Bernoulli **24**(4A), 2018, 2875–2905
https://doi.org/10.3150/17-BEJ948

Large deviations and applications for Markovian Hawkes processes with a large initial intensity

XUEFENG GAO[1] and LINGJIONG ZHU[2]

[1]*Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T. Hong Kong. E-mail: xfgao@se.cuhk.edu.hk*
[2]*Department of Mathematics, Florida State University, 1017 Academic Way, Tallahassee, FL-32306, USA. E-mail: zhu@math.fsu.edu*

Hawkes process is a class of simple point processes that is self-exciting and has clustering effect. The intensity of this point process depends on its entire past history. It has wide applications in finance, insurance, neuroscience, social networks, criminology, seismology, and many other fields. In this paper, we study linear Hawkes process with an exponential kernel in the asymptotic regime where the initial intensity of the Hawkes process is large. We establish large deviations for Hawkes processes in this regime as well as the regime when both the initial intensity and the time are large. We illustrate the strength of our results by discussing the applications to insurance and queueing systems.

Keywords: Hawkes processes; insurance; large deviations; large initial intensity; queueing systems

1. Introduction

Let N be a simple point process on $\mathbb{R}$ and let $\mathcal{F}_t^{-\infty} := \sigma(N(C), C \in \mathcal{B}(\mathbb{R}), C \subset (-\infty, t])$ be an increasing family of σ-algebras. Any nonnegative $\mathcal{F}_t^{-\infty}$-progressively measurable process λ_t with

$$\mathbb{E}\big[N(a,b]|\mathcal{F}_a^{-\infty}\big] = \mathbb{E}\left[\int_a^b \lambda_s\, ds \middle| \mathcal{F}_a^{-\infty}\right] \qquad \text{almost surely,}$$

for all intervals $(a,b]$ is called an $\mathcal{F}_t^{-\infty}$-intensity of N. We use the notation $N_t := N(0,t]$ to denote the number of points in the interval $(0,t]$.

A Hawkes process is a simple point process N admitting an $\mathcal{F}_t^{-\infty}$-intensity

$$\lambda_t := \lambda\left(\int_{-\infty}^{t-} \phi(t-s)\, dN_s\right), \tag{1.1}$$

where $\lambda(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ is locally integrable, left continuous, $\phi(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ and we always assume that $\|\phi\|_{L^1} = \int_0^\infty \phi(t)\, dt < \infty$. In (1.1), $\int_{-\infty}^{t-} \phi(t-s)\, dN_s$ stands for $\sum_{\tau<t} \phi(t-\tau)$, where τ are the occurrences of the points before time t. In the literature, $\phi(\cdot)$ and $\lambda(\cdot)$ are usually referred to as exciting function (or sometimes kernel function) and rate function respectively. A Hawkes process is linear if $\lambda(\cdot)$ is linear and it is nonlinear otherwise.

1350-7265 © 2018 ISI/BS

The linear Hawkes process was first introduced by A.G. Hawkes in 1971 [17,18]. It naturally generalizes the Poisson process and it captures both the self-exciting[1] property and the clustering effect. In addition, Hawkes process is a very versatile model which is amenable to statistical analysis. These explain why it has wide applications in insurance, finance, social networks, neuroscience, criminology and many other fields. For a list of references, we refer to [30].

Throughout this paper, we assume an exponential exciting function $\phi(t) := \alpha e^{-\beta t}$ where $\alpha, \beta > 0$, and a linear rate function $\lambda(z) := \mu + z$ where the base intensity $\mu \geq 0$. That is, we restrict ourselves to the linear Markovian Hawkes process. To see the Markov property, we define

$$Z_t := \int_{-\infty}^{t} \alpha e^{-\beta(t-s)}\, dN_s = Z_0 \cdot e^{-\beta t} + \int_0^t \alpha e^{-\beta(t-s)}\, dN_s.$$

Then, the process Z is Markovian and satisfies the dynamics:

$$dZ_t = -\beta Z_t\, dt + \alpha\, dN_t,$$

where N is a Hawkes process with intensity $\lambda_t = \mu + Z_{t-}$ at time t. In addition, the pair (Z, N) is also Markovian. For simplicity, we also assume $Z_0 = Z_{0-}$, that is, there is no jump at time zero.

In this paper, we consider an asymptotic regime where $Z_0 = n$, and $n \in \mathbb{R}^+$ is sent to infinity. This implies the initial intensity $\lambda_0 = \mu + Z_0$ is large for fixed μ. Our main contribution is to provide the large deviations analysis of Markovian Hawkes processes in this asymptotic regime as well as the regime when both Z_0 and the time are large. The rate functions are found explicitly. Our large deviations analysis here complement our previous results in [13], where we establish functional law of large numbers and functional central limit theorems for Markovian Hawkes processes in the same asymptotic regimes.

For simplicity, the discussions in our paper are restricted to the case when the exciting function ϕ is exponential, that is the Markovian case. Indeed, all the results can be extended to the case when the exciting function ϕ is a sum of exponential functions. For the non-Markovian case, we know that any continuous and integrable function ϕ can be approximated by a sum of exponential functions, see, for example, [36]. In this respect, the Markovian setting in this paper is not too restrictive. From the application point of view, the exponential exciting function and thus the Markovian case, together with the linear rate function, is the most widely used due to the tractability of the theoretical analysis as well as the simulations and calibrations. See, for example, [1,7,17] and the references therein.

To illustrate the strength of our results, we apply them to two examples. In the first example, we develop approximations for finite-horizon ruin probabilities in the insurance setting where claim arrivals are modeled by Hawkes processes. Here, the initial arrival rate of claims could be high, say, right after a catastrophe event. In the second example, we rely on our large deviations results to approximate the loss probability in a multi-server queueing system where the traffic input is given by a Hawkes process with a large initial intensity. Such a queueing system could

[1] Self-exciting refers the phenomenon that the occurrence of one event increases the probability of the occurrence of further events.

be relevant for modeling large scale service systems (e.g., server farms with thousands of servers) with high-volume traffic which exhibits clustering.

We now explain the difference between our work and the existing literature on limit theorems of Hawkes processes, especially the large deviations. The large-time large deviations of Hawkes processes have been extensively studied in the literature, that is the large deviation principle for $\mathbb{P}(N_t/t \in \cdot)$ as $t \to \infty$. Bordenave and Torrisi [5] derived the large deviations when $\lambda(\cdot)$ is linear and obtained a closed-form formula for the rate function. When $\lambda(\cdot)$ is nonlinear, the lack of immigration-birth representation [17] makes the study of large deviations much more challenging mathematically. In the case when $\phi(\cdot)$ is exponential, the large deviations were obtained in Zhu [36] by using the Markovian property, and $\lambda(\cdot)$ is assumed to be sublinear so that a delicate application of minmax theorem can match the lower and upper bounds. For the general non-Markovian case, that is, general $\phi(\cdot)$, the large deviations was obtained at the process-level in Zhu [35]. The large deviations for extensions of Hawkes processes have also been studied in the literature, see, for example, Karabash and Zhu [22] for the linear marked Hawkes process, and Zhu [34] for the Cox–Ingersoll–Ross process with Hawkes jumps and also Zhang *et al.* [29] for affine point processes. Other than the large deviations, the central limit theorems for linear, nonlinear and extensions of Hawkes processes have been considered in, e.g., [3,33,34]. Recently, Torrisi [25,26] studied the rate of convergence in the Gaussian and Poisson approximations of the simple point processes with stochastic intensity, which includes as a special case, the nonlinear Hawkes process. The moderate deviations for linear Hawkes processes were obtained in Zhu [32], that fills in the gap between the central limit theorem and large deviations. Also, the large-time limit theorems for nearly unstable, or nearly critical Hawkes processes have been considered in Jaisson and Rosenbaum [20,21]. The large-time asymptotics for other regimes are referred to Zhu [30]. The limit theorems considered in Bacry *et al.* [3] hold for the multidimensional Hawkes process. Indeed, one can also consider the large dimensional asymptotics for the Hawkes process, that is, mean-field limit, see, for example, Delattre *et al.* [8].

We organize our paper as follows. In Section 2, we will state the main theoretical results in our paper,that is, the large deviations for the linear Markovian Hawkes processes with a large initial intensity. We will then discuss the applications of our results to two examples in Section 3. We prove Theorems 1 and 2 in Section 4. Technical proofs for additional results will be presented in the supplemental article [14] due to space considerations.

2. Main results

In this section, we state our main results. First, let us introduce the notation that will be used throughout the paper and introduce the definition and the contraction principle in the large deviations theory that will be used repeatedly in the paper.

2.1. Notation and background of large deviations theory

We define $\mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}$ and $\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} : x \geq 0\}$. We fix $T > 0$ throughout this paper. Let us first define the following spaces:

- $D[0, T]$ is defined as the space of càdlàg functions from $[0, T]$ to $\mathbb{R}_{\geq 0}$.

- $\mathcal{AC}_x[0,T]$ is defined as the space of absolutely continuous functions from $[0,T]$ to $\mathbb{R}_{\geq 0}$ that starts at x at time 0.
- $\mathcal{AC}_x^+[0,T]$ is defined as the space that consists of all the non-decreasing functions $f \in \mathcal{AC}_x[0,T]$.

We also define $B_\varepsilon(x)$ as the Euclidean ball centered at x with radius $\varepsilon > 0$.

Before we proceed, let us give a formal definition of the large deviation principle and state the contraction principle. We refer readers to Dembo and Zeitouni [9] or Varadhan [27] for general background of large deviations and the applications.

A sequence $(P_n)_{n\in\mathbb{N}}$ of probability measures on a topological space X satisfies the large deviation principle with the speed a_n and the rate function $I : X \to [0,\infty]$ if I lower semicontinuous and for any measurable set A, we have

$$-\inf_{x\in A^o} I(x) \leq \liminf_{n\to\infty} \frac{1}{a_n} \log P_n(A) \leq \limsup_{n\to\infty} \frac{1}{a_n} \log P_n(A) \leq -\inf_{x\in\overline{A}} I(x).$$

Here, A^o is the interior of A and $\overline{A}$ is its closure. The rate function I is said to be good if for any m, the level set $\{x : I(x) \leq m\}$ is compact.

The contraction principle concerns the behavior of large deviation principle under continuous mapping from one space to another. It states that if $(P_n)_{n\in\mathbb{N}}$ satisfies a large deviation principle on X with a good rate function $I(\cdot)$, and F is a continuous mapping from the Polish space X to another Polish space Y, then the family $Q_n = P_n F^{-1}$ satisfies a large deviation principle on Y with a good rate function $J(\cdot)$ given by

$$J(y) = \inf_{x:F(x)=y} I(x).$$

2.2. Large deviation analysis for large initial intensity

In this section, we state a set of results on large deviations behavior of Markovian Hawkes processes when $Z_0 = n$ is sent to infinity. Note that processes Z and N both depend on the initial condition $Z_0 = n$ and we use Z^n, N^n to emphasize the dependence on $Z_0 = n$. We consider the process Z^n first.

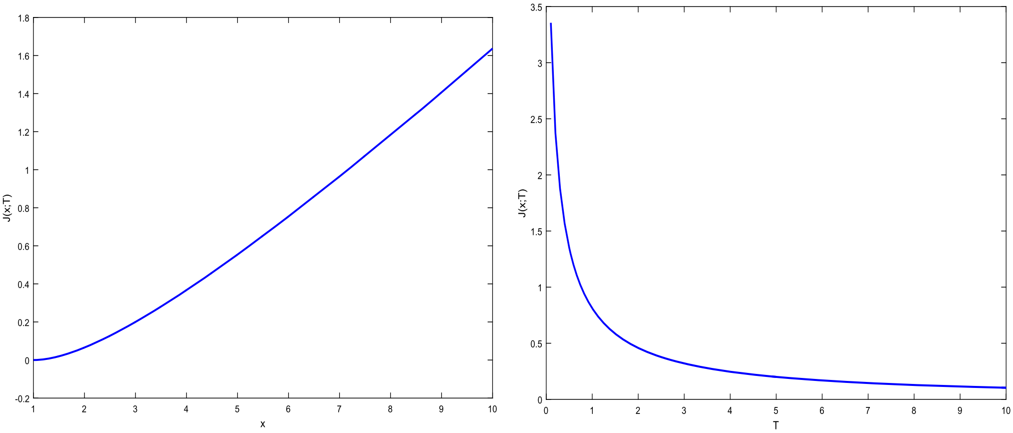
Theorem 1. $\mathbb{P}(\{\frac{1}{n}Z_t^n, 0 \leq t \leq T\} \in \cdot)$ *satisfies a sample-path large deviation principle on* $D[0,T]$ *equipped with uniform topology with the speed n and the good rate function*

$$I_Z(g) = \int_0^T \frac{\beta g(t) + g'(t)}{\alpha} \log\left(\frac{\beta g(t) + g'(t)}{\alpha g(t)}\right) - \left(\frac{\beta g(t) + g'(t)}{\alpha} - g(t)\right) dt, \tag{2.1}$$

if $g \in \mathcal{AC}_1[0,T]$ *and* $g' \geq -\beta g$, *and* $I_Z(g) = \infty$ *otherwise. Moreover,* $\mathbb{P}(\frac{1}{n}Z_T^n \in \cdot)$ *satisfies a scalar large deviation principle on* $\mathbb{R}^+$ *with the good rate function*

$$J(x;T) = \inf_{g(T)=x} I_Z(g) \tag{2.2}$$

$$= \sup_{\theta\in\mathbb{R}} \{\theta x - A(T;\theta)\}, \tag{2.3}$$



(a) $J(x;T)$ as a function of x. $T=5$ is fixed. (b) $J(x;T)$ as a function of T. $x=3$ is fixed.

