Lasso-type estimator and classification algorithm for high-dimensional multivariate Hawkes processes

Christophe Denis, Charlotte Dion-Blanc, Romain Edmond Lacoste, Laure Sansonnet

Lasso-type estimator and classification algorithm for high-dimensional multivariate Hawkes processes

Christophe Denis[(1,2)], Charlotte Dion-Blanc[(1)], Romain E. Lacoste[(3)], Laure Sansonnet[(1,4)]

(1) *Sorbonne Université, Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, F-75005 Paris, France*
(2) *Université Paris 1 Panthéon-Sorbonne, SAMM, Paris, France*
(3) *Université Gustave Eiffel, Université Paris Est Creteil, CNRS, LAMA, F-77454 Marne-la-Vallée, France*
(4) *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France*

Abstract

We propose to deal with high-dimensional events based data using Hawkes processes. Focusing on the Multivariate Hawkes Processes (MHP) in high dimension, an estimation task, followed by a classification task, are addressed in this article. In both cases, we assume to have access to a large number of repeated observations of the process over the same short time interval. MHPs form a versatile class of point processes that model interactions among connected individuals within a network. In this work, we allow the network dimension to be large relative to the number of observations, which necessitates a sparsity assumption on the adjacency matrix. Furthermore, we assume that the observations belong to different classes, discriminated by both the exogenous intensity vector and the adjacency matrix, which encodes the strength of interactions. Specifically, the observed training data consist of labeled, repeated, and independent realizations over a fixed time interval. In this context, we propose a novel methodology comprising an initial interaction recovery step, conducted per class, followed by a refitting step guided by a suitable classification criterion. To recover the support of the adjacency matrix in each class, we introduce a Lasso-type estimator and prove the consistency of the estimated supports under appropriate assumptions on the processes. Leveraging the estimated supports, we then construct a classification procedure based on empirical error minimization. Notably, we provide convergence rates for our classifier. An in-depth numerical study, using both synthetic and real-world datasets, supports our theoretical findings, both for support recovery and for supervised classification.

Keywords Multivariate Hawkes Process · High Dimension · Lasso · Classification · Empirical Risk Minimization

1 Introduction

In recent years, supervised classification of complex data has attracted considerable attention. This statistical problem encompasses a broad class of applications, including the classification of

multivariate time series (Ismail Fawaz et al., 2019). In particular, the development of cutting-edge methodology for classifying event-based data is a matter of great interest. In this paper, we tackle the task of supervised classification of sequences of events into K classes with $K \geq 2$. We assume that each class is characterized by distinct underlying occurrence dynamics, and our objective is to accurately discriminate among them.

In this paper, the object of interest is a multivariate counting process N of dimension M for which the stochastic intensity of the j-th component writes

$$\lambda_{Y,j}(t) = \mu_{Y,j} + \sum_{j'=1}^{M} a_{Y,j,j'} \int_0^{t-} h(t-s)\, \mathrm{d}N_{j'}(s), \tag{1}$$

where Y is a random variable representing the label, taking values in $[K] := \{1, \ldots, K\}$, and distributed according to the probability distribution p^*. In particular we work under the asymptotic framework of repeated observations, assuming that n realizations of the process N are observed on a fixed time interval $[0, T]$, with each class represented among the observations. In this model, described by Equation (1), the classes are discriminated by the parameters of their intensity vectors: the baseline rates $\mu_{Y,j}$ and the adjacency matrix $A_Y = (a_{Y,j,j'})$, which encodes the interactions among process components. The primary challenge we address is the high dimensionality, where the dimension M of the process N can be much larger than the sample size n. To tackle this high-dimensional challenge, we impose a sparsity assumption on the adjacency matrix corresponding to each class. In this context, our contribution is a classification methodology based on two steps. First, we estimate the support of the adjacency matrices corresponding to each class. Next, we leverage this information to design a classification procedure based on the Empirical Risk Minimization (ERM) principle, which benefits from the resulting reduction in the dimensionality of the parameter space. Figure 1 summarizes the contributions of the paper and highlights the two main questions addressed: the estimation of the support of high-dimensional parameters in a multivariate Hawkes model, and the supervised multiclass classification ofmultivariate Hawkes processes.

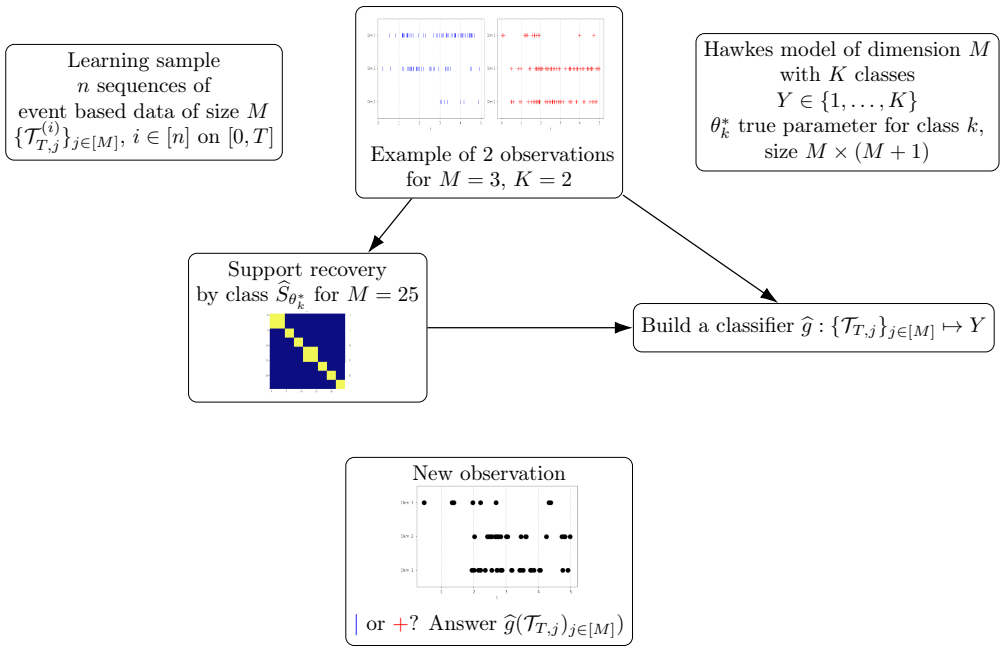

Figure 1: Summary of our classification procedure using support recovery

1.1 Related works

Hawkes processes. Hawkes processes are a family of point processes introduced by Hawkes (1971). Such processes model complex temporal dynamics, where the occurrence of events is impacted by past activity. The multidimensional version of these processes, MHP, is a natural generalization that considerably enriches modeling possibilities. Indeed, in addition to model self-exciting interactions, such a model takes into account interactions between connected individuals within a network. These interactions can be encoded in the adjacency matrix, where each coefficient represents the strength of the interaction of one component on another. Historically applied in seismology (Ogata, 1988), they have since been used in a wide range of applications including genomics (Reynaud-Bouret and Schbath, 2010), finance (Embrechts et al., 2011), urban crime (Mohler et al., 2011), order book in finance (Bacry et al., 2015), football (Baouan et al., 2022) and neuroscience (Bonnet et al., 2022). Another important application is the modeling of social network activity, as in Zhang et al. (2018) and further works. In addition, the Hawkes processes are frequently used as spike-trains models in neurosciences, for example in Reynaud-Bouret et al. (2013); Bonnet et al. (2023); Spaziani et al. (2023). Recently, HP models have also been used in the ecological field in Nicvert et al. (2024) for interaction between species and in Denis et al. (2024) for bats monitoring. Finally, in the context of repeated observations with fixed T, Lotz (2024) provides a likelihood ratio test for testing presence of interaction using the sum of the repeated observations.

High-dimensional setting. In the high-dimensional setting, meaning that the number of components M is large, it is classical to impose a sparsity assumption on the adjacency matrix that characterizes the intensity process. Therefore, it appears that reconstructing the support of the adjacency matrix or the connectivity graph, is a matter of great interest, which is related to the Granger causality (see Eichler et al., 2017; Sulem et al., 2024). In particular, this is a crucial issue for connectivity of neurons, see for example Lambert et al. (2018), or in social network (Carstensen et al., 2010). In particular, penalized method for point processes is the object of the recent review Limnios and Hansen (2025).

Let us focus on the Lasso literature in the classical Gaussian framework. The Lasso procedure has been originally introduced by Tibshirani (1996) and Chen et al. (1998). It is a popular statistical method for high-dimensional problems for which efficient implementation procedure have been developed. Besides, the Lasso procedure has been widely studied from the theoretical point of view (see *e.g.* Meinshausen and Bühlmann, 2006; Tropp, 2006; Bühlmann and Van De Geer, 2011). In particular, support recovery has been investigated by Wainwright (2009), Zou (2006) for the adaptive Lasso, and also in multi-class classification methods (see Abramovich and Grinshtein, 2018). Let us emphasize that an induced and undesired effect of ℓ_1-penalization is the shrinkage of large coefficients. To bypass this issue, refitting strategies are commonly used and well covered in the literature, see for example Chzhen et al. (2019).

For Hawkes process, the work of Donnet et al. (2020) is dedicated to a nonparametric Bayesian procedure to tackle high-dimensionality. Let us mention the work of Zhou et al. (2013) which proposes an efficient algorithm for a Lasso estimator for high-dimension Hawkes process, implemented in the Python library `tick` (Bacry et al., 2018). Later, in the work of Bacry et al. (2020), the authors use a least-squares contrast penalized with an ℓ_1-norm together with a trace norm. The resulting estimator benefits from a sparse structure with low rank. Its construction relies on only one observation over a time interval $[0, T]$, with $T \to +\infty$ that draws the asymptotic for obtaining theoretical results for the intensity process estimation. Under the same observational setup, several sparse support recovery procedure has also been proposed in the literature, for example a likelihood ratio testing procedure in Kim et al. (2011), or a group-Lasso least-squares penalized

estimator in Cai et al. (2024).

Multiclass setting. The present work falls in the supervised classification setting. In particular, we assume that the learning sample of size n consists of i.i.d. labeled data where the features are the jump times of a high-dimensional Hawkes process observed on the fixed time interval $[0, T]$. It is worth noticing that Ditlevsen and Löcherbach (2017) also consider multiclass systems of interacting Hawkes processes to model different groups of neurons. They study the mean-field property of the model without considering statistical issues. In a different observational setup, some works in natural language processing also tackle some similar issues as Lukasik et al. (2016) and later Tondulkar et al. (2022). Closest to ours, the work of Denis et al. (2022) provides a classification procedure for observations coming from a univariate HP where the classes are discriminated by the kernel of the intensity process. In particular, this method is applied in Denis et al. (2024), for modeling echolocation calls of bats recording in several sites throughout France. In the present work, we generalize the approach developed in Denis et al. (2022) in the high-dimensional setting using a Lasso penalty.

1.2 Main contributions

The contributions of this paper are two-fold. First, we introduce a Lasso-type estimator for the Hawkes model parameter within a high-dimensional framework. Second, we propose a supervised classification algorithm, named ERMLR, that leverages this estimator efficiently. Let us now outline the key contributions.

- A key feature associated to a MHP of size M is the matrix of the ℓ_1-norms of the interaction functions, commonly referred as to the adjacency matrix. Its size scales as M^2, which in high-dimensional regimes may exceed the available sample size n. In this context, we assume a sparse representation of the process, and therefore propose a penalized estimator for the parameters, including the baseline of size M and the adjacency matrix. Additionally, we assume the process is observed repeatedly n times over a short time interval. This specific observation framework requires the development of a suitable estimation procedure. More precisely, the estimation of the parameters relies on an empirical least-squares contrast with an ℓ_1-penalty on the parameters. We establish rates of convergence of the estimated support and the estimated coefficients of the MHP. Notably, our theoretical findings show that the established rates of convergence are comparable to those obtained in the classical Gaussian setting. In particular, we extend the results of Bacry et al. (2020) and Cai et al. (2024), which were established for the infinite-time observation regime, to the framework of repeated observations over finite time windows.

- Then we propose to take advantage of these results to provide a supervised classification algorithm dedicated to high-dimensional MHP. Specifically, we consider a classifier based on Empirical Risk Minimization principle, which involves the minimization of a ℓ_2-risk on a set of parameters that depends only on the estimated supports. We show that the rates of convergence of our classification algorithm is, up to a logarithmic factor, of order the square root of the size of the support over the sample size n. Notably, we extend the results obtained in Denis et al. (2022) to the high-dimensional framework.

- Finally, to due the numerical complexity of our problem, the implementation of our classification algorithm constitutes a significant contribution. This implementation leverages cutting-edge optimization algorithms to ensure both efficiency and scalability. In particular, the Lasso-penalized contrast is optimized using the FISTA algorithm, while the calibration

of the ℓ_1-penalty is performed via the EBIC criterion (Chen and Chen, 2008). Then, for the minimization of the empirical risk, we consider a parameter-free projected adaptive gradient descent FREE ADAGRAD recently introduced in Chzhen et al. (2023). We evaluate the performance of our procedure (including estimation and classification) on synthetic data and show the good performance of our algorithm. In particular, it reveals that our algorithm succeeds well for both support recovery and classification accuracy. To further illustrate the practical utility of our inference procedure, we present a social network application using our Lasso approach, which outperforms benchmark methods.

1.3 Outline of the paper

Section 2 describes the model, along with the necessary assumptions. Section 3 proposes a Lasso-type estimator and the classification algorithm named `ERMLR`. Section 4 provides the main theoretical results on both Lasso estimator of the MHP parameters and the classification procedure. Then, full implementation details about the procedure are given in Section 5. Then, Section 6 is devoted to numerical results, both on synthetic data and on real-world data sets. We finally provide a discussion in Section 7. The proofs are relegated in an appendix and details on the numerical part are postponed in a supplementary material.

Notations. For a matrix $A = (a_{j,j'}) \in \mathbb{R}^{M\times M}$, the Frobenius norm is $\|A\|_F^2 = (\mathrm{Tr}(A'A)) = (\sum_{j,j'} a_{j,j'}^2)$, where A' denotes the transposition of the matrix A and Tr is the trace operator that returns the sum of diagonal entries of a square matrix. Recall that $\rho(A) \le \|A\|_2 \le \|A\|_F$ where $\|A\|_2$ is the subordinate norm and $\rho(A)$ is the spectral radius of A, which is the largest eigenvalue of A. Finally, $\|A\|_\infty = \max_i \sum_j |a_{i,j}|$. For a vector $X = (x_j) \in \mathbb{R}^M$ the infinite norm is defined by $\|X\|_\infty = \max_j |x_j|$. We also defined the support of a vector $X \in \mathbb{R}^M$ by $\mathrm{supp}(X) = \{j \in [M],\ \ X_j \neq 0\}$.

2 General framework

Section 2.1 introduces the considered model, some notation, and the considered multiclass classification problem. Section 2.2 is dedicated to the presentation of the main assumptions. Finally, a closed-form expression of the optimal classifier is provided in Section 2.3.

2.1 Formal definitions and notations

Let us first introduce the general linear multivariate Hawkes process and then the considered multiclass classification model.

Multivariate counting process. Throughout the paper, we work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with canonical filtration

$$\mathcal{F}_t := \sigma\Big(N(s), s \le t\Big),$$

where $N = (N_1, \dots, N_M)$ is a M-dimensional counting process. It is observed on a fixed time interval $[0, T]$. More specifically, we assume that the counting process $N = (N_1(t), \dots, N_M(t))_{t\in[0,T]}$ is a linear multivariate Hawkes process, where for each $j \in [M]$, $t \in [0, T]$, $N_j(t)$ denotes the number of events that have occurred before time t for the j-th process. Finally, the set of observed jump times of N over $[0, T]$ is denoted by $\mathcal{T}_T = (\mathcal{T}_{T,1}, \dots, \mathcal{T}_{T,M})$, where for each $j \in [M]$, $\mathcal{T}_{T,j}$ is the observed jump times associated to the process N_j. Each process N_j can be characterized by its intensity function. Heuristically, at a given time, the intensity function gives the infinitesimal

probability of observing an event in the near future, conditionally on the past of the process. For each $j \in [M]$, the predictable intensity of the process N_j is then defined by

$$\lambda_j(t) = \mu_j + \sum_{j'=1}^{M} a_{j,j'} \int_0^{t-} h(t-s)\, \mathrm{d}N_{j'}(s) = \mu_j + \sum_{j'=1}^{M} a_{j,j'} \sum_{T_\ell \in \mathcal{T}_{t-,j'}} h(t - T_\ell), \tag{2}$$

where $\mu = (\mu_j)_{j\in[M]}$ is the vector of exogenous intensities, $A = (a_{j,j'})_{1\leq j,j'\leq M}$ is the matrix of interactions, h is the kernel function. For each $j \in [M]$, the coefficient μ_j models the arrival of spontaneous events for the j-th component. For each $j, j' \in [M]$, the coefficient $a_{j,j'}$ is non-negative and expresses the positive influence of the one-dimensional process $N_{j'}$ on the one-dimensional process N_j.

Finally, the kernel h is a non-negative function supported on $\mathbb{R}_+$ such that $\|h\|_1 = 1$. It dictates how quick these influences vanish over time. In the following, the kernel function h is assumed to be known. Finally, let us define the support of A, or the active set, denoted S. It corresponds to the positions of the non-zero coefficients $a_{j,j'}$, meaning that component j' has an impact on component j.

In the following, we denote the parameters with a star * when they refer to the *true* parameters we aim to estimate.

Multiclass setting. We consider the multiclass classification problem in which each data point is characterized by a couple $(\mathcal{T}_T, Y)$. Here, $\mathcal{T}_T$ denotes the collection of jump times observed for a counting process N over $[0, T]$, and $Y \in [K]$ specifies the class label which is a random variable with distribution $p^* = (p_k^*)_{k\in[K]}$. In particular, we assume that $N = (N_1, \ldots, N_M)$ is a mixture of a M-dimensional linear HPs observed on the time interval $[0, T]$. More precisely, conditional on Y, the counting process N is a M-dimensional linear HP, where for each $j \in [M]$, the predictable intensity of N_j depends on the label Y and is defined at time $t \geq 0$ by Equation (1) (or by Equation (2) with parameters depending on the label Y) with the vector of baselines $\mu_Y = (\mu_{Y,j})_{j\in[M]}$ associated to the class Y, and the matrix $A_Y = (a_{Y,j,j'})_{1\leq j,j'\leq M}$ is the $M \times M$ adjacency matrix of the label Y. This choice of modeling is motivated by the fact that the classes are characterized by different underlying graph behavior, where an edge in the graph matches a non-zero $a_{Y,j,j'}$. Again, the true parameters are denoted with a star * in the following.

We assume in the following that the parameters (μ_Y^*, A_Y^*) are unknown as well as the distribution p^* of Y. Finally, the kernel function h is assumed to be known and for the sake of simplicity, it does not depend on the classes or on the components of the process. Note that in the numerical section, we consider the standard choice of exponential kernel. However, more general choice of the kernel function may be investigated.

