 (Summability). For any \( i, j = 1, \ldots, K, d = 1, \ldots, D \), \( g_{ij}(t; \eta) \) and its partial derivatives \( \partial_{\eta_d}g_{ij} \) are all uniformly summable. In addition

\[
\lim_{t \to \infty} \sum_{k=1}^\infty (t_k - t_{k-1}) \sup_{t' \in [t+t_{k-1}, t+t_k], \eta' \in \Theta_{\eta}} g_{ij}(t'; \eta') = 0,
\]

where the sequence \( \{t_k\}_{k=0}^\infty \) satisfies \( t_0 = 0 \) and \( \sum_{k=1}^\infty (t_k - t_{k-1})^{-1} < \infty \). Here a function \( h(t; \eta), h : \mathbb{R} \times \Theta_{\eta} \to \mathbb{R} \) is uniformly summable if there exists a strictly increasing sequence \( \{t_k\}_{k=0}^\infty \) such that for all \( t \geq 0 \),

\[
\sum_{k=1}^\infty (t_k - t_{k-1}) \sup_{t' \in [t+t_{k-1}, t+t_k], \eta' \in \Theta_{\eta}} |h(t'; \eta')| < E,
\]

for some constant \( E > 0 \), where the sequence \( \{t_k\}_{k=0}^\infty \) satisfies \( t_0 = 0 \) as well as \( \sum_{k=1}^\infty (t_k - t_{k-1})^{-1} < \infty \).
Assumption \(4\) is satisfied for most triggering functions in the literature, including the standard triggering functions of the exponential, the Rayleigh, and the power-law types. For instance, let \(t_0 = 0\), \(t_k - t_{k-1} = k^{1+\epsilon'} (k \geq 1)\) for some \(\epsilon' > 0\), then \(\sum_{k=1}^{\infty} (t_k - t_{k-1})^{-1} = \sum_{k=1}^{\infty} k^{-1-\epsilon'} < \infty\), then Assumption \(4\) holds for all triggering functions with a polynomial-exponential decay of the form \(g_{ij}(t) = \alpha_{ij} t^{m} e^{-\beta t} (i, j = 1, \ldots, K)\) with \(m, \beta > 0\).

### 3.2 Examples

**Example 1.** Consider a one-dimensional MHP with intensity process of the form

\[
\lambda(t; \theta) = \mu + \int_{-\infty}^{t} \frac{K}{(c + t - s)^p} N(ds),
\]

with \(K\) denoting the self-exciting effect, \(p\) denoting the decay, and \(c\) being a regularity constant that avoids explosion, and \(\theta = (\mu, K, c, p) \in \mathbb{R}^4\).

Suppose that the regularization to be used in the MLE is \(P(\theta) = a\|\theta\|_1\). Then Assumptions 1-4 hold if \(K^*(c^*)^{1-p^*}/(p^* - 1) < 1\), and \(\Theta = \{(\mu, K, c, p) \mid l_1 \leq \mu \leq u_1, 0 \leq K \leq u_2, l_3 \leq c \leq u_3, l_4 \leq p \leq u_4, c^2 + (p - p_0)^2 \leq CK\}\), with \(a > 0, l_1, l_3, l_4 > 0, u_1, u_2, u_3, u_4 < \infty, p_0 > 1\) and \(C < \infty\). Here \((K^*, c^*, p^*) = \eta^*\) are true parameters. Notice that the last condition in the definition of \(\Theta\) is to ensure identifiability (Assumption \(3\)), which would otherwise be violated when \(K = 0\).

**Example 2.** Consider a \(K\)-dimensional MHP with the following intensity process:

\[
\lambda_i(t; \theta) = \mu_i + \sum_{j=1}^{K} \int_{-\infty}^{t} \sum_{m=1}^{M} \alpha_{ij}^m e^{-\beta_m (t-s)} N_j(ds),
\]

for \(i = 1, \ldots, K\), with \(\theta = (\mu, \alpha, \beta) \in \mathbb{R}^{K+K^2m+m}\), where \(\mu = (\mu_1, \ldots, \mu_K), \alpha = (\alpha_1, \ldots, \alpha_m), \alpha_m = [\alpha_{ij}^m]_{K \times K}\), and \(\beta = (\beta_1, \ldots, \beta_m)\).

Suppose that the regularization to be used in the MLE is

\[
P(\theta) = \sum_{m=1}^{M} (a_1\|\alpha_m\|_1 + a_2\|\alpha_m\|_1 + a_3\|\beta\|_2^2),
\]

where \(\|\cdot\|_1\) is the nuclear norm and \(\|\cdot\|_1\) is the matrix \(l_1\) norm. Then Assumptions 1-4 hold if \(\rho(\sum_{m=1}^{M} \alpha_m^*/\beta_m^*) < 1\), and \(\Theta = \{(\mu, \alpha, \beta) \mid l_1 \leq \mu \leq u_1, 0 \leq \alpha \leq u_2, l_3 \leq \beta \leq u_3, \min_{m \neq m'} |\beta_m - \beta_{m'}| > \epsilon_0, \|\beta_m - b_0\|_2 \leq C\|\alpha_m\|_F\}\), with \(a_1, a_2, a_3, l_1, l_3, \epsilon_0, b_0 > 0, u_1, u_2, u_3, C < \infty\).

Here \((\alpha^*, \beta^*) = \eta^*\) are true parameters, \(\|\cdot\|_F\) is the Frobenius norm, and similar to the one-dimensional example, the last two conditions ensure the identifiability, which would have been lost if \(\beta_m = \beta_{m'}\) for some \(m \neq m' \in \{1, \ldots, M\}\), or \(\alpha_m = 0\) for some \(m = 1, \ldots, M\).
3.3 Main results

We are now ready to state our main results on the consistency of regularized MLEs for MHPs. All the results are assuming Assumptions 1, 2, 3 and 4.

Theorem 1 (Consistency of regularized MLEs). \( \hat{\theta}_T^{\text{reg}} \), the regularized MLE defined in (2.5), converges to the true parameter \( \theta^* \) in probability as \( T \to \infty \).

Note that this theorem implies the classic result of [Oga78] on the consistency of \( \hat{\theta}_T \), the vanilla MLE.

Theorem 2 (Consistency of approximate regularized MLEs). Suppose that \( \hat{\theta}_T^{\text{approx}} \) is an \( \epsilon_T \)-approximation of \( \hat{\theta}_T^{\text{reg}} \), with \( \lim_{T \to \infty} \epsilon_T / T = 0 \). Then \( \hat{\theta}_T^{\text{approx}} \) converges to \( \theta^* \) in probability as \( T \to \infty \).

Theorem 3 (Consistency of regularized MLEs for missing data). For any \( \epsilon_1, \epsilon_2 > 0 \), \( \exists L_{\epsilon_1,\epsilon_2} > 0 \), s.t. if \( \inf_n (u_n - l_n) \geq L_{\epsilon_1,\epsilon_2} + c \) for some \( c > 0 \), by taking \( I_n := [l_n + L_{\epsilon_1,\epsilon_2}, u_n] \), then

\[
\lim_{N \to \infty} P \left( \hat{\theta}_N^{\text{reg}} \in B(\theta^*, \epsilon_1) \right) \geq 1 - \epsilon_2,
\]

i.e., the regularized MLE \( \hat{\theta}_N^{\text{reg}} \) converges to \( B(\theta^*, \epsilon_1) \) with probability at least \( 1 - \epsilon_2 \) as \( N \to \infty \). Here \( B(x, r) \) is the open ball centered at \( x \) with radius \( r \).

When the triggering functions have compact supports, one can choose \( I_n \) so that \( \hat{\lambda}_i(t; \theta) = \hat{\lambda}_i(t; \theta) \) for \( t \in I_n \), and we have the following consistency result.

Corollary 3.1. Suppose that \( \exists C < \infty \), s.t. \( \supp(g_{ij}) \subseteq [0, C] \) for \( i, j = 1, \ldots, K \). Then if \( \inf_n (u_n - l_n) \geq C + c \) for some \( c > 0 \), by taking \( I_n := [l_n + C, u_n] \), the regularized MLE for missing data \( \hat{\theta}_N^{\text{reg}} \) converges to \( \theta^* \) in probability as \( N \to \infty \).

4 Proofs of the main results

We will focus on the main ideas and key steps for the proof of Theorem 1. The proof ideas for Theorems 2 and 3 are similar to that of Theorem 1 and deferred to the appendix due to space limit.

The proof for the consistency property of regularized MLEs consists of two key components: the branching structure of MHPs, and the proof scheme developed by [Oga78] for vanilla MLEs of point processes.

We first explain the branching presentation of MHPs, which first leads to the cumulant density formulas for MHP and then yields two critical technical lemmas. We then outline the proof of Theorem 1.
4.1 Branching process representation for MHPs.

In addition to its intensity process, a Hawkes process can be equivalently defined as a Poisson cluster process with a certain branching structure. \[\text{[HO74]}\]. More precisely, a \(K\)-dimensional MHP with positive baseline intensity \(\mu\) and nonnegative integrable triggering functions \(g_{ij}\) can be equivalently constructed as the following process. (\[\text{JHR15}\] and \[\text{Ras13}\].)

- For \(k = 1, 2, \ldots, K\), initialize an instance \(I_k\) of a homogeneous Poisson process with rate \(\mu_k\), with its elements called immigrants of type \(k\);

- Immigrants of all types generate independent clusters. More specifically, for each \(k = 1, 2, \ldots, K\), each immigrant \(x \in I_k\) generates a cluster \(C^k_x\) with the following branching structure:
  - Generation 0 consists of the immigrant \(x\);
  - Recursively, given generations \(0, 1, \ldots, n\), for all \(i, j = 1, \ldots, K\), each point \(s\) of type \(j\) in generation \(n\) generates its offsprings of type \(i\) as an instance of an inhomogeneous Poisson process with rate \(\lambda_{ij}(t) := g_{ij}(t - s)\). All these offsprings then constitute generation \(n + 1\).

- The point process is then defined to be the union of all clusters, and the number of type \(i\) points (immigrants and offsprings) within time interval \((a, b]\) is exactly \(N_i(a, b]\).

This branching structure is illustrated in Figure 1 where

- Line 1: three immigrants of type \(k\) (red) from a homogeneous Poisson process instance \(I_k\).

- Line 2: an immigrant \(x \in I_k\) in generation 0 generates two offsprings for generation 1, one of type “blue” and the other of type “purple”.

- Line 3: the “blue” in generation 1 generates a “green” offspring for generation 2, and the “purple” in generation 1 offspring generates two offsprings for generation 2, one of type “red” and the other of type “green”.

- Line 4: Projection of the branching tree onto a time line, \(i.e.,\) taking the union to get the cluster \(C^k_x\).

The branching structure representation of MHPs provides analytical tool for explicitly evaluate cumulants and moments of MHPs via convolution and integration, as shown in \[\text{JHR15}\] (for self-containment, we also summarize related results in the appendix).

This intriguing connection allows us to prove the following two critical lemmas. The first one, based on Assumptions 1 and 2, provides the higher-order (> 2) moment bound on the number of arrivals in a given interval for general linear MHPs. This Lemma 1 is critical for most of the remaining lemmas.
Lemma 1 (Main lemma). Given Assumptions 2 and 2, \( \exists \bar{C} > 0 \) such that
\[
\max_{i=1,\ldots,K} \mathbb{E} \left[ \left| N_i(t, t + h) - \bar{\lambda}_i h \right|^4 \right] \leq \bar{C} h^3,
\]
\( \forall t \in \mathbb{R}, h > 0, \) where \( \bar{\lambda}_i := \mathbb{E}[\lambda_i(t; \theta^*)] \).

Similar results can also be proved for all 2n-th centralized moments, i.e.,
\[
\max_{i=1,\ldots,K} \mathbb{E} \left[ \left| N_i(t, t + h) - \bar{\lambda}_i h \right|^{2n} \right] \leq \bar{C} h^{2n-1}
\]
for all \( n \geq 1 \).

To prove the main lemma, recall first the following definitions of cumulants and cumulant densities of MHPs.

Definition 4 (n-th order cumulant). The cumulant-generating function for a d-dimensional random variable \( X \) is defined as \( K(s) = \log \mathbb{E}e^{\langle s, X \rangle}, s \in \mathbb{R}^d \). And the n-th order cumulant is the n-th order coefficient of the power series expansion of \( K(s) \).

Definition 5 (n-th order cumulant density). Given any time vector \( t = (t_1, \ldots, t_n) \) and type vector \( i = (i_1, \ldots, i_n) \in \{1, \ldots, K\}^n \), the n-th order cumulant density of the MHP is defined as \( k^i(t) := \frac{k(N_1^{i_1}(dt_1), \ldots, N_n^{i_n}(dt_n))}{dt_1 \cdots dt_n} \), where \( k(X_1, \ldots, X_n) \) is the n-th order cumulant of random variables \( X_1, \ldots, X_n \).
Proof outline of Lemma 1. The proof relies on two ingredients: the branching process representation of MHPs and the relation between cumulants and moments.