Figure 1. This figure plots the rate function $J(x;T)$ in (2.3). The parameters are given by: $\alpha=\beta=1$.

where $A(t;\theta)$ *satisfies the ODE* (*Ordinary Differential Equation*):

$$A'(t;\theta) = -\beta A(t;\theta) + e^{\alpha A(t;\theta)} - 1, \tag{2.4}$$

$$A(0;\theta) = \theta. \tag{2.5}$$

Four remarks are in order.

(a) When $g(t)=e^{(\alpha-\beta)t}$ for $t\in[0,T]$, one immediately verifies from (2.1) that $I_Z(g)=0$. This is consistent with the functional law of large numbers for $\{\frac{1}{n}Z_t^n, 0\le t\le T\}$ in [13].

(b) Note that $g'(t)=-\beta g(t)$ for any $0\le t\le T$ corresponds to $Z_t^n=Z_0^n e^{-\beta t}=ne^{-\beta t}$ for any $0\le t\le T$, which is equivalent to $N_T^n=0$. We can compute that $\mathbb{P}(N_T^n=0|Z_0^n=n)=e^{-\int_0^T(\mu+ne^{-\beta t})\,dt}$, which gives $-\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}(Z_t^n=ne^{-\beta t},0\le t\le T)=\int_0^T e^{-\beta t}\,dt$ which is consistent with $I_Z(g)=\int_0^T e^{-\beta t}\,dt$ for $g'(t)=-\beta g(t)$ for any $0\le t\le T$.

(c) We have used $A(t;\theta)$ instead of $A(t)$ to emphasize that A takes value θ at time zero, and the derivative in (2.4) is taken with respect to t.

(d) We have two equivalent expressions for the rate function J: the first expression (2.2) is directly implied by the sample-path large deviation principle together with the contraction principle, and the second expression (2.3) is obtained via Gärtner–Ellis Theorem. See Section 4 for more details. In general, there are no analytical formulas for A and the rate function J. But one can easily numerically solve the ODE for A (e.g., Runge–Kutta methods) and then solve the optimization problem in (2.3) to obtain the rate function J. An illustrative example is given in Figure 1.

Next, we proceed to state a large deviation principle for $\mathbb{P}(\{\frac{1}{n}N_t^n, 0 \leq t \leq T\} \in \cdot)$. To gain some intuition about the result, we note that

$$dZ_t = -\beta Z_t\,dt + \alpha\,dN_t,$$

which implies that

$$N_t = \frac{Z_t - Z_0}{\alpha} + \frac{\beta}{\alpha}\int_0^t Z_s\,ds.$$

Given $Z_0 = n$, equivalently we have

$$\frac{1}{n}N_t^n = \frac{1}{\alpha}\cdot\left(\frac{Z_t^n}{n} - 1\right) + \frac{\beta}{\alpha}\int_0^t \frac{Z_s^n}{n}\,ds. \tag{2.6}$$

Now if we define for $t \in [0, T]$,

$$h(t) = \frac{g(t) - 1}{\alpha} + \frac{\beta}{\alpha}\int_0^t g(s)\,ds,$$

then one readily verifies that the map $g \mapsto h$ is a continuous map from $D[0, T]$ to $D[0, T]$ under the uniform topology. Therefore, by Theorem 1 and the contraction principle, we can obtain the following result. The details of the proof is left to Section 4.

Theorem 2. $\mathbb{P}(\{\frac{1}{n}N_t^n, 0 \leq t \leq T\} \in \cdot)$ *satisfies a sample-path large deviation principle on* $D[0, T]$ *equipped with uniform topology with the speed n and the good rate function*

$$\begin{aligned} I_N(h) = \int_0^T h'(t)\log\left(\frac{h'(t)}{e^{-\beta t} + e^{-\beta t}\int_0^t \alpha e^{\beta s}h'(s)\,ds}\right) \\ - \left(h'(t) - e^{-\beta t} - e^{-\beta t}\int_0^t \alpha e^{\beta s}h'(s)\,ds\right)dt, \end{aligned} \tag{2.7}$$

if $h \in \mathcal{AC}_0^+[0, T]$, *and* $I_N(h) = \infty$ *otherwise. Moreover,* $\mathbb{P}(N_T^n/n \in \cdot)$ *satisfies a scalar large deviation principle on* $\mathbb{R}_{\geq 0}$ *with the good rate function*

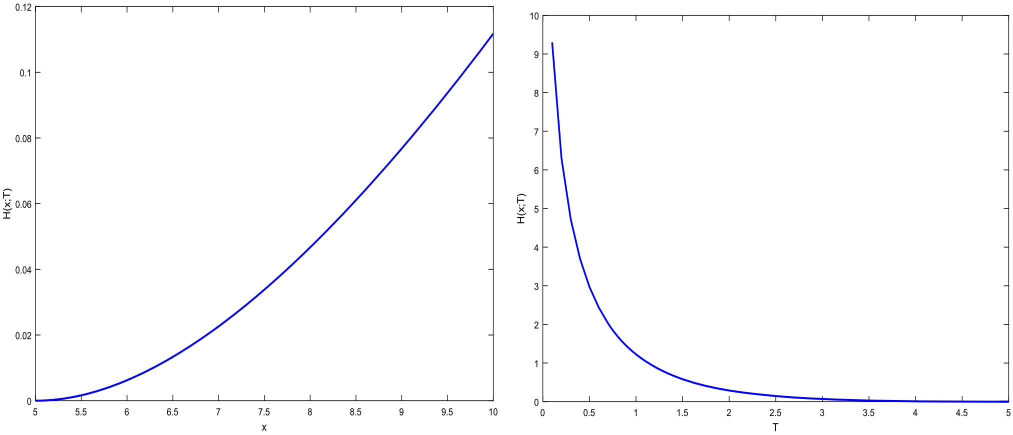
$$H(x; T) = \inf_{h:h(T)=x} I_N(h) \tag{2.8}$$

$$= \sup_{\theta\in\mathbb{R}}\left\{\theta x - C\left(T; \frac{\theta}{\alpha}\right) + \frac{\theta}{\alpha}\right\}, \tag{2.9}$$

where $C(t; \frac{\theta}{\alpha})$ *solves the ODE*

$$C'\left(t; \frac{\theta}{\alpha}\right) = -\beta C\left(t; \frac{\theta}{\alpha}\right) + e^{\alpha\cdot C(t;\frac{\theta}{\alpha})} - 1 + \frac{\beta\theta}{\alpha}, \tag{2.10}$$

$$C\left(0; \frac{\theta}{\alpha}\right) = \frac{\theta}{\alpha}. \tag{2.11}$$



(a) $H(x;T)$ as a function of x. $T=5$ is fixed. (b) $H(x;T)$ as a function of T. $x=5$ is fixed.

Figure 2. This figure plots the rate function $H(x;T)$ in (2.9). The parameters are given by: $\alpha=\beta=1$.

Four remarks are in order.

(a) Notice that $I_N(h)=\int_0^T R(h'(t),e^{-\beta t}+e^{-\beta t}\int_0^t \alpha e^{\beta s}h'(s)\,ds)\,dt$, where $R(x,y):=x\log(\frac{x}{y})-x+y$, for any $x,y>0$. It is easy to see that $R(x,y)\geq 0$ and $R(x,y)=0$ if and only if $x=y$. Therefore, $I_N(h)=0$ if and only if $h'(t)=e^{-\beta t}+e^{-\beta t}\int_0^t \alpha e^{\beta s}h'(s)\,ds$ for any $0\leq t\leq T$. Together with $h'(0)=1$, we get $h'(t)=e^{(\alpha-\beta)t}$. With the initial condition $h(0)=0$, we get $h(t)=\int_0^t e^{(\alpha-\beta)s}\,ds=t$ if $\alpha=\beta$ and $\frac{e^{(\alpha-\beta)t}-1}{\alpha-\beta}$ if $\alpha\neq\beta$. This is consistent with the functional law of large numbers for $\{\frac{1}{n}N_t^n, 0\leq t\leq T\}$ in [13].

(b) Note that $h\equiv 0$ corresponds to $N_T^n=0$. We can compute that $\mathbb{P}(N_T^n=0|Z_0^n=n)=e^{-\int_0^T(\mu+ne^{-\beta t})\,dt}$, which gives $-\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}(N_T^n=0|Z_0^n=n)=\int_0^T e^{-\beta t}\,dt$, which is consistent with $I_N(h)=\int_0^T e^{-\beta t}\,dt$ for $h\equiv 0$.

(c) Similar as in Theorem 1, we use $C(t;\frac{\theta}{\alpha})$ instead of $C(t)$ to emphasize that C takes value $\frac{\theta}{\alpha}$ at time zero. The derivative in (2.10) is taken with respect to t.

(d) Similar as in Theorem 1, we have two equivalent expressions for the rate function H. In general, there is no analytical formula for H. But one can easily numerically solve the ODE for C (e.g., Runge–Kutta methods) and then solve the optimization problem in (2.9) to obtain the rate function H. An illustrative example is given in Figure 2.

2.2.1. *Most likely paths*

In this section, we compute the most likely paths to rare events for Hawkes processes with large initial intensities. More precisely, we are interested to find the minimizer to the variational problems in (2.2) and (2.8).

Fix $x \in \mathbb{R}^+$. Let θ_* be the unique maximizer to the optimization problem (2.3).[2]

Proposition 3. *The minimizer to the variational problem* (2.2) *is given by*

$$g_*(t) = \exp\left(\int_0^t \alpha e^{\alpha A(s;\theta_*)}\, ds - \beta t\right), \tag{2.12}$$

for $0 \leq t \leq T$, *where* $A(s;\theta_*)$ *solves the ODE* (2.4) *with an initial condition* $A(0;\theta_*) = \theta_*$.

Next, we consider the variational problem (2.8). Let $\hat{\theta}_*$ be the unique maximizer to the optimization problem (2.9).[3]

Proposition 4. *The minimizer to the variational problem* (2.8) *is given by*

$$h_*(t) = \int_0^t \exp\left(\alpha \cdot C\left(T - s; \frac{\hat{\theta}_*}{\alpha}\right) + \alpha \int_0^s e^{\alpha C(T-u; \frac{\hat{\theta}_*}{\alpha})}\, du - \beta s\right) ds, \tag{2.13}$$

for any $0 \leq t \leq T$, *where* $C(s; \frac{\hat{\theta}_*}{\alpha})$ *solves the ODE* (2.10) *with the initial condition* $C(0; \frac{\hat{\theta}_*}{\alpha}) = \frac{\hat{\theta}_*}{\alpha}$.

The proofs of these two propositions are relegated to the supplemental article [14].

2.3. Large deviation analysis for large initial intensity and large time

This section is devoted to a set of results on large deviations behavior of Markovian Hawkes processes in the asymptotic regime where both $Z_0 = n$ and the time go to infinity. The proofs of these results are relegated to the supplemental article [14].

When the time is sent to infinity, Hawkes processes behave differently depending on the value of $\|\phi\|_{L^1}$ (see, e.g., Zhu [30]). In our case, the exciting function is exponential: $\phi(t) = \alpha e^{-\beta t}$. So we have the following three different cases: (1) critical: $\alpha = \beta$; (2) super-critical: $\alpha > \beta$; and (3) sub-critical: $\alpha < \beta$. We study each case separately.

2.3.1. *Critical case*

We first consider the critical case, that is, $\alpha = \beta > 0$.

Theorem 5. *Assume that* $\alpha = \beta > 0$. *Let* t_n *be a positive sequence that goes to infinity as* $n \to \infty$ *and* $\lim_{n\to\infty} \frac{t_n}{n} = 0$.

[2] It will be clear from the Proof of Theorem 1 that $A(T;\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}[e^{\theta Z_T} | Z_0 = n]$ if the limit exists. So one readily verifies that $A(T;\theta)$ is convex in θ, and in fact strictly convex in θ from (2.4). Hence, there is a unique optimal θ_* for the optimization problem (2.3).