Objective. In the multiclass setup, the objective is to build a classifier, a measurable function g such that $g(\mathcal{T}_T)$ belongs to $[K]$, and provides an accurate prediction of the label Y. In particular, the misclassification risk assesses the quality of such predictor g. It is defined as

$$\mathcal{R}(g) := \mathbb{P}\left(g(\mathcal{T}_T) \neq Y\right).$$

The set of all classifiers is denoted by $\mathcal{G}$. Naturally, we aim at considering the predictor g^*, namely the Bayes classifier, that achieves the minimum risk over $\mathcal{G}$. In Section 2.3, we provide an explicit formula of the oracle classifier g^*. Nevertheless, since the distribution of the observation $(\mathcal{T}_T, Y)$ is assumed to be unknown, we build a predictor that relies on a training sample of size n which consists of i.i.d. copies of $(\mathcal{T}_T, Y)$. At this step, we draw the reader attention to the fact that the

considered asymptotic is as follows. The horizon time T is fixed, while the sample size n goes to infinity. Recall that the size M of the MHP is actually $M = M_n$ and can increase with n. In the sequel, a predictor built on the training data is denoted by $\hat{g}$. In particular, we require that $\hat{g}$ satisfies the consistency property,

$$\mathbb{E}\left[\mathcal{R}(\hat{g}) - \mathcal{R}(g^*)\right] \underset{n\to\infty}{\longrightarrow} 0,$$

when n tends to infinity. However, in our study, the intrinsic dimension of our problem M^2 may be much larger than the sample size of the learning dataset. In this case, predictor $\hat{g}$ is not consistent. Therefore, as it is usual in this high-dimensional setup, we introduce a sparsity assumption for our model.

2.2 Assumptions

This section is dedicated to the main assumptions that are assumed throughout the paper. In particular, in our multivariate framework, we allow the dimension parameter M to be large, which may induces that M^2 is much larger than the size n of the training sample. To alleviate this issue, we introduce a sparsity assumption on the matrices $(A_k^*)_{k\in[K]}$.

Firstly, we introduce an assumption which ensures that each class occurs with non-zero probability.

Assumption 1 (Non-zero class weight)**.** *There exists $p_0 > 0$ such that $\min_{k\in[K]}(p_k^*) > p_0$.*

We also make the following assumption on the parameters of the process.

Assumption 2 (Boundaries of true MHP parameters)**.**

(i) There exists $0 < \mu_0 < \mu_1$ such that for $j \in [M]$ and $k \in [K]$, $\mu_0 \leq \mu_{k,j}^* \leq \mu_1$.

(ii) $\max_{k\in[K]} \|A_k^*\|_F \leq CM$, with C a positive constant.

Furthermore, we consider the following assumptions, which imply that the process N admits finite exponential moment.

Assumption 3 (Stability condition)**.**

(i) The kernel function h belongs to the set $\mathcal{H} := \{h : \mathbb{R}_+ \to \mathbb{R}_+, \int h(t)\,dt = 1\}$ and is bounded.

(ii) $\max_{k\in[K]} \rho(A_k^*) < 1$.

Assumption 4 (Exponential moment)**.** *There exist $a > 0$, and $C > 0$ that do not depend on M, such that*

$$\sup_{j\in[M]} \mathbb{E}[\exp(aN_j(T))] < C.$$

Finally, we assume that for each $k \in [K]$, the matrix A_k^* is sparse, meaning that a few coefficients are non-zero. For each $k \in [K]$, let us denote by

$$S_k^* := \left\{(j, j') \in [M]^2,\ \ a_{k,j,j'}^* \neq 0\right\} \tag{3}$$

the active sets (or support) of A_k^*, $|S_k^*|$ its cardinality, and S_k^{*c} its complement. Throughout the paper, we consider the following assumption.

Assumption 5 (Sparsity assumption). *There exists a constant $s^* > 0$ such that*

$$\max_{k\in[K]} |S_k^*| \leq s^*.$$

In particular, in our high-dimensional setting, we assume that $s^* << M^2$. Since we do not assume sparsity on the vectors $\mu_k^*,\ k \in [K]$, we consider the following interplay between parameter M and the sample size of the training dataset. The dimension M of the process may depend on n with $M^2 >> n$. In this case, the sparsity assumption is crucial to overcome the high-dimension issues. However, we assume in the following that M satisfies $M/n \to 0$.

Remark 1 (On Assumptions 2 and 4). *Let us highlight that Assumption 2 concerning the Frobenius norm of the matrix A_k^* is not restrictive as soon as n is large enough, due to the sparsity assumption. In particular, if the coefficients are all smaller than 1, which is the considered case in practice, we have that $\|A_k^*\|_F \leq s^*$. And, as the spectral radius is smaller than 1 and granted that A_k^* is symmetric, the norm subordinate to the infinite norm satisfies $\|A_k^*\|_\infty \leq \sqrt{M}$. In supplementary material, we give some examples where Frobenius norm is computed for different setting and never exceeds M.*
Besides, about Assumption 4, note that Leblanc (2024) proves that the exponential moment of the multivariate Hawkes process is finite, under Assumption 3, some additional assumption on $\|(A_k^*)^n\|_\infty$, and when the intensity process is stationary. Nevertheless, in the general case, the bound of the moment depends on M. Hence, we require a stronger condition, such as Assumption 4 which is more suitable in the high-dimensional framework. For instance, this assumption is satisfied if there exists a positive constant C, that depends on M, such that $\sum_{j\in[M]} \mu_{k,j}^* \leq C$. Similar uniform bounds are for example needed in Amorino et al. (2024).

Remark 2 (On Assumption 5). *Note that we only assume sparsity on the adjacency matrix, but not on the vector μ^*. It ensures that all the components of the process are active. However, as in Bacry et al. (2020), it may be possible to consider also sparsity assumption on the vector of exogenous intensities. Nevertheless, this is not the line taken in this work.*

2.3 Bayes rule

In this section, we exhibit a closed form expression of the Bayes classifier g^* that minimizes the misclassification risk over the set $\mathcal{G}$. The Bayes classifier is characterized by,

$$g^*\left(\mathcal{T}_T\right) \in \underset{k\in[K]}{\operatorname{argmax}}\ \pi_k^*(\mathcal{T}_T),$$

with $\pi_k^*\left(\mathcal{T}_T\right) = \mathbb{P}\left(Y = k|\mathcal{T}_T\right)$. The following result is an extension of the result given in Denis et al. (2022). It gives the expression of the conditional probabilities π_k^* and then provides a closed form of the Bayes classifier.

Proposition 1. *Let $T \geq 0$. For each $k \in [K]$, we define,*

$$F_k^*(\mathcal{T}_T) := -\sum_{j=1}^{M} \int_0^T \lambda_{k,j}^*(s)ds + \sum_{j=1}^{M} \sum_{T_\ell \in \mathcal{T}_{T,j}} \log\left(\lambda_{k,j}^*(T_\ell)\right). \tag{4}$$

Therefore, the sequence of conditional probabilities satisfies

$$\pi_k^*\left(\mathcal{T}_T\right) = \frac{p_k^* \mathrm{e}^{F_k^*(\mathcal{T}_T)}}{\sum_{k'=1}^{K} p_{k'}^* \mathrm{e}^{F_{k'}^*(\mathcal{T}_T)}} \quad \mathbb{P}-a.s.$$

This proposition exhibits an explicit link between the unknown parameters $(\mu_k^*, A_k^*)_{k\in[K]}$ and the Bayes classifier. In particular, it suggests that a classification rule can be easily obtained by replacing the unknown parameters by estimators in Equation (4). However, the performance of the resulting classifier strongly depends on the quality of the considered estimators. In the present framework, without taking account Assumption 5, the high-dimension of the problem could lead to bad estimates. To overcome this difficulty, we propose a classification algorithm tailored to our setting, which involves Lasso-type estimators.

3 Lasso-type estimator and classification algorithm

In this section, we present our sparse Lasso-type estimator of the MHP parameters and the proposed classification algorithm that relies on a *refitting* strategy (see *e.g.* Chzhen et al., 2019). The algorithm is referred as `ERMLR` for *Empirical Risk Minimizer with Lasso Refitting*. Since the construction of the prediction rule goes in several steps and involves a splitting of the training dataset, for the sake of the simplicity, we consider a dataset of size $2n$. More specifically, the learning dataset is denoted $\mathcal{D}_n = \{(\mathcal{T}_T^{(i)}, Y^{(i)}), i = 1, \dots, 2n\}$, which consists of $2n$ independent copies of $(\mathcal{T}_T, Y)$. For the estimation purpose, the dataset $\mathcal{D}_n$ is divided into two independent datasets $\mathcal{D}_n^{(1)}$ and $\mathcal{D}_n^{(2)}$ of same size n. For sake of simplicity in the following, we index both samples using $\{1, \dots, n\}$.

To take advantage of Assumption 5, we then consider the following three-stages procedure.

- Based on the first dataset $\mathcal{D}_n^{(1)}$, we estimate the distribution $p^* = (p_k^*)_{k\in[K]}$ by its empirical counterpart $\widehat{p}$.
- Based on the second dataset $\mathcal{D}_n^{(2)}$, and for each $k \in [K]$, we estimate by $\widehat{S}_k$ the active set S_k^*, given in Equation (3), with a Lasso-type criterion, described in Section 3.1.
- Based on the second dataset $\mathcal{D}_n^{(2)}$, then, we build a classifier $\widehat{g}$ that minimizes an empirical L_2-risk on a set of predictors that depends on the estimated support $(\widehat{S}_k)_{k\in[K]}$. This construction is detailed in Section 3.2.

3.1 Estimation of the active sets for each class

Our classification procedure relies on the estimation of the active sets S_k^* for all $k \in [K]$ that are deduced from the estimation of the parameters of the process. To this aim, we consider the least squares contrast with a Lasso penalty for repeated observations and derive the estimators class by class. The considered contrast is an adaptation of the penalized criteria proposed in Bacry et al. (2020) in the context of repeated observations with a fixed horizon time of observations T.

Let $k \in [K]$. We denote $\left(\mathcal{T}_T^{(1)}, \dots, \mathcal{T}_T^{(n_k)}\right)$ the observations from class k coming from $\mathcal{D}_n^{(2)}$, with $n_k = \sum_{i=1}^n \mathbb{1}_{\{Y^{(i)}=k\}}$ the random number of observations from class k. First, we define the considered contrast.

The vector of true parameters for class k is denoted by the vector $\theta_k^* \in \mathbb{R}^{M(M+1)}$. For each $j \in [M]$,

$$\theta_{k,j}^* = \left(\theta_{k,j,\ell}^*\right)_{\ell\in\{0\}\cup[M]} := \left(\mu_{k,j}^*, a_{k,j,1}^*, \dots, a_{k,j,M}^*\right)' \in \mathbb{R}^{M+1}. \tag{5}$$

And, the generic parameter is denoted as the extended vector $\theta = (\theta_1, \dots, \theta_M) \in \mathbb{R}^{M(M+1)}$ with the similar shape as θ^*. Then, for each $\theta \in \mathbb{R}^{M(M+1)}$, and for the observation number $i \in [n_k]$ of class k, we denote the associated intensity process as $\lambda_{k,j,\theta}^{(i)}(t)$, given by Equation (1) for $Y = k$, highlighting here the dependency in the parameter θ in the notation.

The considered empirical contrast is defined, for each $\theta \in \mathbb{R}^{M(M+1)}$, as follows,

$$R_{T,k}(\theta) := \frac{\mathbb{1}_{(n_k \geq 1)}}{n_k} \sum_{i=1}^{n_k} \left(\frac{1}{T} \sum_{j=1}^{M} \int_0^T \lambda_{k,j,\theta}^{(i)^2}(t)\, \mathrm{d}t - \frac{2}{T} \sum_{j=1}^{M} \sum_{T_\ell^{(i)} \in \mathcal{T}_{T,j}^{(i)}} \lambda_{k,j,\theta}^{(i)} \left(T_\ell^{(i)} \right) \right). \tag{6}$$

Note that, if $n_k = 0$, we have $\widehat{\theta} = 0$. The Lasso estimator is then the minimizer of the penalized contrast and defined for some $\kappa > 0$, as

$$\widehat{\theta}_k := \widehat{\theta}_k(\kappa) \in \underset{\theta \in \mathbb{R}^{M(M+1)}}{\operatorname{argmin}} \left\{ R_{T,k}(\theta) + \kappa \sum_{j=1}^{M} \sum_{j'=1}^{M} |\theta_{j,j'}| \right\}. \tag{7}$$

It is worth noting that, beyond the classification framework, this estimation procedure can be applied to estimate high-dimensional MHP parameters in the context of repeated observations. Finally, from the estimator $\widehat{\theta}_k$, we get the estimated support of A_k^*

$$\widehat{S}_k := \{(j, j^{'}) \in [M]^2,\ \widehat{\theta}_{k,j,j'} \neq 0\} = \{(j, j^{'}) \in [M]^2,\ \widehat{a}_{k,j,j'} \neq 0\}.$$

The last equality highlights that $\widehat{S}_k$ represents the estimated active set of A_k^* since it does not involve the first column of $\widehat{\theta}_k$ that contains the vector of estimated baseline $(\mu_j)_{j \in [M]}$.

3.2 ERM classifier with a refitting step

In this section, we present the last step of our estimation procedure, which is dedicated to the construction of the classifier. In particular, it involves the estimation of parameter $\theta^* = (\theta_k^*)_{k \in [K]}$. To this end, we introduce the constrained set of parameters

$$\Theta_n := \left\{ \theta = (\mu, A) \in \mathbb{R}_+^M \times \mathbb{R}_+^{M^2},\ \mu_j \in \left[\frac{1}{n}, \log(n) \right],\ j \in [M],\ \|A\|_F \leq n \right\},$$

where here we have denoted with abuse of notation, $\theta = (\mu, A)$ which is the couple of the baseline vector and the adjacency matrix, instead of using the previous vector notation. Finally the set of interest

$$\widehat{\Theta} := \left\{ \theta = (\theta_1, \ldots, \theta_K) \in \Theta_n^K,\ \ \operatorname{supp}(A_k) = \widehat{S}_k \right\}. \tag{8}$$

Several comments can be made from the definition of the set of parameters $\widehat{\Theta}$. First we observe that conditional on the event $\{\widehat{S}_k = S_k^*\}$, for n large enough, the true parameter θ^* belongs to the set $\widehat{\Theta}$. Indeed, in view of Assumption 2, for n large enough, we may assume that $1/n < \mu_0 < \mu_1 < \log(n)$. Furthermore, we emphasize that the choice of the bounds on the coefficients on the parameters of $\widehat{\Theta}$ allows to get rid of the unknown constants defined in Assumption 2. These choices are also driven by technical aspects. In particular, such bounds are required to apply concentration arguments. Additionally, let us mention that contrary to the previous step, the optimization is performed on $\mathbb{R}_+$ for each coefficient.

Let us present the estimation of the parameter θ^* and then the construction of the resulting classifier $\widehat{g}$. This construction follows the strategy provided in Denis et al. (2022) for $M = 1$, and is based on the dataset $\mathcal{D}_n^{(2)}$. It relies on the empirical risk minimization principle. Specifically, for each $\theta \in \widehat{\Theta}$, we introduce an associated score function $f_\theta = (f_\theta^1, \ldots, f_\theta^K)$. For an observed sequence of events $\mathcal{T}_T$, for $k \in [K]$,

$$f_\theta^k(\mathcal{T}_T) := 2\pi_{k,\widehat{p},\theta}(\mathcal{T}_T) - 1, \quad \pi_{k,\widehat{p},\theta}(\mathcal{T}_T) = \frac{\widehat{p}_k \mathrm{e}^{F_k(\mathcal{T}_T)}}{\sum_{k'=1}^K \widehat{p}_{k'} \mathrm{e}^{F_{k'}(\mathcal{T}_T)}}$$

with

$$F_{k,\theta}(\mathcal{T}_T) = -\sum_{j=1}^{M}\int_0^T \lambda_{k,j,\theta}(s)\,\mathrm{d}s + \sum_{j=1}^{M}\sum_{T_\ell\in\mathcal{T}_{T,j}} \log\left(\lambda_{k,j,\theta}(T_\ell)\right).$$

Note that the form of the score function f_θ is chosen according to the result provided in Proposition 1. Let $\theta \in \widehat{\Theta}$, and f_θ its associated score function, we define its empirical L_2-risk as

$$\widehat{\mathcal{R}}_2\left(f_\theta\right) := \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(Z_k^{(i)} - f_\theta^k(\mathcal{T}_T^{(i)})\right)^2, \quad Z_k^{(i)} = 2\mathbb{1}_{\{Y_i=k\}} - 1.$$

Then, we define the estimator of θ^* as the minimizer of the empirical L_2-risk,

$$\widehat{\theta}^{\mathrm{R}} \in \underset{\theta\in\widehat{\Theta}}{\operatorname{argmin}}\ \widehat{\mathcal{R}}_2(f_\theta). \tag{9}$$

From the estimator $\widehat{\theta}^{\mathrm{R}}$ of parameter θ^*, we define the `ERMLR` classifier as follows

$$\widehat{g}(\mathcal{T}_T) \in \underset{k\in[K]}{\operatorname{argmax}}\ \pi_{k,\widehat{p},\widehat{\theta}^{\mathrm{R}}}(\mathcal{T}_T). \tag{10}$$

Note that for computational purpose, as it is usual in classification, the $0-1$ loss is then replaced by the L_2 convex surrogate (see *e.g* Zhang, 2004). In particular, the L_2-loss is classification calibrated and Zhang's lemma ensures that

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq \frac{1}{\sqrt{2}}\big(\mathbb{E}\left[\mathcal{R}_2(f_{\widehat{\theta}^{\mathrm{R}}}) - \mathcal{R}_2(f_{\theta^*})\right]\big)^{1/2},$$

with $\mathcal{R}_2$ the oracle counterpart of the considered empirical risk $\widehat{\mathcal{R}}_2$ defined as

$$\mathcal{R}_2(f_\theta) = \mathbb{E}\left[\left(Z_k - f_\theta(\mathcal{T}_T)\right)^2\right], \text{ with } Z_k = 2\mathbb{1}_{\{Y=k\}} - 1.$$

One of the main appealing property of our classification algorithm is that we take advantage of the estimated support to perform the minimization of the empirical L_2-risk on a set of parameter whose dimension is much smaller than M^2. Besides, rather than using the estimated parameters obtained at the first step (Lasso-step), we consider the estimator of parameter θ^* as the minimizer of loss adapted to our multiclass classification setting.

4 Theoretical results

In this section, we first provide the consistency of the estimator of the active set in Section 4.1. Then, in Section 4.2, we derive the rate of convergence of our classification procedure with respect to the misclassification risk.

4.1 Support recovery by class

In this section, we present the key result on the Lasso estimator. More precisely, we show that

$$\mathbb{P}\left(\mathrm{supp}(\widehat{\theta}_k) = \mathrm{supp}(\theta_k^*)\right) \underset{n\to+\infty}{\longrightarrow} 1$$

which implies that for each $k \in [K]$, the Lasso estimator $\hat{\theta}_k$ solution of Equation (7) has nonzero entries at the same positions as the true parameter θ_k^*. In particular, for the multivariate Hawkes process, for $j, j' \in [M]^2$, the Lasso step can be interpreted as interaction selection, where the objective is to select whether a component j is impacted by a component j'.

Before, to give our main result, we introduce some notations for the Lagrangian version of the Lasso criterion given in Equation (7).

