Robust identification of controlled Hawkes processes

Michael Mark [*] and Thomas A. Weber [†]
École Polytechnique Fédérale de Lausanne, Station 5, CH-1015 Lausanne, Switzerland

(Received 16 October 2019; accepted 19 February 2020; published 20 April 2020)

The identification of Hawkes-like processes can pose significant challenges. Despite substantial amounts of data, standard estimation methods show significant bias or fail to converge. To overcome these issues, we propose an alternative approach based on an expectation-maximization algorithm, which instrumentalizes the internal branching structure of the process, thus improving convergence behavior. Furthermore, we show that our method provides a tight lower bound for maximum-likelihood estimates. The approach is discussed in the context of a practical application, namely the collection of outstanding unsecured consumer debt.

DOI: 10.1103/PhysRevE.101.043305

I. INTRODUCTION

In contrast to the exogenous intensity of an inhomogeneous Poisson point process, the intensity of a Hawkes process is self-exciting: it depends endogenously on the arrival history [1,2]. Any arrival event induces an intensity jump which dissipates through a memory kernel, and this in turn influences the probability of the next arrival event. The first applications of Hawkes processes appeared in seismology, for the analysis of earthquakes and associated aftershock sequences [3]. Since then, self-exciting processes have proved useful across numerous other fields, including finance [4], marketing [5], and neuroscience [6], to name a few. The performance of the underlying parametric models depends first and foremost on a correct model specification. Here we focus on the identification of a class of linear controlled marked Hawkes processes, where the arrival events include scalar marks and the arrival intensity is regulated by an impulse control. This class of controlled self-exciting processes was considered by Chehrazi and Weber [7] to predict the repayment behavior of unsecured loans placed in credit collections. In this application, the collector disposes of a set of account-treatment actions (e.g., establishing first-party contact or sending a notice letter) to exert pressure on the debtor. A similar class of processes was used by Rambaldi *et al.* [8] to model foreign-exchange price dynamics subject to exogenous deterministic jumps in the form of news about macroeconomic events.

In the credit-collection example, a misspecification of model parameters leads to faulty predictions of account values and suboptimal account-treatment schedules.[1] Although standard identification techniques, such as maximum-likelihood estimation (MLE), may well be asymptotically consistent, the corresponding estimators tend to exhibit a significant bias as soon as the amount of available data is sparse. This is the case in many practical applications such as credit collections where a delinquent account over the collection history usually features only a few repayment events. In addition, it is often difficult to compute the best-fit parameters because of nonconcavities and near-vanishing gradients of the objective function that lead to ill-conditioned iterations. To ameliorate convergence behavior and estimation performance of standard MLE methods, we propose a robust estimation method based on an expectation-maximization (EM) algorithm. The latter exploits the branching structure of the process, featuring a primal-dual type approximation. In each iteration, first the lower bound for the likelihood function is updated ("expectation step") before the parameter estimate is reoptimized ("maximization step"). Using a fairly generic setup (in the context of credit collections, to fix ideas), we show that the EM algorithm achieves substantial improvements in convergence behavior and thus an increased robustness with respect to a broad range of starting values for the parameter vector.

A. Literature

Due to its relative simplicity, MLE is a common inference method for point processes specified via conditional intensity. A semiclosed form for the estimator was derived by Rubin [10] who established a link between the conditional density function of the interarrival times and the intensity for regular point processes.[2] The performance of MLE was tested on seismic data by Ozaki [11], who also introduced a computationally efficient recursive simplification for MLE and derived the Jacobian and Hessian of the corresponding likelihood function. Determining the MLE estimator then amounts to solving a nonconvex program, using appropriate optimization machinery—with the Jacobian and Hessian readily available for the univariate case.

[*] michael.mark@epfl.ch
[†] thomas.weber@epfl.ch

[1] For optimal closed-loop control of repayment processes, see Chehrazi *et al.* [9].

[2] A regular point process, defined on a standard probability space $(\Omega, \mathcal{F}_t, \mathbb{P})$, is nonexplosive [i.e., $N(t) < \infty$ for all finite $t \geqslant 0$].

Although Ogata [12] proved that the associated MLE estimator is asymptotically normal, efficient, and consistent, the amount of data available for fitting is often insufficient for attaining the asymptotic regime.

The resulting estimates tend to be heavily biased or worse, the estimator fails to converge, in many practical applications. For example, such convergence issues were noted in the case of our class of controlled Hawkes processes by Chehrazi and Weber [7] who proceeded, somewhat *ad hoc*, to filter out unlikely local minima using a Cramér-von Mises goodness-of-fit criterion. The generically poor and unreliable convergence of the MLE estimator is further exacerbated by the log-likelihood function's exhibiting frequently multimodal or extremely flat behavior near its critical points, resulting overall in a lack of apparent well-posedness [13], in the sense that close initialization values can produce very different estimation results. Veen and Schoenberg [14] documented these anomalies for the popular seismological spatial-temporal epidemic type aftershock sequence (ETAS) model [15–17] highlighting the low curvature of the ETAS log likelihood which deteriorates the performance of the numerical optimization routine. Furthermore, multimodality of the log likelihood has been empirically confirmed, since for different starting values the optimizer converges generically to different local minima. To overcome the associated computational challenges, the authors suggested to take advantage of the natural branching structure of the process [18] framing the estimation as an incomplete-data problem where the information about which event triggers other events is unobservable. Building on the case presented by Veen and Schoenberg for the ETAS model, we develop an adapted version of the expectation-maximization algorithm suited for our class of controlled Hawkes processes. Furthermore, we improve the original method by an additional term that shifts the EM-objective function (i.e., the "expected complete log likelihood"; see Sec. III B) such that, at the optimum, it becomes a tight lower bound to the log-likelihood function.

B. Outline

The remainder of this paper is organized as follows. In Sec. II, we introduce the class of linear controlled Hawkes processes and showcase its importance in two practical examples. Section III first reviews the MLE estimator pinpointing its shortcomings and then constructs our estimation method based on the EM algorithm. In Sec. IV, we compare the two methods in terms of their respective convergence stability and bias. Section V concludes.

II. CONTROLLED HAWKES PROCESSES

A. Definition

The intensity of a (linear) *controlled Hawkes process* (CHP) is given by

$$\lambda(t|\mathcal{H}_t) = \mu(t) + \sum_{i:\tau_i<t} g(t-\tau_i, m_i) + a(t), \tag{1}$$

where $\tau_i \geqslant 0$ denotes the ith arrival time and $m_i \in \mathbb{R}$ is the corresponding mark, for $i \geqslant 1$. The background intensity rate $\mu(t)$ is a deterministic function of time $t \geqslant 0$, the function

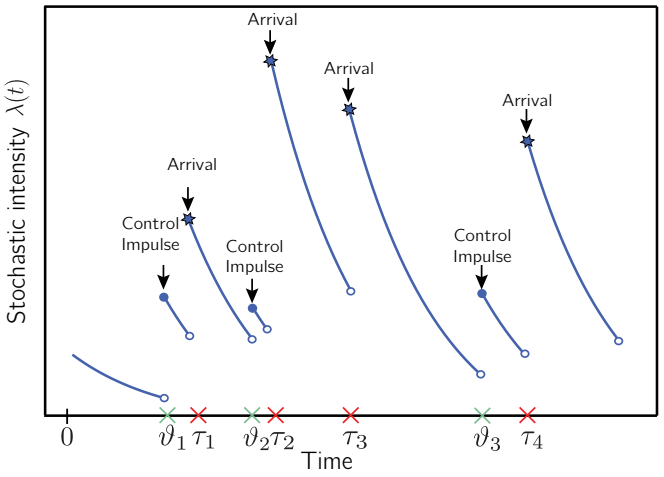

FIG. 1. Dependence of the stochastic intensity $\lambda(t)$ on arrival history and control impulses.

$g : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}_+$ is a (non-negative-valued) memory kernel, and the (open-loop) control $a(t)$ is assumed to be a right-continuous function of the form

$$a(t) = \sum_{j:\vartheta_j<t} \Phi_j(t-\vartheta_j), \tag{2}$$

where each $\Phi_j : \mathbb{R}_+ \to \mathbb{R}_+$ denotes a (non-negative-valued) *exogenous kernel* and each ϑ_j is an instant at which the control variable a undergoes a jump, for $j \geqslant 1$. The corresponding control impulses are dissipated via the (non-negative-valued) exogenous kernels Φ_j as opposed to the *endogenous kernel* g which governs the memory from self-excitation. We assume that on any finite time interval $[0, t]$ the number of impulses is finite, and the intervention times ϑ_j are known in advance. We also assume that both types of kernels satisfy the usual stationarity condition, so $\int_0^\infty \max\{\Phi_j(t), g(t)\}\, dt \leqslant 1$ for all relevant j.[3] The σ algebra $\mathcal{H}_t = \{\{(\tau_i, m_i)\} \times \{\vartheta_j\} : \tau_i < t, \vartheta_j < t\}$ describes the process history, including mark sizes m_i and impulse times ϑ_j. A sample intensity path is shown in Fig. 1.

B. Examples

The practical relevance of CHPs is illustrated by the following two examples.

Example 1: Trading with macroeconomic news. Rambaldi *et al.* [8] analyze the impact of macroeconomic news on market activity, measured by the rate of change in the best quotes. They consider a Hawkes process driven by an endogenous and an exogenous kernel. The self-excitation effect is described by an unmarked endogenous kernel in the form of a linear combination of exponentials,

$$g(t) = \alpha_A e^{-\beta_A t} + \alpha_B e^{-\beta_B t},$$

and the effect of (recurring) macroeconomic news by an exogenous kernel in the form of a single exponential,

$$\Phi_N(t) = \alpha_N e^{-\beta_N t}.$$

[3] In applications with a finite observation horizon, it is sufficient to impose $\int_0^\infty \max\{\Phi_j(t), g(t)\}\, dt < \infty$.

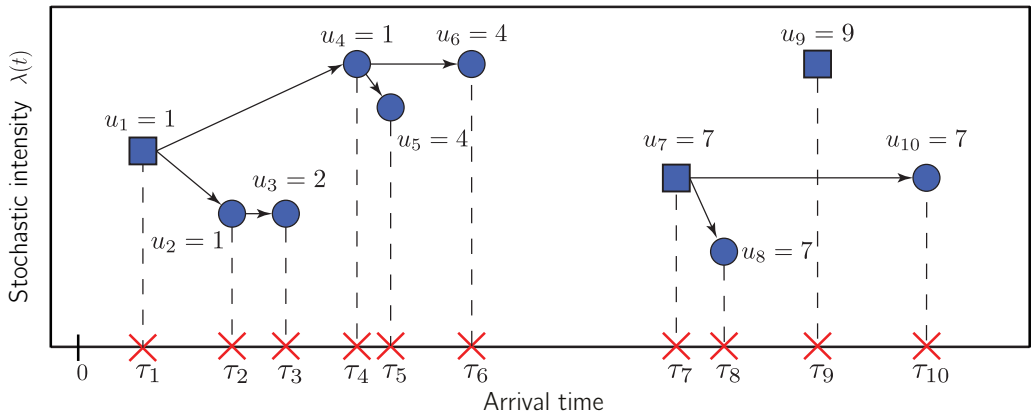

FIG. 2. Branching-structure representation with immigrants (■) and offsprings (●) at the arrival times τ_i (×).

While best-quote changes occur at the random instants τ_i, the arrival times ϑ_j of news releases are known in advance, rendering the exogenous kernel deterministic. Provided the control function $a(t) \equiv \sum_{j:\vartheta_j<t} \Phi_N(t-\vartheta_j)$, the intensity of the controlled Hawkes process is then given by Eq. (1).