[3] It will be clear from the Proof of Theorem 2 that $C(T; \frac{\theta}{\alpha}) - \frac{\theta}{\alpha} = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}[e^{\theta N_T} | Z_0 = n]$ is always convex in θ if the limit exists. Indeed, from the ODE (2.10), the limit must be strictly convex. Hence, there is a unique optimal $\hat{\theta}_*$ for the optimization problem (2.9).

(i) *For any* $T > 0$, $\mathbb{P}(\frac{Z^n_{t_n T}}{n} \in \cdot)$ *satisfies a large deviation principle on* $\mathbb{R}$ *with the speed* $\frac{n}{t_n}$ *and the rate function*

$$\hat{I}_Z(x) = \frac{2(\sqrt{x}-1)^2}{\alpha^2 T} \qquad \text{if } x \geq 0,$$

and $+\infty$ *otherwise.*

(ii) *For any* $T > 0$, $\mathbb{P}(\frac{N^n_{t_n T}}{n t_n} \in \cdot)$ *satisfies a large deviation principle on* $\mathbb{R}$ *with the speed* $\frac{n}{t_n}$ *and rate function*

$$\hat{I}_N(x) = \sup_{\theta \in \mathbb{R}} \bigl\{\theta x - \Lambda(\theta)\bigr\},$$

where

$$\Lambda(\theta) = \begin{cases} \dfrac{\sqrt{-2\theta}}{\alpha} \tanh\left(\dfrac{-\alpha}{\sqrt{2}} \sqrt{-\theta} T\right), & \text{if } \theta \leq 0, \\ \dfrac{\sqrt{2\theta}}{\alpha} \tan\left(\dfrac{\alpha}{\sqrt{2}} \sqrt{\theta} T\right), & \text{if } \theta > 0. \end{cases}$$

The proof of this result relies on Gärtner–Ellis theorem and Gronwall's inequality for nonlinear ODEs (see, e.g., [[10], Theorem 42]) which arise from the characterization of the moment generating functions of Z_t and N_t.

2.3.2. *Super-critical case*

We next state the result for the super-critical case where $\alpha > \beta > 0$. Below, we use the convention that $\infty \cdot 0 = 0$.

Theorem 6. *Assume that* $\alpha > \beta > 0$ *and* $0 < T < 1$. *Let* $t_n = \frac{\log n}{\alpha - \beta}$. *Then,*

(i) $\mathbb{P}(\frac{Z^n_{t_n T}}{n^{1+T}} \in \cdot)$ *satisfies a large deviation principle on* $\mathbb{R}^+$ *with the speed* n^T *and the rate function* $\tilde{I}_Z(x) = 0 \cdot 1_{x=1} + \infty \cdot 1_{x \neq 1}$.

(ii) $\mathbb{P}(\frac{N^n_{t_n T}}{n^{1+T}} \in \cdot)$ *satisfies a large deviation principle on* $\mathbb{R}_{\geq 0}$ *with the speed* n^T *and the rate function* $\tilde{I}_N(x) = 0 \cdot 1_{x = \frac{1}{\alpha-\beta}} + \infty \cdot 1_{x \neq \frac{1}{\alpha-\beta}}$.

We remark that the sequence $\{t_n\}$ in Theorem 6 can be taken to be more general. We choose this particular $\{t_n\}$ for simplicity. Note that when $Z_0 = n \to \infty$, the initial intensity is $\mu + n$ which is of the same order as n, and assuming $\mu = 0$, we have $\mathbb{E}[Z^n_t] = n e^{(\alpha-\beta)t}$. Thus choosing $t_n = \frac{\log n}{\alpha - \beta}$ gives $\mathbb{E}[Z^n_{t_n T}] = n^{1+T}$, which is notation-wise concise.

2.3.3. *Sub-critical case*

Finally, we state the large deviations results for the sub-critical case, i.e., $\beta > \alpha > 0$. Given $Z_0 = z$ where z is a fixed constant and under the assumption $\beta > \alpha > 0$, it is well known that as $t \to \infty$, $\frac{N_t}{t} \to \frac{\mu}{1-\frac{\alpha}{\beta}}$ almost surely and $\mathbb{P}(\frac{N_t}{t} \in \cdot)$ satisfies a large deviation principle, see, for example, [5]. So for $Z_0 = n$, it is natural to study the large deviations for $\frac{N^n_{nT}}{n}$.

Theorem 7. *Assume that $\beta > \alpha > 0$. For any $T > 0$, $\mathbb{P}(\frac{N^n_{nT}}{n} \in \cdot)$ satisfies a scalar large deviation principle on $\mathbb{R}$ with the speed n and the rate function*

$$I(x) = x \log\left(\frac{\beta x}{\alpha x + 1 + \mu\beta T}\right) - x + \frac{\alpha x + 1 + \mu\beta T}{\beta}, \tag{2.14}$$

for $x \geq 0$ and $I(x) = +\infty$ otherwise.

The proof of this result rely on Gärtner–Ellis theorem and asymptotic behavior of the solutions of certain nonlinear ODEs which arise from the characterization of the moment generating function of N_t.

Remark 8. We discuss the connections with existing results on large-time large deviations of Hawkes processes here. Since the dependence on the initial condition should be self-evident here, we omit the superscript n for the processes Z and N. As we have discussed in [13], when $Z_0 = n$, we can decompose $N_t = N^{(0)}_t + N^{(1)}_t$, where $N^{(0)}$ is a simple point process with intensity $Z^{(0)}$, where

$$dZ^{(0)}_t = -\beta Z^{(0)}_t \, dt + \alpha \, dN^{(0)}_t,$$

with $Z^{(0)}_0 = n$ and $N^{(1)}$ is a simple point process with intensity

$$\lambda^{(1)}_t := \mu + \int_0^t \alpha e^{-\beta(t-s)} \, dN^{(1)}_s.$$

That is, we can decompose the Hawkes process N into the sum of $N^{(0)}$ and $N^{(1)}$, where $N^{(0)}$ is a linear Markovian Hawkes process with zero base intensity and initial intensity $Z^{(0)}_0 = n$ and $N^{(1)}$ is a linear Markovian Hawkes process with nonzero base intensity $\mu > 0$ and empty history, that is, $N^{(1)}(-\infty, 0] = 0$. This decomposition is valid due to the immigration-birth representation of linear Hawkes processes [19]. One of the key results from the immigration-birth representation is that the two processes $N^{(0)}$ and $N^{(1)}$ are independent of each other.

By letting $\mu = 0$ in Theorem 7, $\mathbb{P}(\frac{N^{(0)}_{nT}}{n} \in \cdot)$ satisfies a large deviation principle with the rate function

$$I^{(0)}(x) = x \log\left(\frac{\beta x}{\alpha x + 1}\right) - x + \frac{\alpha x + 1}{\beta}.$$

On the other hand, from Bordenave and Torrisi [5], $\mathbb{P}(\frac{N^{(1)}_{nT}}{n} \in \cdot)$ satisfies a large deviation principle with the rate function

$$I^{(1)}(x) = T\left[\frac{x}{T} \log\left(\frac{\frac{x}{T}}{\mu + \frac{x}{T}\frac{\alpha}{\beta}}\right) - \frac{x}{T} + \frac{x}{T}\frac{\alpha}{\beta} + \mu\right].$$

Since $N^{(0)}$ and $N^{(1)}$ are independent, we conclude that $\mathbb{P}(\frac{N_{nT}}{n} \in \cdot)$ satisfies a large deviation principle with the rate function

$$I(x) = \inf_{y+z=x} \{ I^{(0)}(y) + I^{(1)}(z) \}.$$

Notice that $I^{(1)}(x) = \mu T I^{(0)}(\frac{x}{\mu T}) + \mu T(1 - \frac{1}{\beta})$ and $I^{(0)}(x)$ is convex in x. Hence, by Jensen's inequality, we conclude that

$$\begin{aligned} I(x) &= \inf_{0 \le y \le x} \left\{ I^{(0)}(x-y) + \mu T I^{(0)}\left(\frac{y}{\mu T}\right) \right\} + \mu T\left(1 - \frac{1}{\beta}\right) \\ &= (1+\mu T) \inf_{0 \le y \le x} \left\{ \frac{1}{1+\mu T} I^{(0)}(x-y) + \frac{\mu T}{1+\mu T} I^{(0)}\left(\frac{y}{\mu T}\right) \right\} + \mu T\left(1 - \frac{1}{\beta}\right) \\ &= (1+\mu T) I^{(0)}\left(\frac{1}{1+\mu T}(x-y) + \frac{\mu T}{1+\mu T}\frac{y}{\mu T}\right) + \mu T\left(1 - \frac{1}{\beta}\right) \\ &= (1+\mu T) I^{(0)}\left(\frac{x}{1+\mu T}\right) + \mu T\left(1 - \frac{1}{\beta}\right), \end{aligned}$$

which can be easily verified to be consistent with (2.14).

The next result is complementary to Theorem 7.

Theorem 9. *Assume that $\beta > \alpha > 0$ and $\mu > 0$. Let t_n be a positive sequence that goes to infinity as $n \to \infty$.*

(i) *If $\lim_{n\to\infty} \frac{t_n}{n} = 0$, then, for any $T > 0$, $\mathbb{P}(\frac{N^n_{t_n T}}{n} \in \cdot)$ satisfies a large deviation principle on $\mathbb{R}_{\ge 0}$ with the speed n and the rate function*

$$I^{(0)}(x) = x \log\left(\frac{\beta x}{\alpha x + 1}\right) - x + \frac{\alpha x + 1}{\beta}.$$

(ii) *If $\lim_{n\to\infty} \frac{t_n}{n} = \infty$, then, for any $T > 0$, $\mathbb{P}(\frac{N^n_{t_n T}}{t_n} \in \cdot)$ satisfies a large deviation principle on $\mathbb{R}_{\ge 0}$ with the speed t_n and the rate function*

$$I^{(1)}(x) = T\left[\frac{x}{T} \log\left(\frac{\frac{x}{T}}{\mu + \frac{x}{T}\frac{\alpha}{\beta}}\right) - \frac{x}{T} + \frac{x}{T}\frac{\alpha}{\beta} + \mu \right].$$

Let us give some intuition behind the results of Theorem 9. Recall the decomposition $N_t = N_t^{(0)} + N_t^{(1)}$ from Remark 8. Notice that $N^{(1)}_{t_n T}$ is of order t_n and that is because of the large-time law of large numbers of the linear Hawkes process with a fixed initial intensity μ and empty history. Also notice that $N^{(0)}_{t_n T}$ is of order n. Let us explain. Notice that from $Z_0^{(0)} = n$ we obtain $\mathbb{E}[N^{(0)}_{t_n T}] = \int_0^{t_n T} \mathbb{E}[Z_s^{(0)}]\,ds = n \int_0^{t_n T} e^{(\alpha-\beta)s}\,ds$. As $n \to \infty$, we have $t_n T \to \infty$. But $\int_0^\infty e^{(\alpha-\beta)s}\,ds = \frac{1}{\beta-\alpha} < \infty$ for $\beta > \alpha$. Thus, $N^{(0)}_{t_n T}$ is of order n. Hence, when $\lim_{n\to\infty} \frac{t_n}{n} = 0$,

$N^{(0)}$ "dominates" and we have result (i), and when $\lim_{n\to\infty}\frac{t_n}{n}=\infty$, $N^{(1)}$ "dominates" and we obtain (ii).

So far we have discussed the large deviations for the process N^n in the sub-critical case. We next consider the large deviations for the process Z^n in the regime where $Z_0=n$ and the time are both sent to infinity. Below, we use the convention that $0\cdot\infty=0$.

Theorem 10. *Assume that $\beta>\alpha>0$, $0<\gamma<1$, and $t_n:=\frac{\log n}{\beta-\alpha}$. For any $0<T<1-\gamma$, $\mathbb{P}(\frac{Z^n_{t_nT}}{n^{1-T}}\in\cdot)$ satisfies a scalar large deviation principle on $\mathbb{R}^+$ with the speed $n^{1-\gamma-T}$ and the rate function*

$$\bar{I}_Z(x)=0\cdot 1_{x=1}+\infty\cdot 1_{x\neq 1}.$$

We remark that similar as in Theorem 6, here the sequence $\{t_n\}$ in Theorem 10 can be taken to be more general. We choose this particular $\{t_n\}$ for the simplicity of notation.

3. Examples and applications

This section is devoted to two examples that apply the large deviations principle that we have developed in the previous sections. The first example is on ruin probabilities in the insurance setting, and the second example is on the finite-horizon maximum of queue lengths in an infinite-server queue. We assume Markovian Hawkes processes can adequately model the clustering behavior of events occurring in each application. While this assumption may not be completely realistic, it enables us to illustrate the potential strength of our large deviations analysis. Throughout this section, we write $a_n=o(n)$ as $n\to\infty$ if the sequence of numbers a_n satisfies $\lim_{n\to\infty}a_n/n=0$.