Notations. In the rest of the section, we fix a class $k \in [K]$, and for simplicity we drop the dependency on k in all the parameters notations. Besides, throughout this section, we work conditional on the event $n_k \geq 1$. We also remind the reader that n_k is the random number of observations from class k in the dataset $\mathcal{D}_n^{(1)}$ of size n. Then, we define for each $t \in (0, T]$ the random matrix $\mathbb{H}_t \in \mathbb{R}^{n_k \times (M+1)}$, for all $i \in [n_k]$, as follows

$$(\mathbb{H}_t)_{i,j} := \int_0^{t^-} h(t-s)\, \mathrm{d}N_j^{(i)}(s), \text{ for } j \neq 0, \quad \text{and } (\mathbb{H}_t)_{i,0} \equiv 1. \tag{11}$$

From the definition of the matrix $\mathbb{H}_t$, we can observe that $\mathbb{H}_t \theta_j = \left(\lambda_{j,\theta}^{(1)}(t), \ldots, \lambda_{j,\theta}^{(n_k)}(t)\right)'$. We finally define the random matrix $\mathbb{H}$ of size $(M+1) \times (M+1)$ as

$$\mathbb{H} := \frac{1}{T} \int_0^T \mathbb{H}_t' \mathbb{H}_t \, \mathrm{d}t.$$

In particular, when $\kappa = 0$, meaning that the contrast is minimized without the penalization term, the estimator $\widehat{\theta}$ reduces to the classical least-squares estimator, solution of the following equation

$$\mathbb{H}_{j',j'} \widehat{\theta}_{j,j'} = \sum_{i=1}^{n_k} \int_0^T \int_0^{t-} h(t-s)\, \mathrm{d}N_{j'}^{(i)}(s)\, \mathrm{d}N_j^{(i)}(t).$$

This relation highlights that suitable conditions on the matrix $\mathbb{H}$, such as invertibility, are natural requiremements in order to obtain theoretical guarantees for the estimator $\widehat{\theta}$ defined by (7).

Additional assumptions. According to Equation (5), the true parameter is denoted $\theta_j^* = \left(\theta_{j,\ell}^*\right)_{\ell \in \{0,\ldots,M\}} = \left(\mu_j^*, a_{j,1}^*, \ldots, a_{j,M}^*\right)' \in \mathbb{R}^{M+1}$. For each $j \in [M]$, we also denote $S_{\theta_j}^*$ the active set of θ_j^*. Note that, since $\mu_j^* > 0$, it contains at least one element.

Classical conditions in the ℓ_1-constraint framework are considered as, for instance, in Bühlmann and Van De Geer (2011) and references therein.

The first assumption ensures that, almost surely, the submatrix $\mathbb{H}_{S_{\theta_j}^*, S_{\theta_j}^*}$ does not have its columns linearly dependent (in which case it could be impossible to estimate θ^* when the true active set is known). The notation $\Lambda_{\min}$ denotes the minimal eigenvalue.

Assumption 6 (Minimum eigenvalue (ME))**.** *There exists $\Lambda_0 > 0$ such that, the event*

$$\Omega_{\mathrm{ME}} = \left\{ \min_{j=1,\ldots,M} \Lambda_{\min} \left(\frac{\mathbb{H}_{S_{\theta_j}^*, S_{\theta_j}^*}}{n_k} \right) \geq \Lambda_0 \right\},$$

satisfies $\mathbb{P}(\Omega_{\mathrm{ME}}) = 1$.

The following condition is the mutual incoherence, which is also referred as irrepresentability condition. Heuristically, this imposes that, almost surely, the correlation between the non-active and active variables must not be higher than the variations within the actives variables, otherwise the lasso estimator would not be able to dissociate them. It involves an incoherence parameter $\gamma \in (0, 1]$ that must not be too small.

Assumption 7 (Mutual incoherence (MI)). *There exists some* $0 < \gamma \leq 1$ *such that, the event*

$$\Omega_{\mathrm{MI}} = \left\{ \max_{j=1,\ldots,M} \| \mathbb{H}_{S^{*c}_{\theta^*_j}, S^*_{\theta^*_j}} \mathbb{H}^{-1}_{S^*_{\theta^*_j}, S^*_{\theta^*_j}} \|_\infty \leq 1 - \gamma \right\},$$

satisfies $\mathbb{P}(\Omega_{\mathrm{MI}}) = 1$.

Finally, the last condition of minimum signal ensures that the non-zero entries of the true coefficients are large enough to be properly estimated. Specifically, it imposes that the minimum value of the true parameter restricted to the support S^* cannot decay to zero faster than the regularization parameter, κ which is specified in Theorem 1.

Assumption 8 (Minimum signal condition (MS)).

$$\min_{j,j' \in S^*} \left| \theta^*_{j,j'} \right| > \Lambda_0^{-1} \max_{j=1,\ldots,M} \sqrt{\left| S^*_{\theta^*_j} \right|} \frac{\log^4(nM^2)}{\sqrt{n}}.$$

Remark 3. *Let us emphasized that* $1 \leq |S^*_{\theta^*_j}| < (M+1)$ *and* $\sum_{j=1}^{M} |S^*_{\theta^*_j}| \leq M + s^*$. *In particular,*

$$\max_{j=1,\ldots,M} \sqrt{|S^*_{\theta^*_j}|} \leq \sqrt{s^*}.$$

Also, note that our result is still valid under sublinear sparsity assumption, meaning that $s^*(M) := s^*/M \to 0$, *as* $M = o(n)$.

Support recovery result. Under the above assumptions, for each class $k \in [K]$, we establish the uniqueness of the Lasso solution, the consistency of the estimated support, and the uniform consistency of the estimator of θ^*. The proof follows the scheme of Hastie et al. (2015) Chapter 11, in the classical setting, under the three assumptions ME, MI, MS. The result provided by Theorem 1 is the main ingredient to derive rate of convergence of our classification procedure, and is also an interesting result *per se*.

Theorem 1. *Assume that* $n > \frac{2}{p_0}$, *and let* $\kappa = \dfrac{\log^4(nM^2)}{\sqrt{n}}$. *Grant Assumption 1 to Assumption8. There exists an event* Ω_n *with*

$$\mathbb{P}(\Omega_n) = \mathbb{P}\left(\Omega_{\mathrm{MI}} \cap \Omega_{\mathrm{ME}} \cap \Omega_n\right) \geq 1 - \frac{C}{n},$$

such that on $\Omega_{\mathrm{MI}} \cap \Omega_{\mathrm{ME}} \cap \Omega_n$, *we have* $n_k \geq 1$, *and*

$$\min_{\theta \in \mathbb{R}^{M(M+1)}} \left\{ R_{T,k}(\theta) + \kappa \sum_{j=1}^{M} \sum_{j'=1}^{M} |\theta_{j,j'}| \right\},$$

where $R_{T,k}$ *is given in Equation* (6), *admits a unique solution* $\widehat{\theta}$ *which satisfies the following properties*

(i) $\widehat{S}_{\widehat{\theta}} = S_{\theta^*}$,

(ii)

$$\left\| \widehat{\theta} - \theta^* \right\|_\infty \leq \frac{\Lambda_0^{-1} \max_{j=1,\ldots,M} \sqrt{|S^*_{\theta^*_j}|} \log^4(nM^2)}{\sqrt{n}}.$$

Several comments can be made from the above result. First, a straightforward consequence is that for each $k \in [K]$, the probability of the event where the support of $\widehat{A}_k$ is equal to the true support of A_k^* is going to 1 when n increases, meaning that

$$\mathbb{P}\left(\widehat{S}_k = S_k^*\right) \geq 1 - \frac{C}{n}.$$

Hence, our result provides rate of convergence for the estimator of the support $\widehat{S}_k$. Furthermore, in view of Assumption 8, we have that on the event Ω_n, $\hat{\theta}_{j,j'} > 0$. Notably, Theorem 1 extends the result of Bacry et al. (2020) in the context of repeated observations with fix observation time. In particular, the work of Bacry et al. (2020) does not provide support recovery result. However, we emphasize that our result requires stronger assumption than in Bacry et al. (2020). Let us notice that the result holds also for κ larger that $\log^4(nM^2)/\sqrt{n}$ but in this case the rates of convergence is slower.

Second up to logarithmic factor, the condition on the tuning parameter $\kappa = \kappa_n$ is of the same order as in Wainwright (2009). Besides, up to a logarithmic factor, we obtain a rate of convergence of order $\max_{j=1,\dots,M} \sqrt{|S_{\theta_j}^*|}/\sqrt{n}$ in sup-norm for the estimator $\hat{\theta}$, we can note that this rate is of the same order as the one that would expect in the classical Gaussian framework Bühlmann and Van De Geer (2011). We also highlight that in the logarithmic factor, the power of the log term is in part due to the fact that the number of jump-times of the process is not bounded a.s. Note that the result of Theorem 1 may be obtained under weaker condition than Assumption 7 by considering adaptive lasso procedure (see Chapter 7 in Bühlmann and Van De Geer, 2011). However, this time of procedure may have a high computational cost, which is a limitation for the classification task that comes after the support recovery.

Finally, the proof of this result is based on a preliminary lemma, which gives a control in probability of the maximum of the martingale terms $\sum_{i=1}^{n_k} \int_0^T (\mathbb{H}_t)_{i,j} \, \mathrm{d}M_{j'}^{(i)}(t)$, $i \in [n], j, j' \in [M]$. This inequality is obtained using a Bernstein type inequality proven in Bacry et al. (2020). This data-driven inequality and the sub-exponential property of the counting process (see Assumption 4) lead to the concentration result. Then, we follow the primal-dual-witness method of proof (see for instance Tibshirani and Wasserman, 2017).

Remark 4 (On the assumptions (MI)-(ME))**.** *These assumptions are fairly restrictive, since they require that* $\mathbb{P}(\Omega_{\mathrm{ME}})$ *and* $\mathbb{P}(\Omega_{\mathrm{MI}})$ *are exactly equal to one, and therefore hold almost surely. A possible relaxation is to formulate the assumptions in expectation, that is, by replacing every occurrence of the random matrix* $\mathbb{H}$ *with its expectation* $\mathbb{E}[\mathbb{H}]$*.*

The first question that arises is whether there exist cases in which these new, expectation-based assumptions, actually hold, as it is discussed in Cai et al. (2024) for similar assumptions. In particular, Assumption MI in expectation should hold, for example, in the simple case where A^* *is a diagonal matrix. In this case, the covariance matrix is zero and* $S_{\theta_j}^* = \{0, j\}$*, so straightforward conditions on the coefficients ensure that the assumption is satisfied.*

The second question is whether these weaker, expectation-based assumptions are still sufficient to establish the desired results. To the best of our knowledge, within the general framework of linear Hawkes processes, this remains unclear. Let us elaborate on this point. A possible strategy would be to first derive a concentration inequality between the general term of $\mathbb{H}$ *and that of* $\mathbb{E}[\mathbb{H}]$*. This could be done using Hoeffding's inequality, since each entry of* $\mathbb{H}$ *can be written as an empirical mean over repeated observations. However, because some assumptions, such as the bounded-intensity condition used in Cai et al. (2024), do not hold in our setting, it becomes challenging to leverage this relaxed assumption to establish the desired results.*

4.2 Rate of convergence of the ERMLR classifier

In this section, we derive theoretical property of the ERMLR algorithm $\hat{g}$. To establish our result, we take advantage of the support recovery result provided in Section 4.1. On the set $\{\widehat{S}_k = S_k^*\}$, the excess risk of $\hat{g}$ is upper-bounded by applying classical arguments derived from the classification framework. While we use Theorem 1 to bound the excess risk on the event $\{\widehat{S}_k \neq S_k^*\}$. Then, we obtain the following result.

Theorem 2. *Grant Assumption 1- Assumption 8. For n large enough, there exists a constant $C > 0$ such that,*

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\left(\frac{(M + s^*)\log(nM)}{n}\right)^{1/2},$$

where C depends on $T, K, \|h\|_\infty, \mu_0, \mu_1, p_0$.

As expected, we highlight that, thanks to the Lasso step, we manage to obtain, up to a logarithmic factor, a rate of order $\sqrt{(M + s^*)/n}$ rather than $\sqrt{(M + M^2)/n}$. Notably, we show that the proposed algorithm achieves the usual parametric rate.

5 Implementation

In this section, a comprehensive description of the implementation details is specified. As the ERMLR procedure execution involves two minimization problems, these two steps are described separately in Section 5.1 for the Lasso and in Section 5.2 for the classifier. In both cases, each choice is discussed in terms of the state of the art and its relevance in the context of its use. Besides, let us highlight that the implementation of the procedure relies on the Sparklen[1] Python library implemented by Lacoste (2025). This library incorporates C++ source code to handle computationally intensive tasks, making it well-suited for computationally demanding real-world applications.

5.1 Implementation details for the Lasso

For the support recovery step, our strategy consists in the minimization of the least squares contrast with Lasso penalty defined in Equation (7). This objective function is written as the sum of two functions. While the least squares contrast is differentiable, convex and smooth (*i.e.* with a Lipschitz continuous gradient), the ℓ_1-norm is non-differentiable at zero. To this extent, to carry out the minimization of such objective function, we use first-order optimization algorithm based on proximal methods with Nesterov's momentum method, namely FISTA, see Beck and Teboulle (2009). Compared to the classical proximal algorithm, the construction of a new iterate of the descent is based on a specific linear combination of the previous two points. This makes FISTA benefits from a significantly faster rate of convergence. A recommended choice of the descent step is $1/L$ with L the Lipschitz constant of the gradient. We stop the descent after 200 iterations if the stopping criterion, based on relative distance between two successive iterations, is not fulfilled yet.

Another important aspect of the Lasso step concerns the calibration of the penalization constant κ which controls the regularization. As our goal is to recover the true support, κ must be large enough to set all non-active coefficients to zero. To this end, our strategy is the following: different values of κ are explored through a grid of sufficiently fine size, denoted Δ, and the one that minimizes a specific model selection criterion is chosen. The criterion used here is the Extended

[1]GitHub repository https://github.com/romain-e-lacoste/sparklen

Bayesian Information Criteria (EBIC) introduced by Chen and Chen (2008). For some $\gamma \in [0,1]$ and $\kappa \in \Delta$, this criterion takes the following form:

$$\text{EBIC}_\gamma(\kappa) := -2L_{T,n}\left(\widehat{\theta}(\kappa)\right) + \left|S_{\widehat{\theta}(\kappa)}\right| \log(n) + 2\gamma \log\left(\binom{M^2}{\left|S_{\widehat{\theta}(\kappa)}\right|}\right)$$

where $\widehat{\theta}(\kappa)$ is the Lasso estimated with the tuning parameter κ, $L_{T,n}$ is the log-likelihood of the model, $\left|S_{\widehat{\theta}(\kappa)}\right|$ is the size of this support, namely the number of active coefficients of $\widehat{\theta}(\kappa)$.

Compared to a classical BIC criteria (namely $\gamma = 0$), an additional penalization is added to take into account the number of possible active sets of the same size. As this quantity is increasing with the size, it is relevant in a high-dimensional setting. In the following, we choose $\gamma = 1$ and $|\Delta| = 40$ as exploration grid size.

Finally, let us highlight that for both the least squares contrast and the log-likelihood functional, computation such as gradient or loss evaluation are optimized and implemented in `C++` which serves the purpose of rapid computation, see Lacoste (2025) for more details.

5.2 Implementation details for the ERM step

For the classification step, our strategy consists in minimizing the convexified empirical risk defined in Equation (10). According to the definition of the constraint set of parameters defined in Equation (8), each coefficient must be positive. To ensure that each coefficient $a_{k,j,j'}$ remains in $[0, c]$, with $c > 0$, the minimization is done under inequality constraints and we use a projected gradient descent algorithm. Nevertheless, since this objective function is non-smooth and non-convex *w.r.t.* to the coefficients, its minimization requires particular care. In particular, the tuning of the step-size in the descent is very tricky [2]. On the other hand, adaptive gradient methods, such as Adagrad (see Duchi et al., 2011), have been widely used in large-scale optimization due to their ability to adjust the step size for each feature according to the geometry of the problem. In practice, Adagrad is known to be an efficient method in non-convex setting (in particular for training deep neural networks optimization, see Gupta et al. (2014)). In addition, some theoretical guarantees for the convergence of Adagrad for non-convex functions have been provided in the literature (see Ward et al., 2020; Wang et al., 2023). With this in mind, we use a parameter-free projected adaptive gradient descent method, in the inspiration of Adagrad, called Free Adagrad and introduced in Chzhen et al. (2023). Compared with the classical algorithm, its main advantage lies in the fact that it is adaptive to the distance between the initialization and the optimum, and to the sum of the square norm of the gradients. The initial starting point is chosen as the estimate given by the Lasso step, the initial guess for the distance between the starting point and the optimum is taken as $\gamma_0 = 0.1$ and we stop the descent after 1000 iterations if the stopping criteria described before is not fulfilled yet.

6 Numerical results

6.1 Simulation scheme

In this section, we present a scenario in which our Lasso-type estimator and our classifier are evaluated. Note that an additional scenario is provided in the supplementary material for further numerical investigations.

[2] Furthermore, classical method such as backtracking line-search with Armijo-Wolfe condition cannot be used due to the piece-wise constant nature of the projection operator (see Michael W. Ferry and Zhang (2023)).

MHP path generation. Concerning synthetic data generation, each path is simulated using cluster representation algorithm (see Møller and Rasmussen, 2005). This sampling procedure relies on the branching structure of the MHP, that can be viewed as Poisson cluster process. We consider the classical choice of exponential kernel $h(s) = \beta \exp(-\beta s)$ with $\beta = 3$.

Scenario. We consider block-diagonal matrix matrices A^*. In addition to self-exciting interaction, the block structure models interaction between a group of connected components. Coefficient values, which give the intensity of influence, are identical within each block, but vary from one to another. For a larger value of M, the blocks are expanded so that the parsimony rate remains the same for each value of M. The vector of exogenous intensity μ^* is chosen as constant for each component, meaning that spontaneous events occur in the same way for each individual. In Figure 2, a visual representation of the parameters is given, for $M = 25$, in the form of a heat map. In particular, the values of the coefficients of the matrix A^* are given by the color bar. We precise also that the sparsity rate, which is the percentage of zero-coefficients in the matrix, i.e. $|S^{*c}|/M^2$, is 85%.

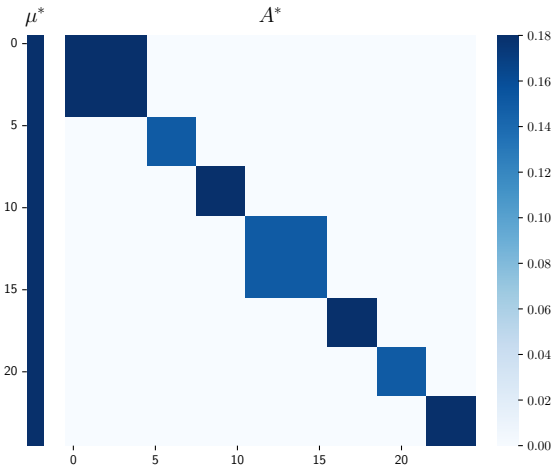

Figure 2: Visualization of $\theta^* = (\mu^*, A^*)$ for $M = 25$. The value of μ^* for each component is 0.4 and the sparsity rate of A^* is 85%.

To illustrate the classification task, we consider the 3-classes classification setting, i.e. $K = 3$. The three classes are created on the basis of the scenario described above. The blocks of different size are interchanged, as well as the values of the coefficients within them. The resulting classes are quite balanced and close from each other. Finally, the exogenous intensity is chosen to be the same for each of the three classes. We provide more details in the supplementary material.

6.2 Study of the support recovery

In this section we focus on the accuracy of the estimation of the active set of A^*. This study goes beyond the context of classification and focuses on a single adjacency matrix.

