A generalization error bound for sparse and low-rank multivariate Hawkes processes

Emmanuel Bacry[1], Stéphane Gaïffas[1], and Jean-Franccois Muzy[1,2]

[1]Centre de Mathématiques Appliquées, École Polytechnique and CNRS
UMR 7641, 91128 Palaiseau, France
[2]Laboratoire Sciences Pour l'Environnement, CNRS, Université de Corse,
UMR 6134, 20250 Corté, France

January 4, 2015

Abstract

We consider the problem of unveiling the implicit network structure of user interactions in a social network, based only on high-frequency timestamps. Our inference is based on the minimization of the least-squares loss associated with a multivariate Hawkes model, penalized by ℓ_1 and trace norms. We provide a first theoretical analysis of the generalization error for this problem, that includes sparsity and low-rank inducing priors. This result involves a new data-driven concentration inequality for matrix martingales in continuous time with observable variance, which is a result of independent interest. A consequence of our analysis is the construction of sharply tuned ℓ_1 and trace-norm penalizations, that leads to a data-driven scaling of the variability of information available for each users. Numerical experiments illustrate the strong improvements achieved by the use of such data-driven penalizations.

1 Introduction

Understanding the dynamics of social interactions is a challenging problem of fastly growing interest [11, 20, 9, 21] because of the large number of applications in web-advertisement and e-commerce, where large-scale logs of event history are available. A common supervised approach consists in the prediction of labels based on declared interactions (friendship, like, follower, etc.) However such supervision is not always available, and it does not always describe accurately the level of interactions between users. Labels are often only binary while a quantification of the interaction is more interesting, declared interactions are often deprecated, and more generally a supervised approach is not enough to infer the latent communities of users, as temporal patterns of actions of users are much more informative.

A recent set of papers [32, 14, 10] consider an approach for recovering latent social groups directly based on the real *actions* or *events* of users, called also nodes in the following, that uses only the timestamps patterns of the considered events. The models assume a structure of data consisting in a sequence of independent cascades, containing timestamps for each nodes. In these works, techniques coming from survival analysis are used to derive a tractable convex likelihood, that allows to infer the latent community structure. However, this model requires that data is already segmented into sets of independent cascades, which is not always realistic. Moreover, it does not allow for recurrent events, namely a node can be infected only once, and it cannot incorporate exogeneous factors, namely influence from the world outside the network.

Another approach is based on self-exciting point processes, such as the Hawkes process [16]. Previously used for geophysics, [28], high-frequency finance [1], crime activity [26], this model has been also recently used for the modelization of users activity in social networks, see for instance [9, 6, 36, 35]. The main point is that the structure of the Hawkes model allows to capture the direct influence of a user's action to the others, based on the recurrency and the patterns of actions timestamps. It encompasses in the same likelihood the decay of the influence over time, the levels of interaction between nodes, which can be seen as a weighted asymmetrical adjacency matrix, and a baseline intensity, that measures the level of exogeny of a user, namely the spontaneous apparition of an action, with no influence from other nodes of the network.

In this paper, we consider a multivariate Hawkes process (MHP), and we combine convex proxies for sparsity and low-rank of the adjacency matrix and the baseline intensities, that are now of common use in low-rank modeling in collaborative filtering problems [7, 8]. Note that this approach is also considered in [36]. We provide a first theoretical analysis of the generalization error for this problem, see [15] for an analysis including only entrywise ℓ_1 penalization. Namely, we prove a sharp oracle inequality for our procedure, that includes sparsity and low-rank inducing priors, see Theorem 1 in Section 4. This result involves a new data-driven concentration inequality for matrix martingales in continuous time, see Theorem 3 in Section 5, which is a result of independent interest, that extends previous non-commutative versions of concentration inequalities for martingales in discrete time, see [33]. A consequence of our analysis is the construction of sharply tuned ℓ_1 and trace-norm penalizations, that leads to a data-driven scaling of the variability of information available for each nodes. We give empirical evidence of the improvements of our data-driven penalizations, by conducting in Section 6 numerical experiments on simulated data. Since the objectives involved are convex with a smooth component, our algorithms build upon standard accelerated batch gradient proximal algorithms.

2 The multivatriate Hawkes model

Consider a finite network with d nodes (each node corresponding to a user in a social network for instance). For each node $j \in \{1, \ldots, d\}$, we observe the timestamps $\{t_{j,1}, t_{j,2}, \ldots\}$ of actions of node j on the network (a message, a click, etc.). To each node j is associated a counting process $N_j(t) = \sum_{i \geq 1} \mathbf{1}_{t_{j,i} \leq t}$ and we consider the d-dimensional counting process $N_t = [N_1(t) \cdots N_d(t)]^\top \in \mathbb{N}^d$, for $t \geq 0$. We observe this process for $t \in [0, T]$. Each N_j has an intensity λ_j, meaning that

$$\mathbb{P}\big(N_j \text{ has a jump in } [t, t+dt] \mid \mathcal{F}_t\big) = \lambda_j(t) dt, \quad j = 1, \ldots, d,$$

where $\mathcal{F}_t$ is the σ-field generated by N up to time t. The multivariate Hawkes model assumes that each N_j has an intensity $\lambda_{j,\theta}$ given by

$$\lambda_{j,\theta}(t) = \mu_j + \sum_{j'=1}^{d} a_{j,j'} \int_{(0,t)} h_{j,j'}(t-s) dN_{j'}(s),$$

where the integral is a Stieljes integral, namely

$$\int_{(0,t)} h_{j,j'}(t-s) dN_{j'}(s) = \sum_{i : t_{j',i} \in [0,t)} h_{j,j'}(t - t_{j',i}),$$

where $\theta = (\mu, \boldsymbol{A})$ with $\mu = [\mu_1, \ldots, \mu_d]^\top$ and $\boldsymbol{A} = [a_{j,j'}]_{1 \leq j,j' \leq d}$, with $\mu_j \geq 0$ which is the baseline intensity of j, where $a_{j,j'} \geq 0$ is a coefficient that quantifies the influence of j' on j, and

$h_{j,j'} : \mathbb{R}^+ \to \mathbb{R}^+$ are decay functions that account for the decay of influence between pairs of nodes in the network. A typical choice for $h_{j,j'}$ is the exponential kernel, i.e., $h_{j,j'}(t) = e^{-\alpha_{j,j'} t}$, where $\alpha_{j,j'} > 0$ is a decay coefficient. We consider these functions fixed and known in this paper. The parameter of interest is the *self-excitement* matrix $\boldsymbol{A}$, which can be viewed as a weighted asymmetrical adjacency matrix of connectivity between nodes.

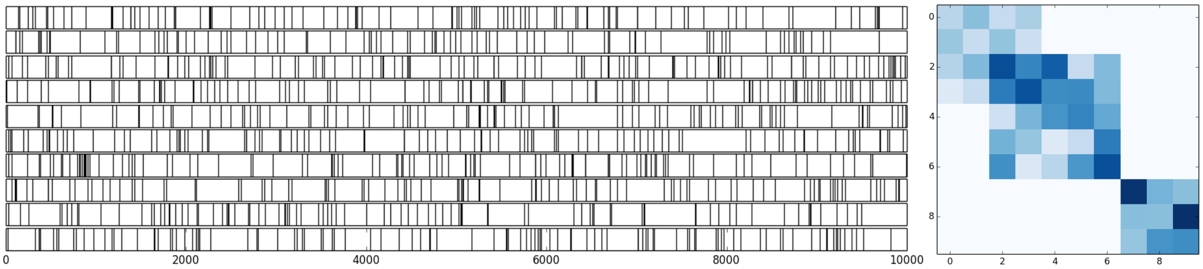

Figure 1: Toy example with $d = 10$ nodes. Based on actions' timestamps of the nodes, represented by vertical bars (left figure), we aim at recovering the matrix $\boldsymbol{A}$ of implicit influence between nodes (right figure).

The Hawkes model is particularly revelant for the modelization of the "microscopic" activity of social networks, and has been considered recently a lot in literature (see [9, 6, 36, 35, 23, 12, 6, 17], among others) for this kind of application, with a particular emphasis on [15] that gives first theoretical results for the Lasso used with Hawkes processes with an application to neurobiology. The main point is that this simple autoregressive structure of the intensity allows to capture the direct influence of a user on to the others, based on the recurrency and the patterns of their actions, by separating the intensity into a baseline and a self-exciting component, hence allowing to filter exogeneity in the estimation of users' influences on each others.

3 The procedure

We want to produce an estimation procedure of $\theta = (\mu, \boldsymbol{A})$ based on data from $\{N_t : t \in [0, T]\}$. The hidden structure underlying the observed actions of nodes will be contained in $\boldsymbol{A}$. A way of achieving this is to minimize the least-squares functional given by

$$R_T(\theta) = \|\lambda_\theta\|_T^2 - \frac{2}{T} \sum_{j=1}^{d} \int_{[0,T]} \lambda_{j,\theta}(t) dN_j(t) \tag{1}$$

with respect to θ, where $\|\lambda_\theta\|_T^2 = \frac{1}{T} \sum_{j=1}^d \int_{[0,T]} \lambda_{j,\theta}(t)^2 dt$ is the norm associated with the inner product

$$\langle \lambda_\theta, \lambda_{\theta'} \rangle_T = \frac{1}{T} \sum_{j=1}^{d} \int_{[0,T]} \lambda_{j,\theta}(t) \lambda_{j,\theta'}(t) dt. \tag{2}$$

This least-squares function is very natural, and comes from the empirical risk minimization principle [34, 25, 18, 3]: assuming that N has an unknown ground truth intensity λ (not necessarily following the Hawkes model), we have easily, using Doob-Meyer's decomposition that

$$\mathbb{E}[R_T(\theta)] = \mathbb{E}\|\lambda_\theta\|_T^2 - 2\mathbb{E}\langle \lambda_\theta, \lambda \rangle_T = \mathbb{E}\|\lambda_\theta - \lambda\|_T^2 - \|\lambda\|_T,$$

so that we expect a minimium $\hat{\theta}$ of $R_T(\theta)$ to lead to a good estimation $\lambda_{\hat{\theta}}$ of λ.

In addition to this goodness-of-fit criterion, we need to use a penalization that allows to reduce the dimensionality of the model. In particular, we want to reduce the dimensionality of

$\boldsymbol{A}$, based on the prior assumption that latent factors explain the connectivity of users in the network. This leads to a low-rank assumption on $\boldsymbol{A}$, which is commonly used in collaborative filtering and matrix completion techniques [30]. Our prior assumptions on μ and $\boldsymbol{A}$ are the following.

Sparsity of μ. Some nodes are basically inactive and react only if stimulated. Hence, we assume that the baseline vector μ is sparse.

Sparsity of $\boldsymbol{A}$. A user interacts only with a fraction of other nodes, meaning that for a fixed node j, only a few $a_{j,j'}$ are non-zero. Hence, we assume that $\boldsymbol{A}$ is a sparse matrix

Low-rank of $\boldsymbol{A}$. Nodes interactions have a community structure. It contains cliques, leading to a block-diagonal adjacency matrix that has the property of being sparse and low-rank.

To induce these prior assumptions on the parameters, we use a penalization based on a mixture of the ℓ_1 and trace norms. These norms are respectively the tightest convex relaxations for sparsity and low-rank, see for instance [7, 8]. They provide state-of-the art results in compressed sensing and collaborative filtering problems, among many other problems. These two norms have been previously combined for the estimation of sparse and low-rank matrices, see for instance [31] and [36] in the context of MHP. We consider indeed the following penalization on the parameter $\theta = (\mu, \boldsymbol{A})$:

$$\operatorname{pen}(\theta) = \|\mu\|_{1,\hat{w}} + \|\boldsymbol{A}\|_{1,\hat{\boldsymbol{W}}} + \hat{\tau}\|\boldsymbol{A}\|_{*}, \tag{3}$$

where each terms are weighted ℓ_1 and trace norm penalizations, given by

$$\|\mu\|_{1,\hat{w}} = \sum_{j=1}^{d} \hat{w}_j |\mu_j|, \quad \|\boldsymbol{A}\|_{1,\hat{\boldsymbol{W}}} = \sum_{1 \le j,k \le d} \hat{\boldsymbol{W}}_{j,k} |a_{j,k}|, \quad \|\boldsymbol{A}\|_{*} = \sum_{j=1}^{d} \sigma_j(\boldsymbol{A}),$$

where the $\sigma_1(\boldsymbol{A}) \geq \cdots \geq \sigma_d(\boldsymbol{A})$ are the singular values of $\boldsymbol{A}$. The weights $\hat{w}$, $\hat{\boldsymbol{W}}$, and the coefficient $\hat{\tau}$ are data-driven tuning parameters described below. The choice of these weights comes from a sharp analysis of the noise terms, see Section 5 below, and they lead to a data-driven scaling of the variability of information available for each nodes. The set of matrices $\boldsymbol{A}$ obtained by minimizing an objective penalized by (3) contains matrices that can be written in a block-diagonal or overlapping block-diagonal form, up to permutations of rows and columns. We consider then

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}_+^d \times \mathbb{R}_+^{d \times d}} \big\{ R_T(\theta) + \operatorname{pen}(\theta) \big\}, \tag{4}$$

which is a solution to the penalized least-squares problem.

