Annals of the Institute of Statistical Mathematics (2020) 72:297–328
https://doi.org/10.1007/s10463-018-0692-7

Asymptotic theory of the adaptive Sparse Group Lasso

Benjamin Poignard[1]

Received: 9 January 2018 / Revised: 3 August 2018 / Published online: 11 October 2018

Abstract
We study the asymptotic properties of a new version of the Sparse Group Lasso estimator (SGL), called adaptive SGL. This new version includes two distinct regularization parameters, one for the Lasso penalty and one for the Group Lasso penalty, and we consider the adaptive version of this regularization, where both penalties are weighted by preliminary random coefficients. The asymptotic properties are established in a general framework, where the data are dependent and the loss function is convex. We prove that this estimator satisfies the oracle property: the sparsity-based estimator recovers the true underlying sparse model and is asymptotically normally distributed. We also study its asymptotic properties in a double-asymptotic framework, where the number of parameters diverges with the sample size. We show by simulations and on real data that the adaptive SGL outperforms other oracle-like methods in terms of estimation precision and variable selection.

Keywords Asymptotic normality · Consistency · Oracle property

1 Introduction

High-dimensional statistical modeling is concerned with the significantly large number of parameters to estimate. For instance, predicting a single outcome is not an easy challenge since the exact functional form used to predict this outcome is rarely known. A consequence is that the researcher/practitioner is faced with a large set of potential variables formed by all the different ways of interacting and altering these underlying variables. There are different methods for developing prediction models within the high-dimensional framework to tackle the over-fitting problem. The key point is model

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-018-0692-7) contains supplementary material, which is available to authorized users.

✉ Benjamin Poignard
poignard@sigmath.es.osaka-u.ac.jp

[1] Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

regularization/penalization—these terms are equivalent—or dimension reduction. In (semi-) parametric models, which require the estimation of many parameters with respect to the sample size, the parameters need to be constrained to avoid the overfitting issue.

A significant literature developed on model penalization. For instance, the Akaike's or Bayesian information criteria aim at selecting the size of a model. However, these methods are unstable, computationally complex, and their sampling properties are difficult to study as Fan and Li (2001) pointed out mainly because they are stepwise and subset selection procedures.

The Lasso procedure of Tibshirani (1996) overcomes these drawbacks as it simultaneously performs variable selection and model estimation. It then fosters sparsity and allows for continuity of the selected models. Other penalties were proposed such as the smoothly clipped absolute deviation (SCAD) of Fan (1997), which modifies the Lasso to shrink large coefficients less severely. The elastic net regularization procedure of Zou and Hastie (2005) was developed to overcome the collinearity between the variables, which hampers the Lasso to perform well. Their idea consists of mixing a l^1 penalty, which performs variable selection, with a l^2 penalty, which stabilizes the solution paths. The Group Lasso of Yuan and Lin (2006) fosters sparsity at a group level. Simon et al. (2013) designed the Sparse Group Lasso to foster sparsity both at a group level and within a group using one regularization parameter. Their penalization involves a combination of a l^1 Lasso-type penalty and a mixed l^1/l^2 penalty for group selection.

Knight and Fu (2000) extensively explored the asymptotic properties of the Lasso penalty for OLS loss functions. Fan and Li (2001) generalized this penalization framework to general likelihood functions and studied the asymptotic properties of the SCAD penalty. They proved that the SCAD estimator satisfies the oracle property, that is, the sparsity-based estimator recovers the true underlying sparse model and is asymptotically normally distributed. This property is actually not satisfied by the Lasso as proposed by Tibshirani. To fix this drawback, Zou (2006) proposed the adaptive Lasso, where adaptive weights are used to penalize different coefficients in the penalty. Nardi and Rinaldo (2008) applied the same methodology for the Group Lasso estimator and studied its oracle property.

These theoretical studies were developed for fixed dimensional models with i.i.d. data, a case where the number of parameters does not depend on the sample size, and for least square-type loss functions, except Fan and Li (2001). The high-dimensional setting was later considered by Fan and Peng (2004), who focused on a penalized likelihood framework when the number of parameters diverges—also called double-asymptotic—with the sample size. In this work, the authors prove that the SCAD penalty satisfies the oracle property. Zou and Zhang (2009) also focused on the oracle property of the adaptive elastic-net within the double-asymptotic framework. Their work highlights that adaptive weights penalizing different coefficients are key quantities to satisfy the oracle property as one can modify the convergence rate of the penalty terms. Nardi and Rinaldo (2008) also proposed within the double-asymptotic setting selection consistency results for the Group Lasso, which states that asymptotically the true set of relevant variables is selected.