Fix $i_0 \in \{1, \ldots, K\}$. By the branching process representation of MHPs, one can derive an explicit formula for the $4$-th order cumulant density $k^4(t) := k^{i_0i_0i_0i_0}(t_1, t_2, t_3, t_4)$, whose integration from $a$ to $b$ gives the $4$-th order cumulant of $N_{i_0}(a, b)$. (See appendix for the derivation and the exact form of the formula.) Moreover, one can derive a compact upper bound for the $4$-th order cumulant $K_4$ of $N_{i_0}(0, h)$, which is a summation of $5$ integrations involving two quantities: $R^{ij}_t := \left[\sum_{n \geq 0} G^{*n}_{ij}(t)\right]_{ij}$ and $\Psi^{ij}_t := \left[\sum_{n \geq 1} G^{*n}(t)\right]_{ij}$ \footnote{Here $G(t) = [g_{ij}(t; \tau^n)]_{n \times n}$ is the matrix of triggering functions, $G^{*n}(t)$ is the $n$-th (self) convolution of $G$ defined recursively as $G^{*0}(t) = I\delta(t)$, $G^{*n}(t) = \int_{-\infty}^0 G^{*(n-1)}(t-s)G(s)ds$, $\delta(t)$ is the Dirac $\delta$ function, and $\int_{-\infty}^0 G(t)dt = G$, where $G$ is the matrix in Assumption 2.} Here by definition, $R^{ij}_t$ and $\Psi^{ij}_t$ are both non-negative, and we have $\int_R R^{ij}_t dt = [(I - G)^{-1}]_{ij} = A_{ij},$ and $\int_R \Psi^{ij}_t dt = [G(I - G)^{-1}]_{ij} := B_{ij}$. To get the idea of how a compact upper bound for $K_4$ is derived, take the example of the first term in the cumulant formula, and it is clear

\[
\begin{align*}
\int_0^h \int_0^h \int_0^h \int_0^h R_{t_1-u}^{i_0j} R_{t_2-u}^{i_0j} R_{t_3-u}^{i_0j} R_{t_4-u}^{i_0j} dudt_1dt_2dt_3dt_4 \\
\leq \int_0^h \left( \int_R R_{t_1-u}^{i_0j} dt_1 \right) \cdot \left( \int_R R_{t_2-u}^{i_0j} dt_2 \right) \cdot \left( \int_R R_{t_3-u}^{i_0j} dt_3 \right) \cdot \left( \int_R R_{t_4-u}^{i_0j} du \right) dt_4 \\
= A_{i_0j}^3 \int_0^h \left( \int_R R_{t_4-u}^{i_0j} du \right) dt_4 = A_{i_0j}^4 h.
\end{align*}
\]

With similar bounds for the other four integration terms, one can establish the following upper bound for $K_4$ of $N(0, h)$:

\[
K_4 := \int_0^h \int_0^h \int_0^h \int_0^h k_4(t_1, t_2, t_3, t_4)dt_1dt_2dt_3dt_4 \\
\leq \sum_{j=1}^K \tilde{\lambda}_j A_{i_0j}^4 h + 4 \sum_{j_1, j_2=1}^K \tilde{\lambda}_{j_2} A_{i_0j_2}^2 A_{i_0j_1}^2 B_{j_1j_2} h^2 \\
+ 6 \sum_{j_1, j_2=1}^K \tilde{\lambda}_{j_2} A_{i_0j_2}^2 A_{i_0j_1}^2 B_{j_1j_2} h^2 + 3 \sum_{j_1, j_2, j_3=1}^K \tilde{\lambda}_{j_3} A_{i_0j_2}^2 B_{j_2j_3} A_{i_0j_1}^2 B_{j_1j_3} h^3 \\
+ 12 \sum_{j_1, j_2, j_3=1}^K \tilde{\lambda}_{j_3} A_{i_0j_3} A_{i_0j_2} B_{j_2j_3} A_{i_0j_1}^2 B_{j_1j_3} h^3 = O(h^3).
\]

Similarly, one can obtain the upper bounds for the first three cumulants of $N(0, h)$, namely $K_1 := \int_0^h k_1^i(t)dt = \tilde{\lambda}_i h = O(h),$ $K_2 := \int_0^h \int_0^h k_1^i(t_1, t_2)dt_1dt_2 = O(h),$ $K_3 := \int_0^h \int_0^h \int_0^h k_1^i(t_1, t_2, t_3)dt_1dt_2dt_3 = O(h^2)$.

Now, using the relation between cumulants and moments, it is not hard to see that for any $i_0 = 1, \ldots, K$,

- $\mathbb{E}[N_{i_0}(0, h)^4] = K_4 + 4K_3K_1 + 3K_2^2 + 6K_2K_1^2 + K_1^4 = \tilde{\lambda}_{i_0}^4 h^4 + O(h^5)$;
• $E[N_{i_0}(0, h)] = K_1 = \bar{\lambda}_{i_0} h$;
• $E[(N_{i_0}(0, h))^2] = K_2 + K_1^2 = \bar{\lambda}_{i_0}^2 h^2 + O(h)$;
• $E[(N_{i_0}(0, h))^3] = K_3 + 3K_2K_1 + K_3^3 = \bar{\lambda}_{i_0}^3 h^3 + O(h^2)$.

Thus, for any $i_0 = 1, \ldots, K$,

$$E[(N_{i_0}(0, h) - \bar{\lambda}_i h)^4] = O(h^3),$$

and the proof is complete by taking a maximum over $i_0 = 1, \ldots, K$.

Now, by the second cumulant formula and the Laplacian transformation (see the appendix), one can establish a second lemma, which is essential for distinguishing the true parameter from others.

**Lemma 2** (Identifiability). Under Assumptions 1, 2 and 3, $\lambda(0; \theta) = \lambda(0; \theta')$ a.s. if and only if $\theta = \theta'$.

### 4.2 Additional technical lemmas

In this section, we present a few lemmas and a proposition that are key to the proof of Theorem 1. The proofs of these lemmas are given in the appendix.

The first three lemmas establish additional moment bounds and continuity properties of the intensity processes. They are derived from Lemma 1.

**Lemma 3.** Under Assumptions 1, 2 and 3, there exist two random variables $\Lambda_0, \Lambda_1$ with finite $(3 + \alpha)$-th moments for any $\alpha \in [0, 1)$, such that for any $1 \leq i \leq K$,

$$\sup_{\theta \in \Theta} \lambda_i(0; \theta') \leq \Lambda_0, \quad \text{and} \quad \sup_{\theta \in \Theta} |\log \lambda_i(0; \theta')| \leq \Lambda_1.$$

**Lemma 4.** Under Assumptions 1, 2 and 3, for all $i = 1, \ldots, K$, for all $t \geq 0$ and some constant $F > 0$,

$$E \left[ \sup_{\theta \in \Theta} \left| \lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta) \right|^2 \right] \leq F.$$

In addition, as $t \to \infty$

$$E \left[ \sup_{\theta \in \Theta} \left| \lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta) \right|^2 \right] \to 0.$$

**Lemma 5.** Under Assumptions 1, 2 and 3, $\lambda(t; \theta)$ is a.s. continuous in $\theta$ for any fixed $t \geq 0$.

The next lemma is derived from Lemma 2. It concerns Kullback-Leibler’s divergence type quantity for the intensity processes $\lambda(t; \theta)$. 

Lemma 6. Under Assumptions 1, 2, 3 and 4
\[
\sum_{i=1}^{K} \mathbb{E} \left[ \lambda_i(0; \theta^*) \left\{ \frac{\lambda_i(0; \theta)}{\lambda_i(0; \theta^*)} - 1 + \log \left( \frac{\lambda_i(0; \theta^*)}{\lambda_i(0; \theta)} \right) \right\} \right] \geq 0,
\]

and the equality holds if and only if \( \theta = \theta^* \). Denoting the left hand side of the inequality as \( D(\theta^*; \theta) \), then \( D(\theta^*; \theta) \) is continuous in \( \theta \) at \( \theta^* \).

The last lemma is derived from the stationarity and ergodicity of the underlying MHPs (as implied from Assumption 2).

Lemma 7 (Ergodicity). Under Assumptions 1, 2 and 4, for all \( 1 \leq i \leq K \) and any subset \( U \subseteq \Theta \), the following limits hold as \( T \to \infty \):
\[
\frac{1}{T} \int_0^T \left( \inf_{\theta \in U} \lambda_i(t; \theta) - \lambda_i(t; \theta^*) \right) dt \xrightarrow{p} \mathbb{E} \left[ \inf_{\theta \in U} \lambda_i(0; \theta) - \lambda_i(0; \theta^*) \right], \tag{4.1}
\]
\[
\frac{1}{T} \int_0^T \log \left( \frac{\lambda_i(t; \theta^*)}{\sup_{\theta \in U} \lambda_i(t; \theta)} \right) N_i(dt) \xrightarrow{p} \mathbb{E} \left[ \lambda_i(0; \theta^*) \log \left( \frac{\lambda_i(0; \theta^*)}{\sup_{\theta \in U} \lambda_i(0; \theta)} \right) \right]. \tag{4.2}
\]

Finally, the following proposition from analysis connects the convergence of a sequence with the convergence of its Cesaro-sum sequence.

Proposition 1. Suppose that \( f(t): \mathbb{R}_+ \to \mathbb{R}_+ \) is Lebesgue measurable, and \( f(t) \leq C \) for all \( t \geq 0 \) and some \( C > 0 \). If \( f(t) \to 0 \) as \( t \to \infty \), then \( \frac{1}{T} \int_0^T f(t) dt \to 0 \) as \( T \to \infty \).

4.3 Proof Outline of Theorem 1

In this section, we outline the proof for Theorem 1. The complete proof is provided in the appendix. With the technical lemmas in the previous sections, the proof of Theorem 1 adopts the proof scheme in [Oga78] for the vanilla MLEs to the regularized MLEs. It consists of two main steps.

Step 1: Consistency for \( L_T \) and \( \theta_T \). The first step is to show the consistency of \( \theta_T \). That is, for any open neighborhood \( U_0 \) of \( \theta^* \), \( \exists \epsilon > 0 \), such that as \( T \to \infty \),
\[
\mathbb{P} \left( \sup_{\theta \in U_0} L_T(\theta) \geq \sup_{\theta \in \Theta \setminus U_0} L_T(\theta) + cT \right) \to 1.
\]

Since \( \theta^* \in U_0 \), it suffices to prove that there exists a finite cover \( \{ U_s \}_{s=1}^N \) of \( \Theta \setminus U_0 \), such that \( \lim_{T \to \infty} \mathbb{P} \left( \frac{1}{T} L_T(\theta^*) - \sup_{\theta \in U_s} \frac{1}{T} L_T(\theta) \geq \epsilon \right) = 1 \).
Notice that
\[
\frac{1}{T} L_T(\theta^*) - \sup_{\theta \in U_s} \frac{1}{T} L_T(\theta) \\
\geq \frac{1}{T} \int_0^T \left( \sum_{i=1}^K \inf_{\theta \in U_s} \lambda_i(t; \theta) - \lambda_i(t; \theta^*) \right) dt + \frac{1}{T} \sum_{i=1}^K \int_0^T \log \left( \frac{\lambda_i(t; \theta^*)}{\sup_{\theta \in U_s} \lambda_i(t; \theta)} \right) N_i(dt),
\]
(4.3)
it then reduces to finding positive lower bounds on the above integrations. By Lemma 7, it can be further reduced to the following expectation lower bounds:
\[
\mathbb{E} \left[ \sum_{i=1}^K \inf_{\theta' \in U_\theta} \lambda_i(0; \theta') - \sum_{i=1}^K \lambda_i(0; \theta^*) + \sum_{i=1}^K \lambda_i(0; \theta^*) \log \left( \frac{\lambda_i(0; \theta^*)}{\sup_{\theta' \in U_\theta} \lambda_i(0; \theta')} \right) \right] \geq 2 \epsilon.
\]
(4.4)
To establish (4.4), note that by Lemma 6, there exists \( \epsilon > 0 \) such that \( D(\theta^*; \theta) \geq 3 \epsilon \) for any \( \theta \in \Theta \setminus U_0 \). Thus by Lemma 3 and Lemma 5, together with the Lebesgue dominated convergence theorem, one can show that there exists a sufficiently small neighborhood \( U_\theta \) for any \( \theta \in \Theta \setminus U_0 \), such that
\[
\mathbb{E} \left[ \sum_{i=1}^K \inf_{\theta' \in U_\theta} \lambda_i(0; \theta') - \sum_{i=1}^K \lambda_i(0; \theta^*) + \sum_{i=1}^K \lambda_i(0; \theta^*) \log \left( \frac{\lambda_i(0; \theta^*)}{\sup_{\theta' \in U_\theta} \lambda_i(0; \theta')} \right) \right] \geq D(\theta^*; \theta) - \epsilon \geq 2 \epsilon.
\]
(4.5)
Finally, since \( \Theta \setminus U_0 \) is compact, the desired convergence is proved by the finite covering theorem.