3.1. Example 1: Ruin probability in insurance risk theory

In this example, we apply our large deviations results to approximate the finite horizon ruin probability in a risk model in insurance mathematics.

Hawkes processes have been applied to insurance settings to accommodate the clustering arrival of claims observed in practice, see, for example, [6,24,31]. When a natural disaster such as an earthquake occurs, the claims typically will not be reported following a constant intensity as in a homogeneous Poisson process. Instead, we expect clustering effect in the claim arrivals after a catastrophe. In addition, the arrival rate of claims is typically high right after a catastrophe event. So one might use Hawkes processes with large initial intensities to model such claim arrival processes, and it is of interest to study the finite horizon ruin probability in a risk model where the claim arrivals are modeled by such Hawkes processes.

To study the ruin probability, let us consider the surplus process of the insurance company:

$$X^n_t=X^n_0+\rho t-\sum_{i=1}^{N^n_t}Y_i.$$

Here, N^n is the claim arrival process modeled as a Hawkes process with an initial intensity $\mu + n$, and an exciting function $\phi(t) = \alpha e^{-\beta t}$; the constant $\rho > 0$ is the premium rate, and we assume it is independent of n for simplicity; $\{Y_i\}$ are the non-negative claim sizes which are independent and identically distributed, and $\{Y_i\}$ is independent of N^n and n. Note that we use N^n to emphasize the dependence on $Z_0 = n$.

We are interested in approximating the finite horizon ruin probability $\mathbb{P}(\tau^n \leq T)$ for fixed $T > 0$ and large n, where τ^n is the ruin time of an insurance company and it is defined as follows:

$$\tau^n := \inf\bigl\{t > 0 : X_t^n \leq 0\bigr\}.$$

We assume that the initial surplus at time 0 is given by $X^n(0) = nx$, which is large, as $n \to \infty$. In the usual setting of the finite horizon ruin probability problem for the classical risk model, the ruin probability is exponentially small when the initial surplus is large, see, for example, [2]. In our example, because N_t^n is of the order n, the ruin will occur at a finite time with probability one.

Notice that N^n satisfies a functional law of large numbers, see [13],

$$\sup_{0 \leq t \leq T} \left| \frac{N_t^n}{n} - \psi(t) \right| \to 0 \qquad \text{almost surely as } n \to \infty,$$

where $\psi(t) := \frac{e^{(\alpha-\beta)t}-1}{\alpha-\beta}$ for $\alpha \neq \beta$, and $\psi(t) := t$ for $\alpha = \beta$. Therefore, as $n \to \infty$,

$$\tau^n \to \tau^\infty := \inf\bigl\{t > 0 : x - \mathbb{E}[Y_1]\psi(t) = 0\bigr\} \qquad \text{almost surely.}$$

It is easy to compute that (assuming that $(\alpha - \beta)\frac{x}{\mathbb{E}[Y_1]} + 1 > 0$; otherwise τ^∞ will be ∞).

$$\tau^\infty = \begin{cases} \dfrac{\log((\alpha-\beta)\frac{x}{\mathbb{E}[Y_1]} + 1)}{\alpha - \beta}, & \text{for } \alpha \neq \beta, \\ \dfrac{x}{\mathbb{E}[Y_1]}, & \text{for } \alpha = \beta. \end{cases}$$

For any $T > \tau^\infty$, $\mathbb{P}(\tau^n \leq T) \to 1$ as $n \to \infty$. For any $T < \tau^\infty$, this probability will go to zero exponentially fast as $n \to \infty$, and falls into the large deviations regime. In the following, we develop approximations for this probability $\mathbb{P}(\tau^n \leq T)$.

Let us assume that $\mathbb{E}[e^{\theta Y_1}] < \infty$ for any $\theta < \theta^+$ and $\mathbb{E}[e^{\theta Y_1}] = \infty$ otherwise, where $\theta^+ > 0$ and we allow it to be $+\infty$. We define $\mathcal{V}^{++}$ as the subspace of $D[0, \infty)$, consisting of unbounded nonnegative increasing functions starting at zero at time zero with finite variation over finite intervals equipped with the vague topology, see [23]. A Mogulskii-type theorem says that, see, for example, Lemma 3.2. [23], $\mathbb{P}(\{\frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} Y_i, 0 \leq t < \infty\} \in \cdot)$ satisfies a large deviation principle on $\mathcal{V}^{++}$ with the good rate function

$$\int_0^\infty \overline{\Lambda}\bigl(g_1'(t)\bigr)\, dt + \theta^+ g_2(\infty) \qquad \text{if } g = g_1 + g_2 \in \mathcal{V}^{++}, g_1 \in \mathcal{AC}_0[0, \infty),$$

where

$$\overline{\Lambda}(x) := \sup_{\theta\in\mathbb{R}}\big\{\theta x - \log\mathbb{E}\big[e^{\theta Y_1}\big]\big\}, \tag{3.1}$$

and $g = g_1 + g_2$ denotes the Lebesgue decomposition of g with respect to Lebesgue measure, where g_2 is the singular component and $g_2(\infty) = \lim_{t\to\infty} g_2(t)$. Note that if $\theta^+ = \infty$, then $g_2 \equiv 0$. Since $\{Y_i\}$ and N^n are independent, then Theorem 2 implies that

$$\mathbb{P}\left(\left\{\left(\frac{1}{n}N_t^n, \frac{1}{n}\sum_{i=1}^{\lfloor ns\rfloor} Y_i\right), 0 \le t \le T, 0 \le s < \infty\right\} \in \cdot\right)$$

satisfies a large deviation principle on $D[0,T] \times \mathcal{V}^{++}$[4] with the good rate function $I_N(h) + \int_0^\infty \overline{\Lambda}(g_1'(t))\,dt + \theta^+ g_2(\infty)$, where the rate function $I_N(h)$ is given in Theorem 2. It is easy to see that $\frac{1}{n}\sum_{i=1}^{N_t^n} Y_i = \frac{1}{n}\sum_{i=1}^{\lfloor n\cdot\frac{1}{n}N_t^n\rfloor} Y_i$. Hence, by the continuity of the first-passage-time map, and the contraction principle, for any fixed $0 < T < \tau^\infty$, we have

$$\begin{aligned}\mathbb{P}\big(\tau^n \le T\big) &= e^{-n\cdot\inf_{h,g:x-g(h(T))\le 0}\{I_N(h)+\int_0^\infty \overline{\Lambda}(g_1'(t))\,dt+\theta^+ g_2(\infty)\}+o(n)} \\ &= e^{-n\cdot\inf_{h,g:x-g(h(T))\le 0}\{I_N(h)+\int_0^{h(T)} \overline{\Lambda}(g_1'(t))\,dt+\theta^+ g_2(h(T))\}+o(n)},\end{aligned} \tag{3.2}$$

as $n \to \infty$. We can replace ∞ by $h(T)$ in (3.2) since $\overline{\Lambda}(x) \ge 0$ for any $x \ge 0$ and it is zero for $x = \mathbb{E}[Y_1]$ and g_2 is also non-decreasing so that $g_2(\infty) \ge g_2(h(T))$, and thus the optimal g satisfies $g_1'(t) = \mathbb{E}[Y_1]$ for $t > h(T)$ so that $\overline{\Lambda}(g_1'(t)) = 0$ for $t > h(T)$ and $g_2(\infty) = g_2(h(T))$.

The expression (3.2) is not very informative, so we next simplify it to obtain a more manageable expression which allows efficient numerical computations. We can first fix $g_2(h(T))$ and then optimize over $g_2(h(T))$. By the convexity of $\overline{\Lambda}(\cdot)$ and using Jensen's inequality, we obtain

$$\int_0^{h(T)} \overline{\Lambda}\big(g_1'(t)\big)\,dt \ge h(T)\overline{\Lambda}\left(\frac{1}{h(T)}\int_0^{h(T)} g_1'(t)\,dt\right) \ge h(T)\overline{\Lambda}\left(\frac{x - g_2(h(T))}{h(T)}\right),$$

where the second inequality is due to $x - g_1(h(T)) - g_2(h(T)) \le 0$ and $\overline{\Lambda}(x)$ is non-decreasing in x for $x > \mathbb{E}[Y_1]$. On the other hand, by considering $g_1^*(t) = \frac{x-g_2(h(T))}{h(T)}t$, we have

$$\int_0^{h(T)} \overline{\Lambda}\big(\big(g_1^*\big)'(t)\big)\,dt = h(T)\overline{\Lambda}\left(\frac{x - g_2(h(T))}{h(T)}\right).$$

This implies that (3.2) can be reduced to the following:

$$\mathbb{P}\big(\tau^n \le T\big) = e^{-n\cdot\inf_{h,z\le x}\{I_N(h)+h(T)\overline{\Lambda}(\frac{x-z}{h(T)})+\theta^+ z\}+o(n)},$$

[4] Here $D[0,T]$ is equipped with Skorokhod topology. In Theorem 1 and Theorem 2 we proved first the large deviation principles hold in the Skorokhod topology.

as $n \to \infty$. Therefore, we have

$$\mathbb{P}\big(\tau^n \leq T\big) = e^{-n \cdot \inf_{y>0, z\leq x} \inf_{h:h(T)=y}\{I_N(h)+y\overline{\Lambda}(\frac{x-z}{y})+\theta^+ z\}+o(n)}.$$

To further simplify the above expression, we note from Theorem 2 that $\mathbb{P}(N_T^n/n \in \cdot)$ satisfies a large deviation principle with the rate function

$$H(x;T) = \inf_{h:h(T)=x} I_N(h) = \sup_{\theta\in\mathbb{R}}\left\{\theta x - C\left(T;\frac{\theta}{\alpha}\right) + \frac{\theta}{\alpha}\right\},$$

where C solves the nonlinear ODE given in (2.10) and (2.11). Hence, we conclude that

$$\mathbb{P}\big(\tau^n \leq T\big) = \exp\big(-n \cdot I_\tau(x;T) + o(n)\big) \qquad \text{as } n \to \infty,$$

where

$$I_\tau(x;T) := \inf_{y>0, z\leq x}\left\{H(y;T) + y\overline{\Lambda}\left(\frac{x-z}{y}\right) + \theta^+ z\right\}.$$

We remark that the function $H(y;T) + y\overline{\Lambda}(\frac{x-z}{y}) + \theta^+ z$ is convex in y. This is because $H(y;T)$ is convex in y and one can also verify directly from the convexity of $\overline{\Lambda}$ that $y\overline{\Lambda}(\frac{x}{y})$ in convex in y. It is also clear that $H(y;T) + y\overline{\Lambda}(\frac{x-z}{y}) + \theta^+ z$ is convex in z. So we can numerically obtain $I_\tau(x;T)$ efficiently.

We now present a numerical example when Y_i has a Poisson distribution with rate 1. Then it is easy to see from (3.1) that $\bar{\Lambda}(v) = v \log v - v + 1$ for $v > 0$ and $\bar{\Lambda}(v) = +\infty$ otherwise. Also in this case $\theta^+ = \infty$. Hence, we obtain

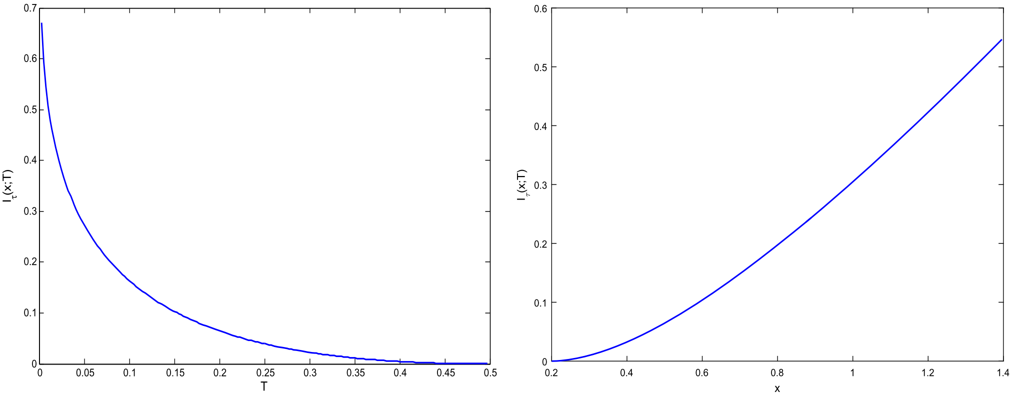
$$I_\tau(x;T) = \inf_{y>0}\left\{H(y;T) + y \cdot \left(\frac{x}{y} - 1 - \log\left(\frac{x}{y}\right)\right)\right\}. \tag{3.3}$$

See Figure 3 for a numerical illustration.