6.2.1 Metric of evaluation

To evaluate the efficiency of the Lasso procedure in recovering support, we introduce the following metrics, each of which measures the distance between the true parameter A^* and the estimation

$\widehat{A}$. First, the Hamming distance is defined as:

$$\text{HammDist}\left(A^*, \widehat{A}\right) = \frac{1}{M^2} \sum_{j,j'=1}^{M} \mathbb{1}_{\left\{A^*_{j,j'} \neq \widehat{A}_{j,j'}\right\}}$$

which measures the binary dissimilarity between the two matrices. Then, the relative error is defined as:

$$\text{RelErr}\left(A^*, \widehat{A}\right) := \frac{1}{M^2} \sum_{j,j'=1}^{M} \left(\frac{|A^*_{j,j'} - \widehat{A}_{j,j'}|}{|A^*_{j,j'}|} \mathbb{1}_{\left\{A^*_{j,j'} \neq 0\right\}} + |\widehat{A}_{j,j'}| \mathbb{1}_{\left\{A^*_{j,j'} = 0\right\}} \right)$$

which quantifies the averaged relative error between the true A^* and the estimated $\widehat{A}$ on non-zero entries of A^* and error on zeros entries. Additionally, the mean Kendall rank correlation is defined as:

$$\text{RankCorr}\left(A^*, \widehat{A}\right) := \frac{1}{M} \sum_{j=1}^{M} \tau(A^*_{j,\cdot}, \widehat{A}_{j,\cdot})$$

which corresponds to the averaged Kendall's rank correlation coefficient between each row of A^* and that of $\widehat{A}$. Finally, the Euclidean distance is defined as:

$$\text{EucliDist}\left(A^*, \widehat{A}\right) := \|A^* - \widehat{A}\|_F$$

which measures the straight-line distance between the two matrices.

6.2.2 Benchmark for support recovery

For the purpose of comparison, we introduce two benchmark approaches. The first one is ADM4 from Zhou et al. (2013). This procedure relies on the log-likelihood and imposes sparse and low-rank regularization on the adjacency matrix. We use the `tick` package implementation of this procedure, with Lasso regularization. Concerning the choice of the penalization constant, we explore a grid of value $C \in \{10, 100, 1000\}$ and select the one that minimizes the score. The second competitor is the nonparametric Lasso estimation procedure from Hansen et al. (2015), implemented in the `nm-bridge`[3] package. The procedure relies on a least-squares criterion combined with a Lasso penalty and involves the choice of the penalization constant γ, the size of bin δ and the number of bins K. We choose $\gamma = 1.0$, $\delta = 0.1$ and $K = 10$. In the following, the estimated adjacency matrices associated with these two methods will be denoted as $\widehat{A}_{\text{ADM4}}$ and $\widehat{A}_{\text{nm-bridge}}$.

6.2.3 Numerical results for support recovery on synthetic data

This section is devoted to the discussion of the obtained results of the Lasso procedure. These results are provided in Table 1 and Figure 3.

First, from Table 1, we observe that the Hamming distance between the true support S^* and the estimated one $\widehat{S}$ is close to zero in all settings, therefore our procedure is able to correctly recover the active set of A^*. This remains true even for small values of n and for high-dimensional networks, for which the Hamming distance is quite small. As expected, the larger n is, the better the support is reconstructed, whether in terms of Hamming distance or ℓ_2 distance. Thus, in addition to reconstructing the support more accurately, a gain is also made in terms of point parameter estimation, illustrating the theoretical result of support consistency and convergence

[3]Website https://neuromod.gitlabpages.inria.fr/nm-bridge

	M	HammDist			EucliDist		
		$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$
$\widehat{A}$	10	0.04 (0.02)	0.03 (0.02)	0.02 (0.02)	0.39 (0.06)	0.18 (0.03)	0.13 (0.02)
	25	0.05 (0.01)	0.04 (0.02)	0.05 (0.03)	0.88 (0.09)	0.42 (0.08)	0.28 (0.06)
	50	0.11 (0.01)	0.04 (0.01)	0.03 (0.01)	1.81 (0.12)	0.70 (0.05)	0.51 (0.05)
	100	0.14 (0.01)	0.10 (0.01)	0.08 (0.00)	2.07 (0.36)	1.15 (0.09)	0.81 (0.06)
$\widehat{A}_{\text{nm-bridge}}$	10	0.00 (0.00)	0.01 (0.01)	0.01 (0.01)	1.24 (0.07)	0.64 (0.03)	0.47 (0.02)
	25	0.07 (0.01)	0.00 (0.01)	0.01 (0.00)	2.30 (0.10)	1.37 (0.03)	1.02 (0.02)
	50	0.13 (0.01)	0.10 (0.00)	0.04 (0.00)	2.34 (0.10)	1.77 (0.03)	1.57 (0.02)
	100	0.15 (0.00)	0.12 (0.00)	0.11 (0.00)	1.88 (0.09)	1.87 (0.04)	1.58 (0.03)

Table 1: Lasso results averaged over 100 repetitions for different values of M and n (standard deviations in parentheses) for our estimator and the one implemented in `nm-bridge`. $T = 5$

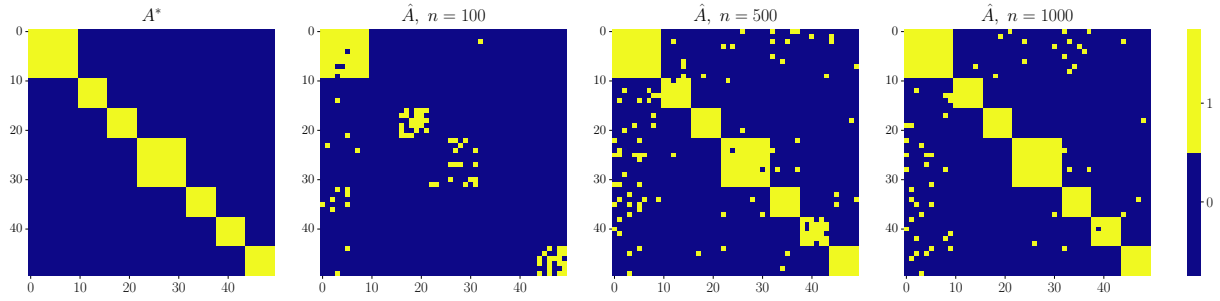

Figure 3: Illustration of our estimator: true support S^* and recovered support $\widehat{S}$, with $M = 50$ for $n \in \{100, 500, 1000\}$.

of the associated estimator established in Section 4. In particular, for large value of M, such as $M = 50$, a clear decrease in the Hamming distance is noticeable for increasing values of n. Finally, it is worth emphasizing that, the Lasso procedure is successful in recovering the underlying block structure of the interaction matrix A^*. This assertion is supported by Figure 3, which visually shows the convergence of the support to the actual structure as the number of observations increases.

We provide further details concerning the computational time of the procedure in the supplementary material. For more details regarding the complexity of the least-square functional, one refers to Lacoste (2025).

6.2.4 Real-world data experiments

In this section, we present a social network application to illustrate our Lasso procedure using the widely studied MemeTracker dataset (Leskovec et al., 2009).

This dataset tracks frequently cited phrases and memes, capturing the flow of information through hyperlinks linking various online sources. Each record includes a timestamp, the source website, and hyperlinks to other posts. In the context of this paper, each website is treated as a component (or node) of a MHP. For a given website, an event corresponds to the publication of a post containing a hyperlink to another site. In this network representation, an edge is assumed to exist between two websites if at least one hyperlink connects them. By leveraging hyperlinks to trace the flow of information, an approximation of the network's true interaction matrix A^* can be obtained, grounded in the causal relationships of MHP (see e.g. Bacry et al., 2015).

To align the data with our framework, we performed the following pre-processing steps. We use posts spanning from August 2008 to December 2008 and extract events from the top $M = 100$ media sites with the highest volume of published content. The event sequence is then segmented by grouping events on a daily basis, resulting in $n = 153$ repeated short-time sequences. Given the large dimensionality of the network relative to the number of observations, the MemeTracker illustration falls within our high-dimensional setting.

To ensure robustness, the evaluation is conducted 25 times on different random sub-samples of the dataset. More precisely, for each repetition, we consider 75% of the data, resulting in $n = 114 > M = 100$, which fits within our framework. From each sub-sample, we run the three procedures and obtain the corresponding estimates $\widehat{A}$, $\widehat{A}_{\text{ADM4}}$ and $\widehat{A}_{\text{nm-bridge}}$. Once the estimates are retrieved, we compute their evaluation metrics.

The obtained results are presented in Table 2. We observe that our procedure outperforms the benchmark approaches across all three metrics. Notably, the estimated interaction matrix produced by our method yields a very low Hamming distance, indicating effective recovery of the support of the ground truth A^*. This suggests that our approach captures the underlying influence network more accurately than the competing methods.

Metric	**HammDist**	**RelErr**	**RankCorr**
$\widehat{A}_{\text{ADM4}}$	0.83 (0.00)	75.93 (3.33)	0.138 (0.002)
$\widehat{A}_{\text{nm-bridge}}$	0.36 (0.01)	45.59 (0.96)	0.199 (0.003)
$\widehat{A}$	**0.06** (0.00)	**37.57** (0.17)	**0.567** (0.005)

Table 2: Averaged values of the metrics over the 25 repetitions (standard deviation in parentheses.)

6.3 Study of the classifier

6.3.1 Benchmark for classification

Let us detail here the different competitors which are compared with our classifier.

Simple plug-in strategy. A full plug-in strategy consists in using the estimators $\widehat{\theta}$ of the parameters, obtained by minimizing the least-squares contrast with Lasso penalty on the adjacency matrix given in Equation (7). Then, we plug $\hat{\theta}$ into the Bayes classifier formula. Consequently, the resulting classifier for a new observation $\mathcal{T}_T$ is

$$\widehat{g}_{\widehat{p},\widehat{\theta}}(\mathcal{T}_T) = \underset{k\in\mathcal{Y}}{\operatorname{argmax}} \frac{\widehat{p}_k e^{F_{\widehat{\mu}_k,\widehat{A}_k}(\mathcal{T}_T)}}{\sum_{k'=1}^{K} \widehat{p}_{k'} e^{F_{\widehat{\mu}_{k'},\widehat{A}_{k'}}(\mathcal{T}_T)}},$$

where $\widehat{p}$ is the estimated distribution of Y. This classifier is learned on the entire training sample $\mathcal{D}_n$ of size $2n$. This classifier is referred as `PI`.

Oracle on estimated support. We are also interested in another predictor, referred to as `OES` for *The Oracle on Estimated Support*, which is defined as follows

$$\widehat{g}_{\widehat{p},\theta^*_{\widehat{S}}}(\mathcal{T}_T) \in \underset{k\in[K]}{\operatorname{argmax}}\, \pi_{k,\widehat{p},\theta^*_{\widehat{S}}}(\mathcal{T}_T).$$

where

$$(\theta^*_{\widehat{S}})_{k,j,j'} := \begin{cases} \theta^*_{k,j,j'} & \text{if } (j,j') \in \widehat{S}_k \\ 0 & \text{otherwise} \end{cases}.$$

It corresponds to the best possible predictor that relies on the support recovered in the Lasso step. Note that if the true support is recovered by the Lasso step, then it exactly corresponds to the Bayes rule. By taking into account this predictor, we can quantify the effect of poor support recovery in terms of classification error, while evaluating the gain that could be obtained by an ERM step.

LSTM. We also consider a classifier that relies on Long Short-Term Memory (LSTM) network. To perform classification, we use a neural network consisting of a single LSTM layer with 64 units to capture temporal dependencies, followed by a dense output layer with softmax activation for multiclass prediction. For training, we use the sparse categorical crossentropy loss and the Adam optimizer, with `epochs = 50` and `batch_size = 16`. The implementation is carried out using the `TensorFlow` library.

Remark 5. *As a natural benchmark, we initially considered Random Forests (RF). However, given the pronounced temporal and cross-component dependencies present in the data, the RF algorithm demonstrates limited performance, even when its hyperparameters are optimized via cross-validation. The results are presented in the supplementary material, and the empirical classification error remains consistently close to 0.6, which is comparable to that of a classifier making random predictions.*

6.3.2 Evaluation scheme

Hereafter, we present the evaluation scheme that relies on Monte-Carlo repetitions. We fix $T = 5$, and $p^* \sim \mathcal{U}_{[3]}$. For each scenario described, each value of $M \in \{10, 25, 50\}$, and each value of $n \in \{400, 800, 2000\}$, we repeat independently 100 times the following steps.

1. Simulate the data set and split it into $\mathcal{D}_{n_{\text{train}}}$ and $\mathcal{D}_{n_{\text{test}}}$, with 75% allocated for training;
2. Based on $\mathcal{D}_{n_{\text{train}}}$, for each $k = 1, \ldots, K$, compute $\widehat{p}_k = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y^i=k\}}$;
3. Based on $\mathcal{D}_{n_{\text{train}}}$, perform the Lasso step:
   (a) For each $k \in [K]$, calibrate the penalization constant using EBIC_1 criteria by exploring values in the grid Δ. For each $\kappa \in \Delta$ do:
      i. using FISTA, compute $\widehat{\theta}_k$ the Lasso estimate with tuning parameter κ;
      ii. based on $\widehat{\theta}_k$, compute $\text{EBIC}_1(\kappa)$;

   and choose $\widehat{\kappa}_k \in \underset{\kappa\in\Delta}{\text{argmin}}\ \text{EBIC}_1(\kappa)$;
   (b) Given $(\widehat{\kappa}_k)_{k\in[K]}$, for each $k = 1, \ldots, K$ do:
      i. using FISTA, compute the Lasso estimates $\widehat{\theta}_k$ with tuning parameter $\widehat{\kappa}_k$;
      ii. get the estimated support $\widehat{S}_k = \left\{(j, j') \in [M],\ \ \widehat{\theta}_{k,j,j'} \neq 0\right\}$.
   (c) From $(\widehat{S}_k)_{k\in[K]}$ compute the classifier $\widehat{g}_{\texttt{OES}}$, from $(\widehat{\theta}_k)_{k\in[K]}$ compute the classifier $\widehat{g}_{\texttt{PI}}$
4. From $\mathcal{D}_{n_{\text{train}}}$, perform the ERM refitting step:
   (a) starting from $(\widehat{\theta}_k)_{k\in[K]}$ as the initial point, minimize the L_2-risk defined in Equation (9) using Free Adagrad to obtain $(\widehat{\theta}^{\text{R}}_k)_{k\in[K]}$;
   (b) from $\widehat{\theta}^{\text{R}}$ and $\widehat{p}$ build the classifiers $\widehat{g}_{\texttt{ERMLR}}$.
5. Based on $\mathcal{D}_{n_{\text{test}}} = \{(\mathcal{T}_T^{(i)}, Y^i), i = 1, \ldots, n_{\text{test}}\}$, evaluate the error rate of the classifiers `PI` and `ERMLR` using

$$\text{Err}_{\texttt{PI}} = \frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}} \mathbb{1}_{\{\widehat{g}_{\texttt{PI}}(\mathcal{T}_T^{(i)})\neq Y^{(i)}\}}, \quad \text{and} \quad \text{Err}_{\texttt{ERMLR}} = \frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}} \mathbb{1}_{\{\widehat{g}_{\texttt{ERMLR}}(\mathcal{T}_T^{(i)})\neq Y^{(i)}\}};$$

6.3.3 Numerical results for classification on synthetic data

This section is devoted to the discussion of the obtained results of the `ERMLR` procedure. These results are provided in Table 3, and Figure 4.

First, from Table 3, we can see that the `ERMLR` is close to the Bayes classifier in terms of error rate, in both scenarios and for each value of M. In particular, note that for $n = 2000$, its error rate is almost equal to that of the Bayes classifier. In fact, as expected the greater the number of data, the closer the classifier comes to the Bayes classifier, which illustrate the consistency of the `ERMLR` procedure established in Section 4. This decreasing trend in the error rate of `ERMLR` is illustrated in Figure 4 for $M = 50$. As shown, the slope of the empirical error (in blue, plain line) closely matches that of the theoretical bound (in red dotted line) on the log-log plot, with only a small gap attributable to conservative constants. This provides empirical support for the theoretical result.

Another important point is the comparison with the `PI` classifier as a benchmark. Overall, it can be seen that the `PI` exhibits good performance. This can be explained by the fact that recovering the true support structure is sufficient for accurate class prediction. On the other hand, poor support recovery also impacts the performance of the `ERMLR` predictor. This gap can be quantified with the `OES` oracle classifier, which gives the gain that could be obtained by an ERM step. For these reasons, it is not expected to see a big gap between the two. Nevertheless, it is

M	Bayes	OES	PI	ERMLR	LSTM
10	0.132	0.132 (0.035)	0.149 (0.035)	**0.147** (0.035)	0.317 (0.068)
25	0.223	0.297 (0.055)	0.366 (0.049)	**0.350** (0.051)	0.561 (0.046)
50	0.081	0.251 (0.054)	0.246 (0.047)	**0.227** (0.044)	0.450 (0.061)

(a) $n = 400$

M	Bayes	OES	PI	ERMLR	LSTM
10	0.132	0.134 (0.027)	0.144 (0.026)	**0.144** (0.028)	0.304 (0.063)
25	0.223	0.237 (0.027)	0.298 (0.030)	**0.293** (0.031)	0.525 (0.039)
50	0.081	0.120 (0.024)	0.176 (0.026)	**0.163** (0.027)	0.412 (0.047)

(b) $n = 800$

M	Bayes	OES	PI	ERMLR	LSTM
10	0.132	0.137 (0.015)	0.141 (0.015)	**0.141** (0.014)	0.319 (0.103)
25	0.223	0.229 (0.020)	0.263 (0.022)	**0.262** (0.020)	0.502 (0.026)
50	0.081	0.086 (0.012)	0.126 (0.013)	**0.118** (0.013)	0.345 (0.062)

(c) $n = 2000$

Table 3: Empirical error averaged over 100 repetitions for each classifier for different values of M and n (standard deviations in parentheses). 75% for train and 25% for test. $T = 5$

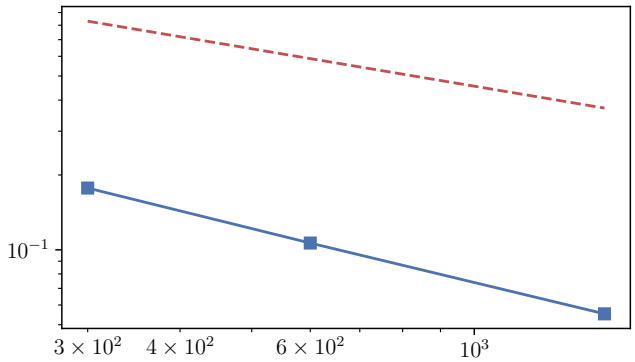

Figure 4: Illustration of the ERMLR procedure for $M = 50$: log-log plot with in red the theoretical rate $\mathcal{O}\left(\sqrt{n^{-1}(M + s^*)}\right)$ as a function of n; in the blue the empirical counterpart of the excess risk $\mathcal{R}(\hat{g}) - \mathcal{R}(g^*)$. The empirical errors of the classifier are averaged over 100 repetitions.

worth noting that, a significant gain by the ERM refitting step can be observed. This assertion is particularly supported by the results obtained for $M = 50$. This suggests that, for some particular structures, a refitting step leading to a finer point estimate of the parameters is relevant and leads to better performance in practice.

Finally, it can be observed that LSTMs exhibit relatively high error rates, particularly as the value of M increases. This may be due to the fact that the data exhibit very complex temporal and cross-component dependencies that are difficult to capture.

7 Discussion

The present work offers some solution to deal with event based data using Hawkes processes. In particular, we propose an innovative classification algorithm tailored to classify Multivariate Hawkes Processes paths in high-dimension. For each class, a first step is dedicated to the sparse estimation of the support of the adjacency matrix. Then, in a second step, we build a classifier that takes of advantage of the estimated support. Specifically, the resulting classifier is based on the minimization of a ERM criterion. We establish rates of convergence for both estimated support and classification algorithm. Finally, we illustrate the numerical performance of our procedure through a comprehensive simulation study.