**Step 2. Consistency for \( \hat{L}_T^{\text{reg}} \) and \( \hat{\theta}_T^{\text{reg}} \).** The second step is to utilize the closeness of \( \lambda \) and \( \hat{\lambda} \) to show that
\[
\mathbb{P} \left( \sup_{\theta \in U_0} \hat{L}_T^{\text{reg}}(\theta) \geq \sup_{\theta \in \Theta \setminus U_0} \hat{L}_T^{\text{reg}}(\theta) + \epsilon T/4 \right) \rightarrow 1.
\]
By the compactness of \( \Theta \) and continuity of \( P \), it suffices to prove that
\[
\lim_{T \to \infty} \mathbb{P} \left( \frac{1}{T} \hat{L}_T(\theta^*) - \sup_{\theta \in U_s} \frac{1}{T} \hat{L}_T(\theta) \geq \epsilon/2 \right) = 1.
\]
(4.6)
Moreover,
\[
\frac{1}{T} \hat{L}_T(\theta^*) - \sup_{\theta \in U_s} \frac{1}{T} \hat{L}_T(\theta) \\
\geq \frac{1}{T} \int_0^T \left( \sum_{i=1}^K \inf_{\theta \in U_s} \hat{\lambda}_i(t; \theta) - \hat{\lambda}_i(t; \theta^*) \right) dt + \frac{1}{T} \sum_{i=1}^K \int_0^T \log \left( \frac{\hat{\lambda}_i(t; \theta^*)}{\sup_{\theta \in U_s} \hat{\lambda}_i(t; \theta)} \right) N_i(dt).
\]
(4.7)
Therefore it reduces to showing that the difference between the right hand sides of inequalities (4.3) and (4.7) converges to 0 in probability, i.e.,

\[
\frac{1}{T} \int_0^T \left[ \left( \inf_{\theta \in U} \lambda_i(t; \theta) - \inf_{\theta \in U} \hat{\lambda}_i(t; \theta) \right) + \left( \hat{\lambda}_i(t; \theta^*) - \lambda_i(t; \theta^*) \right) \right] dt \\
+ \frac{1}{T} \int_0^T \left[ \left( \log \left( \frac{\sup_{\theta \in U} \hat{\lambda}_i(t; \theta)}{\sup_{\theta \in U} \lambda_i(t; \theta)} \right) \right) + \left( \log \left( \frac{\lambda_i(t; \theta^*)}{\hat{\lambda}_i(t; \theta^*)} \right) \right) \right] \, N_i(dt) \to 0 \text{ in probability.}
\] (4.8)

Now, \( \hat{\lambda}_i(t; \theta) \leq \lambda_i(t; \theta) \) for all \( i = 1, \ldots, K \) and for any \( t \) and \( \theta \in \Theta \). By the Markov inequality, to show (4.8) is reduces to proving that as \( T \to \infty \), for any \( U \subseteq \Theta \),

\[
\frac{1}{T} \int_0^T \mathbb{E} \left[ \sup_{\theta \in U} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta)) \right] dt \to 0, \quad (4.9)
\]

\[
\mathbb{E} \left[ \frac{1}{T} \int_0^T \left[ \sup_{\theta \in U} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta)) \right] N_i(dt) \right] \to 0. \quad (4.10)
\]

We then show that the left hand side of (C.18) can be bounded by

\[
\mathbb{E} \left[ \frac{1}{T} \int_0^T \left[ \sup_{\theta \in U} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta)) \right] N_i(dt) \right] \\
\leq \sqrt{\mathbb{E}[\lambda_i^2(0; \theta^*)]} \cdot \frac{1}{T} \int_0^T \sqrt{\mathbb{E}[\sup_{\theta \in U} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta))^2]} dt.
\]

Finally, Lemma 3 implies that \( \mathbb{E}[\lambda_i(0; \theta^*)^2] < \infty \), and Lemma 4 shows that \( \mathbb{E}[\sup_{\theta \in U} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta))^2] \) is uniformly bounded for all \( t \geq 0 \) and that as \( t \to \infty \)

\[
\mathbb{E}[\sup_{\theta \in U} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta))^2] \to 0.
\]

These results further imply that \( \mathbb{E}[\sup_{\theta \in U} |\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta)|] \) is uniformly bounded for all \( t \geq 0 \), and as \( t \to \infty \)

\[
\mathbb{E}[\sup_{\theta \in U} |\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta)|] \to 0.
\]

The proof is then finished by Proposition 1.

5 Conclusions

In this paper, we establish the consistency property of the regularized MLEs and its approximate version for MHPs, including the case with missing data.
References


In this Appendix, to simplify notations, we extend $g_{ij}$ to be defined on $\mathbb{R} \times \Theta_{\eta}$ by setting $g_{ij}(t; \eta) = 0$ for any $t < 0$ and $\eta \in \Theta_{\eta}$.

A Preliminaries

In this section, we introduce necessary preliminary results for the subsequent proofs.

A.1 Preliminary results in analysis

We first recall some basic results in analysis. The first proposition, stated informally in [Oga78], translates a stochastic Lebesgue-Stieltjes integral to a Lebesgue integral.

**Proposition A.1.** Suppose that $\{\xi(t); t \geq 0\}$ is a finite predictable process such that $\int_0^T \mathbb{E}\left[|\xi(t)|\lambda_i(t; \theta^*)\right]dt < \infty$ for any $T \geq 0$, then

$$\mathbb{E}\left[\int_0^T \xi(t)N_i(dt)\right] = \mathbb{E}\left[\int_0^T \xi(t)\lambda_i(t; \theta^*)dt\right].$$

The next proposition controls the mean-square difference between the Lebesgue-Stieltjes integral and the Lebesgue integral. (See [Pro05], Theorem 20 in Section II.5, Theorem 29 and Corollary 3 in Section II.6).

**Proposition A.2.** Suppose that $\{\xi(t); t \geq 0\}$ is a finite predictable process, with

$$\mathbb{E}\left[\int_0^T \xi(t)^2N_i(dt)\right] < \infty$$

for any $T \geq 0$, $i = 1, \ldots, K$. Define $M_i(t) := N_i(t) - \int_0^t \lambda_i(s; \theta^*)ds$ and $X_i(t) := \int_0^t \xi(s)M_i(ds)$. Then

$$\mathbb{E}[X_i(t)^2] = \mathbb{E}\left[\int_0^t \xi(s)^2N_i(ds)\right].$$

A.2 Cumulant density formula via family/category trees

Given any time vector $t = (t_1, \ldots, t_n)$ and type vector $i = (i_1, \ldots, i_n) \in \{1, \ldots, K\}^n$, the $n$-th order cumulant density of the MHP is define as $k^i(t) := k(N_{i_1}(dt_1), \ldots, N_{i_n}(dt_n))$, where $k(X_1, \ldots, X_n)$ is the $n$-th order cumulant of random variables $X_1, \ldots, X_k$. For example, $k(X) = \mathbb{E}[X]$ and $k(X_1, X_2) = \text{Cov}(X_1, X_2)$. Since moments can be expressed as summations and products of cumulants, the desired moments bounds can be derived from bounds on the cumulants.

Under Assumptions 1 and 2, it follows from [JHR15] that

$$k^i(t)dt = \mathbb{P}\left(E_i^t \cap C_i^t\right),$$

21
where

\[ E^i_t = \{ \forall k = 1, \ldots, n, \text{ there is a type } i_k \text{ event at time } t_k \} , \]
\[ C^i_t = \{ \exists \text{ cluster } C \text{ such that, } \forall k = 1, \ldots, n, t_k \in C \} , \]

and \( E^i_t \cap C^i_t \) means that there is a type \( i_k \) event at time \( t_k \) and that all of these events are descendants from a common immigrant. To compute this probability, notice that all possible branching trees in \( E^i_t \cap C^i_t \) can be grouped into a finite number of categories. Therefore \( \mathbb{P}(E^i_t \cap C^i_t) \) reduces to the sum of the probabilities of these categories.

One can define the notion of nearest common ancestor, where \( u \) is called the nearest common ancestor of \( v_1, \ldots, v_k \) if each \( v_i \) (\( i = 1, \ldots, k \)) is either equal to \( u \) or is a descendant of \( u \), and \( u \) is the node with the largest time stamp that has this property. For each rooted branching tree in \( E^i_t \cap C^i_t \) with root \( x \) (immigrant), we keep the root \( x \), the type \( i_k \) event at time \( t_k \) for \( k = 1, \ldots, K \) (represented as \( t_k \) for short), and the nearest common ancestors of all the subsets of \( \{ t_1, \ldots, t_n \} \) (the set of which denoted as \( A \)), and contract the edges which have at least one end point not in \( A \). After this operation, each edge in the resulting family/category tree can represent arbitrary number of generations. The idea of the above operation is marginalization, i.e., integration of the joint probabilities over the intermediate generations over time and types.

One example is illustrated in Figure 2. The left-hand side lists two realizations of branching trees, both of which reduce to the same family/category tree on the right-hand side. Notice that nodes \( w, y \) and \( z \) in the branching trees are removed in the family/category tree because none of them is in the set \( A = \{ u, v \} \) (the nearest common ancestor set).

\[ \text{Figure 2: Left: branching trees; Right: family/category trees} \]

To compute the probability of each family/category tree, we first compute the product of the conditional probability densities along the edges given the types and event time stamps of all the nodes. We then sum over all possible types and integrate over all possible time stamps.
More precisely, define $R^{ij}_t := \left[\sum_{n \geq 0} G^{*n}_{ij}(t)\right]_{ij}$ where $G(t) = [g_{ij}(t;\mathbf{\eta})]_{n \times n}$ is the matrix of triggering functions, and $G^{*n}(t)$ is the $n$-th (self) convolution of $G$ defined recursively as $G^{*0}(t) = I \delta(t)$, $G^{*n}(t) = \int_{-\infty}^{t} G^{*(n-1)}(t-s)G(s)ds$. Here $\delta(t)$ is the Dirac $\delta$ function, and $\int_{-\infty}^{\infty} G(t)dt = G$, where $G$ is the matrix in Assumption 2. It is shown in [JHR15] that

$$R^{ij}_t dt = P(\text{type } j \text{ event at } 0 \text{ causes type } i \text{ event at } t).$$

Suppose that $u$ is of type $m$, then the probability density along edge $(x, u)$ with the time stamp and type of $x$ marginalized is $\lambda_m$, and the probability density along the edge $(u, t_k)$ is $R_{t_k}^{u,m}$. Finally, since two nearest common ancestors can not be identical, the probability density along edge $(u, v)$ which connects two (different) nearest common ancestors with types $j$ and $i$ is $\Psi_{v-u}^{ij} := R_{v-u}^{ij} - \delta_{ij}\delta(v-u) = \left[\sum_{n \geq 1} G^{*n}(v-u)\right]_{ij}$.

For example, the probability corresponding to the family/category tree in Figure 2 is equal to $\sum_{j_1,j_2=1}^K \bar{\lambda}_{j_2} \int \int R_{\theta_T}^{j_1,j_2}(\int R_{\theta_T}^{j_2}R_{\theta_T}^{j_1}R_{\theta_T}^{j_2}d\Psi_{\theta_T})du$.

The above discussion results in the cumulant computation algorithm at the end of Section III in [JHR15].

B  Additional technical lemmas

We first state a few additional lemmas that are used to prove the main theorem (Theorem 1).

Lemma B.1. Under Assumptions 1, 2 and 4, for any $t \geq 0$, define for $i = 1, \ldots, K$ that

$$C_i^{(t)} := \sup_{k \geq 1} N_i(t - t_k, t - t_{k-1})/(t_k - t_{k-1}).$$

Then $E[|C_i^{(t)}|^{3+\alpha}] \leq C$ for any $\alpha \in [0, 1)$. Here $C > 0$ is a positive constant independent of $t$, and $\{t_k\}_{k=0}^\infty$ is the sequence from Assumption 4.

Note that Lemma B.1 follows directly from Lemma 1 and is critical for Lemmas 3, 4 and 5.

Furthermore, the stationarity and ergodicity of the underlying MHP (implied from Assumption 2, see e.g., Theorem 7 in [BM96]), together with Proposition A.1 and Proposition A.2 lead to the following lemma.

Lemma B.2. Suppose that $\xi = \{\xi(t); t \geq 0\}$ is a stationary stochastic process, and suppose that it is adapted and a.s. left continuous in $t \geq 0$, with $E[|\xi(0)|^2] < \infty$. Then the following limits hold:

$$\frac{1}{T} \int_0^T \xi(t)dt \xrightarrow{P} E[\xi(0)], \quad (B.1)$$

$$\frac{1}{T} \int_0^T \frac{\xi(t)}{\lambda_i(t;\mathbf{\theta})} N_i(dt) \xrightarrow{P} E[\xi(0)]. \quad (B.2)$$

23
In contrast to Lemma 2 in [Oga78] which implicitly assumes the almost sure convergence of \( \frac{1}{T} \int_0^T \xi(t) dt \) to 0 to establish the almost sure convergence of (B.1) and (B.2), the above property of convergence in probability suffices for the consistency of MLEs.

Also, one can establish the predictability, stationarity and certain bounds and approximation errors between the log-likelihood functions.

**Lemma B.3.** Under Assumptions 1, 2 and 4, for any \( \theta \in \Theta \), \( \lambda_i(t; \theta) \) is stationary, adapted, and a.s. left continuous in \( t \geq 0 \). Moreover, for any subset \( U \subseteq \Theta \), the following stochastic processes

\[
\begin{align*}
\xi_{U,i}^{(1)}(t) &:= \inf_{\theta \in U} \lambda_i(t; \theta) - \lambda_i(t; \theta^*), \\
\xi_{U,i}^{(2)}(t) &:= \lambda_i(t; \theta^*) (\log (\lambda_i(t; \theta^*)) - \log (\sup_{\theta \in U} \lambda_i(t; \theta))), \\
\xi_{U,i}^{(3)}(t) &:= \sup_{\theta \in U} \left( \lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta) \right)
\end{align*}
\]

are also adapted and a.s. left continuous in \( t \geq 0 \). In addition, \( \xi_{U,i}^{(1)}(t) \) and \( \xi_{U,i}^{(2)}(t) \) are stationary, and for any \( T \geq 0 \) and \( i = 1, \ldots, K \),

\[
E \left[ \left| \xi_{U,i}^{(1)}(0) \right|^2 \right] < \infty, \quad E \left[ \left| \xi_{U,i}^{(2)}(0) \right|^2 \right] < \infty, \quad \int_0^T E \left[ \left| \xi_{U,i}^{(3)}(t) \right| \lambda_i(t; \theta^*) \right] dt < \infty.
\]

It is worth noticing that if \( \{\xi(t); t \geq 0\} \) is an adapted stochastic process defined on the filtered probability space of the (true) MHP, and if the sample paths of \( \xi(t) \) are a.s. left continuous on \((0, \infty)\), then \( \xi(t) \) is predictable. Thus the aforementioned stochastic processes \( \lambda_i(t; \theta) \), \( \xi_{U,i}^{(l)}(t) \), \( l = 1, 2, 3 \) are all predictable.

By plugging \( \xi_{U,i}^{(1)}(t) \) in Eqn. (B.1) and \( \xi_{U,i}^{(2)}(t) \) in Eqn. (B.2), respectively, Lemma B.2 and Lemma B.3 immediately imply the ergodicity property in Lemma 7.

**B.1 Detailed proofs for lemmas (in both main text and appendix)**

**Lemma 1.** Given Assumptions 1 and 2, \( \exists \bar{C} > 0 \) such that

\[
\max_{i=1, \ldots, K} E \left[ \left| N_i(t, t+h) - \bar{\lambda}_i h \right|^4 \right] \leq \bar{C} h^3,
\]

\( \forall t \in \mathbb{R} \), \( h > 0 \), where \( \bar{\lambda}_i := E[\lambda_i(t; \theta^*)] \).