Example 2: Credit collections. A CHP was used by Chehrazi and Weber [7] to predict the repayment behavior by holders of credit-card accounts in default. A delinquent account with outstanding balance $B(0) > 0$ placed into collections at time $t = 0$ is credited with relative repayments r_i at times $\{\tau_i\}_{i\in\mathbb{N}}$ until the outstanding debt is paid in full. The sequence (τ_i, r_i), for $i \geqslant 1$, constitutes a marked point process with intensity dynamics that can be described by a mean-reverting stochastic differential equation together with an initial condition:

$$d\lambda(t) = \underbrace{\kappa[\lambda_\infty - \lambda(t)]\,dt}_{\text{mean-reversion}} + \underbrace{\delta_1^\top dJ(t)}_{\text{self-excitation}} + \underbrace{da(t)}_{\text{control}},$$
$$\lambda(0) = \lambda_0, \tag{3}$$

where the two-dimensional jump process $J(t) = [N(t), R(t)]^\top$ represents the marked and unmarked version of the same counting process; indeed, $N(t) = \sum_i \mathbb{1}_{\{\tau_i<t\}}$ captures the holder's willingness to pay and $R(t) = \sum_i \mathbb{1}_{\{\tau_i<t\}} r_i$ his or her ability to pay. The parameter λ_∞ represents the long-run steady state to which the repayment intensity reverts at the rate κ, while $\delta_1 = [\delta_{10}, \delta_{11}]$ denotes the sensitivity to the self-exciting two-dimensional jumps. The control variable $a(t)$ is assumed to be a deterministic nondecreasing right-continuous and piecewise-constant function, taking values in $\mathbb{R}_+$. The exogenous kernel is

$$a(t; \delta_2, \kappa) = \sum_{j:\vartheta_j<t} \delta_{2l(\vartheta_j)} e^{-\kappa(t-\vartheta_j)},$$

where the parameter vector $\delta_2 = (\delta_{21}, \delta_{22}, \ldots, \delta_{2M})$ contains the sensitivities of the repayment intensity to M different account-treatment actions, and the mapping $l : \mathbb{R}_+ \to \mathbb{N}_+$ describes the type $l(\vartheta_j)$ of the action taken at time ϑ_j. In practice, the impulses map to the available collector actions which can vary from mild (inducing smaller intensity jumps, e.g., by sending a letter of notice or making phone calls) to severe (inducing larger intensity jumps, e.g., by filing a lawsuit). Overall, the repayment intensity evolves according to Eq. (1) for $m_i \equiv r_i$. In this, the deterministic drift,

$$\mu(t) = \lambda_\infty + (\lambda_0 - \lambda_\infty)e^{-\kappa t},$$

is exponentially mean reverting, and the triggering kernel,

$$g(t - \tau_i, r_i) = (\delta_{10} + \delta_{11} r_i)e^{-\kappa(t-\tau_i)},$$

describes the effect of a repayment-event arrival at time τ_i on the repayment intensity, for all $t \geqslant \tau_i$.

C. Branching structure

The branching structure presents an augmented view of a Hawkes point process, consisting of the Poisson cluster-process representation introduced by Hawkes and Oakes [18]. It maps event arrivals to clusters, each of which begins with an immigrant arrival following an inhomogeneous Poisson process of base-rate intensity $\mu(t) + a(t)$. Subsequently, every immigrant generates its own offsprings following an inhomogeneous Poisson process with intensity given by the triggering kernel $g(t)$, and this cascades through all offsprings, thus generically clustering the event arrivals. Conceptually, all events fall into two categories: immigrant arrivals and offspring arrivals. Offspring events are triggered by existing events in the process, while immigrant events arrive independently without being preceded by a parent event. This conceptual separation provides additional inner structure to the process.[4] More specifically, the branching structure of the ith arrival at time τ_i is described by a mapping $u : \mathbb{R}_+ \to \mathbb{N}_+$, so

$$u_i = u(\tau_i), \quad i \geqslant 1, \tag{4}$$

where

$$u_i = \begin{cases} i, & \text{if arrival } i \text{ is an immigrant arrival,} \\ j, & \text{if the immediate ancestor of arrival } i \text{ is arrival } j. \end{cases}$$

Assigning either the immediate ancestor $j < i$ (if it exists), or else the current event i, the variable $u_i \in \{1, \ldots, i\}$ determines the branching structure of the Hawkes process, by means of the marked point process (τ_i, u_i); see Fig. 2.

[4]See Daley and Vere-Jones [20] for details on the theory of branching processes.

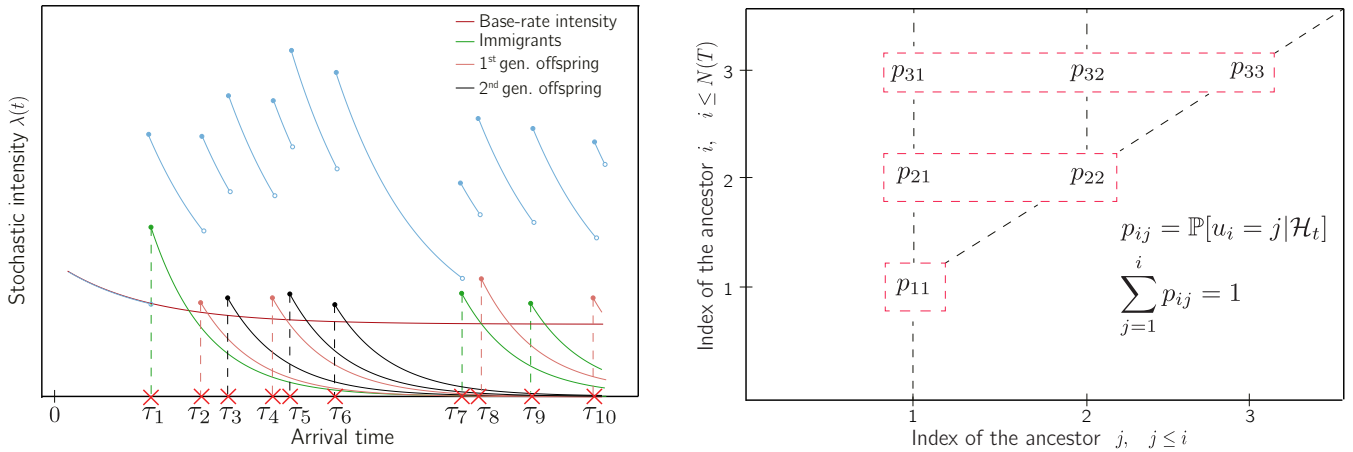

FIG. 3. (a) Intensity decomposition of the Hawkes process from Fig. 2 into a base-rate process and arrival-triggered inhomogeneous Poisson processes. Notice that the ratio of the intensity induced due to a particular arrival to the total intensity determines the probability that the event is an offspring of that particular arrival. (b) Representation of the branching distribution: each row designates the probability mass function for the particular arrival; the diagonal represents probabilities of events being immigrants.

In practice, the branching structure is usually unobservable. Yet, conditional on a set of process parameters and a sample sequence (τ_i, m_i), it is possible to recover its probability distribution,

$$\mathbb{P}[u_i = i|\mathcal{H}_t] = \frac{\mu(\tau_i) + a(\tau_i)}{\lambda(\tau_i)}$$

and

$$\mathbb{P}[u_i = j|\mathcal{H}_t] = \frac{g(\tau_i - \tau_j, m_i)}{\lambda(\tau_i)}, \quad 1 \leqslant j < i. \tag{5}$$

Thus, it is possible to probabilistically assign any arrival i to being an immigrant or offspring [Fig. 3(b)].

As shown in Fig. 3(a), a Hawkes process can be decomposed into a base-rate process and a sum of arrival-triggered inhomogeneous Poisson processes. The resulting (probabilistic) branching structure can be used to perform efficient numerical simulation [19], or as we show in Sec. III, to improve process identification.

Remark 1. The branching structure implies an intuitive isoperimetric constraint on the triggering kernel g that ensures the stability of the system. Indeed, the average number of direct (i.e., first-order) offsprings generated by a single event is the expected branching ratio $\nu = \mathbb{E}_m[\int_0^\infty g(t, m)\, dt]$, whereby the point process remains stable if and only if $\nu < 1$.[5]

III. IDENTIFICATION

Our estimation procedure is presented in the context of Example 2 concerning the collection on defaulted credit-card accounts. Repayments follow a CHP with intensity described by Eq. (1), conditional on the parameter vector $\theta = (\kappa, \lambda_0, \lambda_\infty, \delta_{10}, \delta_{11}, \delta_2)$, a known distribution F of the relative repayments (marks) $m_i = r_i$, and a given sequence of account-treatment times $\{\vartheta_j\}_{j\in\mathbb{N}}$. The information from a realization of such a process then consists of event times $\{\tau_i\}_{i\in\mathbb{N}}$, associated marks $\{r_i\}_{i\in\mathbb{N}}$ (representing a sample draw from the relative-repayment distribution F), and associated account-treatment times $\{\vartheta_j\}_{j\in\mathbb{N}}$. Note that the components $\delta_2^{(j)}$ of the parameter vector δ_2 usually take values in a finite set $\mathcal{D}$ with n_A elements, corresponding to the finitely many available actions, some of which (e.g., phone calls or text messages) may be applied repeatedly to the same account.

In the remainder of this section, we assume that K paths $\mathcal{H}_T^k$ (for $k \in \mathcal{K}$) have been observed over a finite time interval $[0, T]$, corresponding to an account portfolio $\mathcal{K} = \{1, \ldots, K\}$. The joint information is summarized by $\mathcal{H}_T^{\mathcal{K}} = \{\mathcal{H}_T^k : k \in \mathcal{K}\}$.

A. Maximum-likelihood estimation

The conventional MLE procedure directly solves

$$\begin{aligned} &\max_{\theta\in\Theta}, && \ln \mathcal{L}\big(\theta\big|\mathcal{H}_T^{\mathcal{K}}\big), \\ &\text{subject to} && \theta \geqslant 0, \end{aligned} \tag{6}$$

where the *incomplete* data log likelihood is given by

$$\ln \mathcal{L}\big(\theta\big|\mathcal{H}_T^{\mathcal{K}}\big) = \sum_{k=1}^{K} \left(-\int_0^T \lambda\big(s\big|\theta, \mathcal{H}_s^k\big)\, ds + \int_0^T \ln \lambda\big(s\big|\theta, \mathcal{H}_s^k\big)\, dN(s) \right). \tag{7}$$

The descriptor *incomplete* was coined by Veen and Schoenberg [14]; it emphasizes the fact that the estimator does not use additional branching-structure information. The incomplete log-likelihood estimator derived in this manner is asymptotically normal, efficient, and consistent. However, it suffers from the following two notable defects that significantly deteriorate its performance:

[5]Stability is viewed here in the sense that the ratio of total events $N(t)$ to the number $M(t) = \int_0^t [\mu(s) + a(s)]\, ds$ of immigrant events remains bounded with probability 1 for $t \to \infty$. The stability criterion of $\nu < 1$ obtains, since for large t it is $N(t) \approx M(t)/(1 - \nu)$, by the geometric-series formula.