Let us define now the data-driven weights $\hat{w}$, $\hat{\boldsymbol{W}}$ and $\hat{\tau}$ used in (3). From now on, we fix some confidence level $x > 0$, which corresponds to the probability that the oracle inequality from Theorem 1 holds. This can be safely chosen as $x = \log d$ for instance. The weights for ℓ_1-penalization of μ are given by

$$\hat{w}_j = 6\sqrt{2}\sqrt{\frac{(x + \log d + \hat{\ell}_{x,j}(T)) N_j([0,T])/T}{T}} + 27.93 \frac{x + \log d + \hat{\ell}_{x,j}(T)}{T} \tag{5}$$

where $N_j([0,T]) = \int_0^T dN_j(t)$ and $\hat{\ell}_{x,j}(T) = 2 \log\log\big(\frac{6N_j([0,T]) + 56x}{112x} \vee e\big)$. The weighting of each coordinate j in the penalization of μ is natural: it is roughly proportional to the square-root of $N_j([0,T])/T$, which is the average intensity of events on coordinate j. The term $\hat{\ell}_{x,j}(T)$ is a technical term, that can be neglected in practice, see Section 6. The data-driven weights for ℓ_1-penalization of $\boldsymbol{A}$ are given by

$$\hat{\boldsymbol{W}}_{j,k} = 4\sqrt{2}\sqrt{\frac{(x + 2\log d + \hat{\boldsymbol{L}}_{x,j,k}(T))\hat{\boldsymbol{V}}_{j,k}(T)}{T}} + 18.62\frac{(x + 2\log d + \hat{\boldsymbol{L}}_{x,j,k}(T))\boldsymbol{B}_{j,k}(T)}{T} \tag{6}$$

where

$$\begin{aligned} \boldsymbol{B}_{j,k}(t) = \sup_{s\in[0,t]} \boldsymbol{H}_{j,k}(s), \qquad \boldsymbol{H}_{j,k}(t) = \int_{(0,t)} h_{j,k}(t-s)dN_k(s), \\ \hat{\boldsymbol{V}}_{j,k}(t) = \frac{1}{t}\int_0^t \Big(\int_{(0,s)} h_{j,k}(s-u)dN_k(u)\Big)^2 dN_j(s) \end{aligned} \tag{7}$$

and where

$$\hat{\boldsymbol{L}}_{x,j,k}(t) = 2\log\log\Big(\frac{6t\hat{\boldsymbol{V}}_{j,k}(t) + 56x\boldsymbol{B}_{j,k}(t)^2}{112x\boldsymbol{B}_{j,k}(t)^2} \vee e\Big). \tag{8}$$

Once again, this is natural: the variance term $\hat{\boldsymbol{V}}_{j,k}(t)$ in (7) is, roughly, an estimation of the variance of the self-excitements between coordinates j and k. The term $\hat{\boldsymbol{L}}_{x,j,k}(T)$ is a technical term that can be neglected in practice.

The coefficient $\hat{\tau}$ comes from a new concentration inequality for matrix-martingales in continuous time, see Theorem 3 in Section 5 below. We consider indeed

$$\begin{aligned} \hat{\tau} = 8\sqrt{\frac{(x + \log d + \hat{\ell}_x(T))(\|\hat{\boldsymbol{V}}_1(T)\|_{\mathrm{op}} \vee \|\hat{\boldsymbol{V}}_2(T)\|_{\mathrm{op}})}{T}} \\ + \frac{2(x + \log d + \hat{\ell}_x(T))(10.34 + 2.65\sup_{t\in[0,T]}\|\boldsymbol{H}(t)\|_{2,\infty})}{T}, \end{aligned} \tag{9}$$

where $\|\cdot\|_{\mathrm{op}}$ stands for the operator norm, namely the largest singular value, where $\boldsymbol{H}(t)$ is the matrix with entries $\boldsymbol{H}_{j,k}(t)$ given in (7), where $\hat{\boldsymbol{V}}_1(t)$ is the diagonal matrix with entries

$$(\hat{\boldsymbol{V}}_1(t))_{j,j} = \frac{1}{t}\int_0^t \|\boldsymbol{H}(s)\|_{2,\infty}^2 dN_j(s), \tag{10}$$

and where $\hat{\boldsymbol{V}}_2(t)$ is the matrix with entries

$$(\hat{\boldsymbol{V}}_2(t))_{j,k} = \frac{1}{t}\int_0^t \|\boldsymbol{H}(s)\|_{2,\infty}^2 \sum_{l=1}^d \frac{\boldsymbol{H}_{j,l}(s)\boldsymbol{H}_{k,l}(s)}{\|\boldsymbol{H}_{l,\bullet}(s)\|_2^2} dN_l(s), \tag{11}$$

where $\|\cdot\|_2$ is the ℓ_2-norm, $\|\boldsymbol{X}\|_{2,\infty}$ is the maximum ℓ_2 norm of the rows of $\boldsymbol{X}$, and where $\boldsymbol{H}_{l,\bullet}$ is the l-th row of $\boldsymbol{H}$. The extra technical term $\hat{\ell}_x(t)$ is given by

$$\begin{aligned} \hat{\ell}_x(t) = 2\log\log\Big(\frac{2\|\hat{\boldsymbol{V}}_1(t)\|_{\mathrm{op}} + 2(4 + \sup_{s\in[0,t]}\|\boldsymbol{H}(s)\|_{2,\infty}^2/3)x}{x} \vee e\Big) \\ + 2\log\log\Big(\frac{2\|\hat{\boldsymbol{V}}_2(t)\|_{\mathrm{op}} + 2(4 + \sup_{s\in[0,t]}\|\boldsymbol{H}(s)\|_{2,\infty}^2/3)x}{x} \vee e\Big) \\ + 2\log\log\Big(\sup_{s\in[0,t]}\|\boldsymbol{H}(s)\|_{2,\infty}^2 \vee e\Big). \end{aligned} \tag{12}$$

These weights are actually quite natural: the terms $\hat{\boldsymbol{V}}_{j,k}(t)$, $\|\hat{\boldsymbol{V}}_1(t)\|_{\mathrm{op}}$ and $\|\hat{\boldsymbol{V}}_2(t)\|_{\mathrm{op}}$ correspond to estimations of the noise variance, that are the L^2 terms appearing in the empirical Bernstein's inequalities given in Section 5 below. This will allow for a sharp tuning of the penalizations. The terms $\boldsymbol{B}_{j,k}(t)$ and $\sup_{s\in[0,t]}\|\boldsymbol{H}(s)\|_{2,\infty}$ correspond to the L^∞ terms from these Bernstein's inequalities.

4 A sharp oracle inequality

Recall that the inner product $\langle \lambda_1, \lambda_2 \rangle_T$ is given by (2) and recall that $\| \cdot \|_T$ stands for the corresponding norm. Theorem 1 is a sharp oracle inequality on the prediction error measured by $\|\lambda_{\hat\theta} - \lambda\|_T^2$. For the proof of oracle inequalities with a fast rate, one needs a restricted eigenvalue condition on the Gram matrix of the problem [5, 18]. One of the weakest assumptions considered in literature is the Restricted Eigenvalue (RE) condition. In our setting, a natural RE assumption is given in Definition 1 below. We denote by $\| \cdot \|_F$ the Frobenius norm. If $\boldsymbol{X} = \boldsymbol{U\Sigma V}^\top$ is the SVD of $\boldsymbol{X}$, with the columns u_j of $\boldsymbol{U}$ and v_k of $\boldsymbol{V}$ being, respectively, the orthonormal left and right singular vectors of $\boldsymbol{X}$, the projection matrix onto the space spanned by the columns (resp. rows) of $\boldsymbol{X}$ is given by $\boldsymbol{P_U} = \boldsymbol{UU}^\top$ (resp. $\boldsymbol{P_V} = \boldsymbol{VV}^\top$). The operator $\mathcal{P}_{\boldsymbol{X}} : \mathbb{R}^{d \times d} \to \mathbb{R}^{d\times d}$ given by $\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{Y}) = \boldsymbol{P_U Y} + \boldsymbol{Y P_V} - \boldsymbol{P_U Y P_V}$ is the projector onto the linear space spanned by the matrices $u_k x^\top$ and $y v_k^\top$ for $1 \le j, k \le d$ and $x, y \in \mathbb{R}^d$. The projector onto the orthogonal space is given by $\mathcal{P}_{\boldsymbol{X}}^\perp(\boldsymbol{Y}) = (\boldsymbol{I} - \boldsymbol{P_U})\boldsymbol{Y}(\boldsymbol{I} - \boldsymbol{P_V})$. If x is a vector then $\operatorname{supp}(x)$ stands for the support of x (indices of non-zero entries) and for another vector x' the notation $(x')_{\operatorname{supp}(x)}$ stands for the vector with same coordinates as x' where we put 0 at indices outside of $\operatorname{supp}(x)$. We use the same notation $(\boldsymbol{X}')_{\operatorname{supp}(\boldsymbol{X})}$ for matrices $\boldsymbol{X}'$ and $\boldsymbol{X}$. We also use the notation $a \vee b = \max(a, b)$.

Definition 1. *Fix $\theta = (\mu, \boldsymbol{A})$ where $\mu \in \mathbb{R}^d$ and $\boldsymbol{A} \in \mathbb{R}_+^{d\times d}$. We define the constant $\kappa(\theta)$ such that, for any $\theta' = (\mu', \boldsymbol{A}')$ satisfying*

$$\|(\mu')_{\operatorname{supp}(\mu)^\perp}\|_{1,\hat w} \le 5 \|(\mu')_{\operatorname{supp}(\mu)}\|_{1,\hat w},$$

and

$$\|(\boldsymbol{A}')_{\operatorname{supp}(\boldsymbol{A})^\perp}\|_{1,\hat{\boldsymbol{W}}} + \hat\tau \|\mathcal{P}_{\boldsymbol{A}}^\perp(\boldsymbol{A}')\|_* \le 3 \|(\boldsymbol{A}')_{\operatorname{supp}(\boldsymbol{A})}\|_{1,\hat{\boldsymbol{W}}} + 3\hat\tau \|\mathcal{P}_{\boldsymbol{A}}(\boldsymbol{A}')\|_*,$$

we have

$$\|(\mu')_{\operatorname{supp}(\mu)}\|_2 \vee \|(\boldsymbol{A}')_{\operatorname{supp}(\boldsymbol{A})}\|_F \vee \|\mathcal{P}_{\boldsymbol{A}}(\boldsymbol{A}')\|_F \le \kappa(\theta) \|\lambda_{\theta'}\|_T. \tag{13}$$

The constant $1/\kappa(\theta)$ is a restricted eigenvalue depending on the "support" of θ, which is naturally associated with the problem considered here. Roughly, it requires that for any parameter θ' that has a support close to the one of θ (measured by domination of the ℓ_1 norms outside the support of θ by the ℓ_1 norm inside it), we have that the L^2 norm of the intensity given by $\|\lambda_{\theta'}\|_T$ can be compared with the L^2 norm of θ' in the support of θ.

Remark 1. *Under some conditions on the possible set of values for the decay functions $h_{j,j'}$, one can prove that a stronger condition than the one considered here holds with a large probability, see Proposition 4 from [15]. This result is based on a careful analysis of the ergodicity properties of MHP.*

Theorem 1. *Fix $x > 0$, and let $\hat\theta$ be given by (4), with tuning parameters given by (5), (6) and (9). Then, the inequality*

$$\|\lambda_{\hat\theta} - \lambda\|_T^2 \le \inf_\theta \left\{ \|\lambda_\theta - \lambda\|_T^2 + \kappa(\theta)^2 \Big(\frac{5}{4} \|(\hat w)_{\operatorname{supp}(\mu)}\|_2^2 + \frac{9}{8} \|(\hat{\boldsymbol{W}})_{\operatorname{supp}(\boldsymbol{A})}\|_F^2 + \frac{9}{8} \hat\tau^2 \operatorname{rank}(\boldsymbol{A}) \Big) \right\} \tag{14}$$

holds with a probability larger than $1 - 146e^{-x}$.

Note that no assumption is required on the ground truth intensity λ of the multivariate counting process N in Theorem 1. The proof of Theorem 1 is given in Section 8.2. Let us observe that

$$\|(w)_{\mathrm{supp}(\mu)}\|_2^2 \le \|\mu\|_0 \max_{j\in\mathrm{supp}(\mu)} \Big\{ c_1 \frac{(x+\log d+\hat\ell_{x,j}(T))N_j([0,T])/T}{T} + c_2\Big(\frac{x+\log d+\hat\ell_{x,j}(T)}{T}\Big)^2\Big\},$$

where $\|\mu\|_0$ stands for the sparsity of μ, that

$$\|(\boldsymbol{W})_{\mathrm{supp}(\boldsymbol{A})}\|_F^2 \le \|\boldsymbol{A}\|_0 \max_{(j,k)\in\mathrm{supp}(\boldsymbol{A})} \Big\{ c_1 \frac{(x+2\log d+\hat{\boldsymbol{L}}_{x,j,k}(T))\hat{\boldsymbol{V}}_{j,k}(T)}{T} + c_2\Big(\frac{(x+2\log d+\hat{\boldsymbol{L}}_{x,j,k}(T))\boldsymbol{B}_{j,k}(T)}{T}\Big)^2\Big\},$$

where $\|\boldsymbol{A}\|_0$ stands for the sparsity of $\boldsymbol{A}$, and finally that

$$\hat\tau^2 \le c_1 \frac{(x+\log d+\hat\ell_x(T))\|\hat{\boldsymbol{V}}_1(T)\|_{\mathrm{op}} \vee \|\hat{\boldsymbol{V}}_2(T)\|_{\mathrm{op}}}{T} + c_2\Big(\frac{(x+\log d+\hat\ell_x(T))(10.34+2.65\sup_{t\in[0,T]}\|\boldsymbol{H}(t)\|_{2,\infty})}{T}\Big)^2,$$

where $c_1, c_2 > 0$ are numerical constants. Hence, Theorem 1 proves that $\hat\theta$ achieves an optimal tradeoff between approximation and complexity, where the complexity is, roughly, measured by

$$\frac{\|\mu\|_0(x+\log d)}{T}\max_j N_j([0,T])/T + \frac{\|\boldsymbol{A}\|_0(x+2\log d)}{T}\max_{j,k}\hat{\boldsymbol{V}}_{j,k}(T) + \frac{\mathrm{rank}(\boldsymbol{A})(x+\log d)}{T}\|\hat{\boldsymbol{V}}_1(T)\|_{\mathrm{op}} \vee \|\hat{\boldsymbol{V}}_2(T)\|_{\mathrm{op}}.$$

This complexity term depends on both the sparsity and the rank of $\boldsymbol{A}$. The rate of convergence has the "expected" shape $(\log d)/T$, recalling that T is the length of the observation interval of the process, and these terms are balanced by the empirical variance terms coming out of the new concentration results given below.