**Proof of Lemma 1.** The idea is to connect MHPs with the Poisson clustering process outlined above. For the 4-th order cumulant, there are 26 family/category trees, which can be further grouped into 5 generic types by symmetry, as listed in Figure 3.
Hence for \( i = j = k = l = i_0 \),

\[
k_4(t) := k^{i_0i_0i_0i_0}(t_1, t_2, t_3, t_4) = \sum_{j=1}^{K} \lambda_j \int_{\mathbb{R}} \prod_{i=1}^{4} R_{t_{i-u}}^{i_0j} du
\]

\[
+ \sum_{i=1}^{4} \sum_{j_1,j_2=1}^{K} \lambda_{j_2} \int_{\mathbb{R}} R_{t_{1-u}}^{i_0j_2} \left( \int_{\mathbb{R}} \prod_{j \neq i} R_{t_{j-v}}^{i_0j_3} \Psi_{v-u}^{j_2j_3} dv \right) du
\]

\[
+ \sum_{1 \leq i_1 < i_2 \leq 4} \lambda_{j_2} \sum_{j_1,j_2,j_3=1}^{K} \left( \int_{\mathbb{R}} \left( \int_{\mathbb{R}} R_{t_{1-u}}^{i_0j_2} R_{t_{i_2-u}}^{i_0j_2} \Psi_{v-u}^{j_2j_3} dv \right) \left( \int_{\mathbb{R}} \prod_{j \neq i_1,i_2} R_{t_{j-w}}^{i_0j_3} \Psi_{w-u}^{j_2j_3} dw \right) du \right)
\]

\[
+ \sum_{i=1}^{4} \sum_{j_1,j_2,j_3=1}^{K} \lambda_{j_3} \int_{\mathbb{R}} R_{t_{i_3-u}}^{i_0j_3} \left( \int_{\mathbb{R}} \prod_{j \neq i} R_{t_{j-v}}^{i_0j_2} \Psi_{v-u}^{j_3j_3} dv \right) \left( \int_{\mathbb{R}} \prod_{j \neq i_1,i_3} R_{t_{j-w}}^{i_0j_3} \Psi_{w-u}^{j_3j_3} dw \right) du,
\]

where \( t := (t_1, t_2, t_3, t_4) \), and \( k^{ijkl}(t_1, t_2, t_3, t_4) \) is the 4-th order cumulant of \( N(dt) \), as defined in \( \text{[JHR15]} \).

Noticing that \( R_{t}^{ij} \) and \( \Psi_{t}^{ij} \) are both non-negative by definition, \( \int_{\mathbb{R}} R_{t}^{ij} dt = [(I - G)^{-1}]_{ij} := A_{ij} \), and \( \int_{\mathbb{R}} \Psi_{t}^{ij} dt = [G(I - G)^{-1}]_{ij} := B_{ij} \), by integrating from 0 to \( h \) for an arbitrary \( h > 0 \),

\[
K_4 := \int_{0}^{h} \int_{0}^{h} \int_{0}^{h} \int_{0}^{h} k_4(t_1, t_2, t_3, t_4) dt_1 dt_2 dt_3 dt_4
\]

\[
\leq \sum_{j=1}^{K} \lambda_{j} A_{i_0j}^4 h + 4 \sum_{j_1,j_2=1}^{K} \lambda_{j_2} A_{i_0j_2}^3 A_{i_0j_1}^3 B_{j_1j_3} h^2
\]

\[
+ 6 \sum_{j_1,j_2=1}^{K} \lambda_{j_2} A_{i_0j_2}^2 A_{i_0j_1}^2 B_{j_1j_3} h^2 + 3 \sum_{j_1,j_2,j_3=1}^{K} \lambda_{j_3} A_{i_0j_2}^2 B_{j_2j_3} A_{i_0j_1}^2 B_{j_1j_3} h^3
\]

\[
+ 12 \sum_{j_1,j_2,j_3=1}^{K} \lambda_{j_3} A_{i_0j_3} A_{i_0j_2} B_{j_2j_3} A_{i_0j_1}^2 B_{j_1j_2} h^3 = O(h^3).
\]

Notice that here we leave out one of \( t_1, \ldots, t_4 \) to maintain dependence on \( h \). In this way, we have

\[
\int_{0}^{h} \int_{0}^{h} \int_{0}^{h} \int_{0}^{h} R_{t_{1-u}}^{i_0j} R_{t_{2-u}}^{i_0j} R_{t_{3-u}}^{i_0j} R_{t_{4-u}}^{i_0j} du dt_1 dt_2 dt_3 dt_4
\]

\[
\leq \int_{0}^{h} \left( \int_{\mathbb{R}} R_{t_{1-u}}^{i_0j} dt_1 \right) \left( \int_{\mathbb{R}} R_{t_{2-u}}^{i_0j} dt_2 \right) \left( \int_{\mathbb{R}} R_{t_{3-u}}^{i_0j} dt_3 \right) \left( \int_{\mathbb{R}} R_{t_{4-u}}^{i_0j} du \right) dt_4
\]

\[
= A_{i_0j}^3 \int_{0}^{h} \left( \int_{\mathbb{R}} R_{t_{4-u}}^{i_0j} du \right) dt_4 = A_{i_0j}^4 h.
\]
Figure 3: A schematic representation of all possible family category trees. The $\times n$ notation at the top right of each tree indicates that by combination and permutation of $\{t_1, t_2, t_3, t_4\}$, there are $n$ different versions of trees with the same type.

The rest are similar.

Similarly, by Eqn. (14), (37) and (39) in [JHR15], and by setting all the indices to $i_0$, we obtain $K_1 := \int_0^h k_{i_0}(t)dt = \lambda_{i_0} h = O(h)$, $K_2 := \int_0^h \int_0^h k_{i_0i_0}(t_1, t_2)dt_1dt_2 = O(h)$, $K_3 := \int_0^h \int_0^h \int_0^h k_{i_0i_0i_0}(t_1, t_2, t_3)dt_1dt_2dt_3 = O(h^2)$.

Moreover, using the relation between cumulants and moments, we see that for any $i_0 = 1, \ldots, K$, $E[(N_{i_0}(0, h))^4] = K_4 + 4K_3K_1 + 3K_2^2 + 6K_2K_1^2 + K_1^4 = \lambda_{i_0}^4 h^4 + O(h^3)$, $E[N_{i_0}(0, h)] = K_1 = \lambda_{i_0} h$, $E[(N_{i_0}(0, h))^2] = K_2 + K_1^2 = \lambda_{i_0}^2 h^2 + O(h)$, $E[(N_{i_0}(0, h))^3] = K_3 + 3K_2K_1 + K_1^3 = \lambda_{i_0}^3 h^3 + O(h^2)$. Thus, for any $i_0 = 1, \ldots, K$, $E[(N_{i_0}(0, h) - \lambda_{i_0} h)^4] = E[(N_{i_0}(0, h) - \lambda h)^4] - 4E[(N_{i_0}(0, h))^3]\lambda_{i_0} h + 6E[(N_{i_0}(0, h))^2]\lambda_{i_0}^2 h^2 - 4E[(N_{i_0}(0, h))]\lambda_{i_0}^3 h^3 + \lambda_{i_0}^4 h^4 = O(h^3)$.

Finally, the proof is complete by taking a maximum over $i_0 = 1, \ldots, K$.

Lemma 2. Under Assumptions 1, 2 and 3 $\lambda(0; \theta) = \lambda(0; \theta')$ a.s. if and only if $\theta = \theta'$.

Proof of Lemma 2. The proof again relies on the explicit formula for covariance density in [JHR15]. The key is to notice that if $\lambda_i(0; \theta) = \lambda_i(0; \theta')$ a.s., then the expectation and variance of $\lambda_i(0; \theta) - \lambda_i(0; \theta')$ should both be equal to 0. Expanding using Eqn. (2.1), and
defining \( h_{ij}(t) := g_{ij}(t; \eta) - g_{ij}(t; \eta') \), we have

\[
0 = \mathbb{E} [\lambda_i(t; \theta) - \lambda_i(t; \theta')] = \mu_i - \mu_i' + \sum_{j=1}^{K} \mathbb{E} \left[ \int_{-\infty}^{0} h_{ij}(-s) N_j(ds) \right], \quad (B.3)
\]

and

\[
0 = \text{Var} (\lambda_i(t; \theta) - \lambda_i(t; \theta')) = \mathbb{E} \left[ (\lambda_i(t; \theta) - \lambda_i(t; \theta'))^2 \right] = \mathbb{E} \left[ \left( \mu_i - \mu_i' + \sum_{j=1}^{K} \int_{-\infty}^{0} h_{ij}(-s) N_j(ds) \right)^2 \right] = (\mu_i - \mu_i')^2 + 2(\mu_i - \mu_i') \sum_{j=1}^{K} \mathbb{E} \left[ \int_{-\infty}^{0} h_{ij}(-s) N_j(ds) \right] + \sum_{j_1=1}^{K} \sum_{j_2=1}^{K} \int_{-\infty}^{0} \int_{-\infty}^{0} h_{ij_1}(-s_1) h_{ij_2}(-s_2) \mathbb{E}[N_{j_1}(ds_1)N_{j_2}(ds_2)] - (\mu_i - \mu_i')^2
\]

By Eqn. (14) and (37) in [JHR15], this means that for any \( i = 1, \ldots, K \),

\[
0 = \sum_{j_1=1}^{K} \sum_{j_2=1}^{K} \int_{-\infty}^{0} \int_{-\infty}^{0} h_{ij_1}(-s_1) h_{ij_2}(-s_2) \sum_{m=1}^{K} \bar{\lambda}_m \int_{R} R_{s_1-u} R_{s_2-u} ds_1 ds_2
\]

\[
= \sum_{m=1}^{K} \bar{\lambda}_m \int_{R} \left( \sum_{j_1=1}^{K} \sum_{j_2=1}^{K} \int_{-\infty}^{0} \int_{-\infty}^{0} h_{ij_1}(-s_1) h_{ij_2}(-s_2) R_{s_1-u} R_{s_2-u} ds_1 ds_2 \right) du
\]

\[
= \sum_{m=1}^{K} \bar{\lambda}_m \int_{R} \left( \sum_{j=1}^{K} \int_{-\infty}^{0} h_{ij}(-s) R_{s-u} ds \right)^2 du
\]

\[
= \sum_{m=1}^{K} \bar{\lambda}_m \int_{R} \left( \sum_{j=1}^{K} \int_{0}^{\infty} h_{ij}(s) R_{s-u} ds \right)^2 du
\]

Since \( \bar{\lambda}_i > 0 \) for all \( i = 1, \ldots, K \), this implies that \( \sum_{j=1}^{K} h_{ij} \ast R_{s-u} \equiv 0 \) a.e. for all \( u \in \mathbb{R} \) and \( i, m = 1, \ldots, K \). Taking the Laplace transform evaluated at \( t \), we see that

\[
\sum_{j=1}^{K} \hat{h}_{ij}(t) A_{jm}(t) = 0 \text{ a.e., for all } i, m = 1, \ldots, K,
\]

where \( A_{ij}(t) := [(I - G(t))^{-1}]_{ij} = \hat{R}^{ij}(t) \), in which \( G(t) := [g_{ij}(t; \eta')]_{K \times K} \).

27
Rewriting in a matrix form, we see that $\hat{H}(t)(I-\hat{G}(t))^{-1} = 0$ a.e., where $\hat{H}(t) = [\hat{h}_{ij}(t)]_{K \times k}$.

This implies that $\hat{H}(t) \equiv 0$ a.e. for all $t \geq 0$, which holds iff $\hat{H}(t) := [h_{ij}(t)]_{K \times K} \equiv 0$ a.e. Hence $g_{ij}(t; \eta) = g_{ij}(t; \eta')$ a.e. in $t$, and $\eta = \eta'$ by Assumption 3 in Section 2. Finally, plugging in Eqn. (B.3), we see that $\mu_i = \mu'_i$ for $i = 1, \ldots, K$.

**Lemma 3.** Under Assumptions 1, 2 and 4 there exist two random variables $\Lambda_0$, $\Lambda_1$ with finite $(3 + \alpha)$-th moments for any $\alpha \in [0, 1)$, such that for any $1 \leq i \leq K$,

$$
\sup_{\theta' \in \Theta} \lambda_i(0; \theta') \leq \Lambda_0, \quad \text{and} \quad \sup_{\theta' \in \Theta} |\log \lambda_i(0; \theta')| \leq \Lambda_1.
$$

**Lemma 4.** Under Assumptions 1, 2 and 4 for all $i = 1, \ldots, K$, for all $t \geq 0$ and some constant $F > 0$,

$$
\mathbb{E} \left[ \sup_{\theta' \in \Theta} \left| \lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta) \right|^2 \right] \leq F.
$$

In addition, as $t \to \infty$

$$
\mathbb{E} \left[ \sup_{\theta' \in \Theta} \left| \lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta) \right|^2 \right] \to 0.
$$

**Proof of Lemma 3 and Lemma 4.** The proof is based on Lemma B.1 regarding the uniform moment bounds of successive increment of $\mathbf{N}$.

Suppose that $\{t_k\}_{k \geq 0}$ is the sequence in Assumption 4. For notational simplicity, we use the shorthand $C_i := C_i^{0(0)}$, where $C_i^{(t)}$ is the constant defined in Lemma B.1. Then since $\log(x) \leq x - 1 \leq x$ for all $x > 0$ and $\Theta \subset B(0, R)$,

$$
\sup_{\theta' \in \Theta} \lambda_i(0; \theta') \leq R + \sum_{j=1}^{K} C_j \max_{i,j=1,\ldots, K} \sum_{k=1}^{\infty} (t_k - t_{k-1}) \sup_{t \in [t_{k-1}, t_k], \eta' \in \Theta_{\eta}} g_{ij}(t; \eta') := \Lambda_0,
$$

$$
\sup_{\theta' \in \Theta} |\log \lambda_i(0; \theta')| \leq \max \{|\log \mu|, \sup_{\theta' \in \Theta} \lambda_i(0; \theta')\} \leq \Lambda_0 + |\log \mu| := \Lambda_1,
$$

where $\Theta \subset B(0, R)$. In addition,

$$
\sup_{\theta' \in \Theta} |\lambda_i(t; \theta') - \hat{\lambda}_i(t; \theta')| \leq \sum_{j=1}^{K} C_j \sum_{k=1}^{\infty} (t_k - t_{k-1}) \sup_{t' \in [t_{k-1}, t_k], \eta' \in \Theta_{\eta}} g_{ij}(t'; \eta').
$$

Hence it suffices to prove that $\Lambda_0$ has a finite $(3 + \alpha)$-th moment for any $\alpha \in [0, 1)$, and that

$$
\Lambda_0^{(t)} := \sum_{j=1}^{K} C_j \max_{i,j=1,\ldots, K} \sum_{k=1}^{\infty} (t_k - t_{k-1}) \sup_{t' \in [t_{k-1}, t_k], \eta' \in \Theta_{\eta}} g_{ij}(t'; \eta') \to 0
$$

in the mean-square sense as $t \to \infty$, and is uniformly bounded for all $t \geq 0$. Here $\Lambda_0^{(0)} + R = \Lambda_0$. 