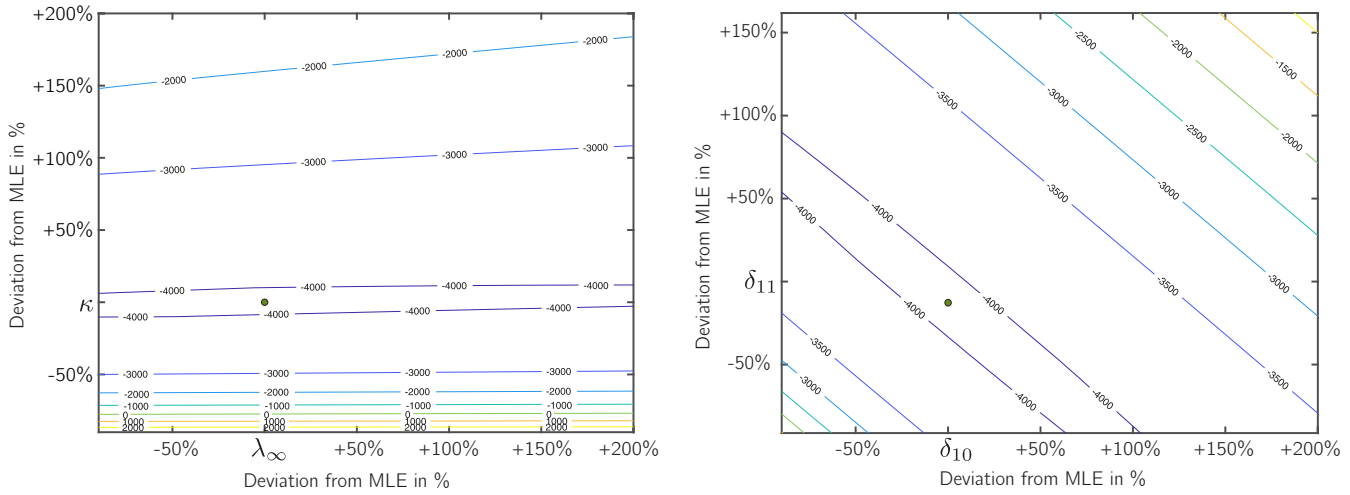

FIG. 4. Flatness of the log likelihood in multidimensional settings. Pairs of components of θ are varied around their MLEs $\hat{\theta}$ (green dot), while all other components remain fixed. (a) Variation of κ. (b) Variation of δ_{11}.

(a) A closed-form solution to the maximization problem (6) is rarely available. Moreover, the efficiency of first- and second-order numerical methods is often poor, as in many cases the log likelihood is extremely flat; see Fig. 4. Along certain trajectories even large disturbances reduce the log likelihood only marginally. For instance, in the (λ_∞, κ) subspace the parameter λ_∞ can be increased by a factor of 2 without significantly impacting the objective function.

(b) Even in the simplest case of a constant-rate exponential Hawkes process, the log likelihood can be multimodal [21]. Specifically, the log likelihood is concave only in the case where κ is fixed. For more complicated models, such as the case of the repayment process in Example 2, the optimization program is guaranteed to be nonconvex. Even if the log likelihood is unimodal, due to the extreme flatness near the MLE estimates $\hat{\theta}$, the objective function can become numerically multimodal as a result of rounding errors.

Although our main focus is to showcase how the branching structure can be employed in the estimation, we note several possible workarounds for MLE-convergence problems; see Fig. 5. The simplest solution to prevent the MLE estimator from getting stuck at a local minimum is to solve the optimization program (6) in parallel for a large batch of starting values and then select the solution that achieves the highest log likelihood. Although effective, the main drawback

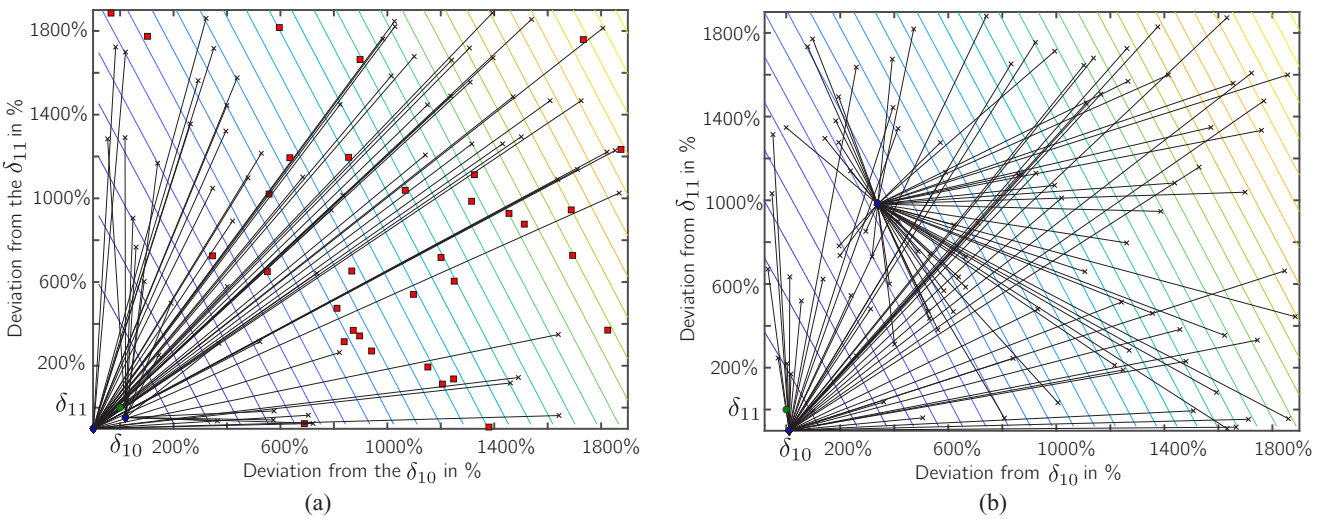

FIG. 5. Convergence problems of the conventional MLE. Except for δ_{10} and δ_{11} the starting values for the estimation procedure are set to their reference values in θ_r; see Table I. Black crosses denote starting values; blue diamonds denote estimation results; the green circle marks the location of the reference parameters. (a) In 39 out of 100 cases, the optimization converged to absurdly large values and was registered as failed by red squares. (b) Although the MLE procedure converged in all 100 cases, the two discovered minima are local, both far from the reference parameters.

of this method is its computational cost, as will be shown in Sec. IV; the advantage of this method, compared to the EM-based algorithm presented below, is negligible. Another highly popular technique relies on the regularization of the estimator, imposing a coefficient penalty in an $\mathscr{L}_1$ or $\mathscr{L}_2$ norm [22,23]. In the context of Hawkes processes, Guo *et al.* [24] proved that the regularized estimator is stable. However, the exact effects of the regularizer on the convergence are still not well understood; that is, despite being functional in practice, it does remain a "black-box solution."

B. Expectation-maximization algorithm

The expectation-maximization algorithm is based on the branching-structure representation introduced in Sec. II C. The idea is to provide the estimator with additional structural information about the process conditional on the observed sample in order to improve the fitting procedure, with the aim of circumventing the problems of ill-conditioning and lack of convergence that are prevalent in the standard MLE procedure.

1. Complete maximum-likelihood estimator

For a known branching structure described with the mapping u in Eq. (4), one obtains the *complete* data log-likelihood function as a sum of two terms, L_1 and L_2.

(i) Log likelihood for immigrant events arriving with base-rate intensity $\lambda_b(t|\mathcal{H}_t^k) = \mu(t) + \sum_{j:\vartheta_j<t} \Phi_j(t-\vartheta_j)$, where $\mu(t) = \lambda_\infty + (\lambda_0 - \lambda_\infty)e^{-\kappa t}$ is the deterministic intensity of the inhomogeneous Poisson process for the immigrants and $\Phi_j(t-\vartheta_j) = \delta_{2l(\vartheta_j)}e^{-\kappa(t-\vartheta_j)}$ is the effect of the action j carried out at time ϑ_j:

$$
\begin{aligned}
&L_1\big(\kappa, \lambda_0, \lambda_\infty, \delta_2 \big| \mathcal{H}_T^k, u\big) \\
&\quad = -\int_0^T \lambda_b\big(s\big|\mathcal{H}_s^k\big)\, ds + \int_0^T \ln \lambda_b\big(s\big|\mathcal{H}_s^k\big)\, dN(s) \\
&\quad = -\left(\int_0^T \mu(s)\, ds + \sum_{j:\vartheta_j<T} \int_{\vartheta_j}^T \Phi_j(s-\vartheta_j)\, ds\right) + \sum_{i:\tau_i\leqslant T} \mathbb{1}_{\{u_i=i\}} \ln \lambda_b\big(\tau_i\big|\mathcal{H}_{\tau_i}^k\big) \\
&\quad = -\left(\lambda_\infty T + \frac{1-e^{-\kappa T}}{\kappa}(\lambda_0-\lambda_\infty) + \sum_{j:\vartheta_j<T} \delta_{2\,l(\vartheta_j)} \frac{1-e^{-\kappa(T-\vartheta_j)}}{\kappa}\right) + \sum_{i:\tau_i\leqslant T} \mathbb{1}_{\{u_i=i\}} \ln\left(\mu(\tau_i) + \sum_{j:\vartheta_j<\tau_i} \delta_{2\,l(\vartheta_j)} e^{-\kappa(\tau_i-\vartheta_j)}\right),
\end{aligned}
$$

for all accounts $k \in \mathcal{K}$.

(ii) Cumulative log likelihood of offspring events generated, respectively, by the different inhomogeneous Poisson processes with intensity $g(t-\tau_i, r_i) = (\delta_{10} + \delta_{11} r_i)e^{-\kappa(t-\tau_i)}$, for $t \in [\tau_i, T]$:

$$
\begin{aligned}
L_2\big(\kappa, \delta_{10}, \delta_{11}\big|\mathcal{H}_T^k, u\big) &= \sum_{i=1}^{N(T)} \left[-\int_{\tau_i}^T g(s-\tau_i, r_i)\, ds + \int_{\tau_i}^T \ln g(s-\tau_i, r_i)\, dN(s)\right] \\
&= \sum_{i=1}^{N(T)} \left[-\int_{\tau_i}^T g(s-\tau_i, r_i)\, ds + \sum_{j=i+1}^{N(T)} \mathbb{1}_{\{u_j=i\}} \ln g(\tau_j-\tau_i, r_i)\right],
\end{aligned}
$$

for all accounts $k \in \mathcal{K}$.

Summing $L_1 + L_2$ over the available sample paths in the account portfolio $\mathcal{K}$, the *complete* log likelihood of the branching process, with intensity in Eq. (1), becomes

$$
\ln \mathcal{L}_C\big(\theta\big|\mathcal{H}_T^{\mathcal{K}}, u\big) = \sum_{k=1}^{K} \big[L_1\big(\kappa, \lambda_0, \lambda_\infty, \delta_2\big|\mathcal{H}_T^k, u\big) + L_2\big(\kappa, \delta_{10}, \delta_{11}\big|\mathcal{H}_T^k, u\big)\big]. \tag{8}
$$

Note that the construction of the complete log likelihood takes into account that the endogenous processes generating the offspring arrivals are mutually independent and independent of the exogenous process generating the immigrant arrivals [18].