5 Data-driven matrix martingale Bernstein's inequalities

The proof of Theorem 1 requires a sharp control of the noise terms. Since we analyze both ℓ_1 and trace-norm penalizations, we need control of this noise term for both the entrywise ℓ_∞ norm and operator norm $\|\cdot\|_{\mathrm{op}}$. The concentration inequalities described below are of independent interest. The noise term is the matrix martingale $\boldsymbol{Z}(t)$ with entries

$$\boldsymbol{Z}_{j,k}(t) = \int_0^t \int_{(0,s)} h_{j,k}(s-u)dN_k(u)dM_j(s), \tag{15}$$

where $M_j(t) = N_j(t) - \int_0^t \lambda_j(s)ds$ are the martingales obtained by compensation of the Hawkes process. A concentration inequality for $\boldsymbol{Z}_{j,k}$ is easily obtained from Bernstein's inequality [24], leading, for any $x > 0$, to

$$\frac{1}{t}(\boldsymbol{Z}(t))_{j,k} \le \sqrt{\frac{2vx}{t}} + \frac{bx}{3t}$$

with a probability larger than $1-e^{-x}$ whenever

$$\frac{1}{t}\langle \boldsymbol{Z}_{j,k}\rangle_t = \frac{1}{t}\int_0^t \Big(\int_{(0,s)} h_{j,k}(s-u)dN_k(u)\Big)^2 \lambda_j(s)ds \leq v$$

and

$$\sup_{s\in[0,t]} \int_{(0,s)} h_{j,k}(s-u)dN_k(u) \leq b.$$

A proof of this fact is implicit in the proof of Theorem 2 below. However, the predictable variation $\langle \boldsymbol{Z}_{j,k}\rangle_t$ depends on the non-observed intensity λ_j, so this inequality in present form is of no use for statistical learning. Morever, this result requires to know an upper bound on $\langle \boldsymbol{Z}_{j,k}\rangle_t$, while we would like an inequality that holds in general.

Hence, we need a new Bernstein's type inequality, that uses an observable empirical variance term, based on the optional variation, instead of the predictable variation. The optional variation is given by $\hat{\boldsymbol{V}}_{j,k}$, see Equation (7) above, and is undersood as an estimation of $\langle \boldsymbol{Z}_{j,k}\rangle_t$. Let us consider also $\boldsymbol{B}_{j,k}(t)$ given by (7) and $\hat{\boldsymbol{L}}_{x,j,k}(t)$ given by (8). The next theorem gives a deviation bound on all the entries of $\boldsymbol{Z}(t)$.

Theorem 2. *We have*

$$\frac{1}{t}\boldsymbol{Z}_{j,k}(t) \leq 2\sqrt{2}\sqrt{\frac{(x+2\log d+\hat{\boldsymbol{L}}_{j,k}(t))\hat{\boldsymbol{V}}_{j,k}(t)}{t}} + 9.31\frac{(x+2\log d+\hat{\boldsymbol{L}}_{j,k}(t))\boldsymbol{B}_{j,k}(t)}{t}$$

for any $1 \leq j,k \leq d$, with a probability larger than $1-30.55e^{-x}$.

The proof of Theorem 2 is given in Section 8.3 below, and has the same flavor as previous inequalities, see [13, 15].

Theorem 3 below gives a non-commutative version of Bernstein's inequality for the noise term, namely a deviation for $\|\boldsymbol{Z}(t)\|_{\mathrm{op}}$. It is based on a concentration inequality by the same authors [2], but it gives a bound with an observable variance term. We consider $\boldsymbol{H}(t)$ given by (7), $\hat{\boldsymbol{V}}_1(t)$ by (10), $\hat{\boldsymbol{V}}_2(t)$ by (11) and $\hat{\ell}_x(t)$ by (12).

Theorem 3. *For any $x>0$, we have*

$$\begin{aligned}\frac{\|\boldsymbol{Z}(t)\|_{\mathrm{op}}}{t} \leq 4\sqrt{\frac{(x+\log d+\hat{\ell}_x(t))\|\hat{\boldsymbol{V}}_1(t)\|_{\mathrm{op}} \vee \|\hat{\boldsymbol{V}}_2(t)\|_{\mathrm{op}}}{t}} \\ + \frac{(x+\log d+\hat{\ell}_x(t))(10.34+2.65\sup_{t\in[0,T]}\|\boldsymbol{H}(t)\|_{2,\infty})}{t}\end{aligned}$$

with a probability larger than $1-84.9e^{-x}$.

The proof of Theorem 3 is given in Section 8.4. This result of independent interest gives a control of the operator norm of the noise term, with an observable variance term. This is the first result of this kind to be found in literature, with [2] that gives a first Bernstein inequality for this kind of probabilistic object.

Once again, let us stress the fact that in both Theorems 2 and 3, all the quantities controlling the noise terms are *observable*, and are used for a sharp data-driven tuning of the penalizations considered in Section 3.

6 Numerical experiments

In this section we conduct experiments on synthetic datasets to evaluate the performance of our method, based on the proposed data-driven weighting of the penalizations, compared to non-weighted penalizations [36]. We generate Hawkes processes using Ogata's thinning algorithm [27], with $d = 100$, baselines μ sampled uniformly in $[0, 0.1]$, $h_{j,k}(t) = e^{-\alpha t}$ with $\alpha = 1$ and an adjacency matrix containing square overlapping boxes (corresponding to overlapping communities) at indexes 1:20, 10:50, 35:56 and 65:100. Each box is filled with uniformly sampled values in $[0, 0.2]$ and the rest of the matrix contains zeros. The matrix is then scaled to have operator norm equal to 0.8, therefore guaranteeing to obtain a stationary process. An instance of this matrix is given on the left side of Figure 2. We then compute several procedures on the generated data, restricting them on a growing interval of length $1000, 2000, 3000, 4000, 5000$, and assessing their performance each time. An overall averaging of the results is done on 10 separate simulations. Note that in this setting, the average number of events on a length 1000 interval is on average equal to 10000, and is, by stationarity, linearly growing with the length of the interval. We consider a procedure based on mimization of the log-likelihood instead of the least-squares used above to derive the theoretical results. This allows to reduce greatly computation times, as the computation of a gradient can be done in parallel and is linear in the number of events and dimension, thanks to recursion formulas that can be used for exponential decays, see [28]. It can be seen that the data-driven weights used in our penalizations are the same when using the log-likelihood loss instead of the least-squares, as the noise term remains the same. This objective is convex, with a goodness-of-fit term locally gradient-lipschitz: we use first-order optimization algorithm, based on proximal accelerated gradient. Namely, we use Fista [4] for problems with a single penalization on $\boldsymbol{A}$ (ℓ_1-norm or trace-norm) and Prisma [29] for mixed ℓ_1 and trace-norm penalizations on $\boldsymbol{A}$. For both procedures we use a linesearch scheme that allows to tune automatically the gradient step at each iteration. We fix a maximum of 100 iterations in all the results given below, for a fair comparison, we observed that it is largely sufficient for convergence to a satisfactory minimum. Note that in [36] an ADMM algorithm with Jensen's maximization minimization principle is used, which is not accelerated, while our algorithms are. We compare the following procedures:

- NoPen : direct minimization of the log-likelihood, with no penalization
- L1: non-weighted L1 penalization of μ and $\boldsymbol{A}$
- wL1: weighted L1 penalization of μ and $\boldsymbol{A}$ given by (16)
- L1Nuclear: non-weighted L1 penalization of μ and $\boldsymbol{A}$, and trace-norm penalization of $\boldsymbol{A}$
- wL1Nuclear: weighted L1 penalization of μ and $\boldsymbol{A}$ given by (16) and trace-norm penalization of $\boldsymbol{A}$

Note that the procedure L1Nuclear is the same as the one considered in [36], however we use a different optimization algorithm, based on an accelerated first-order method (that we expect to be faster than an ADMM based algorithm, although a careful comparison of solvers is beyond the scope of this paper). The data-driven weights used in our procedures are the ones derived from our analysis, see (5) and (6), where we remove negligible terms and where we put $x = \log T$. Namely, we use

$$\hat{w}_j = c_1 \sqrt{\frac{(\log T + \log d) N_j([0, T])/T}{T}} \quad \text{and} \quad \hat{\boldsymbol{W}}_{j,k} = c_2 \sqrt{\frac{(\log T + \log d) \hat{\boldsymbol{V}}_{j,k}(T)}{T}} \tag{16}$$

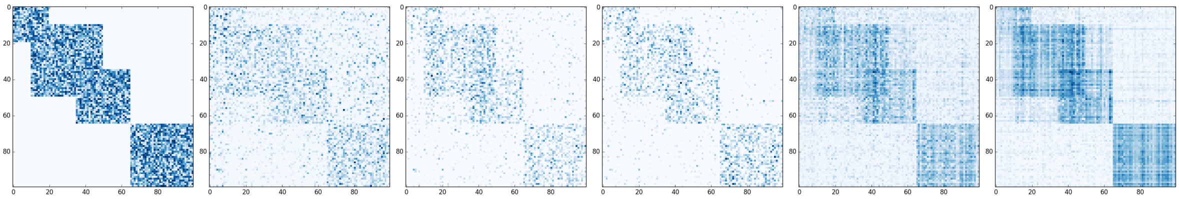

Figure 2: Ground truth matrix $\boldsymbol{A}$; recovered matrix using NoPen ; L1 ; wL1 ; L1Nuclear ; wL1Nuclear. We observe that wL1 and wL1Nuclear leads to better support recovery, as we observe less false positives outside of the node communities.

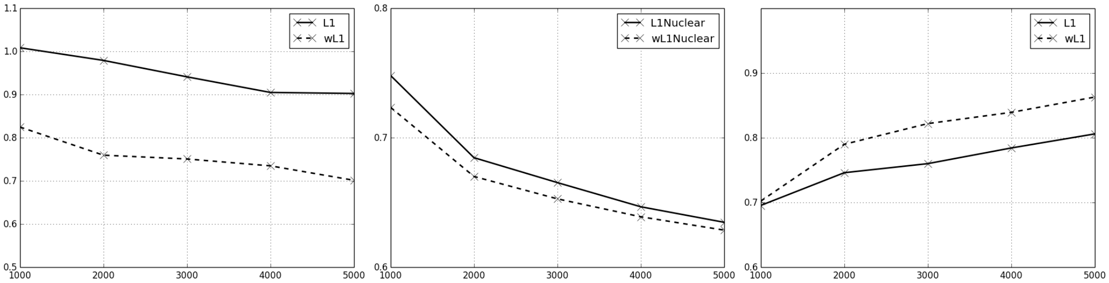

Figure 3: Error for L1 and wL1 ; Error for L1Nuclear and wL1Nuclear ; AUC for L1 and wL1. Abscissa corresponds to the interval length T. Weighted penalizations systematically leads to an improvement, both for L1 and L1 + Nuclear penalization, in terms of error and AUC

for weighted ℓ_1 penalization of μ and weighted ℓ_1 penalization of $\boldsymbol{A}$ respectively. The tuning parameters c_1, c_2 and the parameter for trace-norm penalization of $\boldsymbol{A}$ are tuned using cross-validation, on a testing error measured by the log-likelihood computed on a held-out testing set (we split in half the generated data for training and testing). We use two metrics to assess the procedures:

- error: the relative ℓ_2 estimation error of the parameter θ, given by $\|\hat{\theta} - \theta\|_2^2 / \|\theta\|_2^2$
- AUC: we compute the AUC (area under the roc curve) between the binarized ground truth matrix $\boldsymbol{A}$ and the solution $\hat{\boldsymbol{A}}$ with entries scaled in $[0, 1]$. This allows to quantify the ability of the procedure to detect the support of the connectivity structure between nodes.

In Figures 2 and 3, we compare the procedures in terms of error and AUC. In Figure 2 we can observe, on an instance of the problem, the improvement of wL1 and wL1Nuclear with respect to L1 and L1Nuclear respectively, as we observe less false positives outside the node communities (better viewed on a computer). Figure 3 confirms the fact that weighted penalizations systematically leads to an improvement, both for L1 and L1Nuclear, in terms of error and AUC.

7 Conclusion

In this paper we proposed a careful analysis of the generalization error of a MHP-based modelization of user interactions in a social network. Our theoretical analysis required a new concentration inequality for matrix-martingales in continuous time, with an observable variance term, that is a result of independent interest. This analysis led to a new data-driven tuning of sparsity-inducing penalizations, that we assess on a numerical example. Further work will focus on other matrix factorization techniques for this problem, such as non-negative matrix factorization, and the use of text-mining techniques to incorporate content features for twitter datasets for instance.

Acknowledgements

This research benefited from the support of the “Chair Markets in Transition”, under the aegis of “Louis Bachelier Finance and Sustainable Growth” laboratory, a joint initiative of École Polytechnique, Université d’Évry Val d’Essonne and Fédération Bancaire Francaise.