3.2. Example 2: Finite-horizon maximum of the queue length process in an infinite-server queue

In this example, we use our large deviations results to study certain tail probabilities in an infinite-server queue in heavy traffic where the job arrival process is modeled by a Hawkes process with a large initial intensity. Such a queueing system could be relevant for analyzing the performance of large scale service systems with high-volume traffic which exhibits clustering. For background on infinite-server queues, their engineering applications and related large deviation analysis, see, for example, [4,15,28].

Consider a sequence of queueing systems indexed by n with infinite number of servers. Jobs arrive to the nth system according to a Markovian Hawkes process N^n with an initial intensity



(a) $I_\mathcal{T}(x;T)$ as a function of T. $x = 0.5$ is fixed. (b) $I_\mathcal{T}(x;T)$ as a function of x. $T = 0.2$ is fixed.

Figure 3. This figure plots $I_\tau(x;T)$ in (3.3). The parameters are given by: $\alpha = \beta = 1$.

$\mu + n$, and an exciting function $\phi(t) = \alpha e^{-\beta t}$. We use N^n to emphasize the dependence on $Z_0 = n$. For simplicity, we assume that (a) n is large so the offered load in the system is high; (b) the system is initially empty; (c) the processing time of each job is deterministic given by a constant $c > 0$.

We are interested in the finite-horizon maximum of queue length process in such an infinite-serve queue, similarly as in [4]. Mathematically, we want to develop large deviations approximations for the probability of the event

$$\max_{0\leq s\leq T} Q_s^n \geq nx \tag{3.4}$$

for fixed $T > 0$ and sufficiently large x, as $n \to \infty$. Here Q_s^n is number of jobs (or busy servers) in the nth system at time s. For sufficiently large x, we note that (3.4) is a rare event. This event corresponds precisely to the event of observing a loss in a queue with nx servers, no waiting room, and starting empty.

It is well known that (see, e.g., [16]) for the nth system with deterministic processing time c, the queue length process Q^n can be represented by

$$Q_t^n = N_t^n - N_{t-c}^n,$$

where $N_t^n = 0$ if $t \leq 0$ by convention. It is easy to see that the function Φ mapping y to $\tilde{y}$ where

$$\tilde{y}(t) := \max_{s\leq t}\{y(s) - y(s-c)\},$$

is continuous under the uniform topology. Since Theorem 2 states that $\mathbb{P}(\frac{1}{n}N^n \in \cdot)$ satisfies a sample path large deviation principle with the good rate function I_N, we can apply the contraction principle and obtain:

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big(\max_{0\leq s\leq T} Q_s^n \geq nx\Big) \\
&\quad =\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left(\max_{0\leq s\leq T}\frac{1}{n}\big[N_s^n - N_{s-c}^n\big]\geq x\right) \\
&\quad =-\inf_h\Big\{I_N(h;T):\max_{s\leq T}\big[h(s)-h(s-c)\big]\geq x\Big\},
\end{aligned}\tag{3.5}$$

where we use the notation $I_N(h;T)$ to emphasize the dependence of I_N on T, as can be clearly seen in (2.7).

Therefore, to develop large deviations approximations for $\mathbb{P}(\max_{0\leq s\leq T} Q^n(s)\geq nx)$, it remains to solve the optimization problem in (3.5). For $T\leq c$, since h is a nondecreasing function, then the infimum in (3.5) is simply

$$\inf_{h(T)\geq x} I_N(h;T)=H(x;T).$$

For $T>c$, the infimum in (3.5) is equivalent to:

$$\min\Big\{\inf_{0\leq s\leq c}\inf_{h:h(s)\geq x} I_N(h;s),\inf_{c\leq s\leq T}\inf_{h:h(s)-h(s-c)\geq x} I_N(h;s)\Big\}.$$

Now, let us solve the optimization problem:

$$\inf_{h:h(t)-h(t-c)\geq x} I_N(h;t).$$

Since

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\big(N_t^{ny}/n\in B_\varepsilon(x)|Z_0=ny\big)=-yH(x/y;t),$$

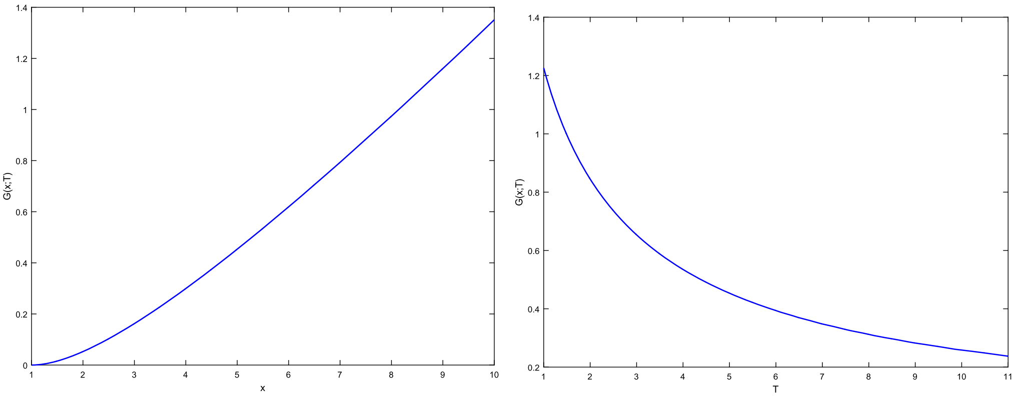
$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\big(Z_t^n/n\in B_\varepsilon(y)|Z_0=n\big)=-J(y;t),$$

and by the Markov property, we get

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\big(\big[N_t^n-N_{t-c}^n\big]/n\in B_\varepsilon(x), Z_{t-c}^n/n\in B_\varepsilon(y)|Z_0=n\big) \\
&\quad =-yH(x/y;c)-J(y;t-c),
\end{aligned}\tag{3.6}$$

and finally for sufficiently large x, by (3.6) and the contraction principle, we obtain

$$\begin{aligned}
\inf_{h:h(t)-h(t-c)\geq x} I_N(h;t) &= -\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\big(\big[N_t^n-N_{t-c}^n\big]/n\in B_\varepsilon(x)|Z_0=n\big) \\
&=\inf_{y>0}\big\{yH(x/y;c)+J(y;t-c)\big\}.
\end{aligned}$$



(a) $G(x;T)$ as a function of x. $T=5$ is fixed. (b) $G(x;T)$ as a function of T. $x=5$ is fixed.

Figure 4. This figure plots $G(x;T)$ in (3.7). We use parameters $\alpha=\beta=1, c=1$.

Hence we conclude that the infimum in (3.5) is equivalent to the following expression:

$$G(x;T) := \min\left\{\inf_{0\le s\le c} H(x;s),\ \inf_{c\le s\le T}\inf_{y>0}\left\{yH(x/y;c)+J(y;s-c)\right\}\right\}, \tag{3.7}$$

where H and J are given in Theorems 1 and 2, respectively. This implies the following approximation for $T>c$ and sufficiently large x:

$$\mathbb{P}\left(\max_{0\le s\le T} Q_s^n \ge nx\right) = \exp\left(-n\cdot G(x;T)+o(n)\right) \quad \text{as } n\to\infty.$$

Since one can solve H and J numerically, we can then also obtain G by solving the optimization problem in (3.7) numerically. We present an example in Figure 4.

4. Proofs of Theorems 1 and 2

4.1. Moment generating functions of Z_t and N_t

We first discuss in this section the moment generating functions of Z_t and N_t for fixed t, conditioned on knowing the value of Z_0. These functions play a critical role in proving our large deviation results.

First, recall from [[13], Section 3.2.1] the moment generating function of Z_t:

$$u(t,z) := \mathbb{E}\left[e^{\theta Z_t}|Z_0=z\right] = e^{A(t;\theta)z+B(t;\theta)}, \tag{4.1}$$

where $A(t;\theta)$, $B(t;\theta)$ satisfy the ODEs:

$$A'(t;\theta) = -\beta A(t;\theta)+e^{\alpha A(t;\theta)}-1, \tag{4.2}$$

$$B'(t;\theta) = \mu\left(e^{\alpha A(t;\theta)}-1\right), \tag{4.3}$$

with initial conditions $A(0;\theta)=\theta$ and $B(0;\theta)=0$. As remarked earlier, we have used $A(t;\theta)$ instead of $A(t)$ to emphasize that A takes value θ at time zero, and the derivative in (4.2) is taken with respect to t. We also write $B(t;\theta)$ instead of $B(t)$ to stress that B depends on the initial condition of A.

Next, we compute the moment generating function of N_t. Recall that $N_t=\frac{Z_t-Z_0}{\alpha}+\frac{\beta}{\alpha}\int_0^t Z_s\,ds$. Thus, $\mathbb{E}[e^{\theta N_t}|Z_0=z]=e^{-\frac{\theta}{\alpha}z}v(t,z)$, where

$$v(t,z):=\mathbb{E}\big[e^{\frac{\theta}{\alpha}Z_t+\frac{\theta\beta}{\alpha}\int_0^t Z_s\,ds}|Z_0=z\big].$$

Recall that Z is a Markov process with the infinitesimal generator

$$\mathcal{A}f(z)=-\beta z\frac{\partial f}{\partial z}+(z+\mu)\big[f(z+\alpha)-f(z)\big].$$

By Feynman–Kac formula, v satisfies the equation:

$$\frac{\partial v}{\partial t}=-\beta z\frac{\partial v}{\partial z}+(\mu+z)\big[v(t,z+\alpha)-v(t,z)\big]+\frac{\theta\beta}{\alpha}zv(t,z),$$

with an initial condition $v(0,z)=e^{\frac{\theta}{\alpha}z}$. Therefore, by the affine structure, see, for example, [11], one deduces that $v(t,z)=e^{C(t;\frac{\theta}{\alpha})z+D(t;\frac{\theta}{\alpha})}$, where $C(t;\frac{\theta}{\alpha})$, $D(t;\frac{\theta}{\alpha})$ satisfy the ODEs:

$$C'\left(t;\frac{\theta}{\alpha}\right)=-\beta C\left(t;\frac{\theta}{\alpha}\right)+e^{\alpha C(t;\frac{\theta}{\alpha})}-1+\beta\cdot C\left(0;\frac{\theta}{\alpha}\right), \tag{4.4}$$

$$D'\left(t;\frac{\theta}{\alpha}\right)=\mu\big(e^{\alpha C(t;\frac{\theta}{\alpha})}-1\big), \tag{4.5}$$

with initial conditions $C(0;\frac{\theta}{\alpha})=\frac{\theta}{\alpha}$ and $D(0;\frac{\theta}{\alpha})=0$. Thus we have

$$\mathbb{E}\big[e^{\theta N_t}|Z_0=z\big]=\exp\left\{\left(C\left(t;\frac{\theta}{\alpha}\right)-C\left(0;\frac{\theta}{\alpha}\right)\right)\cdot z+D\left(t;\frac{\theta}{\alpha}\right)\right\}. \tag{4.6}$$

Finally, we remark that there exists some $\Theta>0$ such that the moment generating functions in (4.1) and (4.6) are both finite for all $\theta\leq\Theta$. See [33].

4.2. Proofs of Theorems 1 and 2

We prove Theorems 1 and 2 in this section. For notational convenience, unless specified explicitly, we use Z and N for Z^n and N^n when $Z_0=n$. We also use $\mathbb{E}[\cdot]$ to denote the conditional expectation $\mathbb{E}[\cdot|Z_0=n]$, and $\mathbb{P}(\cdot)$ for the conditional probability $\mathbb{P}(\cdot|Z_0=n)$.

Proof of Theorem 1. The proof is long, so we split it into four steps.

Step 1. We first establish a scalar large deviation principle for $\mathbb{P}(\frac{1}{n}Z_T\in\cdot)$, using Gärtner–Ellis theorem.

From (4.1), we have

$$u(t,z) := \mathbb{E}\big[e^{\theta Z_t}|Z_0 = z\big] = e^{A(t;\theta)z+B(t;\theta)}.$$

It is easy to see that since Z_t process is positive, $u(t,z)$ is monotonically increasing in θ. Let us recall from Section 4.1 that $A(t;\theta)$, $B(t;\theta)$ satisfy the ODEs:

$$\begin{aligned} A'(t;\theta) &= -\beta A(t;\theta) + e^{\alpha A(t;\theta)} - 1, \\ B'(t;\theta) &= \mu\big(e^{\alpha A(t;\theta)} - 1\big), \end{aligned}$$

with initial conditions $A(0;\theta) = \theta$ and $B(0;\theta) = 0$.