A possible guideline for further research is to consider a more challenging model by including inhibition interaction. From a theoretical aspect, it may be tricky since adding inhibition effect induces complication due to the non linearity of the underlying intensity function. In particular, providing a closed form of the compensator is a key aspect to compute the least-square contrast or the likelihood function. The work of Bonnet et al. (2022) and Bonnet et al. (2023) should form a theoretical basis for this future work. From a practical point of view, a procedure which is able to deal with inhibition, may be applied to generalize the work of Denis et al. (2024). Indeed, the use of MHP allows to model simultaneously different species echolocation calls and then the effects of inter-species cooperation. Furthermore, adding inhibition effects, potentially translates the ecological aspect of inter-species competition.

Another direction could be to investigate a penalized ERM classifier. It would allow to deal with the high-dimensional setting without the prior Lasso step. Indeed, this procedure relies on a global penalized criterion dedicated to the classification task. This direction is left for further investigations.

Acknowledgements

This work has been supported by the Chaire "Modélisation Mathématique et Biodiversité" of Veolia-École polytechnique-Museum national d'Histoire naturelle-Fondation X, through a Ph.D. scholarship. Some of the computations have been performed under the project "hawkes" on the HPC facility Cholesky operated by the IDCS/École polytechnique. This work is also part of the 2022 DAE 103 EMERGENCE(S) - PROCECO project supported by Ville de Paris. Finally, the authors thank Vincent Rivoirard for fruitful discussions and the referees for their relevant comments.

References

Abramovich, F. and Grinshtein, V. (2018). High-dimensional classification by sparse logistic regression. *IEEE Transactions on Information Theory*, 65(5):3068–3079.

Amorino, C., Pina, F., and Podolskij, M. (2024). Sampling effects on lasso estimation of drift functions in high-dimensional diffusion processes. *arXiv preprint arXiv:2408.08638*.

Bacry, E., Bompaire, M., Deegan, P., Gaïffas, S., and Poulsen, S. V. (2018). tick: a python library for statistical learning, with an emphasis on hawkes processes and time-dependent models. *Journal of Machine Learning Research*, 18(214):1–5.

Bacry, E., Bompaire, M., Gaïffas, S., and Muzy, J.-F. (2020). Sparse and low-rank multivariate hawkes processes. *Journal of Machine Learning Research*, 21(50):1–32.

Bacry, E., Mastromatteo, I., and Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005.

Baouan, A., Bismuth, E., Bohbot, A., Coustou, S., Lacome, M., and Rosenbaum, M. (2022). What should clubs monitor to predict future value of football players. *arXiv preprint arXiv:2212.11041*.

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

Bonnet, A., Dion-Blanc, C., Gindraud, F., and Lemler, S. (2022). Neuronal network inference and membrane potential model using multivariate hawkes processes. *Journal of Neuroscience Methods*, 372:109550.

Bonnet, A., Martinez Herrera, M., and Sangnier, M. (2023). Inference of multivariate exponential hawkes processes with inhibition and application to neuronal activity. *Statistics and Computing*, 33(4):91.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cai, B., Zhang, J., and Guan, Y. (2024). Latent network structure learning from high-dimensional multivariate point processes. *Journal of the American Statistical Association*, 119(545):95–108.

Carstensen, L., Sandelin, A., Winther, O., and Hansen, N. (2010). Multivariate hawkes process models of the occurrence of regulatory elements. *BMC bioinformatics*, 11:1–19.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61.

Chzhen, E., Giraud, C., and Stoltz, G. (2023). Parameter-free projected gradient descent. *arXiv preprint arXiv:2305.19605*.

Chzhen, E., Hebiri, M., and Salmon, J. (2019). On Lasso refitting strategies. *Bernoulli*, 25(4A):3175–3200.

Daley, D. and Vere-Jones, D. (2003). Basic properties of the poisson process. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*, pages 19–40.

Denis, C., Dion-Blanc, C., Lacoste, R. E., Sansonnet, L., and Bas, Y. (2024). Bats monitoring: A classification procedure of bats behaviors based on hawkes processes. *Journal of the Royal Statistical Society Series C: Applied Statistics*.

Denis, C., Dion-Blanc, C., and Sansonnet, L. (2022). Multiclass classification for hawkes processes. In *Uncertainty in Artificial Intelligence*, pages 539–547. PMLR.

Ditlevsen, S. and Löcherbach, E. (2017). Multi-class oscillating systems of interacting neurons. *Stochastic Processes and their Applications*, 127(6):1840–1869.

Donnet, S., Rivoirard, V., and Rousseau, J. (2020). Nonparametric bayesian estimation for multivariate hawkes processes. *The Annals of statistics*, 48(5):2698–2727.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Eichler, M., Dahlhaus, R., and Dueck, J. (2017). Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242.

Embrechts, P., Liniger, T., and Lin, L. (2011). Multivariate hawkes processes: an application to financial data. *Journal of Applied Probability*, 48(A):367–378.

Gupta, M. R., Bengio, S., and Weston, J. (2014). Training highly multiclass classifiers. *The Journal of Machine Learning Research*, 15(1):1461–1492.

Hansen, N. R., Reynaud-Bouret, P., and Rivoirard, V. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity The Lasso and Generalizations*. Chapman & Hall/CRC.

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

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. (2019). Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4):917–963.

Kim, S., Putrino, D., Ghosh, ., and Brown, E. (2011). A granger causality measure for point process models of ensemble neural spiking activity. *PLoS computational biology*, 7(3):e1001110.

Lacoste, R. E. (2025). Sparklen: A Statistical Learning Toolkit for High-Dimensional Hawkes Processes in Python. *Preprint arXiv:2502.18979*.

Lambert, R., Tuleau-Malot, C., Bessaih, T., Rivoirard, V., Bouret, Y., Leresche, N., and Reynaud-Bouret, P. (2018). Reconstructing the functional connectivity of multiple spike trains using Hawkes models. *Journal of Neuroscience Methods*, 297:9–21.

Leblanc, T. (2024). Exponential moments for hawkes processes under minimal assumptions. *Electronic Communications in Probability*, 29:1–11.

Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 497–506. ACM.

Limnios, M. and Hansen, N. R. (2025). Selective review of penalized learning methods for event processes. *Stochastic Modeling and Statistical Methods*, pages 159–189.

Lotz, A. (2024). A sparsity test for multivariate Hawkes processes. *arXiv preprint arXiv:2405.08640*.

Lukasik, M., Srijith, P., Vu, D., Bontcheva, K., Zubiaga, A., and Cohn, T. (2016). Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 393–398.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462.

Michael W. Ferry, Philip E. Gill, E. W. and Zhang, M. (2023). A class of projected-search methods for bound-constrained optimization. *Optimization Methods and Software*, 0(0):1–30.

Mohler, G., Short, M., Brantingham, P., Schoenberg, F., and Tita, G. (2011). Self-exciting point process modeling of crime. *Journal of the american statistical association*, 106:100–108.

Møller, J. and Rasmussen, J. G. (2005). Perfect simulation of hawkes processes. *Advances in applied probability*, 37(3):629–646.

Nicvert, L., Donnet, S., Keith, M., Peel, M., Somers, M., Swanepoel, L., Venter, J., Fritz, H., and Dray, S. (2024). Using the multivariate h awkes process to study interactions between multiple species from camera trap data. *Ecology*, page e4237.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27.

Reynaud-Bouret, P. and Schbath, S. (2010). Adaptive estimation for hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5):2781–2822.

Reynaud-Bouret, P., Tuleau-Malot, C., Rivoirard, V., and Grammont, F. (2013). Spike trains as (in) homogeneous poisson processes or hawkes processes: non-parametric adaptive estimation and goodness-of-fit tests. *Journal of Mathematical Neuroscience*.

Spaziani, S., Girardeau, G., Bethus, I., and Reynaud-Bouret, P. (2023). Heterogeneous multiscale multivariate autoregressive model: Existence, sparse estimation and application to functional connectivity in neuroscience. *Annals of Statistics*.

Sulem, D., Rivoirard, V., and Rousseau, J. (2024). Bayesian estimation of nonlinear Hawkes processes. *Bernoulli*, 30(2):1257–1286.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. and Wasserman, L. (2017). Sparsity, the Lasso, and Friends. Technical report, Carnegie Mellon University.

Tondulkar, R., Dubey, M., Srijith, P., and Lukasik, M. (2022). Hawkes process classification through discriminative modeling of text. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory*, 52(3):1030–1051.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l1 constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202.

Wang, B., Zhang, H., Ma, Z., and Chen, W. (2023). Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 161–190. PMLR.

Ward, R., Wu, X., and Bottou, L. (2020). Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(219):1–30.

Zhang, R., Walder, C., Rizoiu, M.-A., and Xie, L. (2018). Efficient non-parametric bayesian hawkes processes. *arXiv preprint arXiv:1810.03730*.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85.

Zhou, K., Zha, H., and Song, L. (2013). Learning triggering kernels for multi-dimensional hawkes processes. In *International Conference on Machine Learning*, pages 1301–1309.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Appendix

In the following we gathers the proofs of the theoretical results of the paper. It is organized as follows. Appendix A provides useful technical results. The proof of the closed-form expression of the Bayes classifier is established in Appendix B. The proof of the support recovery result is given in Appendix C. Finally, the rate of convergence of the `ERMLR` *algorithm is proved in Appendix D.*

Throughout the proofs, the notation C refers to a generic positive constant, which may differ from line to line. In particular, this generic constant C does not depend on n or on the dimension M. However, it may depend on the other parameters.

A Technical results

Proposition A.1. *For any classifier $g \in \mathcal{G}$, we have*

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \mathbb{E}\left[\sum_{1\leq i\neq k\leq K} |\pi_i^*(\mathcal{T}_T) - \pi_k^*(\mathcal{T}_T)| \mathbb{1}_{\{g^*(\mathcal{T}_T)=i, g(\mathcal{T}_T)=k\}}\right].$$

Proof. This result is established by Denis et al. (2022). Let $g \in \mathcal{G}$ a classifier and $\mathcal{T}_T$ an observation of the process on $[0, T]$. We observe that

$$\mathcal{R}(g) = \mathbb{E}\left[\mathbb{1}_{\{g(\mathcal{T}_T)\neq Y\}}\right] = 1 - \mathbb{E}\left[\mathbb{1}_{\{g(\mathcal{T}_T)=Y\}}\right] = 1 - \mathbb{E}\left[\pi^*_{g(\mathcal{T}_T)}\right].$$

Therefore, from the above equation and the definition of the Bayes classifier g^*, we get

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \mathbb{E}\left[\left|\pi^*_{g^*(\mathcal{T}_T)} - \pi^*_{g(\mathcal{T}_T)}\right|\right].$$

Since for each $g \in \mathcal{G}$, $g(\mathcal{T}_T) = \sum_{k=1}^K k\mathbb{1}_{\{g(\mathcal{T}_T)=k\}}$, the above equation yields the result. □

Lemma A.1. *Let $A \in \mathbb{R}^{d\times d}$ (symmetric), and $X \in \mathbb{R}^d$. Then,*

$$\|AX\|_\infty \leq \sqrt{d}\rho(A)\|X\|_\infty$$

Proof. We start by remarking that

$$\|AX\|_\infty \leq \|AX\|_2.$$

Then, by using the definition of the spectral norm as a matrix norm induced by the vector 2-norm, we have

$$\|AX\|_2 \leq \|A\|_2 \|X\|_2.$$

Since the matrix A is assumed to be symmetric, the following equality holds

$$\rho(A) = \|A\|_2.$$

Using that, and remarking that $\|X\|_2 \leq \sqrt{d}\|X\|_\infty$, we get

$$\|AX\|_2 \leq \sqrt{d}\rho(A)\|X\|_\infty.$$

Combining all these inequalities, we finally obtain the desired result. □

Lemma A.2 (Hoeffding)**.** *Let $B \sim \mathcal{B}(n, p)$, with $p \in (0, 1)$. We then have for all $t > 0$ and $n > \frac{t}{p}$,*

$$\mathbb{P}(B \leq t) \leq \exp\left(-2n(p - t/n)^2\right).$$

B Proof for Bayes classifier

We first denote for all $k \in \mathcal{Y}$

$$\Phi_t^k := \frac{\mathrm{d}\mathbb{P}_k|_{\mathcal{F}_t^N}}{\mathrm{d}\mathbb{P}_0|_{\mathcal{F}_t^N}},$$

with $\mathcal{F}_T^N := \sigma\left(\mathcal{T}_T\right) = \sigma\left(N_t, 0 \leq t \leq T\right)$. We classically obtain:

$$\log(\Phi_t^k) = -\sum_{j=1}^{M} \int_0^t (\lambda_{k,j}^*(s) - 1)\, \mathrm{d}s + \int_0^t \log(\lambda_{k,j}^*(s))\, \mathrm{d}N_j(s),$$

by writing *w.r.t.* a Poisson process measure of intensity 1 (see Chapter 13 of Daley and Vere-Jones, 2003). Thus, for $t \geq 0$, we have the following equation for the mixture measure

$$d\mathbb{P}|_{\mathcal{F}_t^N} = \sum_{k=1}^{K} p_k^* \, \mathrm{d}\mathbb{P}_k|_{\mathcal{F}_t^N} = \sum_{k=1}^{K} p_k^* \Phi_t^k \, \mathrm{d}\mathbb{P}_0|_{\mathcal{F}_t^N}$$

and then

$$\frac{\mathrm{d}\mathbb{P}_k|_{\mathcal{F}_t^N}}{\mathrm{d}\mathbb{P}|_{\mathcal{F}_t^N}} = \frac{p_k^* \Phi_t^k \, \mathrm{d}\mathbb{P}_0|_{\mathcal{F}_t^N}}{\sum_{j=1}^K p_j^* \Phi_t^j \, \mathrm{d}\mathbb{P}_0|_{\mathcal{F}_t^N}} = \frac{p_k^* \Phi_t^k}{\sum_{j=1}^K p_j^* \Phi_t^j}.$$

Finally, by using (4), it comes $\pi_k^*\left(\mathcal{T}_T\right) = \frac{p_k^* \mathrm{e}^{F_k^*}}{\sum_{j=1}^K p_j^* \mathrm{e}^{F_j^*}}$, that concludes the proof.

C Proofs for support recovery

In this section, we gather the proof of the result provided in Section 4.1. We first recall and introduce the main notations for the proof of the main result in Section C.1. Then, in section C.2 we establish a Bernstein lemma. This lemma is the cornerstone of the proof of the support recovery which is given in Section C.3.

C.1 Notations

We recall that the learning sample is $D_n = \left\{\left(\mathcal{T}_T^{(1)}, Y^{(1)}\right), \ldots, \left(\mathcal{T}_T^{(n)}, Y^{(n)}\right)\right\}$. Let $k \in [K]$ be a fixed integer. Throughout this section, all the results are established for a generic class k. Let us define the random variables

$$n_k = \sum_{i=1}^{n} \mathbb{1}_{Y^{(i)}=k}.$$

Hence $n_k \sim \mathcal{B}(n, p_k^*)$. We also recall that $\min_{k\in[K]} p_k^* \geq p_0 > 0$.

For sake of simplicity, we remove the dependency *w.r.t.* k. To sum up, our parameters of interests are μ, A, and we have at our disposal a sample of (random) size n_k. In the rest of this section, we work **conditional on** $\{n_k \geq 1\}$.

For observation $i \in [n_k]$ and coordinate $j' \in [M]$, we consider $M_{j'}^{(i)}$ the martingale associated to the counting process $N_{j'}^{(i)}$ through the Doob-Meyer decomposition. We then denote $\mathrm{d}M(t) = \left(\mathrm{d}M_{j'}^{(i)}(t)\right)_{j',i} \in \mathbb{R}^{M\times n_k}$, and define the random martingale matrix Z as

$$Z := \int_0^T \left(\mathrm{d}M(t)\mathbb{H}_t\right)'.$$

Besides, the j-th column of Z is denoted by Z_j. Therefore, for $j, j^{'} \in \{0, \dots, M\} \times [M]$, the main term of Z is the continuous-time martingale,

$$Z_{j,j'} = \sum_{i=1}^{n_k} \int_0^T H_j^{(i)}(t)\, \mathrm{d}M_{j'}^{(i)}(t), \tag{12}$$

where

$$H_j^{(i)}(t) := \int_0^{t-} h(t-s)\, \mathrm{d}N_j^{(i)}(s), \text{ for } j \neq 0, \quad \text{and } H_0^{(i)} \equiv 1.$$

C.2 A Bernstein lemma

Lemma C.1 (Bernstein Lemma)**.** *Assume that* $n \geq \dfrac{2}{p_0}$. *Let us define the event*

$$\Omega_n := \left\{ \frac{1}{n_k} \max_{j,j'} \left| Z_{j,j'} \right| \leq C \frac{\log^3\left(nM^2\right)}{\sqrt{n}} \right\} \bigcap \left\{ n_k \geq \frac{np_k^*}{2} \right\},$$

where $Z_{j,j'}$ *is given in Equation* (12). *There exists* $C_{\|h\|_\infty, p_0} > 0$, *such that* $\mathbb{P}\left(\Omega_n\right) \geq 1 - \dfrac{C_{\|h\|_\infty, p_0}}{n}$.

Proof. For clarity of presentation, the proof is divided in two steps.

First step. In this step we work on the event $\{n_k \geq 1\}$ and conditional on $\mathbb{1}_{\{Y^{(1)}=k\}}, \dots, \mathbb{1}_{\{Y^{(n)}=k\}}$. For $j, j' \in \{0, \dots, M\} \times [M]$, we apply Theorem 4 in Bacry et al. (2020) to the real valued random variable $Z_{j,j'}$. For clarity, we consider the same notations as in Bacry et al. (2020).

To this end, for a fixed $j, j' \in \{0, \dots, M\} \times [M]$ and $t \in [0; T]$, we define the tensor (see Bacry et al. (2020) for its definition and related properties) $\mathbb{T}_t$ of shape $1 \times 1 \times M \times n_k$ as follows

$$(\mathbb{T}_t)_{1,1,k,\ell} = \begin{cases} H_j^{(\ell)}(t) \text{ if } k = j' \\ 0 \text{ else,} \end{cases} \tag{13}$$

for $k \in [M]$ and $\ell \in [n_k]$. We also recall that the matrix $\mathrm{d}M(t)$ is defined by the main term $\mathrm{d}M(t)_{j',i} = \mathrm{d}M_{j'}^{(i)}(t)$. According to Bacry et al. (2020) we have that $Z_{j,j'} = Z_{\mathbb{T}}(T) \in \mathbb{R}$ defined by

$$Z_{\mathbb{T}_t}(T) = \int_0^T \mathbb{T}_t \circ dM_t$$

satisfies

$$Z_{\mathbb{T}_t}(T) = \sum_{k=1}^{M} \sum_{i=1}^{n_k} \int_0^T (\mathbb{T}_t)_{1,1,k,i}\, \mathrm{d}M_{k,i}(t) = \sum_{i=1}^{n_k} \int_0^T H_j^{(i)}(t)\, \mathrm{d}M_{j'}^{(i)}(t). \tag{14}$$

Furthermore, we observe that since the tensor $\mathbb{T}_t$ is symmetric we have

$$\widehat{V}_{\mathbb{T}}(t) := \int_0^t \mathbb{T}_s^2 \circ \mathrm{d}N_s = \sum_{i=1}^{n_k} \int_0^t \left(H_j^{(i)}(s) \right)^2 \mathrm{d}N_{j'}^{(i)}(s),$$

and

$$b_{\mathbb{T}_t} := \sup_{0 \leq s \leq t} \max(\|\mathbb{T}_s\|_{\mathrm{op},\infty} \|\mathbb{T}_s'\|_{\mathrm{op},\infty}) = \sup_{0 \leq s \leq t} \max_{i=1,\dots,n_k} \left| H_j^{(i)}(s) \right|$$

which both depend on (j, j').