In this paper, our first contribution is to propose a generalization of the Sparse Group Lasso (SGL) estimator, initially proposed by Simon et al. (2013), and we perform its asymptotic theory, a work that has not previously been performed. More precisely, our proposed generalization consists in the specification of two regularization parameters controlling for the sparsity degree, one for the l^1 Lasso term and one for the l^1/l^2 Group Lasso term. Our asymptotic results emphasize the trade-off between the group regularization and the within-group regularization. Our second contribution is to perform this theory for dependent data and for any convex loss function with respect to the parameters, for both parametric and semi-parametric models.

Our results show that the SGL as proposed by Simon et al. (2013) does not satisfy the oracle property. We thus propose a new version of the SGL, the adaptive SGL using the same methodology of Zou (2006), which consists of penalizing different coefficients and groups of coefficients using random weights that are positive functions of a first step estimator. This enables to alter the rate of convergence of the penalties to satisfy the oracle property. Our theoretical results also point out the trade-off between these random weights depending if they are related to the l^1 or l^1/l^2 term. Our work is influenced by Fan and Peng (2004) concerning the oracle property for general penalized loss functions and by Zou and Zhang (2009) regarding the modeling of random weights penalizing the coefficients differently. We also prove that the adaptive SGL satisfies the oracle property in a double-asymptotic framework. In this setting, where the number d of parameters diverges with the sample size T, the dimension evolves as $d = O(T^c)$ with $0 < c < 1/5$, a rate that is required to satisfy the oracle property.

The rest of the paper is organized as follows: In Sect. 2, we describe our general framework for penalized convex empirical criteria and the SGL penalty. In Sect. 3, we derive the optimality conditions of the statistical criterion. In Sect. 4, we derive the asymptotic properties of both the SGL and adaptive SGL when the number of parameters is fixed. In Sect. 5, we prove the oracle property of the adaptive SGL in a double-asymptotic setting. In Sect. 6, we use simulations and real data to compare the finite sample performance of the adaptive Sparse Group Lasso with other competitors. Appendix provides some preliminary results and the proofs of Sect. 4.

2 Framework and notations

2.1 Loss function, vector of parameters, sparsity assumption

We observe at time t the vector $\boldsymbol{\epsilon}_t \in \mathbb{R}^N$, $N \geq 1$ and consider a dynamic system in which the criterion is

$$\boldsymbol{\theta} \mapsto \mathbb{G}_T l(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}, \underline{\boldsymbol{\epsilon}}_{t-1}), \tag{1}$$

with $\underline{\boldsymbol{\epsilon}}_{t-1} = (\boldsymbol{\epsilon}_s, s \leq t-1)$ and $\boldsymbol{\theta} \in \mathbb{R}^d, d \geq 1$. $l(.)$ is a generic known loss function on the sample space so that for any process $(\boldsymbol{\epsilon}_t)$, $\boldsymbol{\theta} \mapsto l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}, \underline{\boldsymbol{\epsilon}}_{t-1})$ is convex. This framework includes both parametric and semi-parametric models: for instance, the maximum likelihood method—under the convexity assumption—where the $l(.)$

function corresponds to $l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta},\underline{\boldsymbol{\epsilon}}_{t-1}) = -\log f(\boldsymbol{\epsilon}_t;\boldsymbol{\theta},\underline{\boldsymbol{\epsilon}}_{t-1})$, with $f(\boldsymbol{\epsilon}_t;\boldsymbol{\theta},\underline{\boldsymbol{\epsilon}}_{t-1})$ the density of the observation $(\boldsymbol{\epsilon}_t)$ under $\mathbb{P}_{\boldsymbol{\theta}}$ given the past observations $\underline{\boldsymbol{\epsilon}}_{t-1}$. Alternatively, a linear regression model would be $l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta},\underline{\boldsymbol{\epsilon}}_{t-1}) = \|\boldsymbol{\epsilon}_t - \boldsymbol{\theta}' g(\underline{\boldsymbol{\epsilon}}_{t-1})\|_p$, where $g(\underline{\boldsymbol{\epsilon}}_{t-1})$ corresponds to a transformation of the past observations. For instance, this includes a model where one may want to predict a component of $\boldsymbol{\epsilon}_t$ by its past and by the other components of $\boldsymbol{\epsilon}_{t-1}$ using a linear regression. To simplify the notations, we will keep $l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta})$ instead of $l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta},\underline{\boldsymbol{\epsilon}}_{t-1})$. We denote the empirical score and Hessian of the empirical criterion, respectively, as

$$\dot{\mathbb{G}}_T l(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}),\ \ddot{\mathbb{G}}_T l(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}'} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}).$$