Let us first consider the critical and super-critical case, that is, $\alpha \geq \beta$. When we have $\alpha \geq \beta$, for any $A > 0$, $-\beta A + e^{\alpha A} - 1 > 0$ and thus $A(t;\theta)$ is increasing in t. It is clear that for any $\theta > 0$, $\int_\theta^\infty \frac{dA}{-\beta A + e^{\alpha A} - 1} < \infty$. On the other hand, it is easy to see that $\int_0^\infty \frac{dA}{-\beta A + e^{\alpha A} - 1} = \infty$. Therefore, for any fixed $T > 0$, there exists a unique positive value $\theta_c(T)$ such that

$$\int_{\theta_c(T)}^\infty \frac{dA}{-\beta A + e^{\alpha A} - 1} = T. \tag{4.7}$$

Hence, we conclude that for any fixed $T > 0$, for any $0 < \theta < \theta_c(T)$, $A(T;\theta)$ is the unique positive value greater than θ, that satisfies the equation:

$$\int_\theta^{A(T;\theta)} \frac{dA}{-\beta A + e^{\alpha A} - 1} = T. \tag{4.8}$$

Now let us consider the case $\theta \leq 0$. When $\alpha > \beta$, $-\beta A + e^{\alpha A} - 1 = 0$ when $A = 0$ or when $A = A_c$, for some unique negative value A_c. For $\theta = 0$ or $\theta = A_c$, $A(t;\theta) = 0$ for any t. For $A_c < \theta < 0$, $A(t;\theta)$ is decreasing in t and $A(T;\theta)$ satisfies the equation (4.8). For $\theta < A_c$, $A(t;\theta)$ is increasing in t and $A(T;\theta) < 0$ and satisfies the equation (4.8). When $\alpha = \beta$, $-\beta A + e^{\alpha A} - 1 > 0$ when $A \neq 0$. Thus, for any $\theta < 0$, $A(t;\theta)$ is increasing in t and $A(T;\theta) < 0$ and satisfies the equation (4.8) and also $A(t;0) \equiv 0$. Also, it is easy to see that for $\theta < \theta_c(T)$, $A(t;\theta)$ is continuous and finite in t, and

$$B(T;\theta) = \mu \int_0^T \big(e^{\alpha A(t;\theta)} - 1\big)\, dt$$

is finite. Therefore, for $\theta < \theta_c(T)$

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\big[e^{\theta Z_T}\big] = A(T;\theta).$$

When $\theta \geq \theta_c(T)$, this limit is ∞. By differentiating the equation (4.8) with respect to θ, we get

$$-\frac{1}{-\beta\theta + e^{\alpha\theta} - 1} + \frac{1}{-\beta A(T;\theta) + e^{\alpha A(T;\theta)} - 1} \frac{d}{d\theta} A(T;\theta) = 0. \tag{4.9}$$

It is clear from the equation (4.7) and (4.8) that as $\theta \to \theta_c(T)$, we have $A(T;\theta) \to \infty$. Therefore, from (4.9), we get

$$\frac{\partial}{\partial\theta}A(T;\theta) = \frac{-\beta A(T;\theta) + e^{\alpha A(T;\theta)} - 1}{-\beta\theta + e^{\alpha\theta} - 1} \to \infty \quad \text{as } \theta \to \theta_c(T).$$

Hence, we verified the essential smoothness condition. By Gärtner–Ellis theorem, $\mathbb{P}(\frac{1}{n}Z_T \in \cdot)$ satisfies a large deviation principle on $\mathbb{R}^+$ with the rate function

$$J(x;T) = \sup_{\theta\in\mathbb{R}}\bigl\{\theta x - A(T;\theta)\bigr\}. \tag{4.10}$$

Next, let us consider the sub-critical case, that is, $\alpha < \beta$. In this case, $-\beta A + e^{\alpha A} - 1 = 0$ if and only if $A = 0$ or $A = A_c$, where A_c is a positive constant and it is unique. For $\theta = 0$ or A_c, $A(t;\theta) = 0$ for any t. For $\theta < 0$, $A(t;\theta)$ is increasing in t, and $A(T,\theta) < 0$ satisfies the equation (4.8). For $0 < \theta < A_c$, $A(t,\theta)$ is decreasing in t and satisfies the equation (4.8). For $\theta > A_c$, $A(t,\theta)$ is increasing in t. For any fixed $T > 0$, there exists a unique $\theta_c(T) > A_c$ satisfying the equation (4.7) so that for any $A_c < \theta < \theta_c(T)$, $A(T,\theta)$ is the unique positive value greater than θ that satisfies the equation (4.8) and for $\theta \geq \theta_c(T)$, $A(T,\theta) = \infty$. We can proceed similarly as before and prove that, $\mathbb{P}(\frac{1}{n}Z_T \in \cdot)$ satisfies a large deviation principle on $\mathbb{R}^+$ with the rate function given in (4.10).

Step 2. Next, we need to prove the exponential tightness before we proceed to establish the sample path large deviation principle. To be more precise, we will show that

$$\limsup_{K\to\infty}\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Bigl(\sup_{0\leq t\leq T} Z_t \geq nK\Bigr) = -\infty, \tag{4.11}$$

and for any $\delta > 0$,

$$\limsup_{\varepsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Bigl(\sup_{|t-s|\leq\varepsilon, 0\leq t,s\leq T}|Z_t - Z_s| \geq \delta n\Bigr) = -\infty. \tag{4.12}$$

We will also show that for any $\eta > 0$,

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Bigl(\sup_{0<t\leq T}|Z_t - Z_{t-}| \geq \eta n\Bigr) = -\infty. \tag{4.13}$$

The superexponential estimates (4.11) and (4.12) will guarantee the exponential tightness on $D[0,T]$ equipped with the Skorokhod topology, see, for example, Theorem 4.1. in Feng and Kurtz [12]. Together with Step 3, it will prove the large deviation principle for $\mathbb{P}(\{\frac{1}{n}Z_t, 0 \leq t \leq T\} \in \cdot)$ on $D[0,T]$ equipped with Skorokhod topology. Next, the equation (4.13), that is, the so-called C-exponentially tightness, see, for example, Definition 4.12. in [12] strengthens the large deviation principle for $\mathbb{P}(\{\frac{1}{n}Z_t, 0 \leq t \leq T\} \in \cdot)$ so that it holds on $D[0,T]$ equipped with uniform topology, see, for example, Theorem 4.14. in [12].

Let us first prove (4.11). Notice first that $Z_t - Z_0 \le \alpha N_t$ and $Z_0 = n$. Therefore, for $K > 1$,

$$\begin{aligned}\mathbb{P}\Big(\sup_{0\le t\le T} Z_t \ge nK\Big) &= \mathbb{P}\Big(\sup_{0\le t\le T}(Z_t - Z_0) \ge n(K-1)\Big)\\ &\le \mathbb{P}\Big(\sup_{0\le t\le T} N_t \ge n\frac{K-1}{\alpha}\Big)\\ &= \mathbb{P}\big(\alpha N_T \ge n(K-1)\big)\\ &\le \mathbb{E}\big[e^{\theta N_T}\big]e^{-\theta(K-1)n/\alpha},\end{aligned}$$

where the last inequality follows from Chebychev's inequality. In conjunction with the moment generating function of N_T in (4.6), we hence obtain

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big(\sup_{0\le t\le T} Z_t \ge nK\Big) \le C\Big(T;\frac{\theta}{\alpha}\Big) - \frac{\theta}{\alpha} - \frac{\theta(K-1)}{\alpha},$$

which goes to $-\infty$ as $K \to \infty$. Hence, we proved (4.11).

Next, let us prove (4.12). Note that for $s < t$, $\alpha N(s,t] = Z_t - Z_s + \beta\int_s^t Z_u\,du$. Thus, for $s < t$, we have

$$|Z_t - Z_s| \le \alpha N(s,t] + \beta(t-s)\sup_{s\le u\le t} Z_u.$$

Therefore,

$$\begin{aligned}&\mathbb{P}\Big(\sup_{|t-s|\le\varepsilon, 0\le t,s\le T}|Z_t - Z_s| \ge \delta n\Big)\\ &\quad\le \mathbb{P}\Big(\sup_{|t-s|\le\varepsilon, 0\le s\le t\le T}\Big(\alpha N(s,t] + \beta(t-s)\sup_{s\le u\le t} Z_u\Big) \ge \delta n\Big)\\ &\quad\le \mathbb{P}\Big(\sup_{|t-s|\le\varepsilon, 0\le s\le t\le T}\alpha N(s,t] \ge \frac{\delta}{2}n\Big)\\ &\qquad+ \mathbb{P}\Big(\sup_{|t-s|\le\varepsilon, 0\le s\le t\le T}\beta(t-s)\sup_{s\le u\le t} Z_u \ge \frac{\delta}{2}n\Big).\end{aligned}$$

Note that

$$\mathbb{P}\Big(\sup_{|t-s|\le\varepsilon, 0\le s\le t\le T}\beta(t-s)\sup_{s\le u\le t} Z_u \ge \frac{\delta}{2}n\Big) \le \mathbb{P}\Big(\beta\varepsilon\sup_{0\le u\le T} Z_u \ge \frac{\delta}{2}n\Big).$$

By (4.11), we have

$$\limsup_{\varepsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big(\beta\varepsilon\sup_{0\le u\le T} Z_u \ge \frac{\delta}{2}n\Big) = -\infty.$$

Next, notice that without loss of generality we can assume that $\frac{1}{\varepsilon} \in \mathbb{N}$ and

$$\begin{aligned}\mathbb{P}\left(\sup_{|t-s|\leq\varepsilon, 0\leq s\leq t\leq T} \alpha N(s,t] \geq \frac{\delta}{2}n\right) &\leq \mathbb{P}\left(\exists 1 \leq j \leq T/\varepsilon : \alpha N(t_{j-1}, t_j] \geq \frac{\delta}{4}n\right) \\ &\leq \sum_{j=1}^{T/\varepsilon} \mathbb{P}\left(\alpha N(t_{j-1}, t_j] \geq \frac{\delta}{4}n\right),\end{aligned} \tag{4.14}$$

where $0 = t_0 < t_1 < \cdots < t_{T/\varepsilon} = T$, where $t_j - t_{j-1} = \varepsilon$ for any j. In addition, note that for $\theta > 0$,

$$\begin{aligned}\mathbb{E}\left[e^{\theta\alpha N(t_{j-1},t_j]}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{\theta\alpha N(t_{j-1},t_j]}|Z_{t_{j-1}}\right]\right] \\ &= \mathbb{E}\left[e^{-\theta Z_{t_{j-1}}} e^{C(t_j-t_{j-1};\theta)Z_{t_{j-1}}+D(t_j-t_{j-1};\theta)}\right] \\ &= \exp\left(D(\varepsilon;\theta) + A\left(t_{j-1}; C(\varepsilon;\theta)-\theta\right)n + B\left(t_{j-1}; C(\varepsilon;\theta)-\theta\right)\right),\end{aligned} \tag{4.15}$$

where we have used the moment generating functions of Z_t and N_t in Section 4.1. Hence, using Chebychev's inequality and combining (4.14) and (4.15), we find for fixed $\varepsilon > 0$,

$$\begin{aligned}&\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\sup_{|t-s|\leq\varepsilon, 0\leq s\leq t\leq T} \alpha N(s,t] \geq \frac{\delta}{2}n\right) \\ &\leq \sup_{1\leq j\leq T/\varepsilon} \left\{A\left(t_{j-1}; C(\varepsilon;\theta)-\theta\right) - \theta\frac{\delta}{4}\right\} \\ &\leq \sup_{0\leq t\leq T} \left\{A\left(t; C(\varepsilon;\theta)-\theta\right)\right\} - \theta\frac{\delta}{4}.\end{aligned}$$

So in order to prove (4.12), what remains is to choose θ that depends on ε so that (i) $\theta \to \infty$ as $\varepsilon \to 0$; (ii) $A(t; C(\varepsilon;\theta) - \theta)$ is uniformly bounded for $t \in [0, T]$ and $\varepsilon \to 0$. To this end, let us define $y(t) := C(t;\theta) - C(0;\theta) = C(t;\theta) - \theta$. Then y satisfies the ODE:

$$\begin{aligned}&y'(t) = -\beta y(t) + e^{\alpha\theta} e^{\alpha y(t)} - 1, \\ &y(0) = 0.\end{aligned}$$

For $\theta > 0$, we have $y'(0) = e^{\alpha\theta} - 1 > 0$, which implies y is increasing on $[0, \gamma]$ for some $\gamma > 0$. This suggests that

$$0 < y'(t) \leq e^{\alpha\theta} e^{\alpha y(t)} \quad \text{for } t \in [0, \gamma].$$

By Gronwall's inequality for nonlinear ODEs, we obtain

$$0 \leq y(t) \leq -\frac{1}{\alpha} \cdot \log\left(1 - \alpha e^{\alpha\theta} t\right) \quad \text{for } t \in [0, \gamma]. \tag{4.16}$$