28
Now by Assumption 4 and by setting \( T = 0 \),
\[
\tilde{C} := \max_{i,j=1,\ldots,K} \sum_{k=1}^{\infty} (t_k - t_{k-1}) \sup_{\eta' \in \Theta, \eta \in \Theta} g_{ij}(t'; \eta') < \infty,
\]
\[
\tilde{C}(t) := \max_{i,j=1,\ldots,K} \sum_{k=1}^{\infty} (t_k - t_{k-1}) \sup_{\eta' \in \Theta, \eta \in \Theta} g_{ij}(t'; \eta') \to 0 \quad \text{as} \ t \to \infty,
\]
\[
\tilde{C}(t) \leq E \quad \text{for any} \ t \geq 0 \ \text{and some} \ E > 0.
\]

By taking \( t = 0 \) in Lemma 3.1, we have \( \mathbb{E}||C_j|^3|^{3+\alpha}| < \infty \) for \( j = 1, \ldots, K \) for any \( \alpha \in [0, 1) \).

Since \( \Lambda_0 \leq R + \tilde{C} \sum_{j=1}^{K} C_j \) and \( \Lambda_0^{(t)} \leq \tilde{C}(t) \sum_{j=1}^{K} C_j \), we see that
\[
\begin{align*}
&\bullet \ \mathbb{E}[|\Lambda_0^{(3+\alpha)}|] < \infty; \\
&\bullet \ \mathbb{E}[|\Lambda_0^{(t)}|^{3+\alpha}] \leq F \quad \text{for all} \ t \geq 0 \ \text{and some constant} \ F > 0, \ \mathbb{E}[|\Lambda_0^{(t)}|^{3+\alpha}] \to 0 \quad \text{as} \ t \to \infty,
\end{align*}
\]
then imply that \( \mathbb{E}[|\Lambda_0^{(t)}|^{2}] \leq F^{2/(3+\alpha)} \) for all \( t \geq 0 \) and some constant \( F > 0, \mathbb{E}[|\Lambda_0^{(t)}|^{2}] \to 0 \) as \( t \to \infty. \)

**Lemma 5.** Under Assumptions 1, 2 and 4 \( \lambda(t; \theta) \) is a.s. continuous in \( \theta \) for any fixed \( t \geq 0 \).

**Proof of Lemma 5.** To prove the almost sure continuity of \( \lambda_i(t; \theta) \) as a function of \( \theta \) for each fixed \( t \geq 0 \), it suffices to prove the continuity for each dimension \( i \) and time \( t \). Below we focus on \( \lambda_i(t; \theta) \) for a fixed \( i = 1, \ldots, K \) and \( t \geq 0 \).

By the definition of \( C_{ij}^{(t)} \) in Lemma 3.1, for an arbitrary sequence \( \theta_n \to \theta_0 \), where \( \theta_n = (\mu^n, \eta^n) \) and \( \theta_0 = (\mu^0, \eta^0) \),
\[
|\lambda_i(t; \theta_n) - \lambda_i(t; \theta_0)| \leq |\mu_i^n - \mu_i^0| + \sum_{j=1}^{K} \int_{-\infty}^{t} |g_{ij}(t - s; \eta^n) - g_{ij}(t - s; \eta^0)| \ |
\]
\[
N_j(ds)
\]
\[
\leq |\mu_i^n - \mu_i^0| + \sum_{j=1}^{K} C_{ij}^{(t)} \sum_{k=1}^{\infty} (t_k - t_{k-1}) \sup_{\eta' \in \Theta, \eta \in \Theta} |g_{ij}(t'; \eta^n) - g_{ij}(t'; \eta^0)|.
\]

Let \( L_{tk-1:t_k} := \sum_{d=1}^{D} \sup_{\eta' \in \Theta, \eta \in \Theta} |\partial_{\eta_d} g_{ij}(t'; \eta')| \). Then by Assumption 4
\[
\sum_{k=1}^{\infty} (t_k - t_{k-1}) L_{tk-1:t_k} < \infty. \quad (B.4)
\]

By Assumption 1 \( \exists \epsilon > 0 \) such that \( B(\theta_0, \epsilon) \subseteq \Theta \). Since for a sufficiently large \( n \), \( \theta^n \in B(\theta_0, \epsilon) \), by the mean-value theorem, there exists some \( c \in (0, 1) \), such that
\[
\sup_{t' \in [tk-1, tk]} |g_{ij}(t'; \eta^n) - g_{ij}(t'; \eta^0)| = \sup_{t' \in [tk-1, tk]} \left| \nabla_{\eta} g_{ij}((1 - c) \eta^n + c \eta^n)^T(\eta^n - \eta^0) \right|
\]
\[
\leq \sup_{t' \in [tk-1, tk]} \sup_{\eta' \in \Theta, \eta \in \Theta} \| \nabla_{\eta} g_{ij}(\eta') \|_1 \| \eta^n - \eta^0 \|_{\infty}
\]
\[
\leq L_{tk-1:t_k} \| \eta^n - \eta^0 \|_{\infty},
\]
hence
\[
|\lambda_i(t; \theta_n) - \lambda_i(t; \theta_0)| \leq |\mu_i^n - \mu_i^0| + \frac{K}{T} \sum_{j=1}^{K} C_j^{(t)} \sum_{k=1}^{\infty} (t_k - t_{k-1}) L_{t_{k-1}, t_k} \|\eta^n - \eta^0\|_{\infty}. \tag{B.5}
\]

By Lemma \[B.1\], we have in particular that \(C_j^{(t)}\) is a.s. finite for any \(j = 1, \ldots, K\). Hence as \(\theta_n \to \theta_0\), \(\mu_i^n \to \mu_i^0\) and \(\eta_n \to \eta_0\) as \(n \to \infty\), and hence \(\lambda_i(t; \theta_n) \to \lambda_i(t; \theta_0)\) almost surely as \(n \to \infty\) from Eqns. \[B.4\] and \[B.5\]. \hfill \(\Box\)

**Proof of Lemma \[B.2\]** Since \(N\) is stationary, the time shift operator \(S_1\) through the unit distance is measure-preserving [DVJ07, Chapter 12.2]. Moreover, by the adaptedness, for any \(t \geq 0\), \(\xi(t)\) is a measurable functional of the point process \(\{N_i(s, t'], s < t' < t, i = 1, \ldots, K\}\). Along with the assumption that the underlying (true) MHP is ergodic, the \(\sigma\)-algebra of invariant events under \(S_1\) is trivial. Since \(E\left[\int_0^T \xi(t) dt\right] \leq E[|\xi(0)|] < \infty\), applying Birkhoff’s ergodic theorem [Dur10] to \(X := \int_0^T \xi(t) dt\) and \(S_1\) yields
\[
\lim_{N \to \infty} \frac{1}{N} \int_0^N \xi(t) dt = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} \int_k^{k+1} \xi(t) dt = E[\xi(0)],
\]
where \(N\) takes positive integer values. And since by assumption \(E[|\xi(0)|^2] < \infty\), we have for general \(T > 0\),
\[
\left|\frac{\int_0^T \xi(t) dt - \int_0^{[T]} \xi(t) dt}{[T]}\right| = \left|\int_{[T]} \xi(t) dt \right| \leq \frac{1}{[T]} \int_{T-1}^T |\xi(t)| dt \to 0 \text{ in probability,}
\]
where the last limit follows from \(P(Y(T) \geq \epsilon) \leq \frac{E[Y(T)]}{\epsilon} = \frac{E[|\xi(0)|]}{\epsilon [T]} \to 0\) as \(T \to \infty\), according to the definition \(Y(T) := \frac{1}{[T]} \int_{T-1}^T |\xi(t)| dt\).

Together with the fact that \(\lim_{T \to \infty} \frac{1}{T} \int_0^T \xi(t) dt = \lim_{T \to \infty} \frac{1}{[T]} \int_0^{[T]} \xi(t) dt = E[\xi(0)]\) in probability, we conclude that
\[
\lim_{T \to \infty} \frac{1}{T} \int_0^T \xi(t) dt = \lim_{T \to \infty} \frac{1}{[T]} \int_0^{[T]} \xi(t) dt = E[\xi(0)] \text{ in probability,}
\]
where \(T\) takes general positive real values. This proves Eqn. \[B.1\].

As for Eqn. \[B.2\], consider
\[
\eta(T) := \int_0^T \xi(t) \frac{N_i(dt)}{\lambda_i(t; \theta^*)} - \int_0^T \xi(t) dt = \int_0^T \xi(t) \frac{\lambda_i(t; \theta^*)}{\lambda_i(t; \theta^*)} M_i(dt),
\]
Then since \(\xi(t)\) is stationary and has finite second-order moments, we have
\[
\int_0^T E \left[ \left( \frac{\xi(t)}{\lambda_i(t; \theta^*)} \right)^2 \lambda_i(t; \theta^*) \right] dt \leq \frac{T}{\mu} E \left[ \xi(0)^2 \right] < \infty.
\]
Hence applying Proposition \[A.1\] gives
\[
\mathbb{E} \left[ \int_0^T \left( \frac{\xi(t)}{\lambda_i(t; \theta^*)} \right)^2 N_i(dt) \right] = \mathbb{E} \left[ \int_0^T \left( \frac{\xi(t)}{\lambda_i(t; \theta^*)} \right)^2 \lambda_i(t; \theta^*) dt \right] < \infty.
\]

By Proposition \[A.2\]
\[
\mathbb{E} \left[ \eta_i(T)^2 \right] = \mathbb{E} \left[ \int_0^T \left( \frac{\xi(t)}{\lambda_i(t; \theta^*)} \right)^2 N_i(dt) \right] \leq \frac{T}{\mu} \mathbb{E} \left[ \xi(0)^2 \right],
\]
from which
\[
\lim_{T \to \infty} \mathbb{E} \left[ \left( \frac{1}{T} \eta_i(T) \right)^2 \right] \leq \lim_{T \to \infty} \frac{1}{T \mu} \mathbb{E} \left[ \xi(0)^2 \right] = 0.
\]

Since convergence in expectation implies convergence in probability, \(\lim_{T \to \infty} \eta_i(T)/T = 0\) in probability. This, together with Eqn. \[(B.1)\], implies Eqn. \[(B.2)\].

**Proof of Lemma \[B.3\].** The left continuity of \(\xi_{U,i}^{(l)}(t)\) \((l = 1, 2, 3)\) is directly implied from the left continuity of \(g_{ij}\) in \(t \geq 0\). Now we prove the bounds related to \(\xi_{U,i}^{(l)}(t)\), \(l = 1, \ldots, 3\). First, by Lemma \[3\]
\[
\mathbb{E} \left[ \left| \xi_{U,i}^{(l)}(0) \right|^2 \right] \leq \mathbb{E} \left[ (2\Lambda_0)^2 \right] = 4\mathbb{E}[\Lambda_0^2] < \infty.
\]

Secondly, since \(\log x \leq x - 1\) for any \(x \geq 0\),
\[
-\xi_{U,i}^{(2)}(0) = \lambda_i(0; \theta^*) \log \left( \frac{\sup_{\theta \in U} \lambda_i(0; \theta)}{\lambda_i(0; \theta^*)} \right) \leq \sup_{\theta \in U} \lambda_i(0; \theta) - \lambda_i(0; \theta^*).
\]

Meanwhile, we also have
\[
-\xi_{U,i}^{(2)}(0) = \lambda_i(0; \theta^*) \log \left( \frac{\sup_{\theta \in U} \lambda_i(0; \theta)}{\lambda_i(0; \theta^*)} \right)
= \lambda_i(0; \theta^*) \log(\sup_{\theta \in U} \lambda_i(0; \theta)) - \lambda_i(0; \theta^*) \log \lambda_i(0; \theta^*)
\geq \lambda_i(0; \theta^*) \log \mu - \lambda_i(0; \theta^*) \log \lambda_i(0; \theta^*).
\]

Since \(x \log x = O(x^{1+\alpha})\) for any \(\alpha > 0\), we see that there exists some constant \(c > 0\) such that \(x \log x \leq x^{1+\alpha}\) for all \(x \geq c\). Hence for any \(x \geq \mu\), we have by choosing \(\alpha = 1/2\),
\[
x \log x \leq \max\{\mu \log \mu, c \log c, x^{3/2}\}.
\]

Now replacing \(x\) with \(\lambda_i(0; \theta^*)\), and noticing that \(\lambda_i(0; \theta^*) \geq \mu\), then again by Lemma \[3\]
\[
\mathbb{E} \left[ \left| \xi_{U,i}^{(2)}(0) \right|^2 \right] \leq \mathbb{E} \left[ \left( 2\Lambda_0 + |\log \mu| \Lambda_0 + \mu |\log \mu| + c |\log c| + \Lambda_0^{3/2} \right)^2 \right] < \infty.
\]

31
Finally, by Lemma 4 and stationarity of \( \lambda_i(t; \theta^*) \)

\[
\int_0^T \mathbb{E} \left[ \left| \xi_{U,i}(t) \right| \lambda_i(t; \theta^*) \right] dt \leq \int_0^T \sqrt{\mathbb{E} \left[ \lambda_i^2(t; \theta^*) \right]} \mathbb{E} \sup_{\theta \in U} \left| \lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta) \right|^2 dt 
= \sqrt{\mathbb{E}[\lambda_i^2(0; \theta^*)]} \int_0^T \sqrt{\mathbb{E} \sup_{\theta \in U} \left| \lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta) \right|^2} dt 
\leq T \sqrt{F \mathbb{E} \lambda_i^2(0; \theta^*)} < \infty.
\]

\[\square\]

C Full proof of Theorem 1.