Applying Theorem 4 of Bacry et al. (2020) on the event $\{n_k \geq 1\}$ and conditional on $\mathbb{1}_{\{Y^{(1)}=k\}}$, $\ldots$, $\mathbb{1}_{\{Y^{(n)}=k\}}$, we then obtain that for $x > 0$ with probability at least $1 - C\exp(-x)$ the following holds

$$\left|Z_{j,j'}\right| \leq 2\sqrt{\lambda_{\max}(\widehat{V}_{\mathbb{T}_T})(x + \ell_x(T))} + c(x + \ell_x(T))\left(1 + b_{\mathbb{T}_T}\right), \tag{15}$$

since for all $(j, j^{'}) \in [M] \times \{0, \ldots M\}$,

$$\lambda_{\max}\left(\widehat{V}_{\mathbb{T}}(T)\right) \leq \widehat{V}_{\infty} := n_k \max_{i,j}\left(H_j^{(i)}(T)\right)^2 \max_{i,j} N_j^{(i)}(T), \tag{16}$$

and

$$b_{\mathbb{T}_T} \leq b_{\infty} := \max_{i,j}\left|H_j^{(i)}(T)\right|, \tag{17}$$

from Equation (15), setting $x = \log(nM^2)$, with an union bound on $j, j^{'}$ we obtain that the event

$$\mathcal{Z} := \left\{\max_{j,j'}\left|Z_{j,j'}\right| \leq 2\sqrt{\widehat{V}_{\infty}\left(\log(nM^2) + \ell_{\infty}\right)} + c\left(\log(nM^2) + \ell_{\infty}\right)\left(1 + b_{\infty}\right)\right\},$$

with

$$\ell_{\infty} = 2\log\log\left(\frac{4\widehat{V}_{\infty}}{\log(nM^2)} \vee 2\right) + 2\log\log\left(4b_{\infty} \vee 2\right), \tag{18}$$

satisfies

$$\mathbb{1}_{\{n_k \geq 1\}}\mathbb{P}\left(\mathcal{Z}^c | \mathbb{1}_{\{Y^{(1)}=k\}}, \ldots, \mathbb{1}_{\{Y^{(n)}=k\}}\right) \leq \mathbb{1}_{\{n_k \geq 1\}}\frac{C}{n} \leq \frac{C}{n}.$$

From the above inequality, we deduce that

$$\begin{aligned}
\mathbb{P}\left(\mathcal{Z}^c\right) &= \mathbb{P}\left(\mathcal{Z}^c, n_k \geq 1\right) + \mathbb{P}\left(\mathcal{Z}^c, n_k = 0\right) \\
&\leq \frac{C}{n} + \mathbb{P}\left(n_k = 0\right) \\
&\leq \frac{C}{n} + \exp(n\log(1 - p_k^*)) \\
&\leq \frac{C}{n} + \exp(n\log(1 - p_0). \qquad (19)
\end{aligned}$$

In particular, we have used here that $n \geq 2/p_0$. Thus, for n large enough we have $\mathbb{P}\left(\mathcal{Z}^c\right) \leq \frac{C}{n}$.

Second step. In this step, we provide a bound for $\widehat{V}_{\infty}$, b_{∞}, and ℓ_{∞} respectively defined in Equation (16), (17) and (18). To this end, we introduce the event

$$\Omega := \left\{n_k \geq \frac{np_k^*}{2}\right\} \bigcap \left\{\mathbb{1}_{\{n_k \geq 1\}} \max_{i,j} N_j^{(i)}(T) \leq \log^{5/3}(Mn)\right\}. \tag{20}$$

Note that, in view of the definition of $H_{j'}^{(i)}$, we have that on the event $\{n_k \geq 1\}$, we have

$$\widehat{V}_{\infty} \leq \max\left(n_k \left\|h\right\|_{\infty} \max_{i,j}\left(N_j^{(i)}(T)\right)^3, n_k \max_{i,j}(N_j^{(i)}(T)\right) \leq C_{\left\|h\right\|_{\infty}} n_k \log^5(Mn).$$

With the same idea, we have that $b_{\infty} \leq C_{\left\|h\right\|_{\infty}} \log^{5/3}(n)$. Finally, we observe that $\ell_{\infty} \leq 2\log(nM^2)$ (as $M \geq 2$).

Hence, on the event $\Omega \cap \mathcal{Z}$, it holds that $n_k \geq 1$ (since $n \geq \frac{2}{p_0}$), and by the definition of Ω in Equation (20), $n_k \geq np_k^*/2$ allow to write:

$$\frac{1}{n_k} \max_{j,j'} \left| Z_{j,j'} \right| \leq C \frac{\log^3 \left(nM^2\right)}{\sqrt{n_k}} \leq C \frac{\log^3 \left(nM^2\right)}{\sqrt{np_k^*}} \leq C \frac{\log^3 \left(nM^2\right)}{\sqrt{np_0}}.$$

To conclude the proof, since $\mathbb{P}\left(\Omega_n^c\right) \leq \mathbb{P}\left((\mathcal{Z} \cap \Omega)^c\right)$, it remains to control $\mathbb{P}\left((\mathcal{Z} \cap \Omega)^c\right)$.

Conditional on $\mathbb{1}_{\{Y^{(1)}=k\}}, \ldots, \mathbb{1}_{\{Y^{(n)}=k\}}$, on the event $\{n_k \geq 1\}$, by applying the sub-exponential property of $N_j^{(i)}$, Assumption 4, and Proposition 2.7.1 in Vershynin (2018), we obtain

$$\begin{aligned} \mathbb{P}\left(\max_{i,j} N_j^{(i)}(T) > \log^{5/3}(Mn) \right) &\leq n_k \exp\left(-c \log^{5/3}(nM) \right) \\ &\leq \frac{1}{n}. \end{aligned}$$

Therefore, from Lemma A.2,

$$\begin{aligned} \mathbb{P}\left(\Omega^c\right) &\leq \frac{1}{n} + \mathbb{P}\left(n_k \leq \frac{np_k^*}{2} \right) \leq \frac{1}{n} + \exp\left(-n \frac{p_0}{2} \right) \\ &\leq \frac{C}{n}. \end{aligned}$$

Finally, combining the last equation with Equation (19), we deduce that

$$\mathbb{P}\left((\mathcal{Z} \cap \Omega)^c\right) \leq \frac{C}{n},$$

which yields the result.

□

C.3 Proof of the main result Theorem 1

Since, under Assumptions 6 and 7, we have $\mathbb{P}(\Omega_{\mathrm{MI}} \cap \Omega_{\mathrm{ME}}) = 1$, and then $\mathbb{P}\left(\Omega_{\mathrm{MI}} \cap \Omega_{\mathrm{ME}} \cap \Omega_n\right) = \mathbb{P}(\Omega_n)$. Thus, throughout the proof, we work on the event $\Omega_{\mathrm{MI}} \cap \Omega_{\mathrm{ME}} \cap \Omega_n$, with

$$\Omega_n := \left\{ \frac{1}{n_k} \max_{j,j'} \left| Z_{j,j'} \right| \leq C \frac{\log^3 \left(nM^2\right)}{\sqrt{n}} \right\} \bigcap \left\{ n_k \geq \frac{np_k^*}{2} \right\}.$$

Note that on the event Ω_n, since $n \geq \frac{2}{p_0}$, the random variable n_k satisfies $n_k \geq 1$. Also, let us emphasize that the second event in the definition of Ω_n allows us to get rid of n_k, which is random, in the bound and to replace it by n.

The proof follows the *primal-dual witness method* as in Hastie et al. (2015) Chapter 11, and goes in several steps. Let us consider the penalized contrast

$$\mathcal{C}(\theta) := R_{T,k}(\theta) + \kappa \sum_{j=1}^{M} \sum_{j'=1}^{M} |\theta_{j,j'}|. \tag{21}$$

An element z of the subgradient of $\mathcal{C}$ at some point θ writes as follows

$$\nabla R_{T,k}(\theta) + \kappa z,$$

where the concatenated vector z is $z = (z_1, \ldots, z_M)'$ with $z_{j,0} = 0$ and $z_{j,j'} = \mathrm{sign}(\theta_{j,j'})$ for $j' \geq 1$ (with the convention that $\mathrm{sign}(0) \in [-1, 1]$). We say that a pair $(\widehat{\theta}, \widehat{z})$ is optimal if it satisfies the following zero-subgradient equation

$$\nabla R_{T,k}(\widehat{\theta}) + \kappa \widehat{z} = 0. \tag{22}$$

First step. We first build an "oracle" pair $(\widehat{\theta}, \widehat{z})$ that satisfies Equation (22) and such that $\widehat{\theta}_{S^{*c}} = 0$. First we define $\widehat{\theta}$, and $\widehat{z}_{S^*}$ as follows.

1. $\widehat{\theta}_{S^{*c}} = 0$,
2. $\widehat{\theta}_{S^*} \in \underset{\theta_{S^*}}{\operatorname{argmin}} \widetilde{R}_{T,k}(\theta_{S^*}) + \kappa \sum_{j=1}^{M} \sum_{j' \in S^*_{\theta_j} \setminus \{0\}} |\theta_{j,j'}|$, where

$$\begin{aligned}\widetilde{R}_{T,k}(\theta_{S^*}) \;&:=\; \frac{1}{n_k T} \sum_{i=1}^{n_k} \sum_{j=1}^{M} \int_0^T \left(\sum_{j' \in S^*_{\theta_j}} \theta_{j,j'} H^{(i)}_{j'}(t) \right)^2 \mathrm{d}t \\ &\quad -2 \int_0^T \left(\sum_{j' \in S^*_{\theta_j}} \theta_{j,j'} H^{(i)}_{j'}(t) \right) \mathrm{d}N^{(i)}_j(t).\end{aligned}$$

In view of the above conditions, since $\widehat{\theta}_{S^*}$ is a minimizer, we have for each $j \in [M]$

$$\left(\nabla R_{T,k}(\widehat{\theta}) \right)_{S^*_{\theta_j}} + \kappa \widehat{z}_{S^*_{\theta_j}} = 0.$$

We then have to build for each $j \in [M]$, $\widehat{z}_{S^{*c}_{\theta_j}}$ such that

$$\left(\nabla R_{T,k}(\widehat{\theta}) \right)_{S^{*c}_{\theta_j}} + \kappa \widehat{z}_{S^{*c}_{\theta_j}} = 0.$$

Hence, from the above equations and from the notation given in Equation (11), we deduce that $(\widehat{\theta}, \widehat{z})$ must satisfies

$$\frac{2}{n_k} \mathbb{H}_{S^{*c}_{\theta_j}, S^*_{\theta_j}} \left(\widehat{\theta}_j - \theta^*_j \right)_{S^*_{\theta_j}} - \frac{2}{n_k} (Z_j)_{S^{*c}_{\theta_j}} + \kappa z_{S^{*c}_{\theta_j}} = 0,$$

and

$$\frac{2}{n_k} \mathbb{H}_{S^*_{\theta_j}, S^*_{\theta_j}} \left(\widehat{\theta}_j - \theta^*_j \right)_{S^*_{\theta_j}} - \frac{2}{n_k} (Z_j)_{S^*_{\theta_j}} + \kappa z_{S^*_{\theta_j}} = 0.$$

From the last equation, and as $z_{S^*_{\theta_j}} = \operatorname{sign}((\widehat{\theta}_j)_{S^*_{\theta_j}})$, we observe that

$$\left(\widehat{\theta}_j - \theta^*_j \right)_{S^*_{\theta_j}} = \mathbb{H}^{-1}_{S^*_{\theta_j}, S^*_{\theta_j}} (Z_j)_{S^*_{\theta_j}} - \frac{n_k \kappa}{2} \mathbb{H}^{-1}_{S^*_{\theta_j}, S^*_{\theta_j}} \operatorname{sign}((\widehat{\theta}_j)_{S^*_{\theta_j}}). \tag{23}$$

Therefore, we set for each $j \in [M]$,

$$z_{S^{*c}_{\theta_j}} = -\frac{2}{n_k \kappa} \left(\mathbb{H}_{S^{*c}_{\theta_j}, S^*_{\theta_j}} \mathbb{H}^{-1}_{S^*_{\theta_j}, S^*_{\theta_j}} (Z_j)_{S^*_{\theta_j}} - (Z_j)_{S^{*c}_{\theta_j}} \right) + \mathbb{H}_{S^{*c}_{\theta_j}, S^*_{\theta_j}} \mathbb{H}^{-1}_{S^*_{\theta_j}, S^*_{\theta_j}} \operatorname{sign}((\widehat{\theta}_j)_{S^*_{\theta_j}}). \tag{24}$$

We then have built an optimal solution $(\widehat{\theta}, \widehat{z})$ that satisfies the required condition.

Second step. The goal of the second step is to prove that $\|\widehat{z}_{S^{*c}}\|_\infty < 1$ which implies the following result.

Lemma C.2. *Assume that $\|\widehat{z}_{S^{*c}}\|_\infty < 1$. Then, any solution $\widetilde{\theta}$ of the minimization problem $\min_\theta \mathcal{C}(\theta)$ satisfies $\widetilde{\theta}_{S^{*c}} = 0$.*

Proof. Let $\widetilde{\theta}$ another solution. Then, it holds that

$$R_{T,k}(\widehat{\theta}) + \kappa \left\langle \widehat{z}, \widehat{\theta} \right\rangle = R_{T,k}(\widetilde{\theta}) + \kappa \sum_{j=1}^{M} \sum_{j'=1}^{M} |\widetilde{\theta}_{j,j'}|,$$

we deduce that

$$R_{T,k}(\widehat{\theta}) - \kappa \left\langle \widehat{z}, \widetilde{\theta} - \widehat{\theta} \right\rangle = R_{T,k}(\widetilde{\theta}) + \kappa \left(\sum_{j=1}^{M} \sum_{j'=1}^{M} |\widetilde{\theta}_{j,j'}| - \left\langle \widehat{z}, \widetilde{\theta} \right\rangle \right).$$

Since the pair $(\widehat{\theta}, \widehat{z})$ satisfies Equation (22), we have that

$$\kappa \widehat{z} = -\nabla R_{T,k}(\widehat{\theta}),$$

which leads to

$$R_{T,k}(\widehat{\theta}) - R_{T,k}(\widetilde{\theta}) + \left\langle \nabla R_{T,k}(\widehat{\theta}), \widetilde{\theta} - \widehat{\theta} \right\rangle = \kappa \left(\sum_{j=1}^{M} \sum_{j'=1}^{M} |\widetilde{\theta}_{j,j'}| - \left\langle \widehat{z}, \widetilde{\theta} \right\rangle \right).$$

Hence, from the above equation and the convexity of $R_{T,k}$ we deduce that

$$\kappa \left(\sum_{j=1}^{M} \sum_{j'=1}^{M} |\widetilde{\theta}_{j,j'}| - \left\langle \widehat{z}, \widetilde{\theta} \right\rangle \right) \leq 0.$$

Therefore, we obtain that

$$\sum_{j=1}^{M} \sum_{j'=1}^{M} |\widetilde{\theta}_{j,j'}| \leq \left\langle \widehat{z}, \widetilde{\theta} \right\rangle = \sum_{j=1}^{M} \sum_{j'=1}^{M} \widehat{z}_{j,j'} \widetilde{\theta}_{j,j'}.$$

Since $\|\widehat{z}_{S^{*c}}\|_\infty < 1$, if there exists $\tilde{\theta}_{j,j'} \neq 0$ for $(j,j') \in S^{*c}$ we get

$$\sum_{j=1}^{M} \sum_{j'=1}^{M} |\widetilde{\theta}_{j,j'}| < \sum_{j=1}^{M} \sum_{j'=1}^{M} |\widetilde{\theta}_{j,j'}|,$$

which leads us to a contradiction. Therefore $\widetilde{\theta}_{S^{*c}} = 0$. □

Now we show that for $\kappa \geq \dfrac{\log^4(nM^2)}{\sqrt{n}}$, we have $\|\widehat{z}_{S^{*c}}\|_\infty < 1$ on the event Ω_n. From Equation (24), we deduce that for each $j \in [M]$

$$\begin{aligned} \|\widehat{z}_{S^{*c}_{\theta_j}}\|_\infty \quad \leq \quad & \|\mathbb{H}_{S^{*c}_{\theta_j}, S^*_{\theta_j}} \mathbb{H}^{-1}_{S^*_{\theta_j}, S^*_{\theta_j}}\|_\infty + \|\mathbb{H}_{S^{*c}_{\theta_j}, S^*_{\theta_j}} \mathbb{H}^{-1}_{S^*_{\theta_j}, S^*_{\theta_j}}\|_\infty \frac{2}{n_k \kappa} \|(Z_j)_{S^*_{\theta_j}}\|_\infty \\ & + \frac{2}{n_k \kappa} \|(Z_j)_{S^{*c}_{\theta_j}}\|_\infty. \end{aligned}$$

From Assumption (MI), we get for some $\gamma \in (0,1)$

$$\|\widehat{z}_{S^{*c}}\|_\infty \leq (1-\gamma) \left(1 + \frac{2}{n_k \kappa} \|(Z_j)_{S^*_{\theta_j}}\|_\infty \right) + \frac{2}{n_k \kappa} \|(Z_j)_{S^{*c}_{\theta_j}}\|_\infty. \tag{25}$$

From Lemma C.1 we have with probability larger than $1 - \dfrac{C}{n}$ on an event Ω_n that

$$\frac{1}{n_k}\|(Z_j)_{S^*_{\theta_j}}\|_\infty \leq \frac{C\log^3(nM^2)}{\sqrt{n}}, \quad \frac{1}{n_k}\|(Z_j)_{S^{*c}_{\theta_j}}\|_\infty \leq \frac{C\log^3(nM^2)}{\sqrt{n}}.$$

Hence, from Equation (25), for n large enough, we deduce that, with probability larger than on Ω_n,

$$\|\widehat{z}_{S^{*c}}\|_\infty < 1,$$

provided that $\dfrac{C\log^3(nM^2)}{\kappa\sqrt{n}} \to 0$ as $n \to +\infty$. Therefore, the choice $\kappa \geq \dfrac{\log^4(nM^2)}{\sqrt{n}}$ yields the desired result.

Third step. In the second step, we have shown for n large enough that on Ω_n, any solution of $\min_\theta \mathcal{C}(\theta)$, with $\mathcal{C}$ given in Equation (21), is a solution of

$$\min_{\theta_{S^*}} \widetilde{R}_{T,k}(\theta_{S^*}) + \kappa \sum_{j=1}^{M} \sum_{j' \in S^*_{\theta_j}} |\theta_{j,j'}|.$$

In this step, we establish the following result.