As the branching structure is unobservable, the complete log likelihood is generally unavailable. It is therefore natural to resort to the *expected* complete log likelihood (ECLL), conditional on the observed portfolio history $\mathcal{H}_T^{\mathcal{K}}$:

$$
\begin{aligned}
\mathbb{E}\big[\ln \mathcal{L}_C\big(\theta\big|\mathcal{H}_T^{\mathcal{K}}\big)\big] = \sum_{k=1}^{K} \mathbb{E}\Bigg[&-\int_0^T \lambda_b\big(s\big|\theta, \mathcal{H}_s^k\big)\, ds + \sum_{i=1}^{N(T)} \mathbb{1}_{\{u_i=i\}} \ln \lambda_b\big(\tau_i\big|\theta, \mathcal{H}_s^k\big) \\
&- \sum_{i=1}^{N(T)} \int_{\tau_i}^T g(s-\tau_i, r_i)\, ds + \sum_{i=2}^{N(T)} \sum_{j=1}^{i-1} \mathbb{1}_{\{u_i=j\}} \ln g(\tau_i-\tau_j, r_j)\Bigg].
\end{aligned}
$$

Using the identity $\mathbb{E}[\mathbb{1}_{\{u_i=j\}}|\theta, \mathcal{H}_T^k] = \mathbb{P}[u_i = j|\theta, \mathcal{H}_T^k]$ together with Eq. (5), we obtain the ECLL:

$$\mathbb{E}\big[\mathcal{L}_C\big(\theta\big|\mathcal{H}_T^K\big)\big] = \sum_{k=1}^{K}\Bigg[-\int_0^T \lambda_b\big(s\big|\theta, \mathcal{H}_T^k\big)\,ds + \sum_{i=1}^{N(T)} \mathbb{P}\big[u_i = i|\theta, \mathcal{H}_T^k\big]\ln\lambda_b\big(\tau_i\big|\theta, \mathcal{H}_T^k\big) - \sum_{i=1}^{N(T)}\int_{\tau_i}^T g(s-\tau_i, r_i)\,ds + \sum_{i=2}^{N(T)}\sum_{j=1}^{i-1}\mathbb{P}\big[u_i = j\big|\theta, \mathcal{H}_T^k\big]\ln g(\tau_i - \tau_j, r_j)\Bigg]. \tag{9}$$

2. EM algorithm

The expectation-maximization algorithm is initialized with a parameter value θ_0 obtained by using prior experience or an educated guess. The first step of the two-step iteration procedure (in iteration $n \geqslant 1$) consists of computing the *conditional* ECLL of the branching structure, termed $Q(\theta, \theta_n)$, by conditioning the probability distribution of the branching structure in Eq. (5) on the best available parameter estimate θ_n and the process parameters on the unknown parameter θ and the available portfolio data $\mathcal{H}_T^K$. In the second step, one then performs a maximization of $Q(\theta, \theta_n)$ with respect to θ, resulting in the next iterate: θ_{n+1}.

Expectation step (E step). Using standard notation from the unsupervised-learning literature, where the EM algorithm is sometimes used for clustering purposes [25], the *conditional* ECLL becomes

$$Q(\theta, \theta_n) = \sum_{k=1}^{K}\Bigg[-\int_0^T \lambda_b\big(s\big|\theta, \mathcal{H}_T^k\big)\,ds + \sum_{i=1}^{N(T)} \mathbb{P}\big[u_i = i\big|\theta_n, \mathcal{H}_T^k\big]\ln\lambda_b\big(\tau_i\big|\theta, \mathcal{H}_T^k\big) - \sum_{i=1}^{N(T)}\int_{\tau_i}^T g(s-\tau_i, r_i)\,ds + \sum_{i=2}^{N(T)}\sum_{j=1}^{i-1}\mathbb{P}\big[u_i = j\big|\theta_n, \mathcal{H}_T^k\big]\ln g(\tau_i - \tau_j, r_j)\Bigg]. \tag{10}$$

Note that the endogenous kernel g is computed conditional on the "true" parameter θ.

Maximization step (M step). Based on the current parameter estimate θ_n the next iterate is determined as a result of maximizing the conditional ECLL:

$$\theta_{n+1} \in \arg\max_{\theta\in\Theta} Q(\theta, \theta_n), \tag{11}$$

where the compact parameter set Θ is a subset of the positive orthant, chosen by the user so as to limit the search using standard numerical tools.

Termination. Starting with the initial seed θ_0, one iterates through the expectation and maximization steps until the termination condition,

$$Q(\theta_{n+1}, \theta_{n+1}) - Q(\theta_n, \theta_n) \leqslant \varepsilon, \tag{12}$$

is satisfied for a sufficiently small tolerance $\varepsilon > 0$. The procedure is summarized hereafter.

Initialize seed $\theta_0 \in \Theta$, fix a tolerance $\varepsilon \in (0, 1)$,
and set $n \leftarrow 0$, $\delta \leftarrow 1$;
while $\delta > \varepsilon$ **do**
  E-Step: Calculate $\mathbb{P}[u_i = j|\theta_n, \mathcal{H}_T^K]$ for all $1 \leqslant j \leqslant i \leqslant N(T)$;
  M-Step: Find $\theta_{n+1} \in \arg\max_{\theta\in\Theta} Q(\theta, \theta_n)$;
  $n \leftarrow n + 1$ and $\delta \leftarrow Q(\theta_{n+1}, \theta_{n+1}) - Q(\theta_n, \theta_n)$
end

Convergence. Dempster *et al.* [26] show that the sequence $[Q(\theta_n, \theta_n)]_{n\in\mathbb{N}}$ is increasing and bounded, so that it must converge ([27], p. 55). However, there is no guarantee that the limit of the maximizing sequence (see, e.g., [28], Chap. 8) is indeed associated with a global extremum. Conceptually, the EM estimates are expected MLE estimates. Dempster *et al.* [26] also establish that estimates obtained using the EM algorithm are consistent, just as standard MLE estimates (based on the incomplete log-likelihood function).

Remark 2 (EM produces lower bound for MLE). Solving the MLE-problem (6) numerically usually entails local approximations of the objective function, followed by choosing an appropriate increment in the direction of steepest ascent. By contrast, the EM algorithm produces a local approximation of the objective conditional on the model parameters and the distribution of the branching structure as a latent variable. This local approximation constitutes a lower bound for the incomplete log likelihood [29]. The EM algorithm alternates between updating the lower bound (E step) in Eq. (8) and updating the parameter estimate (M step) in Eq. (10) until the termination condition in Eq. (11) is satisfied. Thus, by construction, $\mathcal{L}(\hat{\theta}|\mathcal{H}_T^K) \geqslant Q(\hat{\theta}, \hat{\theta})$, as shown in Appendix A. Intuitively, direct maximization can be viewed as fitting a single point process with specified intensity function, whereas maximizing the conditional ECLL (via the EM algorithm) *simultaneously* fits $N(T) + 1$ inhomogeneous Poisson processes,[6] each weighted by its corresponding branching-structure probability.

Although the objective function in Eq. (8) minorizes the log likelihood, we note that this lower bound is generally not tight. By taking into account the entropy of the branching

[6]Any immigrant arrival triggers an offspring process and so does each offspring arrival. Hence, there is a process for each arrival [altogether $N(T)$ processes] and one immigrant process. The $N(T) + 1$ processes are coupled by the branching distribution in Eq. (5), which depends on the parameter vector and the observed sample data.

TABLE I. Specification of the reference repayment process (θ_r) for the numerical experiment.

κ	λ_0	λ_∞	δ_{10}	δ_{11}	$\delta_2^{(1)}$	$\delta_2^{(2)}$	$\delta_2^{(3)}$
0.4	0.05	0.03	0.08	0.06	0.03	0.06	0.09

distribution, the following result corrects this shortcoming and provides a tight lower bound, guaranteeing that the approximation of the log likelihood becomes exact at the optimal EM estimate.

Theorem 1 (Representation). For all $\theta \in \Theta$, the incomplete log likelihood can be written in the form

$$\ln \mathcal{L}\big(\theta\big|\mathcal{H}_T^K\big) = Q(\theta,\theta) + \Delta(\theta), \tag{13}$$

where the non-negative defect,

$$\Delta(\theta) = -\sum_{k=1}^{K}\sum_{i=1}^{N(T)}\sum_{j=1}^{i}\mathbb{P}[u_i = j|\theta,\mathcal{H}_T] \times \ln \mathbb{P}[u_i = j|\theta,\mathcal{H}_T] \quad (\geqslant 0), \tag{14}$$

describes the entropy of the branching distribution given the observed history $\mathcal{H}_T$. ■

Proof. See Appendix A.

Theorem 1 implies that the conditional ECLL $Q(\theta,\theta_n)$ can be "adjusted" using the defect Δ to become a tight lower bound for the log likelihood, as follows:

$$\hat{Q}(\theta,\theta_n) = Q(\theta,\theta_n) + \Delta(\theta). \tag{15}$$

This *adjusted* (conditional) ECLL can be written in the form

$$\begin{aligned}\hat{Q}(\theta,\theta_n) &= \sum_{k=1}^{K}\Bigg[-\int_0^T \lambda_b\big(s\big|\theta,\mathcal{H}_T^k\big)\,ds \\ &\quad + \sum_{i=1}^{N(T)}\mathbb{P}\big[u_i = i\big|\theta_n,\mathcal{H}_T^k\big]\ln\frac{\lambda_b\big(\tau_i\big|\theta,\mathcal{H}_T^k\big)}{\mathbb{P}\big[u_i = i\big|\theta_n,\mathcal{H}_T^k\big]} \\ &\quad - \sum_{i=1}^{N(T)}\int_{\tau_i}^{T} g(s-\tau_i,r_i)\,ds \\ &\quad + \sum_{i=2}^{N(T)}\sum_{j=1}^{i-1}\mathbb{P}\big[u_i = j\big|\theta_n,\mathcal{H}_T^k\big]\ln\frac{g(\tau_i-\tau_j,r_j)}{\mathbb{P}\big[u_i = j\big|\theta_n,\mathcal{H}_T^k\big]}\Bigg].\end{aligned}$$

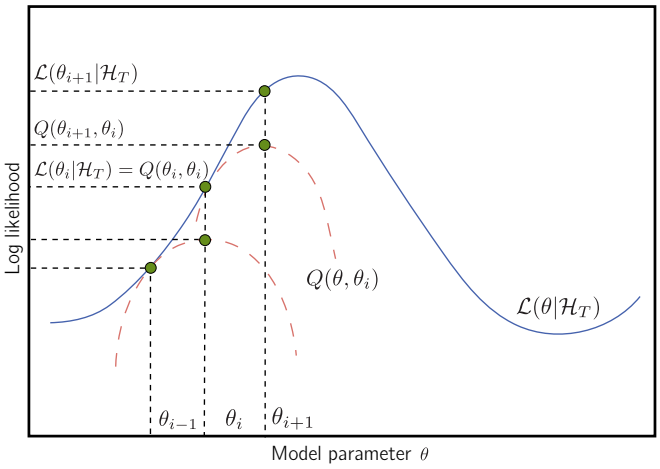

FIG. 6. Illustration of a single EM iteration. Each E step calculates the branching distribution and determines a functional form of the lower bound that is then maximized in the M step.

The adjusted ECLL not only establishes direct comparability with the incomplete log likelihood, but it also significantly reduces the number of iterations needed for convergence. To illustrate the procedure, an iteration of the EM algorithm is sketched in Fig. 6, with additional details provided in Appendix A. Henceforth, all mentions of "ECLL" refer to the *adjusted* ECLL with objective function $\hat{Q}$ (instead of Q).

IV. SIMULATION

For a systematic comparison of the proposed EM algorithm with the standard MLE procedure, we are particularly interested in its convergence performance with respect to randomized initial values θ_0. Convergence performance is key

TABLE II. Asymptotic behavior of the MLE estimator. Each estimate represents the average over 20 independently generated portfolios and 10 random starting values distributed ±25% around the respective (true) reference value.