The parameter vector $\boldsymbol{\theta}$ of size d can be split into m groups $\mathcal{G}_k, k = 1,\ldots,m$, so that $\text{card}(\mathcal{G}_k) = \boldsymbol{c}_k, \boldsymbol{c}_1 + \cdots + \boldsymbol{c}_m = d$ and these groups are non-overlapping. We use the notation $\boldsymbol{\theta}^{(l)}$ as the subvector of $\boldsymbol{\theta}$, that is, the set $\{\theta_k : k \in \mathcal{G}_l\}$. Hence the vector $\boldsymbol{\theta} = (\theta_j,\ j = 1,\ldots,d)$ can be written as $\boldsymbol{\theta} = (\theta_i^{(k)}, k \in \{1,\ldots,m\}, i = 1,\ldots,\boldsymbol{c}_k)$. The size d of $\boldsymbol{\theta}$ corresponds to the number of covariates for predicting the outcome. For instance, in a Cox proportional hazards model, which is semi-parametric—no estimation of the hazard function—the estimation relies on a log-partial likelihood function that is convex with respect to the parameters. The hazard rate at time t can be predicted using m group of covariates, each group containing variables that are correlated. We denote by $\boldsymbol{\theta}_0$ the true parameter vector of interest. Moreover, $\boldsymbol{\theta} \to \mathbb{E}[l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta})]$ is supposed to be a one-to-one mapping and is ***minimized uniquely*** at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

We assume that the true vector of parameters $\boldsymbol{\theta}_0$ is ***sparse***, that is, the number of nonzero components among $\boldsymbol{\theta}_0$ is strictly inferior to d. We denote by $\mathcal{S} := \{k : \boldsymbol{\theta}^{(k)} \neq 0\}$ the set of indices for which the groups are active and $\mathcal{A} := \{j : \theta_{0,j} \neq 0\}$ the true subset model so that $\text{card}(\mathcal{A}) < d$. This set can be decomposed into subgroups of active sets as $l \in \mathcal{S}$, $\mathcal{A}_l = \{(l,i) : \theta_{0,i}^{(l)} \neq 0\}$. Besides, there are inactive indices $\mathcal{G}_l \backslash \mathcal{A}_l = \mathcal{A}_l^c = \{(l,i) : \theta_{0,i}^{(l)} = 0\}$. We have $\{l \notin \mathcal{S}\} \Leftrightarrow \{\forall i = 1,\ldots,\boldsymbol{c}_l, \theta_{0,i}^{(l)} = 0\}$. In this setting, $\mathcal{A} = \underset{l\in\mathcal{S}}{\cup}\mathcal{A}_l$ such that for $k \neq l$, $\mathcal{A}_k \cap \mathcal{A}_l = \emptyset$. Furthermore, $\mathcal{A}^c = \overset{m}{\underset{l=1}{\cup}}\mathcal{A}_l^c$ such that for $k \neq l$, $\mathcal{A}_k^c \cap \mathcal{A}_l^c = \emptyset$.

Based on these notations, we denote $\dot{\mathbb{G}}_T l(\boldsymbol{\theta})_{(k)} \in \mathbb{R}^{\boldsymbol{c}_k}$ the "score" vector of the empirical criterion taken over group k of size $\boldsymbol{c}_k$, $\dot{\mathbb{G}}_T l(\boldsymbol{\theta})_{(k),i} \in \mathbb{R}$ the ith component of this score, and $\dot{\mathbb{G}}_T l(\boldsymbol{\theta})_{\mathcal{A}} \in \mathbb{R}^{\text{card}(\mathcal{A})}$ the score over the set of active indices. $\ddot{\mathbb{G}}_T l(\boldsymbol{\theta})_{(k)(k)} \in \mathcal{M}_{\boldsymbol{c}_k\times\boldsymbol{c}_k}(\mathbb{R})$ (resp. $\mathbb{H}_{(k)(k)}$) is the empirical (resp. theoretical) Hessian taken over the block representing group k, and $\ddot{\mathbb{G}}_T l(\boldsymbol{\theta})_{\mathcal{A}\mathcal{A}} \in \mathcal{M}_{\text{card}(\mathcal{A})\times\text{card}(\mathcal{A})}(\mathbb{R})$ is the Hessian over the set of active indices.