Lemma C.3. *Let $\widehat{\theta}_{S^*}$ defined as*

$$\widehat{\theta}_{S^*} \in \operatorname*{argmin}_{\theta_{S^*}} \left\{ \widetilde{R}_{T,k}(\theta_{S^*}) + \kappa \sum_{j=1}^{M} \sum_{j' \in S^*_{\theta_j}} |\theta_{j,j'}| \right\}.$$

Under Assumption (ME), for $\kappa = \dfrac{\log^4(nM^2)}{\sqrt{n}}$, it holds that on Ω_n

$$\left\|\widehat{\theta}_{S^*} - \theta_{S^*}\right\|_\infty \leq \frac{C\Lambda_0^{-1}\max_j \sqrt{|S^*_{\theta_j}|}\log^4(nM^2)}{\sqrt{n}}.$$

Proof. From Equation (23), we get for each $j \in \{1, \ldots, M\}$

$$\left\|\widehat{\theta}_{S^*_{\theta_j}} - \theta_{S^*_{\theta_j}}\right\|_\infty \leq \left\| \left(\frac{\mathbb{H}_{S^*_{\theta_j}, S^*_{\theta_j}}}{n_k} \right)^{-1} \frac{(Z_j)_{S^*_{\theta_j}}}{n_k} \right\|_\infty + \frac{\kappa}{2} \left\| \left(\frac{\mathbb{H}_{S^*_{\theta_j}, S^*_{\theta_j}}}{n_k} \right)^{-1} \operatorname{sign}((\widehat{\theta}_j)_{S^*_{\theta_j}}) \right\|_\infty.$$

Applying Lemma A.1, and C.1 together with Assumption (ME), we obtain

$$\left\|\widehat{\theta}_{S^*_{\theta_j}} - \theta^*_{S^*_{\theta_j}}\right\|_\infty \leq \Lambda_0^{-1}\sqrt{\left|S^*_{\theta_j}\right|}\left(\frac{C\log^3(nM^2)}{\sqrt{n}} + \kappa \right).$$

Therefore, the choice of $\kappa = \dfrac{\log^4(nM^2)}{\sqrt{n}}$ yields the desired result. □

Fourth step. We deduce from Lemma C.3 and Assumption 8 that

$$\operatorname{sign}(\widehat{\theta}_{S^*}) = \operatorname{sign}(\theta^*_{S^*}).$$

Therefore, from Equation (23), we deduce that on Ω_n for $\kappa = \dfrac{\log^4(nM^2)}{\sqrt{n}}$,

$$\theta_{S^*} \mapsto \min_{\theta_{S^*}} \widetilde{R}_{T,k}(\theta_{S^*}) + \kappa \sum_{j=1}^{M} \sum_{j' \in S^*_{\theta_j}} |\theta_{j,j'}|,$$

admits a unique minimizer $\widehat{\theta}_{S^*}$ which satisfies for each $j \in \{1, \ldots, M\}$,

$$(\widehat{\theta}_j)_{S^*_{\theta_j}} = (\theta_j)_{S^*_{\theta_j}} + \mathbb{H}^{-1}_{S^*_{\theta_j}, S^*_{\theta_j}} (Z_j)_{S^*_{\theta_j}} - \frac{n_k \kappa}{2} \mathbb{H}^{-1}_{S^*_{\theta_j}, S^*_{\theta_j}} \operatorname{sign}\left((\theta^*_j)_{S^*_{\theta_j}}\right).$$

Hence, in view of Steps 2, with the choice of $\kappa = \dfrac{\log^4(nM^2)}{\sqrt{n}}$, we then have shown that there is a unique solution $\widehat{\theta}$ of $\min_\theta \mathcal{C}(\theta)$ which satisfies on Ω_n

$$\widehat{\theta}_{S^{*c}} = 0, \text{ and } \operatorname{sign}(\widehat{\theta}_{S^*}) = \operatorname{sign}(\theta^*_{S^*}),$$

and

$$\left\| \widehat{\theta}_{S^*} - \theta_{S^*} \right\|_\infty \leq \frac{C \Lambda_0^{-1} \max_j \sqrt{|S^*_{\theta_j}|} \log^4(nM^2)}{\sqrt{n}}.$$

D Proofs for the rate of convergence of ERMLR algorithm

We first establish a technical result in Section D.1, then rate of convergence of the ERMLR algorithm is given in Section D.2.

D.1 Technical result

We recall that the set Θ_n is defined as follows

$$\Theta_n := \left\{ \theta = (\mu, A) \in \mathbb{R}_+^M \times \mathbb{R}_+^{M \times M}, \mu_j \in \left[\frac{1}{n}, \log(n) \right], j \in [M], \|A\|_F \leq n \right\}.$$

We also introduce the set Π of conditional probabilities

$$\begin{aligned} \Pi \;:=\; \Bigg\{ & \pi_{p,\theta} = \left(\frac{p_k \mathrm{e}^{F_{\theta_k}(\cdot)}}{\sum_{k'=1}^K p_{k'} \mathrm{e}^{F_{\theta_{k'}}(\cdot)}} \right)_{k \in [K]} : \\ & \theta = (\theta_1, \ldots, \theta_K) \in \Theta_n^K, \; \sum_{k=1}^K p_k = 1, \; \min_k p_k > \frac{p_0}{2} \Bigg\} \end{aligned}$$

The following result provides a bound on ℓ_1-distance between two elements of the set Π. It shows that this distance can be bounded by the distance between the corresponding parameters of the associated model.

Proposition D.1. *Let $\pi = \pi_{p,\theta}$ and $\pi' = \pi_{p',\theta'}$ two elements of Π. Grant Assumptions 3, 2 and 1, the following holds*

$$\begin{aligned}\mathbb{E}\big[\,\big\|\pi - \pi'\big\|_1\,\big] &\leq \frac{K}{p_0}\|p - p'\|_1 \\ &+ CK^2n^2\log(n)\left(\sqrt{M}\max_{k\in[K]}\|\mu_k - \mu_k'\|_1 + M\max_{k\in[K]}\|A_k - A_k'\|_F\right),\end{aligned}$$

where C is a constant depending on T, μ_0, μ_1 and $\|h\|_\infty$.

Proof. Let us consider $\pi, \pi' \in \Pi$ with respective parameters (p,θ), and (p',θ'). We have that for an observation $\mathcal{T}_T$ of the process on $[0,T]$, we have

$$\big\|\pi(\mathcal{T}_T) - \pi'(\mathcal{T}_T)\big\|_1 \leq \big\|\pi(\mathcal{T}_T) - \pi_{p,\theta'}(\mathcal{T}_T)\big\|_1 + \big\|\pi_{p,\theta'}(\mathcal{T}_T) - \pi'(\mathcal{T}_T)\big\|_1. \tag{26}$$

Let $k \in [K]$, for a parameter $p = (p_1, \ldots, p_K)$, we introduce the function ψ_k^p defined as

$$\psi_k^p(x_1,\ldots,x_K) := \frac{p_k\exp(x_k)}{\sum_{j=1}^K p_j\exp(x_j)}$$

Since, $\min_{k\in[K]} p_k \geq \frac{p_0}{2}$, for any k, j and $(x_1,\ldots,x_K)$, we have

$$\left|\frac{\partial\psi_k^p(x_1,\ldots,x_K)}{\partial p_j}\right| \leq \frac{2}{p_0},$$

we deduce by mean value inequality

$$\big\|\pi_{p,\theta'}(\mathcal{T}_T) - \pi'(\mathcal{T}_T)\big\|_1 \leq \frac{2K}{p_0}\,\|p - p'\|_1.$$

Besides for any k, j and p,

$$\left|\frac{\partial\psi_k^p(x_1,\ldots,x_K)}{\partial x_j}\right| \leq 1,$$

we also deduce

$$\big\|\pi(\mathcal{T}_T) - \pi_{p,\theta'}(\mathcal{T}_T)\big\|_1 \leq K\sum_{k=1}^K\left|F_{\theta_k}(\mathcal{T}_T) - F_{\theta_k'}(\mathcal{T}_T)\right|.$$

Therefore, from Equation (26), we obtain

$$\mathbb{E}\left[\big\|\pi(\mathcal{T}_T) - \pi'(\mathcal{T}_T)\big\|_1\right] \leq \frac{2K}{p_0}\,\|p - p'\|_1 + K\sum_{k=1}^K\mathbb{E}\left[\left|F_{\theta_k}(\mathcal{T}_T) - F_{\theta_k'}(\mathcal{T}_T)\right|\right].$$

Hence, it remains to bound the second term in the *r.h.s.* of the above inequality. Using Cauchy-Schwartz inequality, for each k, we have that

$$\begin{aligned}&\mathbb{E}\left[\left|F_{\theta_k}(\mathcal{T}_T) - F_{\theta_k'}(\mathcal{T}_T)\right|\right] \\ &= \mathbb{E}\left[\left|\sum_{j=1}^M\left(\int_0^T\log\left(\frac{\lambda_{j,\theta_k}(t)}{\lambda_{j,\theta_k'}(t)}\right)\mathrm{d}N_j(t) - \int_0^T\left(\lambda_{j,\theta_k}(t) - \lambda_{j,\theta_k'}(t)\right)\mathrm{d}t\right)\right|\right] \\ &\leq \mathbb{E}\left[\left(\sum_{j=1}^M\int_0^T\left|\log\left(\frac{\lambda_{j,\theta_k}(t)}{\lambda_{j,\theta_k'}(t)}\right)\right|\mathrm{d}N_j(t)\right)^2\right]^{1/2} \\ &+ \mathbb{E}\left[\sum_{j=1}^M\int_0^T\left|\lambda_{j,\theta_k}(t) - \lambda_{j,\theta_k'}(t)\right|\mathrm{d}t\right].\end{aligned} \tag{27}$$

Now, we observe that

$$\left|\lambda_{j,\theta_k}(t) - \lambda_{j,\theta'_k}(t)\right| \leq |\mu_{k,j} - \mu'_{k,j}| + \|h\|_\infty \sum_{j'=1}^{M} |a_{k,j,j'} - a'_{k,j,j'}| N_{j'}(T).$$

Therefore, we deduce

$$\begin{aligned}\mathbb{E}\left[\sum_{j=1}^{M}\int_0^T \left|\lambda_{j,\theta_k}(t) - \lambda_{j,\theta'_k}(t)\right| \mathrm{d}t\right] &\leq T\sum_{j=1}^{M}|\mu_{k,j} - \mu'_{k,j}| \\ &\quad + T\|h\|_\infty \sum_{j'=1}^{M}\sum_{j=1}^{M} |a_{k,j,j'} - a'_{k,j,j'}| \mathbb{E}\left[N_{j'}(T)\right]. \quad (28)\end{aligned}$$

Now, we bound the first term in the *r.h.s.* of Equation (27). Since

$$\left|\log\left(\frac{\lambda_{j,\theta_k}(t)}{\lambda_{j,\theta'_k}(t)}\right)\right| = \left|\log\left(\frac{\lambda_{j,\theta_k}(t)}{\mu_{k,j}}\right) - \log\left(\frac{\lambda_{j,\theta'_k}(t)}{\mu'_{k,j}}\right) + \log\left(\frac{\mu_{k,j}}{\mu'_{k,j}}\right)\right|,$$

we deduce since $x \mapsto \log(x)$ is Lipschitz for $x \geq 1$ that

$$\begin{aligned}\left|\log\left(\frac{\lambda_{j,\theta_k}(t)}{\lambda_{j,\theta'_k}(t)}\right)\right| &\leq \left|\log\left(\frac{\mu_{k,j}}{\mu'_{k,j}}\right)\right| + \left|\frac{\lambda_{j,\theta_k}(t)}{\mu_{k,j}} - \frac{\lambda_{j,\theta'_k}(t)}{\mu'_{k,j}}\right| \\ &\leq n\left|\mu_{k,j} - \mu'_{k,j}\right| + n^2\left|\mu'_{k,j}\lambda_{j,\theta_k}(t) - \mu_{k,j}\lambda_{j,\theta'_k}(t)\right| \\ &\leq n\left|\mu_{k,j} - \mu'_{k,j}\right| + n^2\left(|\mu_{k,j} - \mu'_{k,j}|\lambda_{j,\theta_k}(t) + \mu_{k,j}\left|\lambda_{j,\theta_k}(t) - \lambda_{j,\theta'_k}(t)\right|\right) \\ &\leq n\left|\mu_{k,j} - \mu'_{k,j}\right| + n^2\Bigg(|\mu_{k,j} - \mu'_{k,j}|\lambda_{j,\theta_k}(t) \\ &\quad + \log(n)\Big(|\mu'_{k,j} - \mu_{k,j}| + \|h\|_\infty \sum_{j'=1}^{M} N_{j'}(T)|a_{k,j,j'} - a'_{k,j,j'}|\Big)\Bigg). \quad (29)\end{aligned}$$

Besides, applying the Doob's decomposition for the processes $N_j, j \in [M]$, and the Cauchy-Schwartz's inequality, we get

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{j=1}^{M}\int_0^T \left|\log\left(\frac{\lambda_{j,\theta_k}(t)}{\lambda_{j,\theta'_k}(t)}\right)\right| \mathrm{d}N_j(t)\right)^2\right] &\leq M\sum_{j=1}^{M}\mathbb{E}\left[\int_0^T \log^2\left(\frac{\lambda_{j,\theta_k}(t)}{\lambda_{j,\theta'_k}(t)}\right)\lambda^*_{Y,j}(t)\,\mathrm{d}t\right] \\ &\quad + M\sum_{j=1}^{M}\mathbb{E}\left[\left(\int_0^T \left|\log\left(\frac{\lambda_{j,\theta_k}(t)}{\lambda_{j,\theta'_k}(t)}\right)\right| \lambda^*_{Y,j}(t)\,\mathrm{d}t\right)^2\right]. \quad (30)\end{aligned}$$

From Assumption 4, we have $\mathbb{E}\left[\left(\lambda^*_{Y,j}(t)\right)^2\right] < \infty$. Therefore, the first term in the *r.h.s.* in Equation (30) can be bounded as follows

$$\begin{aligned}\mathbb{E}\left[\int_0^T \log^2\left(\frac{\lambda_{j,\theta_k}(t)}{\lambda_{j,\theta'_k}(t)}\right)\lambda^*_{Y,j}(t)\,\mathrm{d}t\right] &\leq \int_0^T \mathbb{E}\left[\log^4\left(\frac{\lambda_{j,\theta_k}(t)}{\lambda_{j,\theta'_k}(t)}\right)\right]^{1/2}\mathbb{E}\left[\left(\lambda^*_{Y,j}(t)\right)^2\right]^{1/2}\mathrm{d}t \\ &\leq CT\sup_{t\in[0,T]}\mathbb{E}\left[\log^4\left(\frac{\lambda_{j,\theta_k}(t)}{\lambda_{j,\theta'_k}(t)}\right)\right]^{1/2}.\end{aligned}$$

Similarly, we obtain:

$$\mathbb{E}\left[\left(\int_0^T \left|\log\left(\frac{\lambda_{j,\theta_k}(t)}{\lambda_{j,\theta'_k}(t)}\right)\right| \lambda^*_{Y,j}(t)\,\mathrm{d}t\right)^2\right] \leq T\mathbb{E}\left[\int_0^T \log^2\left(\frac{\lambda_{j,\theta_k}(t)}{\lambda_{j,\theta'_k}(t)}\right)\left(\lambda^*_{Y,j}(t)\right)^2\,\mathrm{d}t\right]$$
$$\leq CT^2 \sup_{t\in[0,T]} \mathbb{E}\left[\log^4\left(\frac{\lambda_{j,\theta_k}(t)}{\lambda_{j,\theta'_k}(t)}\right)\right]^{1/2}.$$

Then, by Assumption 3, from Equation (29) and Equation (30), we get

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{j=1}^M \int_0^T \left|\log\left(\frac{\lambda_{j,\theta_k}(t)}{\lambda_{j,\theta'_k}(t)}\right)\right| \mathrm{d}N_j(t)\right)^2\right] \\
&\leq CT^2M\sum_{j=1}^M \sup_{t\in[0;T]} \mathbb{E}\Big[|\mu_{k,j}-\mu'_{k,j}|^4\left(n+n^2\lambda_{j,\theta_k}(t)+n^2\log(n)\right)^4 \\
&\quad +n^8\log(n)^4\|h\|_\infty^4\left(\sum_{j'=1}^M N_{j'}(T)|a_{k,j,j'}-a'_{k,j,j'}|\right)^4\Big]^{1/2} \\
&\leq CT^2M\sum_{j=1}^M\left(\left[n^4\sup_{t\in[0;T]}\mathbb{E}\left[(\lambda_{j,\theta_k}(t))^4\right]^{1/2}+n^2+n^4\log(n)^2\right]|\mu_{k,j}-\mu'_{k,j}|^2\right. \\
&\quad \left.+Cn^4\log(n)^2\mathbb{E}\left[\left(\sum_{j'=1}^M N_{j'}(T)|a_{k,j,j'}-a'_{k,j,j'}|\right)^4\right]^{1/2}\right)
\end{aligned}$$

where C is a constant depending on μ_0, μ_1 and $\|h\|_\infty$. In view of Assumption 4 $\mathbb{E}\left[(\lambda_{j,\theta_k}(t))^4\right] \leq C$. Therefore, from the above equation, and Cauchy Schwartz's inequality, we deduce

$$\begin{aligned}
&E\left[\left(\sum_{j=1}^M \int_0^T \left|\log\left(\frac{\lambda_{j,\theta_k}(t)}{\lambda_{j,\theta'_k}(t)}\right)\right| \mathrm{d}N_j(t)\right)^2\right] \\
&\qquad \leq CT^2M\left(n^4+n^2+n^4\log(n)^2\right)\sum_{j=1}^M|\mu_{k,j}-\mu'_{k,j}|^2 \\
&\qquad\quad + CT^2Mn^4\log(n)^2\mathbb{E}\left[\left(\sum_{j'=1}^M N_{j'}(T)^2\right)^2\right]^{1/2}\sum_{j=1}^M\sum_{j'=1}^M|a_{k,j,j'}-a'_{k,j,j'}|^2.
\end{aligned}$$

From Assumption 4 we have that $\mathbb{E}\left[\left(\sum_{j'=1}^M N_{j'}(T)^2\right)^2\right] \leq CM^2$. Thus, gathering Equations (27)

and (28), it comes

$$\begin{aligned}\mathbb{E}[\|\pi-\pi'\|_1] &\leq \frac{2K}{p_0}\|p-p'\|_1 \\ &\quad +KC\sum_{k=1}^{K}\left(\sum_{j=1}^{M}|\mu_{k,j}-\mu'_{k,j}|+\sum_{j=1}^{M}\sum_{j'=1}^{M}|a_{k,j,j'}-a'_{k,j,j'}|\right) \\ &\quad +KCn^2\log(n)\sum_{k=1}^{K}\left(M\sum_{j=1}^{M}|\mu_{k,j}-\mu'_{k,j}|^2\right. \\ &\quad \left.+M^2\sum_{j=1}^{M}\sum_{j'=1}^{M}|a_{k,j,j'}-a'_{k,j,j'}|^2\right)^{1/2}\end{aligned}$$

with C depending on μ_0, μ_1, $\|h\|_\infty$ and T. Finally, using that $\|x\|_2 \leq \|x\|_1 \leq \sqrt{d}\|x\|_2$ for $x \in \mathbb{R}^d$, we obtain

$$\begin{aligned}\mathbb{E}\big[\|\pi-\pi'\|_1\big] &\leq \frac{2K}{p_0}\|p-p'\|_1 \\ &+ K^2C\max_{k\in[K]}\left(\sum_{j=1}^{M}|\mu_{k,j}-\mu'_{k,j}|+\sum_{j=1}^{M}\sum_{j'=1}^{M}|a_{k,j,j'}-a'_{k,j,j'}|\right) \\ &+ K^2Cn^2\log(n)\max_{k\in[K]}\left(\sqrt{M}\sum_{j=1}^{M}|\mu_{k,j}-\mu'_{k,j}|+M\|A_k-A'_k\|_F\right)\end{aligned}$$

thus

$$\begin{aligned}\mathbb{E}\big[\|\pi-\pi'\|_1\big] &\leq \frac{2K}{p_0}\|p-p'\|_1+K^2Cn^2\log(n)\sqrt{M}\max_{k\in[K]}\|\mu_k-\mu'_k\|_1 \\ &\quad +K^2CMn^2\log(n)\max_{k\in[K]}\|A_k-A'_k\|_F.\end{aligned}$$

Finally, combining the above equation, Equations (27) and (28) yields the desired result.