Let us set $\alpha e^{\alpha\theta} = \frac{1}{\sqrt{\varepsilon}}$. Then it is clear that $\theta \to \infty$ as $\varepsilon \to 0$. In addition, we deduce from (4.16) that for $\varepsilon < \gamma$,

$$0 \le C(\varepsilon;\theta) - \theta = y(\varepsilon) \le -\frac{1}{\alpha} \cdot \log(1-\sqrt{\varepsilon}). \tag{4.17}$$

Next we show $\{A(t; C(\varepsilon;\theta)-\theta)\}$ is uniformly bounded for $t \in [0,T]$ and $\varepsilon \to 0$. When $\alpha < \beta$, it is clear that zero is a stable solution for the ODE of A in (4.2). Since $A(0; C(\varepsilon;\theta)-\theta)) = y(\varepsilon) \to 0$ as $\varepsilon \to 0$, so the stability of zero solution implies that when $\varepsilon \to 0$, $\{A(t; C(\varepsilon;\theta)-\theta)\}$ is uniformly small and thus uniformly bounded for all $t \ge 0$. When $\alpha \ge \beta$, since $A(0; C(\varepsilon;\theta) - \theta)) = y(\varepsilon) \ge 0$, one readily checks that A is non-decreasing with respect to time t. Hence, we obtain

$$\sup_{0\le t\le T} \bigl\{A\bigl(t; C(\varepsilon;\theta)-\theta\bigr)\bigr\} = A\bigl(T; y(\varepsilon)\bigr).$$

We have shown in Step 1 that $A(T;\bar{\theta})$ is finite when $\bar{\theta} < \theta_c(T)$, and $A(T;\bar{\theta})$ is continuous as a function of $\bar{\theta}$. Therefore, we deduce from (4.17) that $A(T; y(\varepsilon))$ is uniformly bounded for $\varepsilon \to 0$. Thus, we have proved (4.12).

Finally, the claim in (4.13) trivially holds since for any $0 < t \le T$, $|Z_{t-} - Z_t| = 0$ or α with probability 1.

Step 3. Next, we establish the sample path large deviation principle.

For any $\varepsilon > 0$, let $B_\varepsilon(x)$ denote the open ball centered at x with radius ε. For any $0 =: t_0 < t_1 < t_2 < \cdots < t_{k-1} < t_k := T$ and $x_1, \ldots, x_k \in \mathbb{R}^+$, by the Markov property of the process Z, we have

$$\begin{aligned}
&\mathbb{P}\biggl(\frac{1}{n}Z_{t_1} \in B_\varepsilon(x_1), \frac{1}{n}Z_{t_2} \in B_\varepsilon(x_2), \ldots, \frac{1}{n}Z_{t_k} \in B_\varepsilon(x_k)\biggr) \\
&\quad = \mathbb{P}\biggl(\frac{1}{n}Z_{t_1} \in B_\varepsilon(x_1)\biggr)\mathbb{P}\biggl(\frac{1}{n}Z_{t_2} \in B_\varepsilon(x_2)\bigg|\frac{1}{n}Z_{t_1} \in B_\varepsilon(x_1)\biggr) \\
&\qquad \cdots \mathbb{P}\biggl(\frac{1}{n}Z_{t_k} \in B_\varepsilon(x_k)\bigg|\frac{1}{n}Z_{t_{k-1}} \in B_\varepsilon(x_{k-1})\biggr).
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\lim_{n\to\infty} \frac{1}{n}\log \mathbb{P}\biggl(\frac{1}{n}Z_{t_1} \in B_\varepsilon(x_1), \frac{1}{n}Z_{t_2} \in B_\varepsilon(x_2), \ldots, \frac{1}{n}Z_{t_k} \in B_\varepsilon(x_k)\biggr) \\
&\quad = -J(x_1; t_1) - x_1 J\biggl(\frac{x_2}{x_1}; t_2 - t_1\biggr) \\
&\qquad - \cdots - x_{k-1} J\biggl(\frac{x_k}{x_{k-1}}; t_k - t_{k-1}\biggr),
\end{aligned}$$

where J is given in (4.10). Hence, for any $g \in \mathcal{AC}_1[0,T]$,

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left(\frac{1}{n}Z_{t_1}\in B_\varepsilon\big(g(t_1)\big),\frac{1}{n}Z_{t_2}\in B_\varepsilon\big(g(t_2)\big),\dots,\frac{1}{n}Z_{t_k}\in B_\varepsilon\big(g(t_k)\big)\right)\\
&\quad=-J\big(g(t_1);t_1\big)-g(t_1)J\left(\frac{g(t_2)}{g(t_1)};t_2-t_1\right)-\cdots-g(t_{k-1})J\left(\frac{g(t_k)}{g(t_{k-1})};t_k-t_{k-1}\right).
\end{aligned}$$

For any given positive $g \in \mathcal{AC}_1[0,T]$, we have

$$\begin{aligned}
&J\left(\frac{g(t_j)}{g(t_{j-1})};t_j-t_{j-1}\right)\\
&\quad=\sup_{\theta\in\mathbb{R}}\left\{\theta\frac{g(t_j)}{g(t_{j-1})}-A(t_j-t_{j-1};\theta)\right\}\\
&\quad=\sup_{\theta\in\mathbb{R}}\left\{\theta\left(1+\frac{g'(t_{j-1}^*)}{g(t_{j-1})}(t_j-t_{j-1})\right)-\theta-\int_0^{t_j-t_{j-1}}\big(-\beta A(s;\theta)+e^{\alpha A(s;\theta)}-1\big)\,ds\right\}\\
&\quad=(t_j-t_{j-1})\sup_{\theta\in\mathbb{R}}\left\{\theta\frac{g'(t_{j-1}^*)}{g(t_{j-1})}-\big(-\beta A\big(t_{j-1}^{**};\theta\big)+e^{\alpha A(t_{j-1}^{**};\theta)}-1\big)\right\},
\end{aligned}$$

where $t_{j-1}^* \in [t_{j-1}, t_j]$ is independent of θ and $t_{j-1}^{**} \in [0, t_j - t_{j-1}]$ may depend on θ.

It is easy to see that for any given positive $g \in \mathcal{AC}_1[0,T]$, $\frac{g'(t_j^*)}{g(t_{j-1})}$, is uniformly bounded in j. To see this, notice that g is positive and continuous so $\inf_{0\le t\le T} g(t) > 0$, and since g is absolutely continuous, g' exists almost surely and we can assume that g' exist for any t_j^*. And we can also see that $A(t_{j-1}^{**};\theta)$ is uniformly bounded in j. Therefore, there exists some constant K that may depend on the given g, such that, uniformly in j,

$$\begin{aligned}
&\sup_{\theta\in\mathbb{R}}\left\{\theta\frac{g'(t_{j-1}^*)}{g(t_{j-1})}-\big(-\beta A\big(t_{j-1}^{**};\theta\big)+e^{\alpha A(t_{j-1}^{**};\theta)}-1\big)\right\}\\
&\quad=\sup_{|\theta|\le K}\left\{\theta\frac{g'(t_{j-1}^*)}{g(t_{j-1})}-\big(-\beta A\big(t_{j-1}^{**};\theta\big)+e^{\alpha A(t_{j-1}^{**};\theta)}-1\big)\right\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left|\sup_{\theta\in\mathbb{R}}\left\{\theta\frac{g'(t_{j-1}^*)}{g(t_{j-1})}-\big(-\beta A\big(t_{j-1}^{**};\theta\big)+e^{\alpha A(t_{j-1}^{**};\theta)}-1\big)\right\}\right.\\
&\qquad\left.-\sup_{\theta\in\mathbb{R}}\left\{\theta\frac{g'(t_{j-1}^*)}{g(t_{j-1})}-\big(-\beta\theta+e^{\alpha\theta}-1\big)\right\}\right|\\
&\quad\le\sup_{|\theta|\le K}\sup_{0\le t\le t_j-t_{j-1}}\left|\big(-\beta A(t;\theta)+e^{\alpha A(t;\theta)}-1\big)-\big(-\beta\theta+e^{\alpha\theta}-1\big)\right|\to 0,
\end{aligned}$$

as $t_j - t_{j-1} \to 0$. Hence, we conclude that

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left(\frac{1}{n}Z_t \in B_\varepsilon(g), 0 \le t \le T\right)\\
&\quad= -\int_0^T g(t)\sup_{\theta\in\mathbb{R}}\left\{\theta\frac{g'(t)}{g(t)} - \left(-\beta\theta + e^{\alpha\theta} - 1\right)\right\}dt\\
&\quad= -\sup_{\theta(t):0\le t\le T}\int_0^T \left\{\theta(t)g'(t) - \left(-\beta\theta(t) + e^{\alpha\theta(t)} - 1\right)g(t)\right\}dt.
\end{aligned}$$

Together with the superexponential estimates (4.11) and (4.12), we have proved that, $\mathbb{P}(\{\frac{1}{n}Z_t, 0 \le t \le T\} \in \cdot)$ satisfies a large deviation principle with the rate function

$$I_Z(g) = \sup_{\theta(t):0\le t\le T}\int_0^T \left\{\theta(t)g'(t) - \left(-\beta\theta(t) + e^{\alpha\theta(t)} - 1\right)g(t)\right\}dt,$$

if $g \in \mathcal{AC}_1[0, T]$. Note that the maximization problem

$$\sup_x\left\{xg' - \left(-\beta x + e^{\alpha x} - 1\right)g\right\}$$

has its maximum achieved at $x = \frac{1}{\alpha}\log(\frac{\beta+\frac{g'}{g}}{\alpha})$, provided that $g' \ge -\beta g$. Otherwise, the maximum is $+\infty$. Therefore, we conclude that

$$I_Z(g) = \int_0^T \frac{\beta g(t) + g'(t)}{\alpha}\log\left(\frac{\beta g(t) + g'(t)}{\alpha g(t)}\right) - \left(\frac{\beta g(t) + g'(t)}{\alpha} - g(t)\right)dt,$$

for any $g \in \mathcal{AC}_1[0, T]$ and $g' \ge -\beta g$ and $I_Z(g) = +\infty$ otherwise.

Step 4. Finally let us show that the rate function $I_Z(g)$ is good. That is, we need to show that for any fixed $m > 0$, the level set

$$K_m := \left\{g \in \mathcal{AC}_1[0, T] : I_Z(g) \le m\right\} \tag{4.18}$$

is compact.

Since $Z_t \ge Z_0 e^{-\beta t}$, we have $g(t) \ge g(0)e^{-\beta t} = e^{-\beta t}$ for any t. Therefore, for any $g \in K_m$,

$$e^{-\beta T}\int_0^T \Lambda^*\left(\frac{\beta g(t) + g'(t)}{\alpha g(t)}\right)dt \le m, \tag{4.19}$$

where $\Lambda^*(x) := x\log x - x + 1$ is strictly convex and non-negative. Thus, for any $g \in K_m$,

$$\int_0^T \Lambda^*\left(\frac{\beta}{\alpha} + \frac{1}{\alpha}\frac{g'(t)}{g(t)}\right)dt \le me^{\beta T}. \tag{4.20}$$

Let us define $f(t) = \frac{\beta}{\alpha}t + \frac{1}{\alpha}\log g(t)$. Then $f(0) = 0$ and $f'(t) = \frac{\beta}{\alpha} + \frac{1}{\alpha}\frac{g'(t)}{g(t)}$. From the proof that the rate function for Mogulskii's theorem is good, see, for example, page 183 in Dembo and Zeitouni [9], it follows that the set

$$\left\{ f \in \mathcal{AC}_0[0,T] : \int_0^T \Lambda^*\big(f'(t)\big)\,dt \le me^{\alpha T} \right\} \tag{4.21}$$

is a bounded set of equicontinuous functions. Since $g(t) = e^{\alpha f(t) - \beta t}$, it follows that the set K_m is a bounded set of equicontinuous functions. By Arzelà–Ascoli theorem, the set K_m is compact. Hence, $I_Z(g)$ is a good rate function. The proof is complete. □

Proof of Theorem 2. We apply Theorem 1 and the contraction principle. One then readily obtains from (2.6) that $\mathbb{P}(\{\frac{1}{n}N_t, 0 \le t \le T\} \in \cdot)$ satisfies a large deviation principle with the good rate function

$$I_N(h) = \inf_{h(t) = \frac{g(t)-1}{\alpha} + \frac{\beta}{\alpha}\int_0^t g(s)\,ds, 0 \le t \le T} I_Z(g). \tag{4.22}$$