8 Proofs

8.1 Notations

Denote by $\boldsymbol{X}$ a $d \times d$ matrix and $x \in \mathbb{R}^d$. Then $\operatorname{diag}[x]$ stands for the diagonal matrix with diagonal equal to x, while $\operatorname{diag}[\boldsymbol{X}]$ stands for the diagonal matrix with diagonal equal to the one of $\boldsymbol{X}$. We write the singular value decomposition (SVD) of a rank r matrix $\boldsymbol{X}$ as

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \sum_{j=1}^{r} \sigma_j u_j v_j^\top$$

where $\boldsymbol{\Sigma} = \operatorname{diag}[\sigma(\boldsymbol{X})]$ with $\sigma(\boldsymbol{X}) = [\sigma_1, \ldots, \sigma_r]^\top$ the vector of singular values $\sigma_1 \geq \cdots \geq \sigma_r$ of $\boldsymbol{X}$ and where $\boldsymbol{U} = [u_1 \cdots u_r]$ and $\boldsymbol{V} = [v_1 \cdots v_r]$ are $n \times r$ matrices with columns given by the left and right singular vectors of $\boldsymbol{X}$. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $d \times d$, we denote by $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \operatorname{tr}(\boldsymbol{X}^\top \boldsymbol{Y})$ the Euclidean matrix product, and $\|\boldsymbol{X}\|_F = \sqrt{\langle \boldsymbol{X}, \boldsymbol{X} \rangle}$ the Frobenius norm. We introduce the operator norm $\|\boldsymbol{X}\|_{\mathrm{op}} = \sigma_1(\boldsymbol{X})$ and trace norm $\|\boldsymbol{X}\|_* = \sum_{j=1}^{d} \sigma_j(\boldsymbol{X})$. If $\boldsymbol{W}$ is a $d \times d$ matrix with positive entries, we introduce the weighted entrywise ℓ_1-norm given by $\|\boldsymbol{X}\|_{1,\boldsymbol{W}} = \langle \boldsymbol{W}, |\boldsymbol{X}| \rangle$, where $|\boldsymbol{X}|$ contains the absolute values of the entries of $\boldsymbol{X}$. We denote by $\|\boldsymbol{X}\|_0$ the number of non-zero entries of $\boldsymbol{X}$ and $\boldsymbol{X} \odot \boldsymbol{Y}$ is the entrywise product (Hadamard product) of $\boldsymbol{X}$ and $\boldsymbol{Y}$ with matching dimensions. We use the same notation $x \odot y$ for vectors x and y with matching dimensions. We denote also $\boldsymbol{X}_{\bullet,j}$ for the j-th column of $\boldsymbol{X}$ while $\boldsymbol{X}_{j,\bullet}$ stands for the j-th row. We define

$$\|\boldsymbol{X}\|_{2,\infty} = \max_j \|\boldsymbol{X}_{j,\bullet}\|_2 \quad \text{and} \quad \|\boldsymbol{X}\|_{\infty,2} = \max_j \|\boldsymbol{X}_{\bullet,j}\|_2,$$

where $\| \cdot \|_2$ is the ℓ_2 norm of vectors. If $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is the SVD of $\boldsymbol{X}$ the projection matrix onto the space spanned by the columns (resp. rows) of $\boldsymbol{X}$ is given by $\boldsymbol{P}_{\boldsymbol{U}} = \boldsymbol{U}\boldsymbol{U}^\top$ (resp. $\boldsymbol{P}_{\boldsymbol{V}} = \boldsymbol{V}\boldsymbol{V}^\top$). The operator $\mathcal{P}_{\boldsymbol{X}} : \mathbb{R}^{d \times d} \to \mathbb{R}^{d \times d}$ given by $\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{Y}) = \boldsymbol{P}_{\boldsymbol{U}}\boldsymbol{Y} + \boldsymbol{Y}\boldsymbol{P}_{\boldsymbol{V}} - \boldsymbol{P}_{\boldsymbol{U}}\boldsymbol{Y}\boldsymbol{P}_{\boldsymbol{V}}$ is the projector onto the linear space spanned by the matrices $u_k x^\top$ and $y v_k^\top$ for $1 \leq j, k \leq d$ and $x, y \in \mathbb{R}^d$. The projector onto the orthogonal space is given by $\mathcal{P}_{\boldsymbol{X}}^{\perp}(\boldsymbol{Y}) = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{U}})\boldsymbol{Y}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{V}})$. If x is a vector (or matrix) then $\operatorname{supp}(x)$ stands for the support of x (indices of non-zero entries) and for another vector x' the notation $[x']_{\operatorname{supp}(x)}$ stands the vector with same coordinates as x' where we put 0 at indices outside of $\operatorname{supp}(x)$. We use the same notation $[\boldsymbol{X}']_{\operatorname{supp}(\boldsymbol{X})}$ for matrices $\boldsymbol{X}'$ and $\boldsymbol{X}$. We also use the notation $a \vee b = \max(a, b)$.

8.2 Proof of Theorem 1

The proof is based on the proof of a sharp oracle inequality for trace norm penalization, see [19] and [18]. We endow the space $\mathbb{R}^d \times \mathbb{R}^{d \times d}$ by the inner product $\langle \theta, \theta' \rangle = \langle \mu, \mu' \rangle + \langle \boldsymbol{A}, \boldsymbol{A}' \rangle$ where $\theta = (\mu, \boldsymbol{A})$ and $\theta' = (\mu', \boldsymbol{A}')$ with $\langle \mu, \mu' \rangle = \mu^\top \mu'$ and $\langle \boldsymbol{A}, \boldsymbol{A}' \rangle = \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{A}')$.

For any θ, one has

$$\langle \nabla R_T(\hat{\theta}), \hat{\theta} - \theta \rangle = 2 \sum_{1 \le j \le d} (\hat{\mu}_j - \mu_j) \frac{\partial R_T(\hat{\theta})}{\partial \hat{\mu}_j} + \sum_{1 \le j,k \le d} (\hat{a}_{jk} - a_{j,k}) \frac{\partial R_T(\hat{\theta})}{\partial \hat{a}_{j,k}}.$$

Since

$$\frac{\partial \lambda_{j,\theta}(t)}{\partial \mu_j} = 1 \quad \text{and} \quad \frac{\partial \lambda_{j,\theta}(t)}{\partial a_{j,k}} = \int_{(0,t)} h_{j,k}(t-s) dN_k(s),$$

we have that the derivatives of the empirical risk are given by

$$\frac{\partial R_T(\hat{\theta})}{\partial \mu_j} = \frac{2}{T} \Big(\int_0^T \lambda_{j,\theta}(t) dt - \int_0^T dN_j(t) \Big)$$

and

$$\begin{aligned} \frac{\partial R_T(\hat{\theta})}{\partial a_{j,k}} = \frac{2}{T} \Big(& \int_0^T \int_{(0,t)} h_{j,k}(t-s) dN_k(s) \lambda_{j,\theta}(t) dt \\ & - \int_0^T \int_{(0,t)} h_{j,k}(t-s) dN_k(s) dN_j(t) \Big). \end{aligned}$$

Now, it leads to

$$\begin{aligned} \langle \nabla R_T(\hat{\theta}), \hat{\theta} - \theta \rangle &= \frac{2}{T} \sum_{j=1}^d \int_0^T (\lambda_{j,\hat{\theta}}(t) - dN_j(t))(\hat{\mu}_j - \mu_j) \\ &\quad + \frac{2}{T} \sum_{1 \le j,k \le d} \int_0^T \int_{(0,t)} h_{j,k}(t-s) dN_k(s) (\lambda_{j,\hat{\theta}}(t) - dN_j(t))(\hat{a}_{j,k} - a_{j,k}) \\ &= \frac{2}{T} \sum_{j=1}^d \int_0^T (\lambda_{j,\hat{\theta}}(t) - \lambda_{j,\theta}(t))(\lambda_{j,\hat{\theta}}(t) dt - dN_j(t)). \end{aligned}$$

Using $dM_j(t) = dN_j(t) - \lambda_j(t) dt$ and the recalling that

$$\langle f, g \rangle_T = \frac{1}{T} \sum_{1 \le j \le d} \int_{[0,T]} f_j(t) g_j(t) dt,$$

we obtain the decomposition

$$\langle \nabla R_T(\hat{\theta}), \hat{\theta} - \theta \rangle = 2 \langle \lambda_{\hat{\theta}} - \lambda_\theta, \lambda_{\hat{\theta}} - \lambda \rangle_T - \frac{2}{T} \sum_{j=1}^d \int_0^T (\lambda_{j,\hat{\theta}}(t) - \lambda_{j,\theta}(t)) dM_j(t).$$

Namely, we end up with

$$2 \langle \lambda_{\hat{\theta}} - \lambda_\theta, \lambda_{\hat{\theta}} - \lambda \rangle_T = \langle \nabla R_T(\hat{\theta}), \hat{\theta} - \theta \rangle + \frac{2}{T} \sum_{j=1}^d \int_0^T (\lambda_{j,\hat{\theta}}(t) - \lambda_{j,\theta}(t)) dM_j(t). \tag{17}$$

The parallelogram identity gives

$$2 \langle \lambda_{\hat{\theta}} - \lambda_\theta, \lambda_{\hat{\theta}} - \lambda \rangle_T = \| \lambda_{\hat{\theta}} - \lambda \|_T^2 + \| \lambda_{\hat{\theta}} - \lambda_\theta \|_T^2 - \| \lambda_\theta - \lambda \|_T^2,$$

where we put $\|f\|_T^2 = \langle f, f \rangle_T$.

Let us point out that, in the case $\langle \lambda_{\hat\theta} - \lambda_\theta, \lambda_{\hat\theta} - \lambda \rangle_T < 0$, one obtains

$$\|\lambda_{\hat\theta} - \lambda\|_T^2 \le \|\lambda_\theta - \lambda\|_T^2,$$

which directly implies the inequality of the Theorem.

Thus, from now on, let us assume that

$$\langle \lambda_{\hat\theta} - \lambda_\theta, \lambda_{\hat\theta} - \lambda \rangle_T \ge 0. \tag{18}$$

The first order condition for $\hat\theta \in \operatorname{argmin}_\theta \{R_T(\theta) + \operatorname{pen}(\theta)\}$ gives

$$-\nabla R_T(\hat\theta) \in \partial \operatorname{pen}(\hat\theta).$$

Let $\hat\theta_\partial = -\nabla R_T(\hat\theta)$. Since the subdifferential is a monotone mapping, we have $\langle \hat\theta - \theta, \hat\theta_\partial - \theta_\partial \rangle \ge 0$ for any $\theta_\partial \in \partial \operatorname{pen}(\theta)$. Thus from (17), one gets $\forall \theta_\partial \in \partial \operatorname{pen}(\theta)$,

$$2\langle \lambda_{\hat\theta} - \lambda_\theta, \lambda_{\hat\theta} - \lambda \rangle_T \le -\langle \theta_\partial, \hat\theta - \theta \rangle + \frac{2}{T} \sum_{j=1}^d \int_0^T (\lambda_{j,\hat\theta}(t) - \lambda_{j,\theta}(t)) dM_j(t). \tag{19}$$

We need now to characterize the structure of the subdifferentials involved in $\operatorname{pen}(\theta)$, to describe θ_∂.

If $g_1(\mu) = \sum_{j=1}^d \hat w_j |\mu_j|$, for $\hat w_j \ge 0$, we have

$$\partial g_1(\mu) = \Big\{ \hat w \odot \operatorname{sign}(\mu) + \hat w \odot f : \|f\|_\infty \le 1, \mu \odot f = 0 \Big\}. \tag{20}$$

If $g_2(\boldsymbol{A}) = \sum_{1 \le j,k \le d} \hat{\boldsymbol{W}}_{j,k} |\boldsymbol{A}_{j,k}|$, for $\hat{\boldsymbol{W}}_{j,k} \ge 0$, we have

$$\partial g_2(\boldsymbol{A}) = \Big\{ \hat{\boldsymbol{W}} \odot \operatorname{sign}(\boldsymbol{A}) + \hat{\boldsymbol{W}} \odot \boldsymbol{F} : \|\boldsymbol{F}\|_\infty \le 1, \boldsymbol{A} \odot \boldsymbol{F} = \boldsymbol{0} \Big\}. \tag{21}$$

Let us recall that if $\boldsymbol{A} = \boldsymbol{U\Sigma V}^\top$ is the SVD of $\boldsymbol{A}$, we have $\mathcal{P}_{\boldsymbol{A}}(\boldsymbol{B}) = \boldsymbol{P_U B} + \boldsymbol{B P_V} - \boldsymbol{P_U B P_V}$ and $\mathcal{P}_{\boldsymbol{A}}^\perp(\boldsymbol{B}) = (\boldsymbol{I} - \boldsymbol{P_U})\boldsymbol{B}(\boldsymbol{I} - \boldsymbol{P_V})$ (projection onto the column and row space of $\boldsymbol{A}$ and projection onto its orthogonal space). Now, for $g_3(\boldsymbol{A}) = \hat\tau \|\boldsymbol{A}\|_*$, we have

$$\partial g_3(\boldsymbol{A}) = \Big\{ \hat\tau \boldsymbol{U V}^\top + \hat\tau \mathcal{P}_{\boldsymbol{A}}^\perp(\boldsymbol{F}) : \|\boldsymbol{F}\|_{\mathrm{op}} \le 1 \Big\}, \tag{22}$$

see for instance [22]. Now, write

$$-\langle \theta_\partial, \hat\theta - \theta \rangle = -\langle \mu_\partial, \hat\mu - \mu \rangle - \langle \boldsymbol{A}_{\partial,1}, \hat{\boldsymbol{A}} - \boldsymbol{A} \rangle - \langle \boldsymbol{A}_{\partial,*}, \hat{\boldsymbol{A}} - \boldsymbol{A} \rangle$$

with $\mu_\partial \in g_1(\mu)$, $\boldsymbol{A}_{\partial,1} \in g_2(\boldsymbol{A})$ and $\boldsymbol{A}_{\partial,*} \in g_3(\boldsymbol{A})$. Using Equation (20), (21) and (22), we can write

$$\begin{aligned} -\langle \theta_\partial, \hat\theta - \theta \rangle = &-\langle \hat w \odot \operatorname{sign}(\mu), \hat\mu - \mu \rangle - \langle \hat w \odot f, \hat\mu - \mu \rangle \\ &- \langle \hat{\boldsymbol{W}} \odot \operatorname{sign}(\boldsymbol{A}), \hat{\boldsymbol{A}} - \boldsymbol{A} \rangle - \langle \hat{\boldsymbol{W}} \odot \boldsymbol{F}_1, \hat{\boldsymbol{A}} - \boldsymbol{A} \rangle \\ &- \hat\tau \langle \boldsymbol{U V}^\top, \hat{\boldsymbol{A}} - \boldsymbol{A} \rangle - \hat\tau \langle \boldsymbol{F}_*, \mathcal{P}_{\boldsymbol{A}}^\perp(\hat{\boldsymbol{A}} - \boldsymbol{A}) \rangle, \end{aligned}$$

where by duality between the norms $\| \cdot \|_1$ and $\| \cdot \|_\infty$, and between $\| \cdot \|_*$ and $\| \cdot \|_{\mathrm{op}}$, we can choose $f, \boldsymbol{F}_1$ and $\boldsymbol{F}_*$ such that

$$\langle \hat w \odot f, \hat\mu - \mu \rangle = \|(\hat\mu - \mu)_{\operatorname{supp}(\mu)^\perp}\|_{1,\hat w}, \quad \langle \hat{\boldsymbol{W}} \odot \boldsymbol{F}_1, \hat{\boldsymbol{A}} - \boldsymbol{A} \rangle = \|(\hat{\boldsymbol{A}} - \boldsymbol{A})_{\operatorname{supp}(\boldsymbol{A})^\perp}\|_{1,\hat{\boldsymbol{W}}}$$

and

$$\langle \boldsymbol{F}_*, \mathcal{P}_{\boldsymbol{A}}^{\perp}(\hat{\boldsymbol{A}} - \boldsymbol{A}) \rangle = \| \mathcal{P}_{\boldsymbol{A}}^{\perp}(\hat{\boldsymbol{A}} - \boldsymbol{A}) \|_*,$$

which leads to

$$\begin{aligned} -\langle \theta_\partial, \hat{\theta} - \theta \rangle \leq \; & \| (\hat{\mu} - \mu)_{\mathrm{supp}(\mu)} \|_{1,\hat{w}} - \| (\hat{\mu} - \mu)_{\mathrm{supp}(\mu)^\perp} \|_{1,\hat{w}} \\ & + \| (\hat{\boldsymbol{A}} - \boldsymbol{A})_{\mathrm{supp}(\boldsymbol{A})} \|_{1,\hat{\boldsymbol{W}}} - \| (\hat{\boldsymbol{A}} - \boldsymbol{A})_{\mathrm{supp}(\boldsymbol{A})^\perp} \|_{1,\hat{\boldsymbol{W}}} \\ & + \hat{\tau} \| \mathcal{P}_{\boldsymbol{A}}(\hat{\boldsymbol{A}} - \boldsymbol{A}) \|_* - \hat{\tau} \| \mathcal{P}_{\boldsymbol{A}}^{\perp}(\hat{\boldsymbol{A}} - \boldsymbol{A}) \|_*. \end{aligned}$$