T		$\hat{\kappa}$	$\hat{\lambda}_0$	$\hat{\lambda}_\infty$	$\hat{\delta}_{10}$	$\hat{\delta}_{11}$	$\hat{\delta}_2^{(1)}$	$\hat{\delta}_2^{(2)}$	$\hat{\delta}_2^{(3)}$	Runtime	Mean $[N(T)]$
500	$\hat{\theta}_{\text{MLE}}$	0.3920	0.0457	0.0300	0.0823	0.0509	0.0307	0.0594	0.0982	90 s	23
	$\hat{\theta}_{\text{EM}}$									1456 s	
	bias	(−2.05%)	(−8.58%)	(+0.03%)	(+2.83%)	(−15.22%)	(+2.38%)	(−1.10%)	(+9.11%)		
1000	$\hat{\theta}_{\text{MLE}}$	0.3997	0.0473	0.0298	0.0777	0.0645	0.0332	0.0588	0.0905	155 s	42
	$\hat{\theta}_{\text{EM}}$									2579 s	
	bias	(−0.03%)	(−5.35%)	(−0.57%)	(−2.84%)	(+ 7.53%)	(+10.54%)	(−1.92%)	(+0.55%)		
2000	$\hat{\theta}_{\text{MLE}}$	0.4070	0.0471	0.0301	0.0817	0.0588	0.0359	0.0626	0.0968	426 s	85
	$\hat{\theta}_{\text{EM}}$									6144 s	
	bias	(+1.75%)	(−5.71%)	(+0.41%)	(+2.11%)	(−1.99%)	(+19.71%)	(+4.31%)	(+7.53%)		

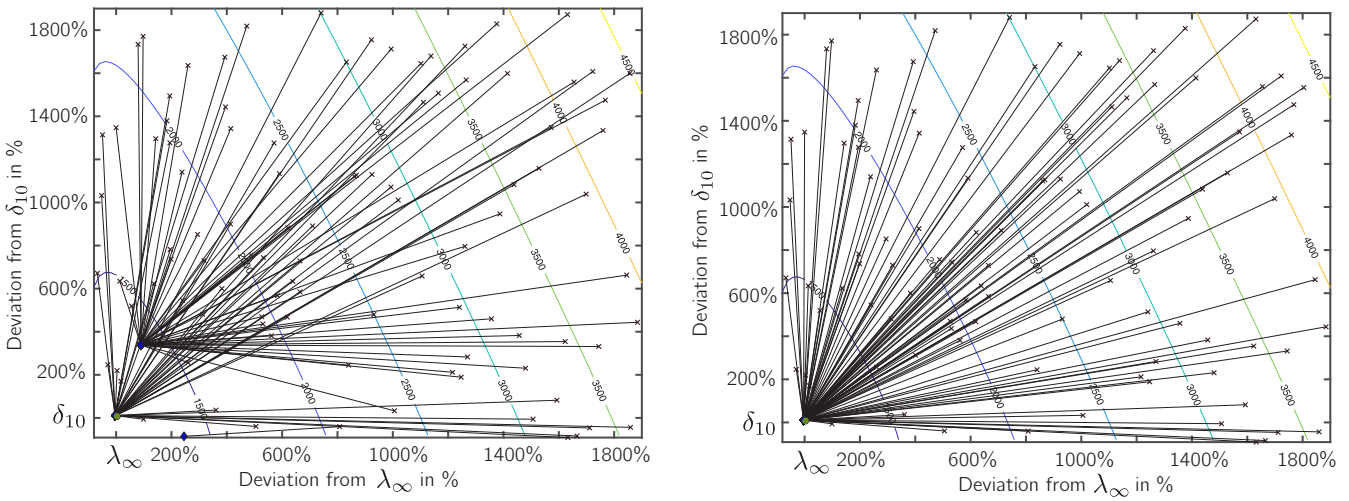

FIG. 7. Comparison of estimator convergence. (a) MLE converging to three distinct points. (b) EM converging close to the reference values. Except for λ_∞ and δ_{10} the parameter starting values are set to their reference values in θ_r; see Table I. The black crosses denote starting values; blue diamonds indicate estimation results; the red circle marks the location of the reference values λ_∞ and δ_{10}.

in practice, since an appropriate parameter range is difficult to determine *ex ante*. Even "educated guesses" for θ_0 are bound to often stray significantly from the ("true") reference value θ_r. The latter is used in our broad numerical experiment to generate synthetic collections data in the context of Example 2 in Sec. II B.

A. Data

It is important to note that credit-collections data by their very nature are relatively sparse. A significant portion of accounts does not exhibit any repayments.[7] This is compensated by the transversal experience across an account portfolio $\mathcal{K}$ containing K sample paths. Throughout the numerical experiment, we consider a CHP driven by the intensity in Eq. (1), generated with the reference parameters specified in Table I. The marks (relative repayments) are assumed to be independent and identically distributed (i.i.d.), uniformly (i.e., $r_i \sim U[0, 1]$).

[7]Even an "empty" sample path conveys valuable information about the underlying process and thus cannot be discarded.

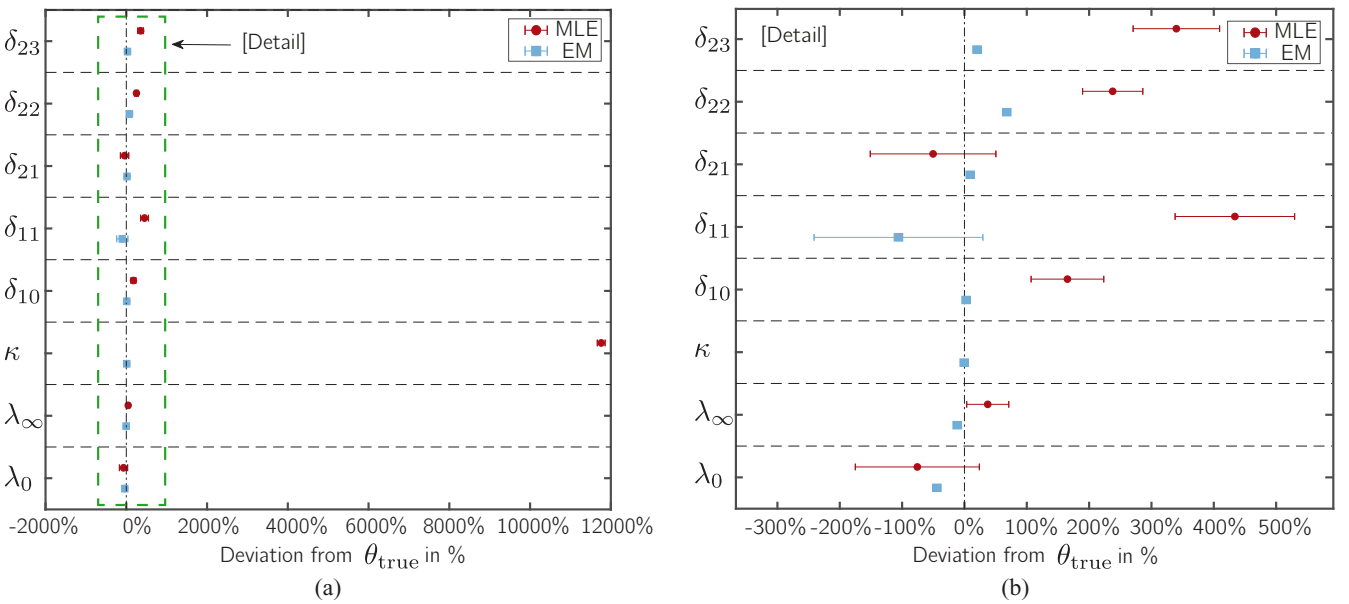

FIG. 8. Estimation results for the scenario in Fig. 7. The center of the error plot designates the average bias over 100 starting values $\hat{\theta}^{(1)} \ldots \hat{\theta}^{(100)}$; error bands are based on one standard deviation. (a) Full display of all parameters, with $\hat{\kappa}$ attaining high local maxima. (b) Detailed view near the zero-bias line.

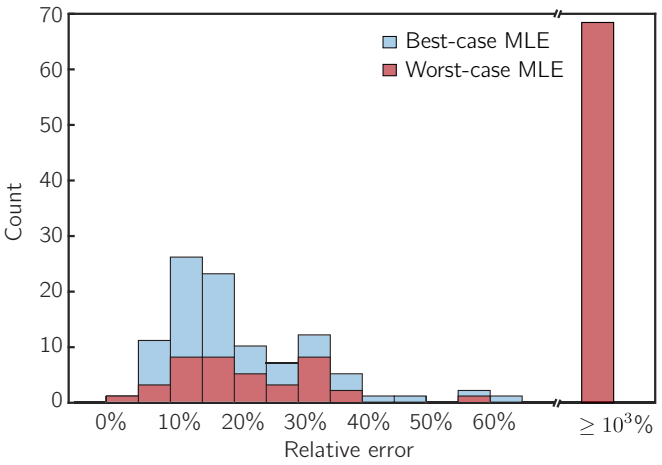

FIG. 9. Sample distribution of the relative error for the best- and worst-case MLE estimates, measured in terms of incomplete log likelihood.

For each of $K = 500$ accounts in the portfolio $\mathcal{K}$, we consider $L = 100$ sample paths, referred to as "scenarios." Each scenario $\ell \in \{1, \ldots, L\}$ generates a history $\mathcal{H}_T^{\mathcal{K}}(\ell)$, based on which the model identification is performed using the two alternative methods (MLE and EM). This is done for $M = 100$ random seeds $\theta_0^{(1)}, \ldots, \theta_0^{(M)}$, obtained as realizations of the random variable $\theta_r \operatorname{diag}(\gamma)$, where $\gamma = (\gamma_g)$ is a vector of the same length as θ_r with entries of the form $\gamma_g = 10^{\beta_g/(20\,\mathrm{dB})}$ describing the gain (positive or negative). In the numerical experiment, gains are considered to be such that $\beta_g \in [-26\,\mathrm{dB}, 26\,\mathrm{dB}]$, corresponding to amplitude distortions γ_g in the interval $[1/20, 20]$.[8]

Each scenario history $\mathcal{H}_T^{\mathcal{K}}(\ell)$ corresponds to data from a portfolio of K treated accounts, with observation horizon $T = 100$, where each account $k \in \mathcal{K}$ is associated with an observed repayment sequence $\{(\tau_i^k, r_i^k)\}$ and a sequence of three control impulses (account treatments) at the i.i.d. times $\vartheta_j^k \sim U[0, T]$ (chosen such that $\vartheta_1^k < \vartheta_2^k < \vartheta_3^k$).

Table II compares the average performance of the MLE estimator $\hat{\theta}_{\mathrm{MLE}}$ and the EM estimator $\hat{\theta}_{\mathrm{EM}}$ over 100 random seeds, distributed uniformly within $\pm 25\%$ of the reference parameter values. It also indicates how the length of the observation horizon T impacts the accuracy of the respective accuracy of the two estimators. Interestingly, both methods produce very similar results, although EM tends to be computationally more expensive. This behavior is somewhat expected, as both methods yield MLE estimates with the EM algorithm relying on additional information related to the branching structure of the repayment process. The real advantage of the EM algorithm over direct MLE maximization becomes apparent when considering initial seeds θ_0 of significant distance to the reference parameter θ_r or when limiting the observation horizon.

B. Results

Consider first the convergence of the estimation in several two-dimensional subspaces of the parameter space Θ. For this, the starting values $\theta_0^{(m)}$, for $m \in \{1, \ldots, M\}$, are set to their reference value θ_r, while the investigated pair of parameter components are randomly varied between -26 dB and $+26$ dB (corresponding to a maximum variation by a factor of $1/20$ or 20) relative to their corresponding reference values as starting values.