2.2 Statistical problem: Sparse Group Lasso penalization

The main problem is to recover $\mathcal{A}$ by the SGL regularization. The statistical problem consists of minimizing over the parameter space Θ a penalized criterion of the form

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \{\mathbb{G}_T \varphi(\boldsymbol{\theta})\}, \tag{2}$$

where $\mathbb{G}_T \varphi(\boldsymbol{\theta}) = \mathbb{G}_T l(\boldsymbol{\theta}) + \boldsymbol{p}_1(\lambda_T, \boldsymbol{\theta}) + \boldsymbol{p}_2(\gamma_T, \boldsymbol{\theta})$ and

$$\boldsymbol{p}_1 : \mathbb{R}_+ \times \mathbb{R}_+^m \times \Theta \to \mathbb{R}_+ \text{ with } (\lambda_T, \alpha, \boldsymbol{\theta}) \mapsto \boldsymbol{p}_1(\lambda_T, \boldsymbol{\theta}) = \frac{\lambda_T}{T} \sum_{k=1}^{m} \alpha_k \|\boldsymbol{\theta}^{(k)}\|_1,$$
$$\boldsymbol{p}_2 : \mathbb{R}_+ \times \mathbb{R}_+^m \times \Theta \to \mathbb{R}_+ \text{ with } (\gamma_T, \xi, \boldsymbol{\theta}) \mapsto \boldsymbol{p}_2(\gamma_T, \boldsymbol{\theta}) = \frac{\gamma_T}{T} \sum_{l=1}^{m} \xi_l \|\boldsymbol{\theta}^{(l)}\|_2.$$

Both α_k and ξ_l are known nonnegative scalar weights; they both can take the same value and can be set as $\sqrt{\boldsymbol{c}_k}$: see Yuan and Lin (2006) or Simon et al. (2013). The regularization parameters (also called tuning parameters) λ_T and γ_T vary with T. This proposed regularization procedure generalizes the SGL as proposed by Simon et al. (2013) as each penalty is specified with its own regularization parameter. Asymptotically, their relative convergence rate plays a key role when deriving consistency and distribution results.

The estimator $\hat{\boldsymbol{\theta}}$ obtained in (2) is not the minimum of the empirical unpenalized criterion $\mathbb{G}_T l(.)$. Our main interest is to analyze the bias generated by the penalties and how the oracle property can be satisfied in the sense of Fan and Li (2001). More precisely, the sparsity-based estimator must satisfy

(i) $\hat{\mathcal{A}} = \{i : \hat{\theta}_i \neq 0\} = \mathcal{A}$ asymptotically—in probability—that is *model selection consistency*.
(ii) $\sqrt{T}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}_0)$ with $\mathbb{V}_0$ a covariance matrix related to the criterion of interest.

We highlight in Proposition 1, Sect. 4 that actually the SGL as proposed in (2) generalizing the SGL of Simon et al. (2013) cannot perform the oracle property. Hence in Sect. 4, we propose a new estimator based on the same idea as Zou (2006), the adaptive Sparse Group Lasso, for which the oracle property is obtained when the weights are randomized, as proved in Theorem 5.

3 Optimality conditions

The statistical problem consists of solving (2). Both $\mathbb{G}_T l(.)$, $\boldsymbol{p}_1(\lambda_T, \alpha, .)$ and $\boldsymbol{p}_2(\gamma_T, \xi, .)$ are convex functions, and there are no inequality constraints. Consequently, by the Karush–Kuhn–Tucker optimality conditions, which are necessary and sufficient since the problem is convex, the estimator $\hat{\boldsymbol{\theta}}$ satisfies for a group k

$$\dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}})_{(k)} + \lambda_T T^{-1} \alpha_k \hat{\boldsymbol{w}}^{(k)} + \gamma_T T^{-1} \xi_k \hat{\boldsymbol{z}}^{(k)} = 0, \tag{3}$$

where $\boldsymbol{w}^{(k)}$ and $\boldsymbol{z}^{(k)}$ are subgradient vectors, respectively, of $\|\hat{\theta}^{(k)}\|_1$ and $\|\hat{\theta}^{(k)}\|_2$ satisfying for $i = 1, \ldots, \boldsymbol{c}_k$

$$\hat{\boldsymbol{w}}_i^{(k)} = \text{sgn}(\hat{\theta}_i^{(k)}) \text{ if } \hat{\theta}_i^{(k)} \neq 0 \text{ and } \hat{\boldsymbol{w}}_i^{(k)} \in \{\hat{\boldsymbol{w}}_i^{(k)} : |\hat{\boldsymbol{w}}_i^{(k)}| \leq 1\} \text{ if } \hat{\theta}_i^{(k)} = 0,$$
$$\hat{z}^{(k)} = \hat{\boldsymbol{\theta}}^{(k)}/\|\hat{\boldsymbol{\theta}}^{(k)}\|_2 \text{ if } \hat{\boldsymbol{\theta}}^{(k)} \neq \mathbf{0} \text{ and } \hat{z}^{(k)} \in \{\hat{z}^{(k)} : \|\hat{z}^{(k)}\|_2 \leq 1\} \text{ if } \hat{\boldsymbol{\theta}}^{(k)} = \mathbf{0}.$$

If $\hat{\boldsymbol{\theta}}^{(k)} \neq 0$, the criterion function $\mathbb{G}_T\varphi(.)$ is partially differentiable with respect to $\boldsymbol{\theta}^{(k)}$ and it is necessary and sufficient that these partial derivatives are zero due to the convexity.