Now, decompose the noise term of (19) :

$$\begin{aligned} \frac{2}{T} \sum_{j=1}^d \int_0^T & (\lambda_{j,\hat{\theta}}(t) - \lambda_{j,\theta}(t)) dM_j(t) \\ & = \frac{2}{T} \sum_{j=1}^d (\hat{\mu}_j - \mu_j) \int_0^T dM_j(t) \\ & \quad + \frac{2}{T} \sum_{1 \leq j,k \leq d} (\hat{a}_{j,k} - a_{j,k}) \int_0^T \int_{(0,t)} h_{j,k}(t-s) dN_k(s) dM_j(t) \\ & = \frac{2}{T} \langle \hat{\mu} - \mu, M_T \rangle + \frac{2}{T} \langle \hat{\boldsymbol{A}} - \boldsymbol{A}, \boldsymbol{Z} \rangle, \end{aligned}$$

where

$$M_T = \Big[\int_0^T dM_1(t), \cdots, \int_0^T dM_d(t) \Big]^\top$$

and where we recall that $\boldsymbol{Z}$ is given by (15), see Section 5. We have

$$2 |\langle \hat{\mu} - \mu, M_T \rangle| \leq 2 \sum_{j=1}^d |\hat{\mu}_j - \mu_j| |M_j([0,T])|, \quad |\langle \hat{\boldsymbol{A}} - \boldsymbol{A}, \boldsymbol{Z} \rangle| \leq \sum_{1 \leq j,k \leq d} |\hat{\boldsymbol{A}}_{j,k} - \boldsymbol{A}_{j,k}| |\boldsymbol{Z}_{j,k}|$$

and

$$|\langle \hat{\boldsymbol{A}} - \boldsymbol{A}, \boldsymbol{Z} \rangle| \leq \| \boldsymbol{Z} \|_{\mathrm{op}} \| \hat{\boldsymbol{A}} - \boldsymbol{A} \|_*,$$

where we used again duality between trace norm and operator norm.

We need now to use the concentration inequalities given in Section 5, that are proved in Sections 8.3 and 8.4 below. Using Theorem 2 (see Section 8.3 below) with $h \equiv 1$ and an union bound on $j = 1, \ldots, d$ gives that

$$\frac{1}{T} |M_j([0,T])| \leq 2\sqrt{2} \sqrt{\frac{(x + \log d + \hat{\ell}_{x,j}(T)) N_j([0,T])/T}{T}} + 9.31 \frac{x + \log d + \hat{\ell}_{x,j}(T)}{T}$$

for any $1 \leq j \leq d$ with a probability larger than $1 - 30.55 e^{-x}$. Using Theorem 2 from Section 5 entails that

$$\begin{aligned} \frac{1}{T} |\boldsymbol{Z}_{j,k}| \leq 2\sqrt{2} & \sqrt{\frac{(x + 2\log d + \hat{\boldsymbol{L}}_{x,j,k}(T)) \hat{\boldsymbol{V}}_{j,k}(T)}{T}} \\ & + 9.31 \frac{(x + 2\log d + \hat{\boldsymbol{L}}_{x,j,k}(T)) \boldsymbol{B}_{j,k}(T)}{T} \end{aligned}$$

for any $1 \leq j, k \leq d$ with a probability larger than $1 - 30.55e^{-x}$ and finally Theorem 3 (see Section 5) entails that

$$\frac{\|\boldsymbol{Z}(T)\|_{\mathrm{op}}}{T} \leq 4\sqrt{\frac{(x + \log d + \hat{\ell}_x(T))\|\hat{\boldsymbol{V}}_1(T)\|_{\mathrm{op}} \vee \|\hat{\boldsymbol{V}}_2(T)\|_{\mathrm{op}}}{T}} \\ + \frac{(x + \log d + \hat{\ell}_x(T))(10.34 + 2.65 \sup_{t \in [0,T]} \|\boldsymbol{H}(t)\|_{2,\infty})}{T},$$

with a probability larger than $1 - 84.9e^{-x}$. Hence, the choice of weights (5), (6) and (9) entails

$$\frac{2}{T}|\langle \mu - \mu, M_T\rangle| \leq \frac{2}{3}\|\hat{\mu} - \mu\|_{1,\hat{w}}, \quad |\langle \hat{\boldsymbol{A}} - \boldsymbol{A}, \boldsymbol{Z}\rangle| \leq \frac{1}{2}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{1,\hat{\boldsymbol{W}}}$$

and

$$|\langle \hat{\boldsymbol{A}} - \boldsymbol{A}, \boldsymbol{Z}\rangle| \leq \frac{\hat{\tau}}{2}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_*$$

on an event with a probability larger than $1 - 146e^{-x}$. This entails

$$\begin{aligned} 0 \leq -\langle \theta_\partial, \hat{\theta} - \theta\rangle &+ \frac{2}{T}\sum_{j=1}^{d}\int_0^T (\lambda_{j,\hat{\theta}}(t) - \lambda_{j,\theta}(t))dM_j(t) \\ &\leq \frac{5}{3}\|(\hat{\mu} - \mu)_{\mathrm{supp}(\mu)}\|_{1,\hat{w}} - \frac{1}{3}\|(\hat{\mu} - \mu)_{\mathrm{supp}(\mu)^\perp}\|_{1,\hat{w}} \\ &\quad + \frac{3}{2}\|(\hat{\boldsymbol{A}} - \boldsymbol{A})_{\mathrm{supp}(\boldsymbol{A})}\|_{1,\hat{\boldsymbol{W}}} - \frac{1}{2}\|(\hat{\boldsymbol{A}} - \boldsymbol{A})_{\mathrm{supp}(\boldsymbol{A})^\perp}\|_{1,\hat{\boldsymbol{W}}} \\ &\quad + \frac{3}{2}\hat{\tau}\|\mathcal{P}_{\boldsymbol{A}}(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_* - \frac{1}{2}\hat{\tau}\|\mathcal{P}_{\boldsymbol{A}}^\perp(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_*. \end{aligned}$$

Taking $\boldsymbol{A} = \hat{\boldsymbol{A}}$ gives a cone constraint on $\hat{\mu} - \mu$:

$$\|(\hat{\mu} - \mu)_{\mathrm{supp}(\mu)^\perp}\|_{1,\hat{w}} \leq 5\|(\hat{\mu} - \mu)_{\mathrm{supp}(\mu)}\|_{1,\hat{w}},$$

while taking $\mu = \hat{\mu}$ gives a cone constraint on $\hat{\boldsymbol{A}} - \boldsymbol{A}$:

$$\begin{aligned} &\|(\hat{\boldsymbol{A}} - \boldsymbol{A})_{\mathrm{supp}(\boldsymbol{A})^\perp}\|_{1,\hat{\boldsymbol{W}}} + \hat{\tau}\|\mathcal{P}_{\boldsymbol{A}}^\perp(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_* \\ &\qquad \leq 3\|(\hat{\boldsymbol{A}} - \boldsymbol{A})_{\mathrm{supp}(\boldsymbol{A})}\|_{1,\hat{\boldsymbol{W}}} + 3\hat{\tau}\|\mathcal{P}_{\boldsymbol{A}}(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_*. \end{aligned}$$

Namely, we have now using Assumption 1 that

$$\|(\hat{\mu} - \mu)_{\mathrm{supp}(\mu)}\|_2 \vee \|(\hat{\boldsymbol{A}} - \boldsymbol{A})_{\mathrm{supp}(\boldsymbol{A})}\|_F \vee \|\mathcal{P}_{\boldsymbol{A}}(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_F \leq \kappa(\theta)\|\lambda_{\hat{\theta}} - \lambda_\theta\|_T. \tag{23}$$

Putting all this together gives

$$\begin{aligned} -\langle \theta_\partial, &\hat{\theta} - \theta\rangle + \frac{2}{T}\langle \hat{\mu} - \mu, M_T\rangle + \frac{2}{T}\langle \hat{\boldsymbol{A}} - \boldsymbol{A}, \boldsymbol{Z}\rangle \\ &\leq \frac{5}{3}\|(\hat{\mu} - \mu)_{\mathrm{supp}(\mu)}\|_{1,\hat{w}} - \frac{1}{3}\|(\hat{\mu} - \mu)_{\mathrm{supp}(\mu)^\perp}\|_{1,\hat{w}} \\ &\quad + \frac{3}{2}\|(\hat{\boldsymbol{A}} - \boldsymbol{A})_{\mathrm{supp}(\boldsymbol{A})}\|_{1,\hat{\boldsymbol{W}}} - \frac{1}{2}\|(\hat{\boldsymbol{A}} - \boldsymbol{A})_{\mathrm{supp}(\boldsymbol{A})^\perp}\|_{1,\hat{\boldsymbol{W}}} \\ &\quad + \frac{3}{2}\hat{\tau}\|\mathcal{P}_{\boldsymbol{A}}(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_* - \frac{1}{2}\hat{\tau}\|\mathcal{P}_{\boldsymbol{A}}^\perp(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_* \\ &\leq \frac{5}{3}\|(\hat{w})_{\mathrm{supp}(\mu)}\|_2\|(\hat{\mu} - \mu)_{\mathrm{supp}(\mu)}\|_2 + \frac{3}{2}\|(\hat{\boldsymbol{W}})_{\mathrm{supp}(\boldsymbol{A})}\|_F\|(\hat{\boldsymbol{A}} - \boldsymbol{A})_{\mathrm{supp}(\boldsymbol{A})}\|_F \\ &\quad + \frac{3}{2}\hat{\tau}\sqrt{\mathrm{rank}(\boldsymbol{A})}\|\mathcal{P}_{\boldsymbol{A}}(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_F, \end{aligned}$$

where we used Cauchy-Schwarz's inequality. This finally gives

$$\|\lambda_{\hat\theta} - \lambda\|_T^2 \le \|\lambda_\theta - \lambda\|_T^2 - \|\lambda_{\hat\theta} - \lambda_\theta\|_T^2 + \kappa(\theta)\Big(\frac{5}{3}\|(\hat w)_{\text{supp}(\mu)}\|_2 + \frac{3}{2}\|(\hat{\boldsymbol{W}})_{\text{supp}(\boldsymbol{A})}\|_F + \frac{3}{2}\hat\tau\sqrt{\text{rank}(\boldsymbol{A})}\Big)\|\lambda_{\hat\theta} - \lambda_\theta\|_T$$

where we used (23). The conclusion of the proof of Theorem 1 follows from the fact that $ax - x^2 \le a^2/4$ for any $a, x > 0$.

8.3 Proof of Theorem 2

We want to control all the entries of the matrix $\boldsymbol{Z}$ given by

$$\boldsymbol{Z}_{j,k}(t) = \int_0^t \int_{(0,s)} h_{j,k}(s-u)dN_k(u)dM_j(s).$$

We use the next Theorem, which is based on Theorem 3 from [13].