**Theorem 1.** Given Assumptions 1, 2, 3, and 4, \( \hat{\theta} \) reg, the regularized MLE defined in (2.5), converges to the true parameter \( \theta^* \) in probability as \( T \to \infty \).

**Step 1: Consistency for \( L_T \) and \( \theta_T \).** Let \( U \) be any neighborhood of \( \theta \). When \( U \) shrinks to \( \{ \theta \} \), by Lemma 3 and Lemma 5, we can apply the Lebesgue dominated convergence theorem to obtain

\[
\mathbb{E} \left[ \sum_{i=1}^K \inf_{\theta \in U} \lambda_i(0; \theta') \right] \to \mathbb{E} \left[ \sum_{i=1}^K \lambda_i(0; \theta) \right],
\]

and

\[
\mathbb{E} \left[ \sum_{i=1}^K \lambda_i(0; \theta^*) \log \left( \frac{\lambda_i(0; \theta^*)}{\sup_{\theta' \in \Theta} \lambda_i(0; \theta')} \right) \right] \to \mathbb{E} \left[ \sum_{i=1}^K \lambda_i(0; \theta^*) \log \left( \frac{\lambda_i(0; \theta^*)}{\lambda_i(0; \theta)} \right) \right].
\]

Here we use the fact that for any \( U \subseteq \Theta \),

\[
\left| \sum_{i=1}^K \inf_{\theta \in U} \lambda_i(0; \theta') \right| \leq K \Lambda_0, \quad \sum_{i=1}^K \lambda_i(0; \theta^*) \log \left( \frac{\lambda_i(0; \theta^*)}{\sup_{\theta' \in \Theta} \lambda_i(0; \theta')} \right) \leq 2K \Lambda_0 \Lambda_1,
\]

\( \mathbb{E}[\Lambda_0] < \infty \), and \( \mathbb{E}[2K \Lambda_0 \Lambda_1] \leq 2K \sqrt{\mathbb{E}[\Lambda_0^2] \mathbb{E}[\Lambda_1^2]} < \infty \) implied from Lemma 3.

Let \( U_0 \) be an open neighborhood of \( \theta^* \). Then by Lemma 6 there exists \( \epsilon > 0 \) such that \( D(\theta^*; \theta) \geq 3 \epsilon \) for any \( \theta \in \Theta \setminus U_0 \). Now for any \( \theta \in \Theta \setminus U_0 \), one can choose a sufficiently small open neighborhood \( \bar{U}_\theta \) of \( \theta \) such that

\[
\mathbb{E} \left[ \sum_{i=1}^K \inf_{\theta' \in \bar{U}_\theta} \lambda_i(0; \theta') - \sum_{i=1}^K \lambda_i(0; \theta^*) + \sum_{i=1}^K \lambda_i(0; \theta^*) \log \left( \frac{\lambda_i(0; \theta^*)}{\sup_{\theta' \in \bar{U}_\theta} \lambda_i(0; \theta')} \right) \right] 
\geq D(\theta^*; \theta) - \epsilon \geq 2 \epsilon.
\]

By the finite covering theorem, one can select a finite number of \( \theta_s \in \Theta \setminus U_0 \), \( 1 \leq s \leq N \) such that the union of the sets \( U_s = U_{\theta_s} \) covers \( \Theta \setminus U_0 \). By Lemma 7 for \( s = 1, 2, \cdots, N \),
\[ \lim_{T \to \infty} \mathbb{P} \left( \frac{1}{T} \int_0^T \left( \sum_{i=1}^K \inf_{\theta \in U_i} \lambda_i(t; \theta) - \lambda_i(t; \theta^*) \right) dt + \frac{1}{T} \sum_{i=1}^K \int_0^T \log \left( \frac{\lambda_i(t; \theta^*)}{\sup_{\theta \in U_i} \lambda_i(t; \theta)} \right) N_i(dt) \right) \geq \mathbb{E} \left[ \sum_{i=1}^K \inf_{\theta \in U_i} \lambda_i(0; \theta') - \sum_{i=1}^K \lambda_i(0; \theta^*) + \sum_{i=1}^K \lambda_i(0; \theta^*) \log \left( \frac{\lambda_i(0; \theta^*)}{\sup_{\theta \in U_i} \lambda_i(0; \theta')} \right) \right] - \epsilon \left( \Theta \right) \] 

which implies

\[ \lim_{T \to \infty} \mathbb{P} \left( \frac{1}{T} \int_0^T \left( \sum_{i=1}^K \inf_{\theta \in U_i} \lambda_i(t; \theta) - \lambda_i(t; \theta^*) \right) dt + \frac{1}{T} \sum_{i=1}^K \int_0^T \log \left( \frac{\lambda_i(t; \theta^*)}{\sup_{\theta \in U_i} \lambda_i(t; \theta)} \right) N_i(dt) \right) \geq \epsilon \left( \Theta \right) \] 

(C.1)

Since

\[ \frac{1}{T} \hat{L}_T(\theta^*) - \sup_{\theta \in U_0} \frac{1}{T} L_T(\theta) \]

\[ \geq \frac{1}{T} \int_0^T \left( \sum_{i=1}^K \inf_{\theta \in U_i} \lambda_i(t; \theta) - \lambda_i(t; \theta^*) \right) dt + \frac{1}{T} \sum_{i=1}^K \int_0^T \log \left( \frac{\lambda_i(t; \theta^*)}{\sup_{\theta \in U_i} \lambda_i(t; \theta)} \right) N_i(dt), \]

(C.2)

\[ \lim_{T \to \infty} \mathbb{P} \left( \frac{1}{T} L_T(\theta^*) - \sup_{\theta \in U_0} \frac{1}{T} L_T(\theta) \geq \epsilon \right) = 1. \] Moreover, since \( \theta^* \in U_0, N \) is finite and the union of \( U_\theta \) covers \( \Theta \setminus U_0 \),

\[ \lim_{T \to \infty} \mathbb{P} \left( \sup_{\theta \in U_0} L_T(\theta) \geq \sup_{\theta \in \Theta \setminus U_0} L_T(\theta) + \epsilon T \right) = 1. \] 

(C.3)

This implies \( \lim_{T \to \infty} \mathbb{P} (\theta_T \in U_0) = 1 \) and the consistency of \( \theta_T \) is then established by the arbitrariness of \( U_0 \).

**Step 2. Consistency for \( \hat{L}_T^{reg} \) and \( \hat{\theta}_T^{reg} \).** As in the first step, to establish the consistency of \( \hat{\theta}_T^{reg} \) is to show

\[ \lim_{T \to \infty} \mathbb{P} \left( \frac{1}{T} \hat{L}_T^{reg}(\theta^*) - \sup_{\theta \in U} \frac{1}{T} \hat{L}_T^{reg}(\theta) \geq \epsilon/4 \right) = 1. \] 

(C.4)

Since \( \Theta \) is compact and \( P \) is continuous, we have \( \sup_{\theta \in \Theta} |P(\theta)| < \infty \). Hence it suffices to prove that

\[ \lim_{T \to \infty} \mathbb{P} \left( \frac{1}{T} \hat{L}_T(\theta^*) - \frac{1}{T} \sup_{\theta \in U} \hat{L}_T(\theta) - \epsilon/4 \geq \frac{1}{T} P(\theta^*) + \frac{1}{T} \sup_{\theta \in U} |P(\theta)| \right) = 1. \] 

(C.5)
By noticing that \( \lim_{T \to \infty} \left| \frac{1}{T} P(\theta^*) + \frac{1}{T} \sup_{\theta \in \Theta} |P(\theta)| \right| \leq \lim_{T \to \infty} \frac{2}{T} \sup_{\theta \in \Theta} |P(\theta)| = 0 \), we see that limit (C.6) can be further reduced to the following:

\[
\lim_{T \to \infty} P \left( \frac{1}{T} \hat{L}_T(\theta^*) - \sup_{\theta \in U_s} \frac{1}{T} \hat{L}_T(\theta) \geq \epsilon/2 \right) = 1. \tag{C.7}
\]

To prove this, notice that

\[
\frac{1}{T} \hat{L}_T(\theta^*) - \sup_{\theta \in U_s} \frac{1}{T} \hat{L}_T(\theta)
\geq \frac{1}{T} \int_0^T \left( \sum_{i=1}^K \inf_{\theta \in U_s} \hat{\lambda}_i(t; \theta) - \hat{\lambda}_i(t; \theta^*) \right) dt + \frac{1}{T} \sum_{i=1}^K \int_0^T \log \left( \frac{\hat{\lambda}_i(t; \theta^*)}{\sup_{\theta \in U_s} \hat{\lambda}_i(t; \theta)} \right) N_i(dt).
\tag{C.8}
\]

By (C.2), we only need to show

\[
\lim_{T \to \infty} P \left( \frac{1}{T} \int_0^T \left( \sum_{i=1}^K \inf_{\theta \in U_s} \hat{\lambda}_i(t; \theta) - \hat{\lambda}_i(t; \theta^*) \right) dt + \frac{1}{T} \sum_{i=1}^K \int_0^T \log \left( \frac{\hat{\lambda}_i(t; \theta^*)}{\sup_{\theta \in U_s} \hat{\lambda}_i(t; \theta)} \right) N_i(dt)
\geq \frac{1}{T} \int_0^T \left( \sum_{i=1}^K \inf_{\theta \in U_s} \lambda_i(t; \theta) - \lambda_i(t; \theta^*) \right) dt
\]

\[
+ \frac{1}{T} \sum_{i=1}^K \int_0^T \log \left( \frac{\lambda_i(t; \theta^*)}{\sup_{\theta \in U_s} \lambda_i(t; \theta)} \right) N_i(dt) - \epsilon/2 \right) = 1. \tag{C.9}
\]

Recall that \( \hat{\lambda}_i(t; \theta) \leq \lambda_i(t; \theta) \) for all \( i = 1, \ldots, K \) and for any \( t \) and \( \theta \in \Theta \), it suffices to show that the following limits hold as \( T \to \infty \):

\[
E \left[ \frac{1}{T} \int_0^T \left[ \inf_{\theta \in U_s} \lambda_i(t; \theta) - \inf_{\theta \in U_s} \hat{\lambda}_i(t; \theta) \right] dt \right] \to 0, \tag{C.10}
\]

\[
E \left[ \frac{1}{T} \int_0^T \left[ \lambda_i(t; \theta^*) - \hat{\lambda}_i(t; \theta^*) \right] dt \right] \to 0, \tag{C.11}
\]

\[
E \left[ \frac{1}{T} \int_0^T \left[ \log \left( \frac{\sup_{\theta \in U_s} \lambda_i(t; \theta)}{\sup_{\theta \in U_s} \hat{\lambda}_i(t; \theta)} \right) \right] N_i(dt) \right] \to 0, \tag{C.12}
\]

\[
E \left[ \frac{1}{T} \int_0^T \left[ \log \left( \frac{\lambda_i(t; \theta^*)}{\hat{\lambda}_i(t; \theta^*)} \right) \right] N_i(dt) \right] \to 0. \tag{C.13}
\]

By the fact that for any \( f \) and \( g \),

\[
\inf_{x \in U} f(x) - \inf_{x \in U} g(x) = -\sup_{x \in U} (-f(x)) + \sup_{x \in U} (-g(x)) \leq \sup_{x \in U} (f(x) - g(x)),
\]

\[
\sup_{x \in U} f(x) - \sup_{x \in U} g(x) \leq \sup_{x \in U} (f(x) - g(x)),
\]

34
and the facts that $\log(1 + x) \leq x$ for $x \geq -1$ and $\lambda_i(t; \theta) \geq \mu$ for any $t$ and $\theta$, one can see that the following inequality holds:

$$
\log \left( \frac{\sup_{\theta \in U_s} \lambda_i(t; \theta)}{\sup_{\theta \in U_s} \hat{\lambda}_i(t; \theta)} \right) = \log \left( \frac{\sup_{\theta \in U_s} \lambda_i(t; \theta) - \sup_{\theta \in U_s} \hat{\lambda}_i(t; \theta)}{\sup_{\theta \in U_s} \hat{\lambda}_i(t; \theta)} + 1 \right) \leq \frac{\sup_{\theta \in U_s} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta))}{\mu}.
$$

Hence (C.10) and (C.12) can be reduced to the following limits as $T \to \infty$,

$$
E \left[ \frac{1}{T} \int_0^T \sup_{\theta \in U_s} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta)) \, dt \right] \to 0, \quad \text{(C.14)}
$$

$$
E \left[ \frac{1}{T} \int_0^T \sup_{\theta \in U_s} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta)) \, N_i(dt) \right] \to 0. \quad \text{(C.15)}
$$

Similarly, (C.13) can be reduced to the following limit as $T \to \infty$:

$$
E \left[ \frac{1}{T} \int_0^T (\lambda_i(t; \theta^*) - \hat{\lambda}_i(t; \theta^*)) \, N_i(dt) \right] \to 0. \quad \text{(C.16)}
$$

Therefore, it suffices to prove that as $T \to \infty$, for any $U \subseteq \Theta$,

$$
E \left[ \frac{1}{T} \int_0^T \sup_{\theta \in U} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta)) \, dt \right] = \frac{1}{T} \int_0^T E \left[ \sup_{\theta \in U} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta)) \right] \, dt \to 0, \quad \text{(C.17)}
$$

$$
E \left[ \frac{1}{T} \int_0^T \sup_{\theta \in U} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta)) \, N_i(dt) \right] \to 0. \quad \text{(C.18)}
$$

To this end, we see by Lemma B.3 and by applying Proposition A.1 to $\xi_i^{(3)}(t)$,

$$
E \left[ \frac{1}{T} \int_0^T \sup_{\theta \in U} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta)) \, N_i(dt) \right] = E \left[ \frac{1}{T} \int_0^T \lambda_i(t; \theta^*) \sup_{\theta \in U} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta)) \right] \, dt \\
\leq \frac{1}{T} \int_0^T \sqrt{E[\lambda_i^2(t; \theta^*)]} \sqrt{E[\sup_{\theta \in U} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta))^2]} \, dt \\
= \sqrt{E[\lambda_i^2(0; \theta^*)]} \frac{1}{T} \int_0^T \sqrt{E[\sup_{\theta \in U} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta))^2]} \, dt.
$$

Moreover, Lemma 3 implies $E[\lambda_i(0; \theta^*)^2] < \infty$, and Lemma 4 shows that $E[\sup_{\theta \in U} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta))^2]$ is uniformly bounded for all $t \geq 0$ and goes to zero as $t \to \infty$. These results further imply that $E[\sup_{\theta \in U} |\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta)|]$ is uniformly bounded for all $t \geq 0$ and goes to zero as $t \to \infty$.