Figure 7 shows that MLE fails to converge to the reference values for more than half of the initial seeds (in 52 of 100

[8]Here we consider a uniform distribution (i.i.d.) of γ_g on $[1/20, 20]$. We have also run the entire study for a uniform distribution in the dB space, i.e., for β_g uniformly distributed (i.i.d.) on the interval $[-26\,\mathrm{dB}, 26\,\mathrm{dB}]$, with similar results.

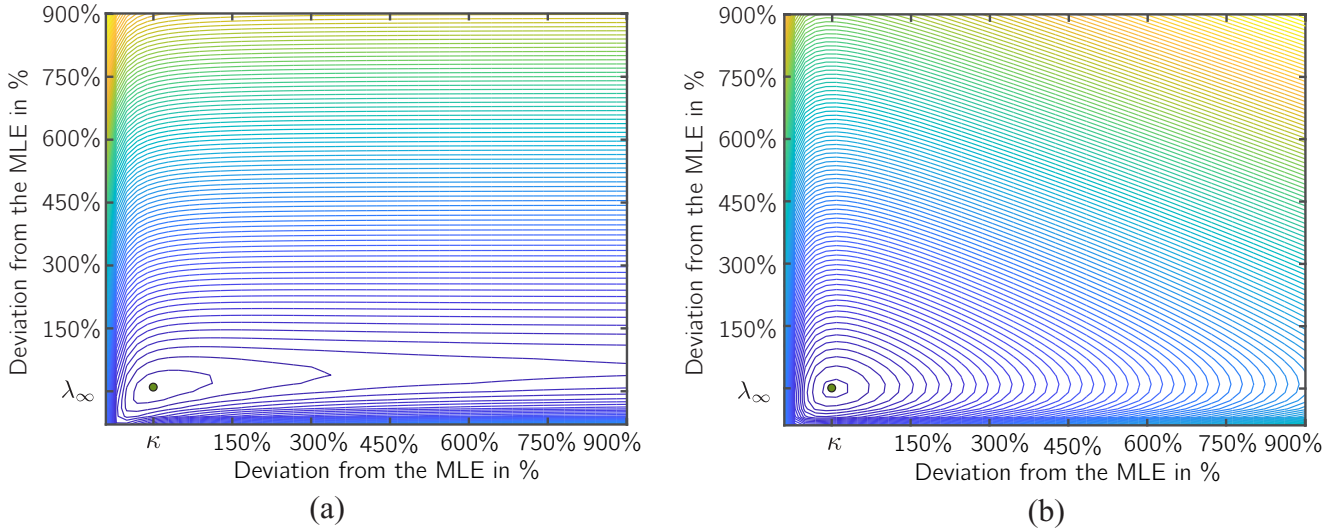

FIG. 10. Comparison of the curvature of the objective functions. (a) Incomplete log likelihood exhibits a long valley around λ_∞. (b) ECLL $Q(\theta, \theta_{n'})$, where n' is the last iterate before attaining the termination condition.

TABLE III. Comparison of the best MLE estimate to the worst EM estimate; results rounded to four significant digits; relative errors (bias) in parentheses.

	$\hat{\kappa}$	$\hat{\lambda}_0$	$\hat{\lambda}_\infty$	$\hat{\delta}_{11}$	$\hat{\delta}_{12}$	$\hat{\delta}_{21}$	$\hat{\delta}_{22}$	$\hat{\delta}_{23}$
$\bar{\theta}_{\mathrm{MLE}}$	0.3703	0.0434	0.0312	0.0652	0.0726	0.0238	0.0554	0.0695
bias	(−7.44%)	(−13.15%)	(+4.12%)	(−18.53%)	(+20.93%)	(−20.74%)	(−7.71%)	(−22.75%)
$\underline{\theta}_{\mathrm{EM}}$	0.3702	0.0434	0.0312	0.0652	0.0725	0.0238	0.0554	0.0695
bias	(−7.45%)	(−13.14%)	(+4.11%)	(−18.53%)	(+20.91%)	(−20.69%)	(−7.71%)	(−22.76%)

instances). We note that the optimizer designated all of the estimation results (blue diamonds) as local minima (implying that a step in any direction would not improve the objective function). To visualize the performance of the estimator in the complete parameter space we use error plots investigating the relationship $\theta_0^{(m)} \to \hat{\theta}^{(m)}$. Figure 8 indicates that while the EM algorithm succeeds in bypassing erroneous local minima, it does so with reduced variance in the estimation results.

Another encountered deficiency is that the MLE estimator failed to converge entirely for a subset of starting values, as can be seen in Figs. 5 and 9.[9] Again, this behavior was not registered for the EM algorithm, except for cases with starting points deviating by more than +10 000% (corresponding to exactly 60 dB) from the reference values.

Remark 3 (Numerical Conditioning). The superior convergence performance of the EM algorithm has two possible sources. First, the properties derived for the MLE estimator hold only asymptotically. Although both $\hat{\theta}_{\mathrm{MLE}}$ and $\hat{\theta}_{\mathrm{EM}}$ are consistent, in a limited sample both estimates can differ, as

$$Q(\hat{\theta}_{\mathrm{EM}}, \hat{\theta}_{\mathrm{EM}}) \geqslant Q(\hat{\theta}_{\mathrm{MLE}}, \hat{\theta}_{\mathrm{MLE}})$$

and

$$\mathcal{L}(\hat{\theta}_{\mathrm{MLE}}; \mathcal{H}_T^{\mathcal{K}}) \geqslant \mathcal{L}(\hat{\theta}_{\mathrm{EM}}; \mathcal{H}_T^{\mathcal{K}}).$$

In our application, the number of repayments may not be large enough to attain an asymptotic regime in the numerical maximization of the incomplete log likelihood. On the other hand, it might be enough for the EM algorithm to produce accurate results, due to the additional branching-structure information captured by the EM estimator.

Furthermore, given the EM construction as a lower bound for the MLE, intuitively, it is expected that the EM-objective function will exhibit a larger "curvature" compared to the incomplete log likelihood. Indeed, as shown in Fig. 10, the objective function for the EM algorithm appears to be a better conditioned objective function for the same problem. We showcase this property using the condition number of the Hessian matrix, which is intricately linked to the convergence performance. In particular, we focus on the difference between the condition numbers for the MLE and EM surface. Computationally, we obtain that the EM objective shows a better conditioned Hessian on average in 80% of all points in the search space; see Fig. 11. Nevertheless, it is important to remember that both methods are local techniques, so neither can provide any guarantee for attaining a global maximizer.

Classical benchmark. As indicated in Sec. III A the erroneous local minima and hence the convergence issues can be circumvented using certain heuristic techniques. Disregarding for a moment the computational burden of repeating the optimization for M initial guesses, we characterize every batch of starting values by a single vector of estimates $\bar{\theta}$ that produces the largest incomplete log likelihood for MLE and EM, respectively:

$$\bar{\theta}_{\mathrm{MLE}} \in \arg\max_{\hat{\theta} \in \hat{\Theta}_M} \mathcal{L}\big(\hat{\theta} \big| \mathcal{H}_T^{\mathcal{K}}\big),$$

and

$$\bar{\theta}_{\mathrm{EM}} \in \arg\max_{\hat{\theta} \in \hat{\Theta}_M} Q(\hat{\theta}, \hat{\theta}),$$

where $\hat{\Theta}_M = \{\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(M)}\}$ denotes the set of estimates coming from M initial values. The best MLE estimates $\bar{\theta}_{\mathrm{MLE}}$ are then compared to the worst (lowest ECLL) EM estimates $\underline{\theta}_{\mathrm{EM}}$ for all scenarios. Table III presents the results for the scenario with the best MLE performance measured as a relative distance from the reference parameter values. Clearly, even the benefit of $M = 100$ different starting values is not enough to outperform a single run (here, the worst case) of the EM. This puts the run times in Table II into perspective. Despite MLE being significantly faster per single run, a large number of runs is needed in order to ensure convergence to the global maximum.

To evaluate the accuracy of the estimation with a single number, we define the (aggregate) relative error for the best (respectively, worst) case using the 2-norm $\|\cdot\|$ as

$$\bar{e} = \frac{\|\bar{\theta} - \theta_r\|}{\|\theta_r\|} \quad \text{and} \quad \underline{e} = \frac{\|\underline{\theta} - \theta_r\|}{\|\theta_r\|}.$$

The evidence from our data indicates that the worst-case and the best-case EM estimates measured in the complete log-likelihood function value are almost indistinguishable. The difference between highest and lowest value of the complete log likelihood, over the batch of initial guesses, was on average in the order of 10^{-2}. This means that the sample distribution of the relative error for the worst-case and best-case EM estimates are almost identical. This dramatically contrasts to

[9] An estimation run is counted as "failed" if any of the estimates exceeds a value of 10^3; see Fig. 5(a).

TABLE IV. Empirical relationship between EM estimates and MLE estimates.

Best of 100		Worst of 100
$\bar{e}_{\mathrm{EM}}$	$\approx$	$\underline{e}_{\mathrm{EM}}$
$\bar{e}_{\mathrm{MLE}}$	$\ll$	$\underline{e}_{\mathrm{MLE}}$
$\bar{e}_{\mathrm{MLE}}$	$\approx$	$\underline{e}_{\mathrm{EM}}$

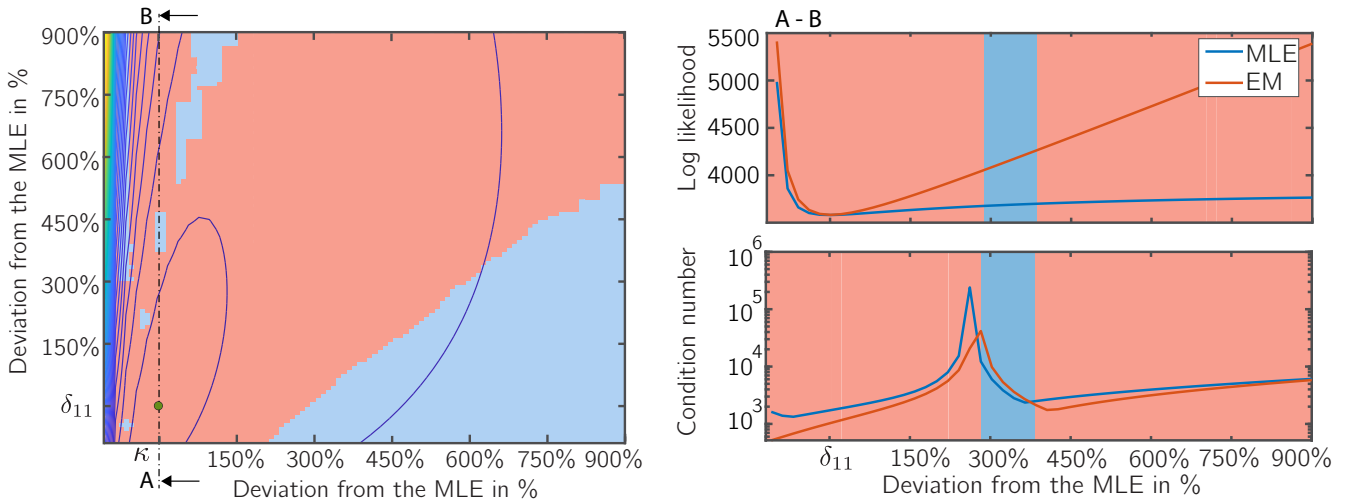

FIG. 11. Comparison of condition numbers for Hessian matrices. (a) Red tiles represent a better-conditioned Hessian for the ECLL,whereas blue tiles represent better-conditioned incomplete log likelihood. (b) Comparison of the condition numbers in κ direction.

the MLE relative-error distribution presented in Fig. 9, where the difference between the best and worst estimates can be extremely large. These empirical relationships are captured in Table IV. The superior convergence performance of the EM algorithm and negligible difference in the best-to-worst comparison speaks for EM as a more robust method of the two. Throughout the numerical experiment *we have not observed a single instance of the direct MLE procedure outperforming the EM in terms of convergence*. Given the substantial number of scenarios tested, we believe that this is a representative and significant result.