□

D.2 Proof of Theorem 2

We begin this section by a lemma that provides a bound on the ε-covering number of the set $\hat{\Theta}$ defined in Equation (8).

Lemma D.1. *Let $\varepsilon > 0$. There exists an ε-net $\mathcal{M}_\varepsilon \subset \hat{\Theta}$ with*

$$|\mathcal{M}_\varepsilon| \leq \left(\frac{(\log(n)-1/n)M}{\varepsilon}\right)^{MK}\left(\frac{3n}{\varepsilon}\right)^{\sum_k \widehat{S}_k}.$$

In particular, for all $(\mu, A) \in \widehat{\Theta}$ there exists $(\mu_\varepsilon, A_\varepsilon) \in \mathcal{M}_\varepsilon$ s.t. $\max_{k\in[K]}\|\mu_k-\mu_{k,\varepsilon}\|_1 \leq \varepsilon$ and $\|A_k-A_{k,\varepsilon}\|_F \leq \varepsilon$.

Proof of Lemma D.1. First, we observe that the set

$$\left\{\frac{1}{n}+k\frac{(\log(n)-1/n)}{\lceil\frac{M(\log(n)-1/n)}{\varepsilon}\rceil},\ k\in\left\{1,\dots,\frac{M(\log(n)-1/n)}{\varepsilon}-1\right\}\right\}$$

is and ε/M-cover of the interval $[1/n, \log(n)]$. Therefore, we deduce that there exists $\mathcal{M}_{\varepsilon,\mu}$ an ε-cover of $\{\mu \in \mathbb{R}^M,\ s.t.\ \mu \in \Theta_n\}$ for $\|\cdot\|_1$, such that

$$|\mathcal{M}_{\varepsilon,\mu}| \leq \left(\frac{(\log(n) - 1/n)M}{\varepsilon}\right)^M. \tag{31}$$

Let $k \in [K]$. For $\varepsilon > 0$, the covering number of the Euclidean ball centered in 0 and with radius n in $\mathbb{R}^{\widehat{S}_k}$, satisfies

$$\mathcal{N}(\varepsilon, \bar{\mathcal{B}}(0,n), \|.\|_2) \leq \left(\frac{3n}{\varepsilon}\right)^{\widehat{S}_k}.$$

Hence, we deduce that there exists $\mathcal{M}_{\varepsilon,A,k}$ an ε-cover of $\{A \in \Theta_n,\ s.t.\ \mathrm{supp}(A) = \widehat{S}_k\}$, for $\|\cdot\|_F$, such that

$$|\mathcal{M}_{\varepsilon,A,k}| \leq \left(\frac{3n}{\varepsilon}\right)^{\widehat{S}_k}. \tag{32}$$

From Equation (31) and (32) we obtain the desired result. □

Proof of Theorem 2. We first recall that the construction of the `ERMLR` algorithm is based on a dataset $\mathcal{D}_n = \{(\mathcal{T}_T^{(i)}, Y^{(i)}), i = 1, \ldots, 2n\}$ of size $2n$ which is split into two independent dataset of same size n that are denoted respectively $\mathcal{D}_n^{(1)}$ and $\mathcal{D}_n^{(2)}$.

Based on the first sample $\mathcal{D}_n^{(1)}$, we estimate the vector of weights p^* by its empirical frequencies $\hat{p}$. Hence for each k, we have

$$\hat{p}_k = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y^{(i)}=k\}}$$

Then, based on sample $\mathcal{D}_n^{(2)}$, we build the estimator $\widehat{S} := \left(\widehat{S}_1, \ldots, \widehat{S}_K\right)$ as described in Section 3.1. Besides, we also build the estimator of the vector of score function $\hat{f} = f_{\hat{\theta}^R}$, and $\hat{g}$ its associated classifier. Since $\mathcal{D}_n^{(1)}$ and $\mathcal{D}_n^{(2)}$ are independent, we have that $\hat{p}$ is independent of $\widehat{f}$ and $\widehat{g}$.

Let us introduce the set $\mathcal{A} := \left\{\widehat{p} :\ \min(\widehat{p}) \geq \frac{p_0}{2}\right\}$. Note that on $\mathcal{A}^c$ we have

$$|\min(p^*) - \min(\widehat{p})| \geq \frac{p_0}{2},$$

which implies that there exists $k \in [K]$ s.t. $|p_k^* - \widehat{p}_k| \geq \frac{p_0}{2}$. Thus, using Hoeffding's inequality we get

$$\begin{aligned}
\mathbb{P}(\mathcal{A}^c) &\leq \sum_{k=1}^{K} \mathbb{P}\left(|p_k^* - \widehat{p}_k| \geq \frac{p_0}{2}\right) \\
&\leq 2K\mathrm{e}^{-np_0^2/2}.
\end{aligned} \tag{33}$$

Now, let us work on $\Omega := \mathcal{A} \bigcap \{\hat{S} = S^*\}$, and denote

$$\Delta_n := \sum_{k=1}^{K} (\widehat{p}_k - p_k^*)^2, \tag{34}$$

which is a random variable independent from $\mathcal{D}_n^{(2)}$. We also recall that for each $\theta \in \hat{\Theta}$, the score function f_θ is defined as follows

$$f_\theta^k(\mathcal{T}_T) = 2\pi_{k,\widehat{p},\theta}(\mathcal{T}_T) - 1,\ k \in [K].$$

We introduce the oracle counterpart of $\hat{f}$

$$\tilde{\theta} = \underset{\theta \in \hat{\Theta}}{\operatorname{argmin}}\, \mathcal{R}_2(f_\theta).$$

Our aim is to control the following excess risk:

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_2(\widehat{f}) - \mathcal{R}_2(f^*)\right] &= \mathbb{E}\left[\left(\mathcal{R}_2(\widehat{f}) - \mathcal{R}_2(f^*)\right)\mathbb{1}_{\{\Omega\}}\right] \\ &\quad + \mathbb{E}\left[\left(\mathcal{R}_2(\widehat{f}) - \mathcal{R}_2(f^*)\right)\mathbb{1}_{\{\Omega^c\}}\right]. \end{aligned} \tag{35}$$

Since for each $\theta \in \hat{\Theta}$ defined by (8), $\mathcal{R}_2(f_\theta)$ is bounded, from Theorem 1, and Equation (33), we deduce that

$$\mathbb{E}[\left(\mathcal{R}_2(\widehat{f}) - \mathcal{R}_2(f^*)\right)\mathbb{1}_{\{\Omega^c\}}] \leq C\mathbb{P}\left(\Omega^c\right) \leq C\left(\frac{1}{n} + \exp\left(-np_0^2/2\right)\right). \tag{36}$$

Therefore, it remains to bound the first term in the *r.h.s.* of Equation (35). Hence, we work on the set Ω. We consider the following decomposition

$$\mathcal{R}_2(\hat{f}) - \mathcal{R}_2(f^*) = \left(\mathcal{R}_2(\hat{f}) - \mathcal{R}_2(f_{\tilde{\theta}})\right) + \left(\mathcal{R}_2(f_{\tilde{\theta}}) - \mathcal{R}_2(f^*)\right) \tag{37}$$

In a first step, we control the second term in the *r.h.s.* of the above equation. For n large enough, we observe that on Ω, $\theta^* \in \hat{\Theta}$. Therefore, from the definition of $\tilde{\theta}$, we deduce

$$\begin{aligned}\mathcal{R}_2(f_{\tilde{\theta}}) - \mathcal{R}_2(f^*) &= \mathcal{R}_2(f_{\tilde{\theta}}) - \mathcal{R}_2(f_{\theta^*}) + \mathcal{R}_2(f_{\theta^*}) - \mathcal{R}_2(f^*) \\ &\leq \mathcal{R}_2(f_{\theta^*}) - \mathcal{R}_2(f^*).\end{aligned}$$

Then on Ω, we deduce from the mean value theorem that

$$\mathcal{R}_2(f_{\tilde{\theta}}) - \mathcal{R}_2(f^*) \leq \mathcal{R}_2(f_{\theta^*}) - \mathcal{R}_2(f^*) \leq C\Delta_n, \tag{38}$$

with Δ_n given in Equation (34). Since $\mathbb{E}\left[\Delta_n\right] \leq \dfrac{C}{n}$, from Equation (37) we deduce that

$$\mathbb{E}\left[\left(\mathcal{R}_2(\widehat{f}) - \mathcal{R}_2(f^*)\right)\mathbb{1}_{\{\Omega\}}\right] \leq \mathbb{E}\left[\mathcal{R}_2(\hat{f}) - \mathcal{R}_2(f_{\tilde{\theta}})\mathbb{1}_{\{\Omega\}}\right] + \frac{C}{n}. \tag{39}$$

Now, we focus on the first term in the r.h.s. of Equation (37). We denote

$$D_f := \mathcal{R}_2(f) - \mathcal{R}_2(f_{\tilde{\theta}}), \quad \text{and} \quad \widehat{D}_f := \widehat{\mathcal{R}}_2(f) - \widehat{\mathcal{R}}_2(f_{\tilde{\theta}}).$$

And we want to control $\mathbb{E}[D_{\widehat{f}}]$. By Lemma D.1, there exists a subset $\mathcal{M}_\varepsilon \subset \hat{\Theta}$ such that for $\hat{\theta}^R = \left(\hat{\mu}, \hat{A}\right)$, there exists $\theta_\varepsilon = (\mu_\varepsilon, A_\varepsilon) \in \mathcal{M}_\varepsilon$ satisfying

$$\max_{k\in[K]} \|\mu_{k,\varepsilon} - \hat{\mu}_k\|_1 \leq \varepsilon \quad \text{and} \quad \max_{k\in[K]} \|A_{k,\varepsilon} - \hat{A}_k\|_F \leq \varepsilon.$$

Then, the following decomposition holds

$$\begin{aligned}D_{\widehat{f}} &\leq D_{\widehat{f}} - 2\widehat{D}_{\widehat{f}} \\ &= (D_{\widehat{f}} - D_{f_{\theta_\varepsilon}}) + (2\widehat{D}_{f_{\theta_\varepsilon}} - 2\widehat{D}_{\widehat{f}}) + (D_{f_{\theta_\varepsilon}} - 2\widehat{D}_{f_{\theta_\varepsilon}}) \\ &=: T_1 + T_2 + T_3.\end{aligned}$$

Applying Proposition D.1 combined with Lemma D.1 with $\varepsilon := 1/(n^3 M \log(n))$ we get

$$\mathbb{E}\left[T_i\right] \leq \frac{C}{n}, \quad \text{for } i = 1, 2.$$

Besides,

$$T_3 \leq \max_{\theta \in \mathcal{M}_\varepsilon} (D_{f_\theta} - 2\widehat{D}_{f_\theta}).$$

Therefore, gathering Equation (35), (36), (38), and (39), we deduce that

$$\mathbb{E}[\mathcal{R}_2(\widehat{f}) - \mathcal{R}_2(f^*)] \;\leq\; \mathbb{E}\left[\max_{\theta \in \mathcal{M}_\varepsilon} (D_{f_\theta} - 2\widehat{D}_{f_\theta}) \mathbb{1}_{\{\Omega\}}\right] + \frac{C}{n}. \tag{40}$$

To finish the proof, it remains to control the first term in the *r.h.s.* of Inequality (40). Conditional on $\mathcal{D}_n^{(1)}$, we have that

$$\mathbb{E}\left[\max_{\theta \in \mathcal{M}_\varepsilon} (D_{f_\theta} - 2\widehat{D}_{f_\theta}) \mathbb{1}_{\{\Omega\}} | \mathcal{D}_n^{(1)}\right] = \mathbb{1}_{\{\mathcal{A}\}} \mathbb{E}\left[\max_{\theta \in \mathcal{M}_\varepsilon} (D_{f_\theta} - 2\widehat{D}_{f_\theta}) \mathbb{1}_{\{\hat{S} = S^*\}} | \mathcal{D}_n^{(1)}\right].$$

Note that on the set $\{\hat{S} = S^*\}$ (as we are on Ω), the set $\mathcal{M}_\varepsilon$ is an ε-net of the *deterministic* set

$$\tilde{\Theta} = \left\{\theta = (\theta_1, \ldots, \theta_K) \in \Theta_n^K, \;\; \text{supp}(A_k) = S_k^*\right\},$$

and then is also deterministic. Besides, from Lemma D.1, we deduce that for $\varepsilon = \dfrac{1}{n^3 M \log(n)}$

$$\log\left(|\mathcal{M}_\varepsilon|\right) \leq CK\left(M + s^*\right) \frac{\log(nM)}{n}.$$

Furthermore, for $u > 0$ conditional on $\mathcal{D}_n^{(1)}$, it holds that

$$\begin{aligned}\mathbb{E}\left[\max_{\theta \in \mathcal{M}_\varepsilon} (D_{f_\theta} - 2\widehat{D}_{f_\theta}) \mathbb{1}_{\{\hat{S} = S^*\}}\right] &\leq u + \int_u^\infty \mathbb{P}\left(\max_{\theta \in \mathcal{M}_\varepsilon} (D_{f_\theta} - 2\widehat{D}_{f_\theta}) \mathbb{1}_{\{\hat{S} = S^*\}} \geq t\right) \mathrm{d}t \\ &\leq u + \int_u^\infty \mathbb{P}\left(\max_{\theta \in \mathcal{M}_\varepsilon} (D_{f_\theta} - 2\widehat{D}_{f_\theta}) \geq t\right) \mathrm{d}t. \end{aligned} \tag{41}$$

Now, we have to bound the last term in the above equation. Let $\theta \in \mathcal{M}_\varepsilon$, and $f := f_\theta$. Let us introduce the least squares function

$$\ell_f(Z, \mathcal{T}_T) := \sum_{k=1}^K (Z_k - f^k(\mathcal{T}_T))^2.$$

Since for each $\theta \in \tilde{\Theta}$, f_θ is uniformly bounded by 1, we get from Bernstein's inequality that, conditionally on $\mathcal{D}_n^{(1)}$, for $t \geq 0$

$$\begin{aligned}\mathbb{P}\left(D_f - 2\widehat{D}_f \geq t\right) &\leq \mathbb{P}\left(2(D_f - \widehat{D}_f) \geq t + D_f\right) \\ &\leq \exp\left(\frac{-n(t + D_f)^2/8}{B_f + (t + D_f)4K/3}\right), \end{aligned} \tag{42}$$

with

$$B_f := \mathbb{E}\left[\left(\ell_f(Z, \mathcal{T}_T) - \ell_{f_{\tilde{\theta}}}(Z, \mathcal{T}_T)\right)^2\right].$$

From the Cauchy-Schwartz inequality, we observe that conditionally on $\mathcal{D}_n^{(1)}$

$$\begin{aligned}\mathbb{E}\left[(\ell_f(Z,\mathcal{T}_T)-\ell_{f^*}(Z,\mathcal{T}_T))^2\right] &\leq C_K\sum_{k=1}^{K}\mathbb{E}\left[(f^k(\mathcal{T}_T)-f^{*k}(\mathcal{T}_T))^2\right]\\ &= C_K\left(\mathcal{R}_2(f)-\mathcal{R}_2(f^*)\right).\end{aligned}$$

Thus, since

$$B_f \leq 2\mathbb{E}\left[\left(\ell_f(Z,\mathcal{T}_T)-\ell_{f^*}(Z,\mathcal{T}_T)\right)^2\right]+2\mathbb{E}\left[\left(\ell_{f_{\tilde{\theta}}}(Z,\mathcal{T}_T)-\ell_{f^*}(Z,\mathcal{T}_T)\right)^2\right],$$

we deduce that

$$B_f \leq C_K\left(\mathcal{R}_2(f)-\mathcal{R}_2(f^*)+\mathcal{R}_2(f_{\tilde{\theta}})-\mathcal{R}_2(f^*)\right).$$

Then, as $\mathcal{R}_2(f)-\mathcal{R}_2(f^*)=\mathcal{R}_2(f)-\mathcal{R}_2(f_{\tilde{\theta}})+\mathcal{R}_2(f_{\tilde{\theta}})-\mathcal{R}_2(f^*)$, on the event $\mathcal{A}$ and conditionally on $\mathcal{D}_n^{(1)}$, we deduce from the above inequality and Equation (38) that

$$B_f \leq C_K\left(D_f+\Delta_n\right).$$

Therefore, for $t\geq\Delta_n$, we have that $B_f \leq_K (D_f+\Delta_n)$. then, since $D_f>0$, we deduce that

$$\frac{-n(t+D_f)^2/8}{B_f+(t+D_f)4K/3}\leq -C_K n(t+D_f)\leq -C_K nt.$$

Hence, from Inequality (42), we get for $t\geq\Delta_n$,

$$\mathbb{P}\left(D_f-2\widehat{D}_f\geq t\right)\leq\exp\left(-C_K nt\right),$$

which leads to

$$\mathbb{P}\left(\max_{\theta\in\mathcal{M}_\varepsilon}(D_f-2\widehat{D}_f)\geq t\right)\leq|\mathcal{M}_\varepsilon|\exp\left(-C_K nt\right).$$

In view of Equation (41), we then obtain that, conditionally on $\mathcal{D}_n^{(1)}$,

$$\begin{aligned}\mathbb{E}\left[\max_{\theta\in\mathcal{M}_\varepsilon}(D_{f_\theta}-2\widehat{D}_{f_\theta})\mathbb{1}_{\{\Omega\}}|\mathcal{D}_n^{(1)}\right] &\leq \max\left(\Delta_n,\frac{C\log(|\mathcal{M}_\varepsilon|)}{n}\right)\\ &\quad+\int_{C\log(\mathcal{M}_\varepsilon)/n}^{+\infty}|\mathcal{M}_\varepsilon|\exp(-Cnt)\,\mathrm{d}t.\end{aligned}$$

As before, we use that $\mathbb{E}\left[\Delta_n\right]\leq C/n$, and we deduce from the above inequality by integrating over $\mathcal{D}_n^{(1)}$ that

$$\mathbb{E}\left[\max_{\theta\in\mathcal{M}_\varepsilon}(D_{f_\theta}-2\widehat{D}_{f_\theta})\mathbb{1}_{\{\Omega\}}\right]\leq\frac{C\log(|\mathcal{M}_\varepsilon|)}{n}.$$

Since for $\varepsilon=1/(\log(n)n^3M)$ we have that $\log(|\mathcal{M}_\varepsilon|)\leq C(M+s^*)\log(nM)$, we obtain from the above inequality and Equation (40) that

$$\mathbb{E}[\mathcal{R}_2(\widehat{f})-\mathcal{R}_2(f^*)]\leq C\frac{(M+s^*)\log(nM)}{n}.$$

From the above inequality, we get the desired by applying the Zhang's lemma

$$\mathbb{E}[\mathcal{R}(\hat{g})-\mathcal{R}(g^*)]\leq\frac{1}{\sqrt{2}}\left(\mathbb{E}[\mathcal{R}_2(\widehat{f})-\mathcal{R}_2(f^*)]\right)^{1/2}.$$

□