The proof is finished by Proposition 1 and by letting $f(t) = E[\sup_{\theta \in U} (\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta))^2]$ for (C.18) and $f(t) = E[\sup_{\theta \in U} |\lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta)|]$ for (C.17). □
D Proofs for the practical variants: Theorem 2 and Theorem 3

D.1 Proof of Theorem 2

**Theorem 2.** Given Assumptions 1, 2, 3 and 4. Suppose that \( \hat{\theta}_T^{\text{approx}} \) is an \( \epsilon_T \)-approximation of \( \hat{\theta}_T^{\text{reg}} \), with \( \lim_{T \to \infty} \epsilon_T/T = 0 \). Then \( \hat{\theta}_T^{\text{approx}} \) converges to \( \theta^* \) in probability as \( T \to \infty \).

Looking into the proof of Theorem 1, Theorem 2 is indeed a straightforward corollary. Recall that we have proved that

\[
\lim_{T \to \infty} \mathbb{P} \left( \frac{1}{T} \hat{L}_T^{\text{reg}}(\theta^*) - \sup_{\theta \in U} \frac{1}{T} \hat{L}_T^{\text{reg}}(\theta) \geq \frac{\epsilon}{4} \right) = 1. \tag{D.1}
\]

And thus

\[
\lim_{T \to \infty} \mathbb{P} \left( \sup_{\theta \in U_0} \frac{1}{T} \hat{L}_T^{\text{reg}}(\theta) \geq \sup_{\theta \in \Theta \setminus U_0} \frac{1}{T} \hat{L}_T^{\text{reg}}(\theta) + \frac{\epsilon}{4} \right) = 1. \tag{D.2}
\]

By definition, \( \hat{L}_T^{\text{reg}}(\hat{\theta}_T^{\text{approx}}) \geq \max_{\theta \in \Theta} \hat{L}_T^{\text{reg}}(\theta) - \epsilon_T \geq \sup_{\theta \in U_0} \hat{L}_T^{\text{reg}}(\theta) - \epsilon_T \). This implies that

\[
\mathbb{P} \left( \hat{\theta}_T^{\text{approx}} \in \Theta \setminus U_0 \right) \leq \mathbb{P} \left( \sup_{\theta \in \Theta \setminus U_0} \hat{L}_T^{\text{reg}}(\theta) \geq \hat{L}_T^{\text{reg}}(\hat{\theta}_T^{\text{approx}}) \right) \leq \mathbb{P} \left( \sup_{\theta \in \Theta \setminus U_0} \frac{1}{T} \hat{L}_T^{\text{reg}}(\theta) > \sup_{\theta \in U_0} \frac{1}{T} \hat{L}_T^{\text{reg}}(\theta) - \frac{\epsilon_T}{T} \right) \tag{D.3}
\]

\[
= \mathbb{P} \left( \sup_{\theta \in U_0} \frac{1}{T} \hat{L}_T^{\text{reg}}(\theta) < \sup_{\theta \in \Theta \setminus U_0} \frac{1}{T} \hat{L}_T^{\text{reg}}(\theta) + \frac{\epsilon_T}{T} \right) \to 0,
\]

as \( T \) goes to infinity because of (D.2) and the fact that \( \epsilon_T/T \to 0 \).

Then we have \( \mathbb{P} \left( \hat{\theta}_T^{\text{approx}} \in U_0 \right) \to 1 \). The proof is then finished. \( \square \)

D.2 Proof of Theorem 3

Throughout this section, we denote \( \{I_n\}_{n=0}^{\infty} \) as a sequence of (integration) intervals, where \( I_n \) are disjoint closed intervals and are arranged in increasing orders. Define \( |I_n| \) to be the length of \( I_n \).

**Theorem 3.** Given Assumptions 1, 2, 3 and 4. For any \( \epsilon_1, \epsilon_2 > 0 \), \( \exists L_{\epsilon_1, \epsilon_2} > 0 \), s.t. if \( \inf_n (l_n - u_n) \geq L_{\epsilon_1, \epsilon_2} + c \) for some \( c > 0 \), by taking \( I_n := [l_n + L_{\epsilon_1, \epsilon_2}, u_n] \), then

\[
\lim_{N \to \infty} \mathbb{P} \left( \hat{\theta}_N^{\text{reg}} \in B(\theta^*, \epsilon_1) \right) \geq 1 - \epsilon_2,
\]

i.e., the regularized MLE \( \hat{\theta}_N^{\text{reg}} \) converges to \( B(\theta^*, \epsilon_1) \) with a probability at least \( 1 - \epsilon_2 \) as \( N \to \infty \). Here \( B(x, r) \) is the open ball centered at \( x \) with radius \( r \).
The key to proving the above theorem is a fictitious estimator $\hat{\theta}_{\text{reg}}^N := \arg\max_{\theta \in \Theta} \hat{L}_N(\theta) - P(\theta)$, where $\hat{L}_N$ is defined by replacing $\tilde{\lambda}$ with $\hat{\lambda}$ in the definition of $\tilde{L}_N$, i.e.,

$$\hat{L}_N(\theta) := \arg\max_{\theta \in \Theta} \left[ -\int_{\cup_{n=0}^{N} I_n} \lambda_i(t; \theta) dt + \int_{\cup_{n=0}^{N} I_n} \log(\hat{\lambda}_i(t; \theta)) N_i(dt) \right].$$ (D.4)

The above estimator is called “fictitious” as it is computationally infeasible in practice, since the computation of $\hat{\lambda}$ requires obtaining the missing data.

The proof of Theorem 3 is then divided into two steps. Firstly, we establish a new ergodicity lemma similar to Lemma 7 for integration over a sequence of disjoint closed intervals. This enables us to prove the consistency of $\hat{\theta}_{\text{reg}}^N$. Secondly, we show that the difference between $\hat{\lambda}$ and $\tilde{\lambda}$ can be well controlled in the mean-square sense. And this finally allows us to establish the approximate consistency of $\tilde{\theta}_{\text{reg}}^N$ following a similar proof as in the second step of the proof for Theorem 1.

Step 1: consistency of $\hat{\theta}_{\text{reg}}^N$. We begin by generalizing Propositions 1, A.1 and A.2 to deal with a general sequence of disjoint integration intervals.

Proposition D.1. Suppose that $f(t) : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is Lebesgue measurable, and $f(t) \leq C$ for all $t \geq 0$ and some $C > 0$. If $f(t) \rightarrow 0$ as $t \rightarrow \infty$. Also suppose that $\sum_{n=0}^{\infty} |I_n| = \infty$. Then

$$\frac{1}{\sum_{n=0}^{N} |I_n|} \int_{\cup_{n=0}^{N} I_n} f(t) dt \rightarrow 0 \quad \text{as} \quad T \rightarrow \infty.$$ 

Proposition D.2. Suppose that $\{\xi(t); \ t \geq 0\}$ is a finite predictable process such that $\int_{\cup_{n=0}^{N} I_n} \mathbb{E}[|\xi(t)| \lambda_i(t; \theta^*)] dt < \infty$ for any $N \geq 0$, then

$$\mathbb{E} \left[ \int_{\cup_{n=0}^{N} I_n} \xi(t) N_i(dt) \right] = \mathbb{E} \left[ \int_{\cup_{n=0}^{N} I_n} \xi(t) \lambda_i(t; \theta^*) dt \right].$$

Proposition D.3. Suppose that $\{\xi(t); \ t \geq 0\}$ is a finite predictable process, with

$$\mathbb{E} \left[ \int_{\cup_{n=0}^{N} I_n} \xi(t)^2 N_i(dt) \right] < \infty$$

for any $T \geq 0$, $i = 1, \ldots, K$. Define $M_i(t) := N_i(t) - \int_{0}^{t} \lambda_i(s; \theta^*) ds$ and

$$X_i(t) := \int_{\cup_{n=0}^{N} I_n} \xi(s) M_i(ds).$$

Then

$$\mathbb{E}[X_i(t)^2] = \mathbb{E} \left[ \int_{\cup_{n=0}^{N} I_n} \xi(s)^2 N_i(ds) \right].$$
Here Propositions \textcolor{red}{[D.1]} and \textcolor{red}{[D.2]} are straightforward by linearity, and Proposition \textcolor{red}{[D.3]} can be shown by observing that $\mathbb{E}[(M_t - M_s)(M_{t'} - M_{s'})] = 0$ for any martingale $M$ and $s' < t' < s < t$.

Now we can establish the following ergodicity lemma for an arbitrary sequence of integration intervals $\{I_n\}_{n=0}^{\infty}$, where $I_n$ are disjoint and arranged in increasing orders. Define $|I_n|$ to be the length of $I_n$.

\textbf{Lemma D.1 (Generalized ergodicity).} Suppose that $\sum_{n=0}^{\infty} |I_n| = \infty$. Under Assumptions \textcolor{red}{[A.1]} and \textcolor{red}{[A.2]} replaced by Lemma \textcolor{red}{D.1}, Proposition \textcolor{red}{D.1}, Proposition \textcolor{red}{D.2} and we take

$$\lim_{N \to \infty} \frac{1}{\sum_{n=0}^{N} |I_n|} \int_{\bigcup_{n=0}^{N} I_n} \left( \inf_{\theta \in U} \lambda_i(t; \theta) - \lambda_i(t; \theta^*) \right) dt \xrightarrow{p} \mathbb{E} \left[ \inf_{\theta \in U} \lambda_i(0; \theta) - \lambda_i(0; \theta^*) \right], \quad (D.5)$$

$$\lim_{N \to \infty} \frac{1}{\sum_{n=0}^{N} |I_n|} \int_{\bigcup_{n=0}^{N} I_n} \log \left( \frac{\lambda_i(t; \theta^*)}{\sup_{\theta \in U} \lambda_i(t; \theta)} \right) N_i(dt) \xrightarrow{p} \mathbb{E} \left[ \lambda_i(0; \theta^*) \log \left( \frac{\lambda_i(0; \theta^*)}{\sup_{\theta \in U} \lambda_i(0; \theta)} \right) \right]. \quad (D.6)$$

The proof is done by replacing the classic Birkhoff’s ergodic theorem used to prove Lemma \textcolor{red}{7} with its non-uniform generalization (\textcolor{red}{[KT03, Kor10]}), which allows for an arbitrary probability density $\rho$ supported on $[0, T]$. Here to apply the non-uniform Birkhoff’s theorem, we take

$$T := \sum_{n=0}^{N} |I_n|, \quad \rho(t) := 1/ \sum_{n=0}^{N} |I_n| \text{ for } t \in \bigcup_{n=0}^{N} I_n \text{ and } \rho(t) := 0 \text{ otherwise.}$$

Note that when $\rho(t) := 1/T$ for $t \in [0, T]$ and $\rho(t) := 0$ otherwise, the non-uniform Birkhoff theorem reduces to the classical uniform version as we used in proving Lemma \textcolor{red}{7}.

One can now repeat the proof of Theorem \textcolor{red}{1}, with Lemma \textcolor{red}{7} Proposition \textcolor{red}{1} Proposition \textcolor{red}{A.1} and Proposition \textcolor{red}{A.2} replaced by Lemma \textcolor{red}{D.1}, Proposition \textcolor{red}{D.1}, Proposition \textcolor{red}{D.2} and Proposition \textcolor{red}{D.3}, respectively. And the following consistency result for the fictitious estimator can be proved.

\textbf{Theorem D.1 (Consistency of the fictitious estimator).} Given Assumptions \textcolor{red}{7} \textcolor{red}{2} \textcolor{red}{3} \textcolor{red}{5} \textcolor{red}{6} and \textcolor{red}{4} \textcolor{red}{-reg}

Suppose that $\sum_{n=0}^{\infty} |I_n| = \infty$. Then the fictitious estimator $\hat{\theta}_N$ converges to $\theta^*$ in probability as $N \to \infty$.

In addition, for any open neighborhood $U_0$ of $\theta^*$, $\exists \epsilon_0 > 0$ and a finite cover $\{U_i\}_{i=1}^{M}$ of $\Theta \setminus U_0$, such that as $N \to \infty$,

$$\lim_{N \to \infty} \mathbb{P} \left( \frac{1}{\sum_{n=0}^{N} |I_n|} \int_{\bigcup_{n=0}^{N} I_n} \left( \sum_{i=1}^{K} \inf_{\theta \in U_i} \lambda_i(t; \theta) - \hat{\lambda}_i(t; \theta^*) \right) dt + \frac{1}{\sum_{n=0}^{N} |I_n|} \sum_{i=1}^{K} \int_{\bigcup_{n=0}^{N} I_n} \log \left( \frac{\hat{\lambda}_i(t; \theta^*)}{\sup_{\theta \in U_i} \hat{\lambda}_i(t; \theta)} \right) N_i(dt) \geq \inf_{\theta \in \Theta \setminus U_0} D(\theta^*; \theta)/6 \right) = 1. \quad (D.7)$$
Notice that here the limit \([D.7]\) is exactly a counterpart of \([C.2]\) and \([C.9]\) in the missing data setting here, with \(\epsilon\) replaced with \(D(\theta^*; \theta)/6\) (since \(\epsilon\) comes from \(D(\theta^*; \theta) \geq 3\epsilon\)), and will be used to ultimately establish the consistency of \(\hat{\theta}_N^{reg}\) below.