Theorem 4. *Let N be a counting process with predictable intensity λ and compensator Λ. Let g be a predictable function bounded a.s. Put $M = N - \Lambda$ the martingale obtained by compensation. Then, the martingale given by*

$$Z_t = \int_0^t g(s)dM(s)$$

satisfies

$$|Z(t)| \le 2\sqrt{2}\sqrt{(x + \hat\ell_x(t))[Z]_t} + 9.31(x + \hat\ell_x(t))\|g\|_\infty$$

with a probability larger than $1 - 30.55e^{-x}$, for $\|g\|_\infty = \sup_{s\in[0,t]}|g(s)|$ for the optional variation

$$[Z]_t = \int_0^t g(s)^2 dN(s),$$

and for

$$\hat\ell_x(t) = 2\log\log\Big(\frac{6t[Z]_t + 56x\|g\|_\infty^2}{112x\|g\|_\infty^2} \vee e\Big).$$

We fix $(j,k) \in \{1,\dots,d\}^2$ and choose $M = M_j$, $N = N_j$

$$g(t) = \int_{(0,t)} h_{j,k}(t-s)dN_k(s)$$

in Theorem 4, which leads to

$$\frac{1}{T}|\boldsymbol{Z}_{j,k}(T)| \le 2\sqrt{2}\sqrt{\frac{(x + \hat{\boldsymbol{L}}_{j,k}(T))\hat{\boldsymbol{V}}_{j,k}(T)}{T}} + 9.31\frac{(x + \hat{\boldsymbol{L}}_{j,k}(t))\boldsymbol{B}_{j,k}(T)}{T}$$

with a probability larger than $1 - 30.55e^{-x}$ for any j,k. Now, using an union bound over $(j,k) \in \{1,\dots,d\}^2$ with this inequality gives the same inequality with the same probability, where we increase x by $2\log d$. □

8.4 Proof of Theorem 3

8.4.1 Notations

Let us introduce

$$\mathscr{D}_{\boldsymbol{X}}^{(r)}(t) = \operatorname{diag}[\boldsymbol{X}(t)\boldsymbol{X}(t)^\top] \quad \text{and} \quad \mathscr{D}_{\boldsymbol{X}}^{(c)}(t) = \operatorname{diag}[\boldsymbol{X}(t)^\top \boldsymbol{X}(t)], \tag{24}$$

where we note that

$$\begin{aligned} &\mathscr{D}_{\boldsymbol{X}}^{(r)}(t) = \operatorname{diag}\left[\|\boldsymbol{X}_{1,\bullet}(t)\|_2^2, \cdots, \|\boldsymbol{X}_{d,\bullet}(t)\|_2^2\right] \quad \text{and} \\ &\mathscr{D}_{\boldsymbol{X}}^{(c)}(t) = \operatorname{diag}\left[\|\boldsymbol{X}_{\bullet,1}(t)\|_2^2, \cdots, \|\boldsymbol{X}_{\bullet,d}(t)\|_2^2\right]. \end{aligned} \tag{25}$$

8.4.2 Preliminary results

Let us introduce a process of the form

$$\boldsymbol{U}_{\boldsymbol{A},\boldsymbol{B}}(t) = \int_0^t \boldsymbol{A}_s \operatorname{diag}[dM_s]\boldsymbol{B}_s, \tag{26}$$

where $\{\boldsymbol{A}_t\}_{t\geq 0}$ and $\{\boldsymbol{B}_t\}_{t\geq 0}$ are arbitrary $(\mathcal{F}_t)$-predictable $d \times d$ processes, so that the entries of $\boldsymbol{U}_{\boldsymbol{A},\boldsymbol{B}}(t)$ are given by

$$(\boldsymbol{U}_{\boldsymbol{A},\boldsymbol{B}}(t))_{i,j} = \sum_{k=1}^d \int_0^t (\boldsymbol{A}_s)_{i,k}(\boldsymbol{B}_s)_{k,j}(dM_s)_k.$$

Recalling that $\boldsymbol{H}_t$ is the matrix with entries $\boldsymbol{H}_{j,k}(t) = \int_{(0,t)} h_{j,k}(t-s)dN_k(s)$, see (7) and that

$$\boldsymbol{Z}_{j,k}(t) = \int_0^t \int_{(0,s)} h_{j,k}(s-u)dN_k(u)dM_j(s),$$

we have

$$\boldsymbol{Z}_t = \int_0^t \operatorname{diag}[dM_s]\boldsymbol{H}_s = \boldsymbol{U}_{\boldsymbol{I},\boldsymbol{H}}(t).$$

Let us recall that we want to control $\|\boldsymbol{Z}(t)\|_{\mathrm{op}}$. The next Theorem is given in [2], see Theorem 4 herein, and is a core ingredient for the proof of Theorem 3.

Theorem 5. *Let $\boldsymbol{U}_{\boldsymbol{A},\boldsymbol{B}}(t)$ be given by (26) with $M_j(t) = N_j(t) - \int_0^t \lambda_j(s)ds$ that are martingales obtained by compensation of the counting processes N_j for $j = 1, \ldots, d$. Define the matrix*

$$\boldsymbol{V}_{\boldsymbol{A},\boldsymbol{B},\lambda}(t) = \int_0^t \|\boldsymbol{A}(s)\|_{\infty,2}^2 \|\boldsymbol{B}(s)\|_{2,\infty}^2 \boldsymbol{W}_{\boldsymbol{A},\boldsymbol{B},\lambda}(s)ds, \tag{27}$$

where

$$\boldsymbol{W}_{\boldsymbol{A},\boldsymbol{B},\lambda}(t) = \begin{bmatrix} \boldsymbol{A}_t \operatorname{diag}[\boldsymbol{A}_t^\top \boldsymbol{A}_t]^{-1} \operatorname{diag}[\lambda_t]\boldsymbol{A}_t^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}_t^\top \operatorname{diag}[\boldsymbol{B}_t\boldsymbol{B}_t^\top]^{-1} \operatorname{diag}[\lambda_t]\boldsymbol{B}_t \end{bmatrix}, \tag{28}$$

and introduce also

$$b_{\boldsymbol{A},\boldsymbol{B}}(t) = \sup_{s\in[0,t]} \|\boldsymbol{A}(s)\|_{\infty,2}\|\boldsymbol{B}(s)\|_{2,\infty}.$$

Then, for any $v, x > 0$, the following holds:

$$\mathbb{P}\left[\|\boldsymbol{U}_{\boldsymbol{A},\boldsymbol{B}}(t)\|_{\mathrm{op}} \geq \sqrt{2vx} + \frac{bx}{3}, \quad b_{\boldsymbol{A},\boldsymbol{B}}(t) \leq b, \quad \lambda_{\max}(\boldsymbol{V}_{\boldsymbol{A},\boldsymbol{B},\lambda}(t)) \leq v\right] \leq 2de^{-x}.$$

An immediate corollary of Theorem 5 is given in Corollary 1 below. For $0 \leq v_1 < v_2$ and $0 \leq b_1 < b_2$, we introduce the events

$$\mathcal{V}_{v_1,v_2} = \{v_1 < \|\boldsymbol{V}_{\boldsymbol{A},\boldsymbol{B},\lambda}(t)\|_{\mathrm{op}} \leq v_2\} \text{ and } \mathcal{B}_{b_1,b_2} = \{b_1 < b_{\boldsymbol{A},\boldsymbol{B}}(t) \leq b_2\}.$$

These events give lower and upper bounds of the random quantitives involved in this theorem. A peeling argument, given below, will allow to remove these events from the concentration, by slightly enlarging the concentration bound by a poly-logarithmic factor.

Corollary 1. *Fix any $\epsilon, b, v > 0$ and $x > 0$. The following deviation inequalities hold.*

$$\mathbb{P}\Big[\|\boldsymbol{U}_{\boldsymbol{A},\boldsymbol{B}}(t)\|_{\mathrm{op}} \geq \sqrt{2vx} + \frac{b}{3}x \cap \mathcal{V}_{0,v} \cap \mathcal{B}_{0,b}\Big] \leq 2de^{-x}, \tag{29}$$

$$\mathbb{P}\Big[\|\boldsymbol{U}_{\boldsymbol{A},\boldsymbol{B}}(t)\|_{\mathrm{op}} \geq \sqrt{2(1+\epsilon)\|\boldsymbol{V}_{\boldsymbol{A},\boldsymbol{B},\lambda}(t)\|_{\mathrm{op}}x} + \frac{b}{3}x \cap \mathcal{V}_{v,(1+\epsilon)v} \cap \mathcal{B}_{0,b}\Big] \leq 2de^{-x}, \tag{30}$$

$$\mathbb{P}\Big[\|\boldsymbol{U}_{\boldsymbol{A},\boldsymbol{B}}(t)\|_{\mathrm{op}} \geq \sqrt{2(1+\epsilon)vx} + \frac{b_{\boldsymbol{A},\boldsymbol{B}}(t)}{3}x \cap \mathcal{V}_{0,v} \cap \mathcal{B}_{b,(1+\epsilon)b}\Big] \leq 2de^{-x}, \tag{31}$$

$$\mathbb{P}\Big[\|\boldsymbol{U}_{\boldsymbol{A},\boldsymbol{B}}(t)\|_{\mathrm{op}} \geq (1+\epsilon)\sqrt{2\|\boldsymbol{V}_{\boldsymbol{A},\boldsymbol{B},\lambda}(t)\|_{\mathrm{op}}x} + \frac{b_{\boldsymbol{A},\boldsymbol{B}}(t)}{3}x \cap \mathcal{V}_{v,(1+\epsilon)v} \cap \mathcal{B}_{b,(1+\epsilon)b}\Big] \leq 2de^{-x}. \tag{32}$$

8.4.3 First concentration inequalities for $\|\boldsymbol{Z}_t\|_{\mathrm{op}}$

As explained above, if we choose $\boldsymbol{A} = \boldsymbol{I}$ and $\boldsymbol{B} = \boldsymbol{H}$, we have $\boldsymbol{U}_{\boldsymbol{A},\boldsymbol{B}}(t) = \boldsymbol{Z}_t$. For this choice, we have

$$b_{\boldsymbol{I},\boldsymbol{H}}(t) = b_{\boldsymbol{H}}(t) = \sup_{s \in [0,t]} \|\boldsymbol{H}(s)\|_{2,\infty}$$

and

$$\boldsymbol{V}_{\boldsymbol{I},\boldsymbol{H},\lambda}(t) = \begin{bmatrix} \boldsymbol{V}_{\boldsymbol{H},\lambda,1}(t) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{\boldsymbol{H},\lambda,2}(t) \end{bmatrix}$$

where we defined

$$\begin{aligned} \boldsymbol{V}_{\boldsymbol{H},\lambda,1}(t) &= \int_0^t \|\boldsymbol{H}(s)\|_{2,\infty}^2 \boldsymbol{D}_\lambda(s) ds \\ \boldsymbol{V}_{\boldsymbol{H},\lambda,2}(t) &= \int_0^t \|\boldsymbol{H}(s)\|_{2,\infty}^2 \boldsymbol{H}(s)^\top \mathscr{D}_{\boldsymbol{H}}^{(r)}(s)^{-1} \boldsymbol{D}_\lambda(s) \boldsymbol{H}(s) ds. \end{aligned}$$

Note that

$$\|\boldsymbol{V}_{\boldsymbol{I},\boldsymbol{H},\lambda}(t)\|_{\mathrm{op}} = \|\boldsymbol{V}_{\boldsymbol{H},\lambda,2}(t)\|_{\mathrm{op}} \vee \|\boldsymbol{V}_{\boldsymbol{H},\lambda,2}(t)\|_{\mathrm{op}}.$$

We will denote for short (t is fixed throughout)

$$Z = \|\boldsymbol{Z}_t\|_{\mathrm{op}}, \quad V_1 = \|\boldsymbol{V}_{\boldsymbol{H},\lambda,1}(t)\|_{\mathrm{op}}, \quad V_2 = \|\boldsymbol{V}_{\boldsymbol{H},\lambda,2}(t)\|_{\mathrm{op}}, \quad B = b_{\boldsymbol{H}}(t)$$

until the end of the proof. Introduce also, for $v_1, v_2, b_1, b_2 \geq 0$, the events

$$\mathcal{V}_{v_1,v_2}^{(1)} = \{v_1 < V_1 \leq v_2\}, \quad \mathcal{V}_{v_1,v_2}^{(2)} = \{v_1 < V_2 \leq v_2\}, \quad \mathcal{B}_{b_1,b_2} = \{b_1 < B \leq b_2\}.$$

Corollary 1 entails the following inequalities.

$$\mathbb{P}\Big[Z \geq \sqrt{2vx} + \frac{bx}{3} \cap \mathcal{V}^{(1)}_{0,v_1} \cap \mathcal{V}^{(2)}_{0,v_2} \cap \mathcal{B}_{0,b}\Big] \leq 2de^{-x} \tag{33}$$

$$\mathbb{P}\Big[Z \geq \sqrt{2(1+\epsilon)(V_1 \vee v_2)x} + \frac{bx}{3} \cap \mathcal{V}^{(1)}_{v_1,(1+\epsilon)v_1} \cap \mathcal{V}^{(2)}_{0,v_2} \cap \mathcal{B}_{0,b}\Big] \leq 2de^{-x}, \tag{34}$$

$$\mathbb{P}\Big[Z \geq \sqrt{2(1+\epsilon)(v_1 \vee V_2)x} + \frac{bx}{3} \cap \mathcal{V}^{(1)}_{0,v_1} \cap \mathcal{V}^{(2)}_{v_2,(1+\epsilon)v_2} \cap \mathcal{B}_{0,b}\Big] \leq 2de^{-x}, \tag{35}$$

$$\mathbb{P}\Big[Z \geq \sqrt{2(1+\epsilon)(V_1 \vee V_2)x} + \frac{bx}{3} \cap \mathcal{V}^{(1)}_{v_1,(1+\epsilon)v_1} \cap \mathcal{V}^{(2)}_{v_2,(1+\epsilon)v_2} \cap \mathcal{B}_{0,b}\Big] \leq 2de^{-x}, \tag{36}$$

$$\mathbb{P}\Big[Z \geq \sqrt{2(1+\epsilon)vx} + \frac{Bx}{3} \cap \mathcal{V}^{(1)}_{0,v_1} \cap \mathcal{V}^{(2)}_{0,v_2} \cap \mathcal{B}_{b,(1+\epsilon)b}\Big] \leq 2de^{-x}, \tag{37}$$

$$\mathbb{P}\Big[Z \geq (1+\epsilon)\sqrt{2(V_1 \vee v_2)x} + \frac{Bx}{3} \cap \mathcal{V}^{(1)}_{v_1,(1+\epsilon)v_1} \cap \mathcal{V}^{(2)}_{0,v_2} \cap \mathcal{B}_{b,(1+\epsilon)b}b\Big] \leq 2de^{-x}, \tag{38}$$

$$\mathbb{P}\Big[Z \geq (1+\epsilon)\sqrt{2(v_1 \vee V_2)x} + \frac{Bx}{3} \cap \mathcal{V}^{(1)}_{0,v_1} \cap \mathcal{V}^{(2)}_{v_2,(1+\epsilon)v_2} \cap \mathcal{B}_{b,(1+\epsilon)b}\Big] \leq 2de^{-x}, \tag{39}$$

$$\mathbb{P}\Big[Z \geq (1+\epsilon)\sqrt{2(V_1 \vee V_2)x} + \frac{Bx}{3} \cap \mathcal{V}^{(1)}_{v_1,(1+\epsilon)v_1} \cap \mathcal{V}^{(2)}_{v_2,(1+\epsilon)v_2} \cap \mathcal{B}_{b,(1+\epsilon)b}\Big] \leq 2de^{-x}. \tag{40}$$

These inequalities looks like the desired Bernstein's inequality. But they have two major problems. First, the variance terms V_1 and V_2 depend on λ, hence are non-observable. Second, we need to remove the events $\mathcal{V}^{(1)}_{\cdot,\cdot}$, $\mathcal{V}^{(2)}_{\cdot,\cdot}$ and $\mathcal{B}_{\cdot}$ to end-up with a usable inequality. Natural estimators of $\boldsymbol{V}_{\boldsymbol{H},\lambda,1}(t)$ and $\boldsymbol{V}_{\boldsymbol{H},\lambda,2}(t)$ are given, respectively, by

$$\hat{\boldsymbol{V}}_{\boldsymbol{H},1}(t) = \int_0^t \|\boldsymbol{H}(s)\|^2_{2,\infty} \operatorname{diag}[dN_s]$$

$$\hat{\boldsymbol{V}}_{\boldsymbol{H},2}(t) = \int_0^t \|\boldsymbol{H}(s)\|^2_{2,\infty} \boldsymbol{H}(s)^\top \mathscr{D}^{(r)}_{\boldsymbol{H}}(s)^{-1} \operatorname{diag}[dN_s] \boldsymbol{H}(s).$$

We introduce for short

$$\hat{V}_1 = \|\hat{\boldsymbol{V}}_{\boldsymbol{H},1}(t)\|_{\mathrm{op}}, \quad \hat{V}_2 = \|\hat{\boldsymbol{V}}_{\boldsymbol{H},2}(t)\|_{\mathrm{op}}.$$

The next step is to prove that we can replace the non-observable variance terms V_1 and V_2, that involve the quadratic variation, by the observable variance terms $\hat{V}_1$ and $\hat{V}_2$ involving the optional variation.