Observe that differentiating the integral equation $h(t) = \frac{g(t)-1}{\alpha} + \frac{\beta}{\alpha}\int_0^t g(s)\,ds$, we get

$$h'(t) = \frac{1}{\alpha}g'(t) + \frac{\beta}{\alpha}g(t),$$

which is a first-order linear ODE for $g(t)$ with initial condition $g(0) = 1$. Thus, we can solve this ODE and get

$$g(t) = e^{-\beta t} + e^{-\beta t}\int_0^t \alpha e^{\beta s}h'(s)\,ds.$$

Hence, we infer from (4.22) and the expression of $I_Z(g)$ in (2.1) that

$$\begin{aligned} I_N(h) &= \int_0^T h'(t)\log\frac{h'(t)}{g(t)} - \big(h'(t) - g(t)\big)\,dt \\ &= \int_0^T h'(t)\log\left(\frac{h'(t)}{e^{-\beta t} + e^{-\beta t}\int_0^t \alpha e^{\beta s}h'(s)\,ds}\right) \\ &\quad - \left(h'(t) - e^{-\beta t} - e^{-\beta t}\int_0^t \alpha e^{\beta s}h'(s)\,ds\right)dt. \end{aligned}$$

Using this sample path large deviations result and applying the contraction principle, we can also obtain that, $\mathbb{P}(N_T/n \in \cdot)$ satisfies a scalar large deviation principle on $\mathbb{R}^+$ with the good rate function

$$H(x;T) = \inf_{h:h(T)=x} I_N(h). \tag{4.23}$$

Next, we prove that the rate function H in (4.23) can be equivalently given by (2.9). Recall the moment generating function of N_t in (4.6),

$$\mathbb{E}\big[e^{\theta N_t}|Z_0=n\big]=\exp\left\{\left(C\left(t;\frac{\theta}{\alpha}\right)-\frac{\theta}{\alpha}\right)n+D\left(t;\frac{\theta}{\alpha}\right)\right\},$$

where

$$C'\left(t;\frac{\theta}{\alpha}\right)=-\beta C\left(t;\frac{\theta}{\alpha}\right)+e^{\alpha C(t;\frac{\theta}{\alpha})}-1+\frac{\beta\theta}{\alpha},\qquad C\left(0;\frac{\theta}{\alpha}\right)=\frac{\theta}{\alpha}.$$

Let us first consider the critical and super-critical case, that is, $\alpha\geq\beta$. When we have $\alpha\geq\beta$, for any $C>0$ and $\theta>0$, $-\beta C+e^{\alpha C}-1+\frac{\beta\theta}{\alpha}>0$ and thus $C(t;\frac{\theta}{\alpha})$ is increasing in t. It is clear that for any $\theta>0$, $\int_{\frac{\theta}{\alpha}}^{\infty}\frac{dC}{-\beta C+e^{\alpha C}-1+\frac{\beta\theta}{\alpha}}<\infty$. On the other hand, it is easy to see that $\int_0^{\infty}\frac{dC}{-\beta C+e^{\alpha C}-1}=\infty$. Therefore, for any fixed $T>0$, there exists a unique positive value $\theta_d(T)$ such that

$$\int_{\frac{\theta_d(T)}{\alpha}}^{\infty}\frac{dC}{-\beta C+e^{\alpha C}-1+\frac{\beta\theta_d(T)}{\alpha}}=T. \tag{4.24}$$

Hence, we conclude that for any fixed $T>0$, for any $0<\theta<\theta_d(T)$, $C(T;\frac{\theta}{\alpha})$ is the unique positive value greater than $\frac{\theta}{\alpha}$, that satisfies the equation:

$$\int_{\frac{\theta}{\alpha}}^{C(T;\frac{\theta}{\alpha})}\frac{dC}{-\beta C+e^{\alpha C}-1+\frac{\beta\theta}{\alpha}}=T. \tag{4.25}$$

The case for $\theta\leq 0$ is similar. Also, it is easy to see that for $\theta<\theta_d(T)$, $C(t;\frac{\theta}{\alpha})$ is continuous and finite in t, and

$$D\left(T;\frac{\theta}{\alpha}\right)=\mu\int_0^T\left(e^{\alpha C(t;\frac{\theta}{\alpha})}-1\right)dt$$

is finite. Therefore, for $\theta<\theta_d(T)$,

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}\big[e^{\theta N_T}\big]=C\left(T;\frac{\theta}{\alpha}\right)-\frac{\theta}{\alpha}.$$

When $\theta\geq\theta_d(T)$, this limit is ∞. By differentiating the equation (4.25) with respect to θ, we get

$$\begin{aligned}&-\frac{1}{e^{\theta}-1}-\frac{\beta}{\alpha}\int_{\frac{\theta}{\alpha}}^{C(T;\frac{\theta}{\alpha})}\frac{dC}{(-\beta C+e^{\alpha C}-1+\frac{\beta\theta}{\alpha})^2}\\&\quad+\frac{1}{-\beta C(T;\frac{\theta}{\alpha})+e^{\alpha C(T;\frac{\theta}{\alpha})}-1+\frac{\beta\theta}{\alpha}}\frac{d}{d\theta}C\left(T;\frac{\theta}{\alpha}\right)=0.\end{aligned} \tag{4.26}$$

It is clear from the equation (4.24) and (4.25) that as $\theta \to \theta_d(T)$, we have $C(T; \frac{\theta}{\alpha}) \to \infty$. Therefore, from (4.26), we get

$$\begin{aligned}\frac{\partial}{\partial\theta} C\left(T; \frac{\theta}{\alpha}\right) &= \left(-\beta C\left(T; \frac{\theta}{\alpha}\right) + e^{\alpha C(T; \frac{\theta}{\alpha})} - 1 + \frac{\beta\theta}{\alpha}\right) \\ &\quad \times \left(\frac{1}{e^{\theta} - 1} + \frac{\beta}{\alpha} \int_{\frac{\theta}{\alpha}}^{C(T; \frac{\theta}{\alpha})} \frac{dC}{(-\beta C + e^{\alpha C} - 1 + \frac{\beta\theta}{\alpha})^2}\right) \to \infty,\end{aligned}$$

as $\theta \to \theta_d(T)$. Hence, we verified the essential smoothness condition. By Gärtner–Ellis theorem, we get the desired result. The proof for the sub-critical case is similar and is omitted here. □

Acknowledgements

The authors are grateful to the Editor and an anonymous referee for their helpful suggestions that greatly improved the quality of the paper. Xuefeng Gao acknowledges support from Hong Kong RGC ECS Grant 2191081 and CUHK Direct Grants 4055035 and 4055054. Lingjiong Zhu is grateful to the support from NSF Grant DMS-1613164. We thank Xiang Zhou for help with the figures.

Supplementary Material

Supplement to "Large deviations and applications for Markovian Hawkes processes with a large initial intensity" (DOI: 10.3150/17-BEJ948SUPP; .pdf). We provide proofs for additional results in the paper in the supplemental article [14].

References

[1] Abergel, F. and Jedidi, A. (2015). Long-time behavior of a Hawkes process-based limit order book. *SIAM J. Financial Math.* **6** 1026–1043. MR3418223

[2] Asmussen, S. and Albrecher, H. (2010). *Ruin Probabilities*, 2nd ed. *Advanced Series on Statistical Science & Applied Probability* **14**. Hackensack, NJ: World Scientific. MR2766220

[3] Bacry, E., Delattre, S., Hoffmann, M. and Muzy, J.F. (2013). Some limit theorems for Hawkes processes and application to financial statistics. *Stochastic Process. Appl.* **123** 2475–2499. MR3054533

[4] Blanchet, J., Chen, X. and Lam, H. (2014). Two-parameter sample path large deviations for infinite-server queues. *Stoch. Syst.* **4** 206–249. MR3353218

[5] Bordenave, C. and Torrisi, G.L. (2007). Large deviations of Poisson cluster processes. *Stoch. Models* **23** 593–625. MR2362700

[6] Dassios, A. and Zhao, H. (2012). Ruin by dynamic contagion claims. *Insurance Math. Econom.* **51** 93–106. MR2928746

[7] Dassios, A. and Zhao, H. (2013). Exact simulation of Hawkes process with exponentially decaying intensity. *Electron. Commun. Probab.* **18** 13. MR3084573

[8] Delattre, S., Fournier, N. and Hoffmann, M. (2016). Hawkes processes on large networks. *Ann. Appl. Probab.* **26** 216–261. MR3449317

[9] Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*, 2nd ed. *Applications of Mathematics* (*New York*) **38**. New York: Springer. MR1619036

[10] Dragomir, S.S. (2003). *Some Gronwall Type Inequalities and Applications*. Hauppauge, NY: Nova Science. MR2016992

[11] Errais, E., Giesecke, K. and Goldberg, L.R. (2010). Affine point processes and portfolio credit risk. *SIAM J. Financial Math*. **1** 642–665. MR2719785

[12] Feng, J. and Kurtz, T.G. (2006). *Large Deviations for Stochastic Processes*. *Mathematical Surveys and Monographs* **131**. Providence, RI: Amer. Math. Soc. MR2260560

[13] Gao, X. and Zhu, L. (2015). Limit theorems for Markovian Hawkes processes with a large initial intensity. Available at arXiv:1512.02155.

[14] Gao, X. and Zhu, L. Supplement to "Large deviations and applications for Markovian Hawkes processes with a large initial intensity." DOI:10.3150/17-BEJ948SUPP.

[15] Glynn, P.W. (1995). Large deviations for the infinite server queue in heavy traffic. In *Stochastic Networks*. *IMA Vol. Math. Appl.* **71** 387–394. New York: Springer. MR1381021

[16] Glynn, P.W. and Whitt, W. (1991). A new view of the heavy-traffic limit theorem for infinite-server queues. *Adv. in Appl. Probab.* **23** 188–209. MR1091098

[17] Hawkes, A.G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58** 83–90. MR0278410

[18] Hawkes, A.G. (1971). Point spectra of some mutually exciting point processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **33** 438–443. MR0358976

[19] Hawkes, A.G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *J. Appl. Probab.* **11** 493–503. MR0378093

[20] Jaisson, T. and Rosenbaum, M. (2015). Limit theorems for nearly unstable Hawkes processes. *Ann. Appl. Probab.* **25** 600–631. MR3313750

[21] Jaisson, T. and Rosenbaum, M. (2016). Rough fractional diffusions as scaling limits of nearly unstable heavy tailed Hawkes processes. *Ann. Appl. Probab.* **26** 2860–2882. MR3563196

[22] Karabash, D. and Zhu, L. (2015). Limit theorems for marked Hawkes processes with application to a risk model. *Stoch. Models* **31** 433–451. MR3395721

[23] Puhalskii, A. (1995). Large deviation analysis of the single server queue. *Queueing Syst.* **21** 5–66. MR1372048

[24] Stabile, G. and Torrisi, G.L. (2010). Risk processes with non-stationary Hawkes claims arrivals. *Methodol. Comput. Appl. Probab.* **12** 415–429. MR2665268

[25] Torrisi, G.L. (2016). Gaussian approximation of nonlinear Hawkes processes. *Ann. Appl. Probab.* **26** 2106–2140. MR3543891

[26] Torrisi, G.L. (2017). Poisson approximation of point processes with stochastic intensity, and application to nonlinear Hawkes processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **53** 679–700. MR3634270

[27] Varadhan, S.R.S. (1984). *Large Deviations and Applications*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **46**. Philadelphia, PA: SIAM. MR0758258

[28] Whitt, W. (2002). *Stochastic-Process Limits*: *An Introduction to Stochastic-Process Limits and Their Application to Queues*. *Springer Series in Operations Research*. New York: Springer. MR1876437

[29] Zhang, X., Blanchet, J., Giesecke, K. and Glynn, P.W. (2015). Affine point processes: Approximation and efficient simulation. *Math. Oper. Res.* **40** 797–819. MR3423739

[30] Zhu, L. (2013). Nonlinear Hawkes Processes. Ph.D. thesis, New York University.

[31] Zhu, L. (2013). Ruin probabilities for risk processes with non-stationary arrivals and subexponential claims. *Insurance Math. Econom.* **53** 544–550. MR3130449

[32] Zhu, L. (2013). Moderate deviations for Hawkes processes. *Statist. Probab. Lett.* **83** 885–890. MR3040318

[33] Zhu, L. (2013). Central limit theorem for nonlinear Hawkes processes. *J. Appl. Probab.* **50** 760–771. MR3102513

[34] Zhu, L. (2014). Limit theorems for a Cox–Ingersoll–Ross process with Hawkes jumps. *J. Appl. Probab.* **51** 699–712. MR3256221

[35] Zhu, L. (2014). Process-level large deviations for nonlinear Hawkes point processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** 845–871. MR3224291

[36] Zhu, L. (2015). Large deviations for Markovian nonlinear Hawkes processes. *Ann. Appl. Probab.* **25** 548–581. MR3313748

Received April 2016 and revised February 2017