Remark 4 (Action Sensitivity). It is worth pointing out that EM may improve the estimation performance of MLE even in settings with limited significance of the control process $a(t)$.[10] Figure 12 showcases the estimation performance, measured in terms of the relative errors of both estimation methods for various δ_2. We consider a similar setup as in the previous section (i.e., ten initial guesses for the solver and ten independently generated portfolios for each value δ_2). In addition, we employ the same filtering technique using the log-likelihood function value to separate the best MLE and

[10]For any given realization of $a(t)$, for $t \geqslant 0$, the importance of the control process for the evolution of the arrival intensity is fully described by the (non-negative) sensitivity parameter δ_2. The case of an *autonomous* Hawkes process (*without* control) corresponds to a degenerate situation with $\delta_2 = 0$.

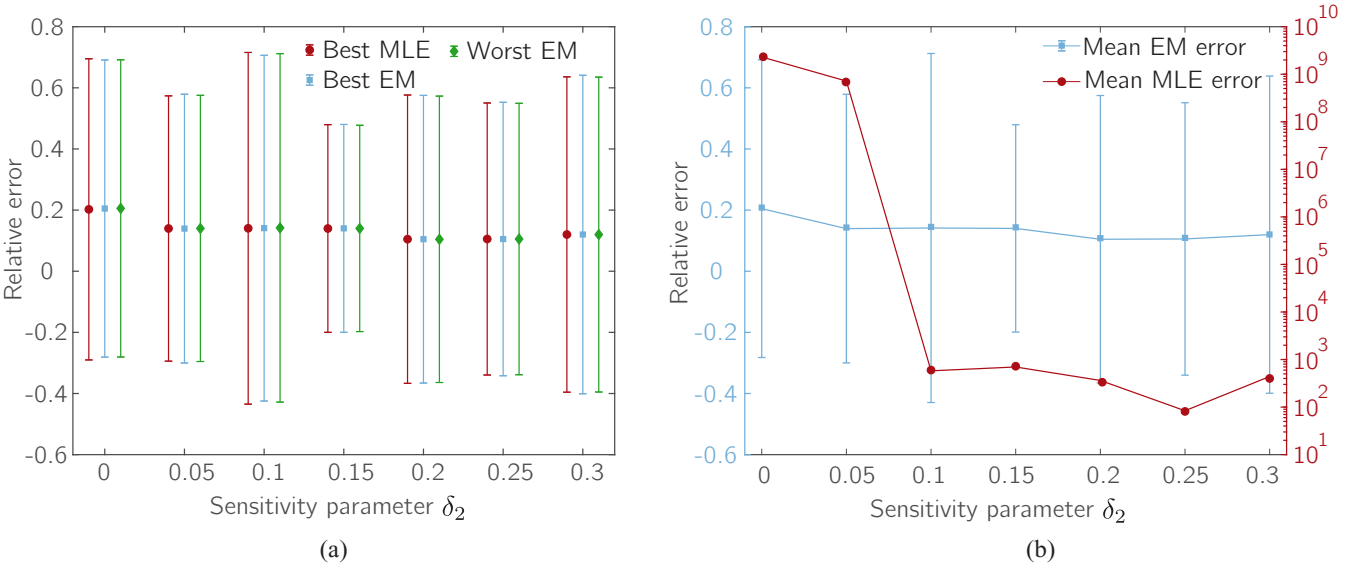

FIG. 12. Impact of sensitivity parameter δ_2 on the relative error. Reported values are averages over ten independently generated portfolios. (a) Relative error of the estimates producing the highest log likelihood over ten initial guesses. (b) Average relative error of all estimates over ten initial guesses. Error bands mark the applicable coefficients of variation.

the worst EM estimates. As expected, the relative error of the best MLEs closely corresponds to the EM estimates; see Fig. 12(a). However, when considering the average relative error of all solutions (not just the best), we observe the previously recorded behavior of MLE's divergence as a result of disparate local likelihood minima; see Fig. 12(b). This suggests that EM can be a preferred estimation method even for uncontrolled (i.e., autonomous) Hawkes processes.

V. CONCLUSION

We have constructed an alternative estimation method for (linear) controlled Hawkes processes based on the EM algorithm. Compared to conventional maximum-likelihood maximization, the presented method exhibits a substantially more robust behavior in terms of convergence and choice of the initial guesses. The robustness was tested based on extensive synthetic credit-collections data mirroring sparse repayment observations as encountered in practice. The EM algorithm performed reliably well across all scenarios and produced maximum-likelihood estimates with a variance significantly below that produced by the standard MLE method. The bias of the EM method was assessed on the best-case MLE to the worst-case EM measured in the value of the log-likelihood function. The difference in the estimates produced was inconsequential ($\pm 0.02\%$) suggesting that the EM algorithm provides a significant stability gain and would therefore be the advised method for the estimation of linear CHP. Our findings suggest that EM is a viable alternative to the conventional MLE, and in applications where a rich history of observations is unavailable, it is a superior estimation method. In cases where direct maximization is preferred, the EM algorithm may be used to obtain bootstrapped initial seeds of the model parameters in question. On the theoretical side, we have shown that the (non-negative) difference between the incomplete log likelihood and the expected complete log likelihood (ECLL) is given by the entropy of the branching distribution, thus establishing a lower bound for the incomplete log likelihood which at the optimal EM estimator becomes tight.

ACKNOWLEDGMENTS

The authors wish to thank N. Chehrazi, two anonymous referees, as well as participants of the INFORMS Annual Meeting in Phoenix, Arizona, for helpful comments and suggestions. Support for this research by the Swiss National Science Foundation (under Grant No. 105218-179175) is gratefully acknowledged.

APPENDIX A: ANALYTICAL DETAILS

ECLL forms lower bound for incomplete log likelihood. For simplicity we assume that $a = 0$ (or else consider $\hat{\mu} = \mu + a$ instead of μ). Comparing the classical incomplete log likelihood and the expected complete log likelihood (as in a log-likelihood-ratio test) yields

$$\begin{aligned}
\ln \mathcal{L} - \mathbb{E}[\ln \mathcal{L}_C] &= \sum_{i=1}^{N(T)} \ln \lambda(\tau_i) - \sum_{i=1}^{N(T)} \frac{\mu(\tau_i)}{\lambda(\tau_i)} \ln \mu(\tau_i) - \sum_{i=2}^{N(T)} \sum_{j=1}^{i-1} \frac{g(\tau_i - \tau_j, m_j)}{\lambda(\tau_i)} \ln g(\tau_i - \tau_j, m_j) \\
&= \sum_{i=2}^{N(T)} \left[\ln \left(\mu(\tau_i) + \sum_{j=1}^{i-1} g(\tau_i - \tau_j, m_j) \right) - \frac{\mu(\tau_i)}{\lambda(\tau_i)} \ln \mu(\tau_i) - \sum_{j=1}^{i-1} \frac{g(\tau_i - \tau_j, m_j)}{\lambda(\tau_i)} \ln g(\tau_i - \tau_j, m_j) \right] \\
&= \sum_{i=2}^{N(T)} \left[\ln \left(\mu(\tau_i) + \sum_{j=1}^{i-1} g(\tau_i - \tau_j, m_j) \right) - \ln \left(\mu(\tau_i)^{\mathbb{P}[u_i=i]} \prod_{j=1}^{i-1} g(\tau_i - \tau_j, m_j)^{\mathbb{P}[u_i=j]} \right) \right] \\
&= \sum_{i=2}^{N(T)} \ln \left(\frac{\mu(\tau_i) + \sum_{j=1}^{i-1} g(\tau_i - \tau_j, m_j)}{\mu(\tau_i)^{\mathbb{P}[u_i=i]} \prod_{j=1}^{i-1} g(\tau_i - \tau_j, m_j)^{\mathbb{P}[u_i=j]}} \right) \geqslant 0.
\end{aligned}$$

To obtain the last inequality, note first that μ and g have non-negative values, and $\mathbb{P}[u_i = i] + \sum_{j=1}^{i-1} \mathbb{P}[u_i = j] = 1$, for all $i \in \{1, \ldots, N(T)\}$. Furthermore, it is

$$\mu(\tau_i) + \sum_{j=1}^{i-1} g(\tau_i - \tau_j) \geqslant \mathbb{P}[u_i = i]\mu(\tau_i) + \sum_{j=1}^{i-1} g(\tau_i - \tau_j, m_j)\mathbb{P}[u_i = j].$$

By the concavity of the natural logarithm and Jensen's inequality we get

$$\ln \left(\mathbb{P}[u_i = i]\mu(\tau_i) + \sum_{j=1}^{i-1} g(\tau_i - \tau_j, m_j)\mathbb{P}[u_i = j] \right) \geqslant \mathbb{P}[u_i = i] \ln \mu(\tau_i) + \sum_{j=1}^{i-1} \mathbb{P}[u_i = j] \ln g(\tau_i - \tau_j, m_j).$$

The right-hand side can then be rewritten in the form

$$\ln \mu(\tau_i)^{\mathbb{P}[u_i=j]} + \sum_{j=1}^{i-1} \ln g(\tau_i - \tau_j, m_j)^{\mathbb{P}[u_i=j]} = \ln \left(\mu(\tau_i)^{\mathbb{P}[u_i=i]} \prod_{j=1}^{i-1} g(\tau_i - \tau_j, m_j)^{\mathbb{P}[u_i=j]} \right).$$

Finally, given that the logarithm is an increasing function, it is

$$\mu(\tau_i) + \sum_{j=1}^{i-1} g(\tau_i - \tau_j, m_j) \geqslant \mu(\tau_i)^{\mathbb{P}[u_i=j]} \prod_{j=1}^{i-1} g(\tau_i - \tau_j, m_j)^{\mathbb{P}[u_i=j]},$$

which implies the inequality in question.