8.4.4 Replacing V_2 by $\hat{V}_2$

First, note that

$$\begin{aligned}\boldsymbol{V}_{\boldsymbol{H},\lambda,2}(t) &= \int_0^t \|\boldsymbol{H}(s)\|^2_{2,\infty} \boldsymbol{H}(s)^\top \mathscr{D}^{(r)}_{\boldsymbol{H}}(s)^{-1} \boldsymbol{D}_\lambda(s) \boldsymbol{H}(s) ds \\ &= \hat{\boldsymbol{V}}_{\boldsymbol{H},2}(t) - \int_0^t \|\boldsymbol{H}(s)\|^2_{2,\infty} \boldsymbol{H}(s)^\top \mathscr{D}^{(r)}_{\boldsymbol{H}}(s)^{-1} \operatorname{diag}[dM_s] \boldsymbol{H}(s) \\ &= \hat{\boldsymbol{V}}_{\boldsymbol{H},2}(t) - \boldsymbol{U}_{\boldsymbol{Q}^\top,\boldsymbol{Q}}(t),\end{aligned}$$

where

$$\boldsymbol{Q}(t) = \|\boldsymbol{H}(t)\|_{2,\infty} \mathscr{D}^{(r)}_{\boldsymbol{H}}(t)^{-1/2} \boldsymbol{H}(t).$$

Hence, we use again Proposition 1 with $\boldsymbol{A} = \boldsymbol{Q}^\top$ and $\boldsymbol{B} = \boldsymbol{Q}$. Note that

$$\begin{aligned}b_{\boldsymbol{Q}^\top,\boldsymbol{Q}}(t) &= \sup_{s\in[0,t]} \|\boldsymbol{Q}^\top(s)\|_{\infty,2} \|\boldsymbol{Q}(s)\|_{\infty,2} = \sup_{s\in[0,t]} \|\boldsymbol{Q}(s)\|^2_{\infty,2} \\ &= \sup_{s\in[0,t]} \|\boldsymbol{H}(s)\|^2_{2,\infty} \|\mathscr{D}^{(r)}_{\boldsymbol{H}}(s)^{-1/2} \boldsymbol{H}(s)\|^2_{2,\infty},\end{aligned}$$

but using (25) gives

$$\|\mathscr{D}_{\boldsymbol{H}}^{(r)}(t)^{-1/2}\boldsymbol{H}(t)\|_{2,\infty}^2 = \max_j \|\big(\mathscr{D}_{\boldsymbol{H}}^{(r)}(t)^{-1/2}\boldsymbol{H}(t)\big)_{j,\bullet}\|_2^2$$
$$= \max_j \|\|\boldsymbol{H}_{j,\bullet}(t)\|_2^{-1}\boldsymbol{H}_{j,\bullet}(t)\|_2^2 = 1,$$

so that

$$b_{\boldsymbol{Q}^\top,\boldsymbol{Q}}(t) = b_{\boldsymbol{H}}(t)^2 = \sup_{s\in[0,t]} \|\boldsymbol{H}(s)\|_{2,\infty}^2.$$

Moreover, note that

$$\mathscr{D}_{\boldsymbol{Q}^\top}^{(c)}(t) = \|\boldsymbol{H}(t)\|_{2,\infty}^2 \boldsymbol{I},$$

so that

$$\boldsymbol{Q}(t)^\top \mathscr{D}_{\boldsymbol{Q}^\top}^{(c)}(t)^{-1}\boldsymbol{D}_\lambda(t)\boldsymbol{Q}(t) = \boldsymbol{H}(t)^\top \mathscr{D}_{\boldsymbol{H}}^{(r)}(t)^{-1}\boldsymbol{D}_\lambda(t)\boldsymbol{H}(t).$$

Recalling (**??**) and (27), this means that

$$\boldsymbol{V}_{\boldsymbol{Q}^\top,\boldsymbol{Q},\lambda}(t) = \begin{bmatrix} \boldsymbol{V}_{\boldsymbol{H},\lambda,2}(t) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{\boldsymbol{H},\lambda,2}(t) \end{bmatrix}.$$

Hence, Proposition 1 gives, for the choice $\boldsymbol{A} = \boldsymbol{Q}^\top$ and $\boldsymbol{B} = \boldsymbol{Q}$

$$\mathbb{P}\Bigg[\|\boldsymbol{U}_{\boldsymbol{Q}^\top,\boldsymbol{Q}}(t)\|_{\mathrm{op}} \geq (1+\epsilon)\sqrt{2\|\boldsymbol{V}_{\boldsymbol{H},\lambda,2}(t)\|_{\mathrm{op}}x} + \frac{b_{\boldsymbol{H}}(t)^2}{3}x,$$
$$v \leq \|\boldsymbol{V}_{\boldsymbol{H},\lambda,2}(t)\|_{\mathrm{op}} < (1+\epsilon)v, \quad b^2 \leq b_{\boldsymbol{H}}(t)^2 \leq (1+\epsilon)b^2\Bigg] \leq 2de^{-x}.$$

We obtain that with a probability larger than $1 - 2de^{-x}$, we have

$$\|\boldsymbol{V}_{\boldsymbol{H},\lambda,2}(t) - \hat{\boldsymbol{V}}_{\boldsymbol{H},2}(t)\|_{\mathrm{op}} \leq (1+\epsilon)\sqrt{2\|\boldsymbol{V}_{\boldsymbol{H},\lambda,2}(t)\|_{\mathrm{op}}x} + \frac{b_{\boldsymbol{H}}(t)^2}{3}x$$

on the event $\{v \leq \|\boldsymbol{V}_{\boldsymbol{H},\lambda,2}(t)\|_{\mathrm{op}} < (1+\epsilon)v\} \cap \{b^2 \leq b_{\boldsymbol{H}}(t)^2 \leq (1+\epsilon)b^2\}$. So, using the so-called "square-root trick", namely the fact that $A \leq b + \sqrt{aA}$ entails $A \leq a + 2b$ for any $a, A, b > 0$, we obtain

$$V_2 \leq 2\hat{V}_2 + 2((1+\epsilon)^2 + b_{\boldsymbol{H}}(t)^2/3)x \tag{41}$$

and

$$\hat{V}_2 \leq 2V_2 + ((1+\epsilon)^2/2 + b_{\boldsymbol{H}}(t)^2/3)x \tag{42}$$

on this event, with a probability larger than $1 - 2de^{-x}$.

8.4.5 Replacing V_1 by $\hat{V}_1$

The exact same strategy as for V_2 is used. We write

$$\boldsymbol{V}_{\boldsymbol{H},\lambda,1}(t) = \hat{\boldsymbol{V}}_{\boldsymbol{H},1}(t) - \boldsymbol{U}_{\boldsymbol{Q}^\top,\boldsymbol{Q}}(t),$$

where this time $\boldsymbol{Q} = \|\boldsymbol{H}(t)\|_{2,\infty}\boldsymbol{I}$. We have

$$b_{\boldsymbol{Q}^\top,\boldsymbol{Q}}(t) = b_{\boldsymbol{H}}(t)^2 = \sup_{s\in[0,t]} \|\boldsymbol{H}(s)\|_{2,\infty}^2$$

and

$$\boldsymbol{V}_{\boldsymbol{Q}^\top,\boldsymbol{Q},\lambda}(t) = \begin{bmatrix} \boldsymbol{V}_{\boldsymbol{H},\lambda,1}(t) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{\boldsymbol{H},\lambda,1}(t) \end{bmatrix}.$$

Hence, we use again Proposition 1 with $\boldsymbol{A} = \boldsymbol{B} = \boldsymbol{Q}$, which gives

$$\|\boldsymbol{V}_{\boldsymbol{H},\lambda,1}(t) - \hat{\boldsymbol{V}}_{\boldsymbol{H},1}(t)\|_{\mathrm{op}} \leq (1+\epsilon)\sqrt{2\|\boldsymbol{V}_{\boldsymbol{H},\lambda,1}(t)\|_{\mathrm{op}} x} + \frac{b_{\boldsymbol{H}}(t)^2}{3} x$$

on the event $\{v \leq \|\boldsymbol{V}_{\boldsymbol{H},\lambda,1}(t)\|_{\mathrm{op}} < (1+\epsilon)v\} \cap \{b^2 \leq b_{\boldsymbol{H}}(t)^2 \leq (1+\epsilon)b^2\}$. Now, using the same trick as before, we obtain

$$V_1 \leq 2\hat{V}_1 + 2((1+\epsilon)^2 + b_{\boldsymbol{H}}(t)^2/3)x \tag{43}$$

and

$$\hat{V}_1 \leq 2V_1 + ((1+\epsilon)^2/2 + b_{\boldsymbol{H}}(t)^2/3)x \tag{44}$$

on this event, with a probability larger than $1 - 2de^{-x}$.

8.4.6 Concluding the proof

Plugging Equations (41) and (43) with Equations (33)–(40), we can replace V_1 and V_2 by $\hat{V}_1$ and $\hat{V}_2$ respectively. Now, we use a peeling argument to remove the events that lower and upper bound V_1, V_2 and B. Fix $\epsilon, c_0, x, b_0 > 0$ and introduce

$$v_{1,j} = v_{2,j} = c_0 x(1+\epsilon)^j \text{ and } b_j = b_0(1+\epsilon)^j$$

for $j \geq 0$, and take

$$v_{1,-1} = v_{2,-1} = b_{-1} = 0.$$

Put again for short $Z = \|\boldsymbol{Z}_t\|_{\mathrm{op}}$, $V_1 = \|\boldsymbol{V}_{\boldsymbol{H},\lambda,1}(t)\|_{\mathrm{op}}$, $V_2 = \|\boldsymbol{V}_{\boldsymbol{H},\lambda,2}(t)\|_{\mathrm{op}}$ and $B = b_{\boldsymbol{H}}(t)$, and introduce the events

$$\mathcal{V}_1 = \{V_1 > v_{1,0}\}, \quad \mathcal{V}_2 = \{V_2 > v_{2,0}\} \text{ and } \mathcal{B} = \{B > b_0\}$$

We partition the whole probability space in the following way:

$$(\mathcal{V}_1^{\complement} \cup \mathcal{V}_1) \cap (\mathcal{V}_2^{\complement} \cup \mathcal{V}_2) \cap (\mathcal{B}^{\complement} \cup \mathcal{B}) = \bigcup_{j,k,l \geq -1} \mathcal{V}_{1,j} \cap \mathcal{V}_{2,k} \cap \mathcal{B}_l,$$

where

$$\mathcal{V}_{1,j} = \{v_{1,j} < V_1 \leq v_{1,j+1}\}, \quad \mathcal{V}_{2,k} = \{v_{2,k} < V_2 \leq v_{2,k+1}\}, \quad \mathcal{B}_l = \{b_l < B \leq b_{l+1}\}.$$

On each event $\mathcal{V}_{1,j} \cap \mathcal{V}_{2,k} \cap \mathcal{B}_l$, we have a deviation on Z. Using (33) gives

$$\mathbb{P}\Big[Z \geq (\sqrt{2c_0} + \frac{b_0}{3})x \cap \mathcal{V}_{1,-1} \cap \mathcal{V}_{2,-1} \cap \mathcal{B}_{-1}\Big] \leq 2de^{-x}. \tag{45}$$

Using (34) with (43), together with the fact that on this event, $V_2 \leq v_{2,0} = v_{1,0} \leq v_{1,j} \leq V_1$, gives

$$\mathbb{P}\Big[Z \geq \sqrt{2c_{1,\epsilon}\hat{V}_1 x} + (c_{2,\epsilon} + c_{3,\epsilon}b_0)x \cap \mathcal{V}_{1,j} \cap \mathcal{V}_{2,-1} \cap \mathcal{B}_{-1}\Big] \leq 4de^{-x} \tag{46}$$

for any $j \geq 0$, where we introduced the constants

$$c_{1,\epsilon} = 2(1+\epsilon), \; c_{2,\epsilon} = 2(1+\epsilon)^{3/2}, \; c_{3,\epsilon} = 2\sqrt{\frac{1+\epsilon}{3}} + \frac{1}{3}.$$