Now assuming that the triggering functions have compact supports, \(i.e., \text{supp}(g_{ij}) \subseteq [0, C]\) for \(i, j = 1, \ldots, K\) for some \(C \in [0, \infty)\), and supposing that \(u_n - l_n \geq C + c\) for all \(n \geq 0\) and some constant \(c \in [0, \infty)\), by taking \(I_n := [l_n + C, u_n]\), we have \(\hat{\lambda}(t; \theta) = \hat{\lambda}(t; \theta)\) for any \(t \in \bigcup_{n=0}^{\infty} I_n\). Then from the above theorem, this immediately leads to Corollary 3.1.

**Step 2: bounding the truncation errors.** Now we consider generalizing the previous results to allow for triggering functions with general supports (\(e.g.,\) exponential and power-law triggering functions).

The key ingredient is the following bound on the truncation error between \(\hat{\lambda}\) and \(\tilde{\lambda}\), which can be derived from Lemma 1 (or more directly, Lemma B.1) following a similar approach by which we prove Lemma A.

**Lemma D.2 (Truncation error).** Under Assumptions 1, 2 and 4, for any \(\epsilon > 0\), \(\exists L_{\epsilon} > 0\), s.t. if \(\inf_n (u_n - l_n) \geq L_{\epsilon} + c\) for some \(c > 0\), by taking \(I_n := [l_n + L_{\epsilon}, u_n]\), we have

\[
\mathbb{E} \left[ \sup_{\theta \in \Theta} \left| \hat{\lambda}_i(t; \theta) - \hat{\lambda}_i(t; \theta) \right|^2 \right] \leq \epsilon
\]

for all \(i = 1, \ldots, K\) and \(t \in \bigcup_{n=0}^{\infty} I_n\).

**Proof.** Define \(C_j^n := C_j^{(l_n)}\). Then by Lemma B.1, \(\mathbb{E}[|C_j^n|^2] \leq (\mathbb{E}[|C_j^{(3+\alpha)}|^2])^{2/(3+\alpha)} \leq C^{2/(3+\alpha)}\) for all \(n \geq 0\) and \(\alpha \in [0, 1)\). Denote \(\hat{C} := C^{2/(3+\alpha)}\) for simplicity.

Now given \(\epsilon > 0\), by Assumption 4, \(\exists L_{\epsilon} > 0\), such that for all \(t \geq L_{\epsilon}\),

\[
\sum_{k=1}^{\infty} (t_k - t_{k-1}) \sup_{t' \in [t + t_{k-1}, t + t_k], \eta' \in \Theta} g_{ij}(t' \eta') \leq \frac{\sqrt{\epsilon}}{K \sqrt{\hat{C}}}
\]

Then notice that if \(t \in I_n\) for some \(n \geq 0\), then by defining \(\delta t := t - l_n - L_{\epsilon}\), we have

\[
\sup_{\theta \in \Theta} \left| \hat{\lambda}_i(t; \theta') - \hat{\lambda}_i(t; \theta') \right| \leq \sum_{j=1}^{K} C_j^n \sum_{\{k: t_k \leq t_n\}} (t_k - t_{k-1}) \sup_{t' \in [\delta t + L_{\epsilon} + t_{k-1}, \delta t + L_{\epsilon} + t_k], \eta' \in \Theta} g_{ij}(t' \eta')
\]

\[
\leq \sum_{j=1}^{K} C_j^n \sum_{k=1}^{\infty} (t_k - t_{k-1}) \sup_{t' \in [\delta t + L_{\epsilon} + t_{k-1}, \delta t + L_{\epsilon} + t_k], \eta' \in \Theta} g_{ij}(t' \eta').
\]

where \(\{t_k\}_{k \geq 0}\) is the sequence in Assumption 4.

Finally, we have

\[
\mathbb{E} \left[ \sup_{\theta \in \Theta} \left| \hat{\lambda}_i(t; \theta') - \hat{\lambda}_i(t; \theta') \right| ^2 \right] \leq K \sum_{j=1}^{K} E[|C_j^n|^2] \epsilon / (K^2 \hat{C}) \leq \epsilon,
\]

which completes the proof. \(\square\)

39
We also extend the part of the results in Lemma 4 and Lemma D.3 to the difference between \( \lambda \) and \( \hat{\lambda} \).

**Lemma D.3.** Under Assumptions 1, 2 and 4 for any \( \theta \in \Theta \), the stochastic processes \( \xi_{U,i}^{(3)}(t) := \sup_{\theta \in U} (\hat{\lambda}_i(t;\theta) - \tilde{\lambda}_i(t;\theta)) \) \((i = 1, \ldots, K)\) are predictable, and for all \( t \geq 0 \) and some constant \( F > 0 \), \( \mathbb{E} \left[ |\xi_{U,i}^{(3)}(t)|^2 \right] \leq F \). In addition, for any \( T \geq 0 \) and \( i = 1, \ldots, K \),

\[
\int_0^T \mathbb{E} \left[ |\xi_{U,i}^{(3)}(t)| \lambda_i(t;\theta^*) \right] dt < \infty.
\]

The proof is mostly a repetition of the proofs for Lemma 4 and Lemma D.3 with different notations, and the main difference in the proof is also exhibited in the proof of Lemma D.2 above. We thus omit the proof of Lemma D.3 here.

**Finishing proof of Theorem 3: consistency of \( \hat{\theta}_{reg} \).** Finally, combining the consistency of \( \hat{\theta}_{reg} \) (Theorem D.1), Lemmas D.2 and D.3 and Propositions D.1, D.2 and D.3, we arrive at the consistency result in Theorem 3, as shown below.

**Proof.** We are now ready to put together the established results to finish the final piece towards proving Theorem 3.

Given \( \epsilon_1, \epsilon_2 > 0 \), let \( U_0 := B(\theta^*, \epsilon_1) \). Then by Lemma 6, there exists \( \epsilon' > 0 \) such that \( D(\theta^*; \theta) \geq 6\epsilon' \) for any \( \theta \in \Theta \setminus U_0 \). We then notice that to establish that \( \hat{\theta}_{reg} \) converges to \( U_0 = B(\theta^*, \epsilon_1) \) with probability at least \( 1 - \epsilon_2 \), it suffices to show that

\[
\lim_{N \to \infty} \mathbb{P} \left( \tilde{L}_{reg}(\theta^*) - \sup_{\theta \in U_s} \tilde{L}_{reg}(\theta) \geq \epsilon' \frac{1}{4} \sum_{n=0}^{N} |I_n| \right) \geq 1 - \epsilon_2. \tag{D.8}
\]

for the finite cover \( \{U_s\}_{s=1}^M \) in Theorem D.1. Notice that here \( \tilde{L}_{reg}(\theta) = \tilde{L}_{reg}(\theta) - P(\theta) \).

Now following the same derivations from (C.6) to (C.9) in step 2 in the proof of Theorem 1 and by (D.7) and the fact that \( \lim_{\theta \in \Theta \setminus U_0} D(\theta^*; \theta)/6 \geq \epsilon' \), one only needs to show that

\[
\lim_{N \to \infty} \mathbb{P} \left( \frac{1}{\sum_{n=0}^{N} |I_n|} \int_{\bigcup_{n=0}^{N} I_n} \sum_{i=1}^{K} \inf_{\theta \in U_s} \Lambda_i(t;\theta) - \Lambda_i(t;\theta^*) \right) dt \\
+ \frac{1}{\sum_{n=0}^{N} |I_n|} \sum_{i=1}^{K} \int_{\bigcup_{n=0}^{N} I_n} \log \left( \frac{\Lambda_i(t;\theta^*)}{\sup_{\theta \in U_s} \Lambda_i(t;\theta)} \right) N_i(dt) \\
\geq \frac{1}{\sum_{n=0}^{N} |I_n|} \int_{\bigcup_{n=0}^{N} I_n} \left( \sum_{i=1}^{K} \inf_{\theta \in U_s} \Lambda_i(t;\theta) - \Lambda_i(t;\theta^*) \right) dt \\
+ \frac{1}{\sum_{n=0}^{N} |I_n|} \sum_{i=1}^{K} \int_{\bigcup_{n=0}^{N} I_n} \log \left( \frac{\Lambda_i(t;\theta^*)}{\sup_{\theta \in U_s} \Lambda_i(t;\theta)} \right) N_i(dt) - \epsilon'/2 \right) \geq 1 - \epsilon_2. \tag{D.9}
\]
Noticing that $\tilde{\lambda}_i(t; \theta) \leq \hat{\lambda}_i(t; \theta)$ for all $i = 1, \ldots, K$, again following the same derivations from (C.10) to (C.18), and utilizing the Markov inequality, one can reduce (D.9) to the following two inequalities for $i = 1, \ldots, K$:

$$\frac{1}{\sum_{n=0}^{N} |I_n|} \int_{\bigcup_{n=0}^{N} I_n} \mathbb{E} \left[ \sup_{\theta \in U} (\hat{\lambda}_i(t; \theta) - \tilde{\lambda}_i(t; \theta)) \right] dt \leq \frac{\epsilon' \epsilon_2}{32}, \quad (D.10)$$

$$\frac{1}{\sum_{n=0}^{N} |I_n|} \int_{\bigcup_{n=0}^{N} I_n} \mathbb{E} \left[ \sup_{\theta \in U} (\tilde{\lambda}_i(t; \theta) - \tilde{\lambda}_i(t; \theta)) \right] N_i(dt) \leq \frac{\epsilon' \epsilon_2}{32}, \quad (D.11)$$

for any $U \subseteq \Theta$.

To prove (D.10) and (D.11), notice that by Lemma D.3 and by applying Proposition D.2 to $\epsilon_i(t)$, we arrive at

$$\frac{1}{\sum_{n=0}^{N} |I_n|} \mathbb{E} \left[ \int_{\bigcup_{n=0}^{N} I_n} \left[ \sup_{\theta \in U} (\hat{\lambda}_i(t; \theta) - \tilde{\lambda}_i(t; \theta)) \right] N_i(dt) \right] = \frac{1}{\sum_{n=0}^{N} |I_n|} \mathbb{E} \left[ \int_{\bigcup_{n=0}^{N} I_n} \lambda_i(t; \theta^*) \sup_{\theta \in U} (\hat{\lambda}_i(t; \theta) - \tilde{\lambda}_i(t; \theta)) dt \right] \leq \frac{1}{\sum_{n=0}^{N} |I_n|} \int_{\bigcup_{n=0}^{N} I_n} \sqrt{\mathbb{E} [\lambda^2_i(t; \theta^*)] \mathbb{E} [\sup_{\theta \in U} (\hat{\lambda}_i(t; \theta) - \tilde{\lambda}_i(t; \theta))^2] dt}$$

Similarly, we also have

$$\frac{1}{\sum_{n=0}^{N} |I_n|} \int_{\bigcup_{n=0}^{N} I_n} \mathbb{E} \left[ \sup_{\theta \in U} (\tilde{\lambda}_i(t; \theta) - \tilde{\lambda}_i(t; \theta)) \right] dt \leq \frac{1}{\sum_{n=0}^{N} |I_n|} \int_{\bigcup_{n=0}^{N} I_n} \sqrt{\mathbb{E} [\sup_{\theta \in U} (\hat{\lambda}_i(t; \theta) - \tilde{\lambda}_i(t; \theta))^2] dt}.$$ 

Now by Lemma D.3 $\mathbb{E} [\lambda^2_i(0; \theta^*)] \leq D$ for some constant $D > 0$. Hence by Lemma D.2, $L_{\epsilon_1, \epsilon_2} > 0$, such that if $\inf_n (u_n - l_n) > L_{\epsilon_1, \epsilon_2}$, by taking $I_n := [l_n + L_{\epsilon_1, \epsilon_2}, u_n]$, we have

$$\mathbb{E} \left[ \sup_{\theta \in \Theta} (\tilde{\lambda}_i(t; \theta) - \hat{\lambda}_i(t; \theta))^2 \right] \leq \left( \frac{\epsilon' \epsilon_2}{32} \right)^2 \cdot \min \{1, 1/D\} \quad (D.12)$$

for all $i = 1, \ldots, K$ and $t \in \bigcup_{n=0}^{N} I_n$.

This finally establishes inequalities (D.10) and (D.11), which completes the proof.

\section{Extensions and future work}

\begin{itemize}
\item \textbf{Consistency with repeated finite-time experiments} This paper is focused on the MLEs defined on a single simulation path, and the asymptotic behavior is studied as the
simulation time $T$ goes to infinity. However, in practice people also utilize several independent finite-time simulation paths resulting from repeated experiments (e.g., [XFZ16]). It would be interesting to see the asymptotic behavior as the repetitions goes to infinity. Some quantification (e.g., central limit theorem type convergence) of the relation between the two limits (i.e., the simulation time $T$ and the number of repetitions) may also shed light on some practical trade-offs in the data collection/usage.

Towards more flexible modeling. It is worth noticing that the Assumptions [14] that lead to our asymptotic convergence results may include a wide class of deep neural networks (DNN) based triggering functions as well, and it would be interesting to empirically test the power of the end-to-end DNN-based parametric MLE compared to non-parametric methods that are commonly used in the literature. It is also natural to consider extensions of our results to marked and/or non-linear MHPs. In addition, one may consider a coupled system of an MDP and a state-dependent MHP ([MPP18]), and apply it in a reinforcement learning scenario. Here the inclusion of MHP may serve as a natural exploration bonus (or risk component).

Missing data with entity-dependent observation intervals. Finally, we remark that it would be interesting to consider further extensions of Theorem 3 that allow one to deal with different observed intervals for different entities $k = 1, \ldots, K$, as considered empirically in [Le18].