Proof of Theorem 1. Without any loss of generality, we set $K = 1$, so that there is only a single data path in a singleton portfolio, with the superscript k dropped for notational convenience. Assume a realization $\boldsymbol{X} = \{(\tau_1, r_1), (\tau_2, r_2), \ldots, (\tau_n, r_n)\}$ of a CHP given by Eq. (1) with a branching structure described with a latent variable $\boldsymbol{Y} = \{y_1, y_2, \ldots, y_n\}$ (i.e., y_i denotes the ancestor of the ith arrival).[11] That is, $\boldsymbol{X}$ is the incomplete data with complete data given by $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})$. Furthermore, we assume a density of the observed variable $p(\boldsymbol{X}|\theta)$, an arbitrary density of the latent variable $q(\boldsymbol{Y})$, and a joint density $p(\boldsymbol{X}, \boldsymbol{Y}|\theta)$ between the observed and hidden variables. In the setting of CHPs, we can identify the first and the last of the densities with the incomplete and complete log likelihoods, i.e.,

$$p(\boldsymbol{X}|\theta) = \mathcal{L}(\theta|\boldsymbol{X}), \tag{A1}$$

$$p(\boldsymbol{X}, \boldsymbol{Y}|\theta) = \mathcal{L}_C(\theta|\boldsymbol{X}, \boldsymbol{Y}). \tag{A2}$$

Let G be a lower bound to the log-likelihood function parametrized by a parameter θ and the density $q(\boldsymbol{Y})$, such that

$$G(\theta, q) = \ln \mathcal{L}(\theta|\boldsymbol{X}) - D[q \parallel p(\cdot|\boldsymbol{X}, \theta)] \leqslant \ln \mathcal{L}(\theta|\boldsymbol{X}), \tag{A3}$$

where $D[q \parallel p(\cdot|\boldsymbol{X}, \theta)]$ denotes the Kullback-Leibler divergence (relative entropy) of q with respect to $p(\cdot|\boldsymbol{X}, \theta)$.[12] Clearly, the bound G becomes tight if and only if the two distributions are identical. The tight lower bound G can therefore be expressed [for $q(\cdot) = p(\cdot|X, \theta)$] as

$$\begin{aligned} G(\theta, q) &= \ln p(\boldsymbol{X}|\theta) - \mathbb{E}_{\boldsymbol{Y}}\left[\ln \frac{p(\boldsymbol{Y}|\boldsymbol{X}, \theta)}{p(\boldsymbol{Y}|\boldsymbol{X}, \theta)}\right] = \mathbb{E}_{\boldsymbol{Y}}\left[\ln \frac{p(\boldsymbol{Y}|\boldsymbol{X}, \theta)}{p(\boldsymbol{Y}|\boldsymbol{X}, \theta)} + \ln p(\boldsymbol{X}|\theta)\right] \\ &= \mathbb{E}_{\boldsymbol{Y}}\left[\ln \frac{p(\boldsymbol{Y}|\boldsymbol{X}, \theta) p(\boldsymbol{X}|\theta)}{p(\boldsymbol{Y}|\boldsymbol{X}, \theta)}\right] = \mathbb{E}_{\boldsymbol{Y}}\left[\ln \frac{p(\boldsymbol{X}, \boldsymbol{Y}|\theta)}{p(\boldsymbol{Y}|\boldsymbol{X}, \theta)}\right] \\ &= \mathbb{E}_{\boldsymbol{Y}}\big[\ln p(\boldsymbol{X}, \boldsymbol{Y}|\theta)\big] - \mathbb{E}_{\boldsymbol{Y}}[\ln p(\boldsymbol{Y}|\boldsymbol{X}, \theta)], \end{aligned}$$

where the penultimate equality is obtained using the law of total probability. The two terms correspond to the ECLL in Eq. (8) and the adjustment term $\Delta(\theta)$ in Eq. (13), respectively. Consequently, the branching distribution $p(\cdot|\boldsymbol{X}, \theta)$ is given by

$$p(\boldsymbol{Y}|\boldsymbol{X}, \theta) = \prod_{i=1}^{n} \mathbb{P}[u_i = y_i|\theta, \boldsymbol{X}],$$

for any branching-structure realization $\boldsymbol{Y}$ (see also Fig. 3). Finally, we recover

$$\begin{aligned} \Delta(\theta) &= -\mathbb{E}_{\boldsymbol{Y}}[\ln p(\boldsymbol{Y}|\theta, \boldsymbol{X})] = -\mathbb{E}_{\boldsymbol{Y}}\left[\ln \prod_{i=1}^{n} \mathbb{P}[u_i = y_i|\theta, \boldsymbol{X}]\right] = -\mathbb{E}_{\boldsymbol{Y}}\left[\sum_{i=1}^{n} \ln \mathbb{P}[u_i = y_i|\theta, \boldsymbol{X}]\right] \\ &= -\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{Y}}[\ln \mathbb{P}[u_i = y_i|\theta, \boldsymbol{X}]] = -\sum_{i=1}^{n} \sum_{j=1}^{i} \mathbb{P}[u_i = j|\boldsymbol{X}, \theta] \ln \mathbb{P}[u_i = j|\theta, \boldsymbol{X}], \end{aligned}$$

which concludes the proof. ■

EM algorithm. Building on the proof of Theorem 1 (preserving the notation used there), the EM algorithm can be described as follows.

[11] The index n describes the total number of arrivals, i.e., $n = N(T)$.

[12] While not being a proper metric, the Kullback-Leibler divergence is non-negative (Gibbs' inequality), and it vanishes if and only if the two distributions in its argument coincide almost everywhere.

Initialize seed $\theta_0 \in \Theta$, fix a tolerance $\varepsilon \in (0, 1)$, and set $n \leftarrow 0$, $\delta \leftarrow 1$;
while $\delta > \varepsilon$ **do**
  E-step: Calculate $q_{n+1} \in \arg\max_q G(\theta_n, q) = \{p(\boldsymbol{Y}|\boldsymbol{X}, \theta_n)\}$ **do**;
  M-step: Find $\theta_{n+1} \in \arg\max_{\theta\in\Theta} G(\theta, q_{n+1})$;
  $n \leftarrow n + 1$ and $\delta \leftarrow G(\theta_{n+1}, q_{n+1}) - G(\theta_n, q_n)$;
end

The E step determines the next density q_{n+1} of the latent variable $\boldsymbol{Y}$ (i.e., the branching distribution), based on the current parameter estimate θ_n, by maximizing the adjusted (conditional) ECLL $G(\theta_n, \cdot)$; by Eq. (A3) the maximum is equal to the incomplete log likelihood and is achieved at $q_{n+1} = p(\cdot|\boldsymbol{X}, \theta_n)$. The M step then provides the next parameter estimate θ_{n+1} by maximizing the adjusted ECLL $G(\cdot, q_{n+1})$ on the compact set Θ.

APPENDIX B: NOTATION

Symbol	Description	Range
$a(t)$	Account treatment schedule	$\mathbb{R}_+$
$g(t, m)$	Triggering kernel	$\mathbb{R}_+$
$\mathcal{L}(\cdot)$	Incomplete log-likelihood function	$\mathbb{R}$
$\mathcal{L}_C(\cdot)$	Complete log-likelihood function	$\mathbb{R}$
m_i	Relative repayment at time $t = \tau_i$	$\mathbb{R}_+$
$\mathcal{H}_t$	Available information at time t	–
$J(t) = [N(t), R(t)]$	Repayment process	$\mathbb{N} \times \mathbb{R}_+$
$N(t)$	Repayment counting process	$\mathbb{N}$
$Q(\theta, \theta_n)$	Expected complete log-likelihood function	$\mathbb{R}$
$R(t)$	Cumulative relative-repayment process	$\mathbb{R}_+$
r_i	Relative repayment at time $t = \tau_i$	[0,1]
t	Current time	$\mathbb{R}_+$
T	Observation period	$\mathbb{R}_{++}$
δ_1	Sensitivity of intensity with respect to J	$\mathbb{R}_{++}^{\dim(J)}$
δ_2	Sensitivity of intensity with respect to a	$\mathbb{R}_{++}^{\mathbb{N}}$
κ	Mean-reversion parameter	$\mathbb{R}_{++}$
$\lambda(t)$	Intensity process	$\mathbb{R}_+$
λ_0	Initial value of intensity $[\lambda(0) = \lambda_0]$	$\mathbb{R}_{++}$
λ_∞	Long-run stationary value of intensity	$\mathbb{R}_+$
$\mu(t)$	Base-rate intensity	$\mathbb{R}$
θ	Vector of process parameters $[\theta = (\kappa, \lambda_0, \lambda_\infty, \delta_1, \delta_2)]$	Θ
ϑ_j	Time of jth account-treatment action ($j \geqslant 1$)	$\mathbb{R}_{++}$
τ_i	Arrival time of ith repayment ($i \geqslant 1$)	$\mathbb{R}_{++}$

[1] A. G. Hawkes, Spectra of some self-exciting and mutually exciting point processes, Biometrika **58**, 83 (1971).
[2] A. G. Hawkes, Point spectra of some mutually exciting point processes, J. R. Stat. Soc.: Ser. B **33**, 438 (1971).
[3] A. Hawkes and L. Adamopoulos, Cluster models for earthquakes-regional comparisons, Bull. Int. Stat. Inst. **45**, 454 (1973).
[4] E. Bacry, I. Mastromatteo, and J.-F. Muzy, Hawkes processes in finance, Mark. Microstruct. Liquidity **1**, 1550005 (2015).
[5] L. Xu, J. A. Duan, and A. B. Whinston, Path to purchase: A mutually exciting point process model for online advertising and conversion, Manage. Sci. **60**, 1392 (2014).
[6] W. Truccolo, From point process observations to collective neural dynamics: Nonlinear Hawkes process GLMs, low-dimensional dynamics and coarse graining, J. Physiol. Paris **110**, 336 (2016).
[7] N. Chehrazi and T. A. Weber, Dynamic valuation of delinquent credit-card accounts, Manage. Sci. **61**, 3077 (2015).
[8] M. Rambaldi, P. Pennesi, and F. Lillo, Modeling foreign exchange market activity around macroeconomic news: Hawkes-process approach, Phys. Rev. E **91**, 012819 (2015).
[9] N. Chehrazi, P. Glynn, and T. A. Weber, Dynamic credit-collections optimization, Manage. Sci. **65**, 2737 (2019).
[10] I. Rubin, Regular point processes and their detection, IEEE Trans. Inf. Theory **18**, 547 (1972).
[11] T. Ozaki, Maximum likelihood estimation of Hawkes' self-exciting point processes, Ann. Inst. Stat. Math. **31**, 145 (1979).

[12] Y. Ogata, The asymptotic behaviour of maximum likelihood estimators for stationary point processes, Ann. Inst. Stat. Math. **30**, 243 (1978).

[13] J. Hadamard, Sur les problèmes aux dérivées partielles et leur signification physique, *Princeton University Bulletin*, No. 23 (1902), pp. 49–52.

[14] A. Veen and F. P. Schoenberg, Estimation of space-time branching process models in seismology using an EM-type algorithm, J. Am. Stat. Assoc. **103**, 614 (2008).

[15] Y. Ogata, Space-time point-process models for earthquake occurrences, Ann. Inst. Stat. Math. **50**, 379 (1998).

[16] Y. Ogata, Statistical models for earthquake occurrences and residual analysis for point processes, J. Am. Stat. Assoc. **83**, 9 (1988).

[17] Y. Ogata, Space-time modeling of earthquake occurrences, Bull. Int. Stat. Inst. **55**, 249 (1993).

[18] A. G. Hawkes and D. Oakes, A cluster process representation of a self-exciting process, J. Appl. Prob. **11**, 493 (1974).

[19] J. Møller and J. G. Rasmussen, Approximate simulation of Hawkes processes, Method. Comput. Appl. Probab. **8**, 53 (2006).

[20] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes, Vol. 2: General Theory and Structure* (Springer, New York, 2007).

[21] Y. Ogata and H. Akaike, On linear intensity models for mixed doubly stochastic Poisson and self-exciting point processes, J. R. Stat. Soc.: Ser. B **44**, 102 (1982).

[22] I. Valera and M. Gomez-Rodriguez, Modeling adoption and usage of competing products, in *Proceedings of the 2015 IEEE International Conference on Data Mining (ICDM 2015), Atlantic City, NJ* (IEEE Computer Society, Washington, DC, 2015), pp. 409–418.

[23] K. Zhou, H. Zha, and L. Song, Learning triggering kernels for multi-dimensional Hawkes processes, in *Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA* (PLMR, 2013), pp. 1301–1309.

[24] X. Guo, A. Hu, R. Xu, and J. Zhang, Consistency and computation of regularized mles for multivariate Hawkes processes, arXiv:1810.02955.

[25] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, New York, 2009).

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc.: Ser. B **39**, 1 (1977).

[27] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. (McGraw-Hill, New York, 1976).

[28] I. M. Gelfand and S. V. Fomin, *Calculus of Variations* (Prentice-Hall, Englewood Cliffs, NJ, 1963).

[29] T. P. Minka, Expectation-maximization as lower bound maximization, Working Paper, Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, 1998.