Using (35) with (41), together with the fact that on this event, $V_1 \le v_{1,0} = v_{2,0} \le v_{2,k} \le V_2$, gives

$$\mathbb{P}\Big[Z \ge \sqrt{2c_{1,\epsilon}\hat{V}_2 x} + (c_{2,\epsilon} + c_{3,\epsilon}b_0)x \cap \mathcal{V}_{1,-1} \cap \mathcal{V}_{2,k} \cap \mathcal{B}_{-1}\Big] \le 4de^{-x} \tag{47}$$

for any $k \ge 0$. Using (36) with (41) and (43) gives

$$\mathbb{P}\Big[Z \ge \sqrt{2c_{1,\epsilon}(\hat{V}_1 \vee \hat{V}_2)x} + (c_{2,\epsilon} + c_{3,\epsilon}b_0)x \cap \mathcal{V}_{1,j} \cap \mathcal{V}_{2,k} \cap \mathcal{B}_{-1}\Big] \le 4de^{-x} \tag{48}$$

for any $j, k \ge 0$. Using (37) gives

$$\mathbb{P}\Big[Z \ge (\sqrt{2(1+\epsilon)c_0} + \frac{B}{3})x \cap \mathcal{V}_{1,-1} \cap \mathcal{V}_{2,-1} \cap \mathcal{B}_l\Big] \le 2de^{-x} \tag{49}$$

for any $l \ge 0$. Using (38) with (43), together with the fact that on this event, $V_2 \le v_{2,0} = v_{1,0} \le v_{1,j} \le V_1$, gives

$$\mathbb{P}\Big[Z \ge c_{1,\epsilon}\sqrt{\hat{V}_1 x} + (c_{4,\epsilon} + c_{5,\epsilon}B)x \cap \mathcal{V}_{1,j} \cap \mathcal{V}_{2,-1} \cap \mathcal{B}_l\Big] \le 4de^{-x} \tag{50}$$

for any $j, l \ge 0$, where we introduced the constants

$$c_{4,\epsilon} = 2(1+\epsilon)^2, \quad c_{5,\epsilon} = \frac{2(1+\epsilon)}{\sqrt{3}} + \frac{1}{3}.$$

Using (39) with (41), together with the fact that on this event, $V_1 \le v_{1,0} = v_{2,0} \le v_{2,k} \le V_2$, gives

$$\mathbb{P}\Big[Z \ge c_{1,\epsilon}\sqrt{\hat{V}_2 x} + (c_{4,\epsilon} + c_{5,\epsilon}B)x \cap \mathcal{V}_{1,-1} \cap \mathcal{V}_{2,k} \cap \mathcal{B}_l\Big] \le 4de^{-x} \tag{51}$$

for any $k, l \ge 0$. Using (40) with (41) and (43) gives

$$\mathbb{P}\Big[Z \ge c_{1,\epsilon}\sqrt{(\hat{V}_1 \vee \hat{V}_2)x} + (c_{4,\epsilon} + c_{5,\epsilon}B)x \cap \mathcal{V}_{1,j} \cap \mathcal{V}_{2,k} \cap \mathcal{B}_l\Big] \le 4de^{-x} \tag{52}$$

for any $j, k, l \ge 0$. Taking the largest term upper bounding Z in these inequalities, we obtain that

$$\mathbb{P}\Big[Z \ge c_{1,\epsilon}\sqrt{(\hat{V}_1 \vee \hat{V}_2)x} + (c_{6,\epsilon} + c_{5,\epsilon}B)x \cap \mathcal{V}_{1,j} \cap \mathcal{V}_{2,k} \cap \mathcal{B}_l\Big] \le 4de^{-x} \tag{53}$$

for any $j, k, l \ge -1$, where we introduced

$$c_{6,\epsilon} = \sqrt{2(1+\epsilon)c_0} + b_0/3 + 2(1+\epsilon)^2.$$

So, Z is controlled by an observable term in all cases. It remains to remove all the events $\mathcal{V}_{1,j} \cap \mathcal{V}_{2,k} \cap \mathcal{B}_l$ for $j, k, l \ge -1$. This is done by increasing x by a very small observable term, and by using an union bound on all the possible combinations $j, k, l \ge -1$. Introduce for some $c_\ell > 0$

$$\begin{aligned}\hat{\ell} = c_\ell\Big(&\log\log\Big(\frac{2\hat{V}_1 + 2((1+\epsilon)^2 + B^2/3)x}{c_0 x} \vee e\Big) \\ &+ \log\log\Big(\frac{2\hat{V}_2 + 2((1+\epsilon)^2 + B^2/3)x}{c_0 x} \vee e\Big) + \log\log\Big(\frac{B}{b_0} \vee e\Big)\Big).\end{aligned}$$

Note that $\hat{\ell} \geq 0$. Now, write

$$\mathbb{P}\Big[Z \geq c_{1,\epsilon}\sqrt{(\hat{V}_1 \vee \hat{V}_2)(x+\hat{\ell})} + (c_{6,\epsilon} + c_{5,\epsilon}B)(x+\hat{\ell})\Big] = \sum_{j,k,l \geq -1} \mathbb{P}_{j,k,l},$$

where

$$\mathbb{P}_{j,k,l} = \mathbb{P}\Big[Z \geq c_{1,\epsilon}\sqrt{(\hat{V}_1 \vee \hat{V}_2)(x+\hat{\ell})} + (c_{6,\epsilon} + c_{5,\epsilon}B)(x+\hat{\ell}) \cap \mathcal{V}_{1,j} \cap \mathcal{V}_{2,k} \cap \mathcal{B}_l\Big],$$

and decompose in the following way

$$\begin{aligned}\sum_{j,k,l \geq -1} \mathbb{P}_{j,k,l} = \mathbb{P}_{-1,-1,-1} + \sum_{j\geq 0}\mathbb{P}_{j,-1,-1} + \sum_{k\geq 0}\mathbb{P}_{-1,k,-1} + \sum_{l \geq 0}\mathbb{P}_{-1,-1,l} \\ + \sum_{j,k\geq 0}\mathbb{P}_{j,k,-1} + \sum_{j,l\geq 0}\mathbb{P}_{j,-1,l} + \sum_{k,l\geq 0}\mathbb{P}_{-1,k,l} + \sum_{j,k,l\geq 0}\mathbb{P}_{j,k,l}.\end{aligned}$$

For $\mathbb{P}_{-1,-1,-1}$, we use the fact that $c_{1,\epsilon}\sqrt{(\hat{V}_1 \vee \hat{V}_2)(x+\hat{\ell})} + (c_{6,\epsilon}+c_{5,\epsilon}B)(x+\hat{\ell}) \geq c_{6,\epsilon}x \geq (\sqrt{2c_0}+b_0/3)x$, and then (45) to obtain that

$$\mathbb{P}_{-1,-1,-1} \leq 2de^{-x}.$$

For $\mathbb{P}_{j,-1,-1}$, we use (43) to get that on $\mathcal{V}_{1,j} \cap \mathcal{V}_{2,-1} \cap \mathcal{B}_{-1}$

$$\begin{aligned}c_{1,\epsilon}\sqrt{(\hat{V}_1 \vee \hat{V}_2)(x+\hat{\ell})} + (c_{6,\epsilon}+c_{5,\epsilon}B)(x+\hat{\ell}) \\ \geq \sqrt{2c_{1,\epsilon}\hat{V}_1(x+\ell_j^{(1)})} + (c_{2,\epsilon}+c_{3,\epsilon}b_0)(x+\ell_j^{(1)}),\end{aligned}$$

where we put

$$\ell_j^{(1)} = c_\ell \log\log\Big(\frac{v_{1,j}}{c_0 x} \vee e\Big) = \log\big((j\log(1+\epsilon) \vee 1)^{c_\ell}\big). \tag{54}$$

So, using (46), we obtain

$$\sum_{j\geq 0}\mathbb{P}_{j,-1,-1} \leq 4de^{-x}\sum_{j\geq 0} e^{-\ell_j^{(1)}} = 4de^{-x}\Big(1 + \log(1+\epsilon)^{c_\ell}\sum_{j\geq 1} j^{-c_\ell}\Big).$$

We obtain in the exact same way using (41) and (47) that

$$\sum_{k\geq 0}\mathbb{P}_{-1,k,-1} \leq 4de^{-x}\Big(1 + \log(1+\epsilon)^{c_\ell}\sum_{k\geq 1} k^{-c_\ell}\Big),$$

and also

$$\sum_{l\geq 0}\mathbb{P}_{-1,-1,l} \leq 4de^{-x}\Big(1 + \log(1+\epsilon)^{c_\ell}\sum_{l\geq 1} l^{-c_\ell}\Big)$$

using (49). For $\mathbb{P}_{j,k,-1}$, we use (48) with (41) and (43), together with the fact that on $\mathcal{V}_{1,j} \cap \mathcal{V}_{2,k} \cap \mathcal{B}_{-1}$ we have

$$\begin{aligned}c_{1,\epsilon}\sqrt{(\hat{V}_1 \vee \hat{V}_2)(x+\hat{\ell})} + (c_{6,\epsilon}+c_{5,\epsilon}B)(x+\hat{\ell}) \\ \geq \sqrt{2c_{1,\epsilon}(\hat{V}_1 \vee \hat{V}_2)(x+\ell_j^{(1)}+\ell_k^{(2)})} + (c_{2,\epsilon}+c_{3,\epsilon}b_0)(x+\ell_j^{(1)}+\ell_k^{(2)}).\end{aligned}$$

This gives

$$\sum_{j,k\geq 0}\mathbb{P}_{j,k,-1} \leq 4de^{-x}\sum_{j\geq 0} e^{-\ell_j^{(1)}}\sum_{k\geq 0} e^{-\ell_k^{(2)}} = 4de^{-x}\Big(1 + \log(1+\epsilon)^{c_\ell}\sum_{j\geq 1} j^{-c_\ell}\Big)^2.$$

We obtain in the same way, using (50), (51), (52) and (41), (43), that

$$\sum_{k,l\geq 0} \mathbb{P}_{-1,k,l} \leq 4de^{-x}\Big(1+\log(1+\epsilon)^{c_\ell}\sum_{j\geq 1} j^{-c_\ell}\Big)^2$$

$$\sum_{j,l\geq 0} \mathbb{P}_{j,-1,l} \leq 4de^{-x}\Big(1+\log(1+\epsilon)^{c_\ell}\sum_{j\geq 1} j^{-c_\ell}\Big)^2$$

$$\sum_{j,k,l\geq 0} \mathbb{P}_{j,k,l} \leq 4de^{-x}\Big(1+\log(1+\epsilon)^{c_\ell}\sum_{j\geq 1} j^{-c_\ell}\Big)^3.$$

Put $c_{\ell,\epsilon} = 1 + \log(1+\epsilon)^{c_\ell}\sum_{j\geq 1} j^{-c_\ell}$. We finally have that

$$\mathbb{P}\Big[Z \geq c_{1,\epsilon}\sqrt{(\hat{V}_1 \vee \hat{V}_2)(x+\hat{\ell})} + (c_{6,\epsilon} + c_{5,\epsilon}B)(x+\hat{\ell})\Big] \\ \leq 2(1+6(c_{\ell,\epsilon}+6c_{\ell,\epsilon}^2)+2c_{\ell,\epsilon}^3)de^{-x}.$$

Now, choose $\epsilon = c_0 = b_0 = 1$ and $c_\ell = 2$. For this choice we have $2(1+6(c_{\ell,\epsilon}+6c_{\ell,\epsilon}^2)+2c_{\ell,\epsilon}^3) \leq 84.9$, $c_{1,\epsilon} = 4$, $c_{5,\epsilon} = 4/\sqrt{3}+1/3 \leq 2.65$ and $c_{6,\epsilon} = 2+1/3+8 \leq 10.34$. This concludes the proof of Theorem 3.

References

[1] E. Bacry, S. Delattre, M. Hoffmann, and J.-F. Muzy. Modelling microstructure noise with mutually exciting point processes. *Quantitative Finance*, 13(1):65–77, 2013.

[2] E. Bacry, S. Gaïffas, and J.-F. Muzy. Concentration for matrix martingales in continuous time and microscopic activity of social networks. *arxiv preprint arXiv:1412.7705*, 2014.

[3] P. L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.

[4] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal of Imaging Sciences*, 2(1):183–202, 2009.

[5] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[6] C. Blundell, J. Beck, and K. A. Heller. Modelling reciprocating relationships with hawkes processes. In *Advances in Neural Information Processing Systems*, pages 2600–2608, 2012.

[7] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 12(51):4203–4215, 2004.

[8] E.J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5), 2009.

[9] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41), 2008.

[10] N. Daneshmand, M. Rodriguez, L. Song, and B. Schölkpof. Estimating diffusion network structure: Recovery conditions, sample complexity, and a soft-thresholding algorithm. *ICML*, 2014.

[11] M. Argollo de Menezes and A.-L. Barabási. Fluctuations in network dynamics. *Phys. Rev. Lett.*, 92:028701, Jan 2004.

[12] C. DuBois, C. Butts, and P. Smyth. Stochastic blockmodeling of relational event dynamics. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 238–246, 2013.

[13] S. Gaïffas and A. Guilloux. High-dimensional additive hazards models and the lasso. *Electronic Journal of Statistics*, 6:522–546, 2012.

[14] M. Gomez-Rodriguez, J. Leskovec, and B. Schölkopf. Modeling information propagation with survival theory. *ICML*, 2013.

[15] N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. Technical report, Arvix preprint, 2012.

[16] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[17] T. Iwata, A. Shah, and Z. Ghahramani. Discovering latent influence in online social activities via shared cascade poisson processes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 266–274. ACM, 2013.

[18] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Saint-Flour XXXVIII-2008*, volume 2033. Springer, 2011.

[19] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[20] J. Leskovec. *Dynamics of large networks.* PhD thesis, Machine Learning Department, Carnegie Mellon University, 2008.

[21] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD.* ACM, 2009.

[22] A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1):173–183, 1995.

[23] S. W. Linderman and R. P. Adams. Discovering latent network structure in point process data. *arXiv preprint arXiv:1402.0914*, 2014.

[24] R. S. Liptser and A. N. Shiryayev. *Theory of martingales.* Springer, 1989.

[25] P. Massart. *Concentration inequalities and model selection*, volume 1896. Springer, 2007.

[26] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 2011.

[27] Y. Ogata. On lewis' simulation method for point processes. *Information Theory, IEEE Transactions on*, 27(1):23–31, 1981.

[28] Y. Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402, 1998.

[29] F. Orabona, A. Argyriou, and N. Srebro. Prisma: Proximal iterative smoothing algorithm. *arXiv preprint arXiv:1206.2372*, 2012.

[30] F. Ricci, L. Rokach, and B. Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[31] E. Richard, S. Gaïffas, and N. Vayatis. Link prediction in graphs with autoregressive features. *Journal of Machine Learning Research*, 2014.

[32] M. Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. *ICML*, 2011.

[33] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[34] S. Van De Geer. *Empirical Processes in M-estimation*, volume 105. Cambridge university press Cambridge, 2000.

[35] S.-H. Yang and H. Zha. Mixture of mutually exciting processes for viral diffusion. In *ICML*, 2013.

[36] K. Zhou, H. Zha, and L. Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. *AISTATS*, 2013.