Now with $\hat{\boldsymbol{\theta}}^{(k)} = \mathbf{0}$, from (3), we obtain

$$\sum_{i=1}^{\boldsymbol{c}_k} \left(\dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}})_{(k),i} + \lambda_T T^{-1} \alpha_k \hat{\boldsymbol{w}}_i^{(k)}\right)^2 = \sum_{i=1}^{\boldsymbol{c}_k} \left(\gamma_T T^{-1} \xi_k \hat{z}_i^{(k)}\right)^2 \leq \gamma_T^2 T^{-2} \xi_k^2 \|\hat{z}^{(k)}\|_2^2.$$

The subgradient equations are satisfied with $\hat{\boldsymbol{\theta}}^{(k)} = 0$ if

$$\|\dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}})_{(k)} + \lambda_T T^{-1} \alpha_k \hat{\boldsymbol{w}}^{(k)}\|_2 \leq \gamma_T T^{-1} \xi_k.$$

The subgradient equations also provide understanding into the sparsity within a group that is partially a nonzero group. With $\hat{\boldsymbol{\theta}}^{(k)} \neq \mathbf{0}$, the subgradient condition for $\hat{\boldsymbol{\theta}}_i^{(k)}$ becomes

$$\forall i = 1, \ldots, \boldsymbol{c}_k, -\dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}})_{(k),i} = \lambda_T T^{-1} \alpha_k \hat{\boldsymbol{w}}_i^{(k)} + \gamma_T T^{-1} \xi_k \frac{\hat{\theta}_i^{(k)}}{\|\hat{\boldsymbol{\theta}}^{(k)}\|_2}.$$

This subgradient equation is satisfied for $\hat{\theta}_i^{(k)} = 0$ when

$$|\dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}})_{(k),i}| \leq \lambda_T T^{-1} \alpha_k.$$

Bertsekas (1995) proposed the use of subdifferential calculus to characterize necessary and sufficient solutions for problems such as (2). The conditions we derived are close to those of Simon et al. (2013) (obtained for a least squares loss function). They will be extensively used in the rest of the paper.

4 Asymptotic properties: fixed *d*

In this section, we consider the *fixed* dimensional case only. All the proofs of this section can be found in Appendix. To prove the asymptotic results, we make the following assumptions.

Assumption 1 $(\boldsymbol{\epsilon}_t)$ is a strictly stationary and ergodic process.

Assumption 2 The parameter set $\Theta \subset \mathbb{R}^d$ is convex and not necessarily compact.

Assumption 3 For any $(\boldsymbol{\epsilon}_t)$, the function $\boldsymbol{\theta} \mapsto l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta})$ is convex and $C^3(\mathbb{R}, \Theta)$.

Assumption 4 $(\nabla l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0))$ is a square integrable martingale difference.

Assumption 5 $\mathbb{H} := \mathbb{E}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}'} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)]$ and $\mathbb{M} := \mathbb{E}[\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}'} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)]$ exist and are positive definite.

Assumption 6 Let $\upsilon_t(C) = \sup_{k,l,m=1,\ldots,d} \{ \sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2 \leq \nu_T C} |\partial^3_{\theta_k\theta_l\theta_m} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta})|\}$, where $C > 0$ is a fixed constant and $\nu_T \underset{T\to\infty}{\longrightarrow} 0$, a quantity that will be made explicit in the Appendix. Then

$$\eta(C) := \frac{1}{T^2}\sum_{t,t'=1}^{T} \mathbb{E}[\upsilon_t(C)\upsilon_{t'}(C)] < \infty.$$

4.1 Non-adaptive version of the Sparse Group Lasso estimator

We focus on the large sample properties of the estimator given by (2).

Theorem 1 *Under Assumptions 1–3, if $\lambda_T/T \to \lambda_0 \geq 0$ and $\gamma_T/T \to \gamma_0 \geq 0$, then for any compact set $\boldsymbol{B} \subset \Theta$ such that $\boldsymbol{\theta}_0 \in \boldsymbol{B}$,*

$$\hat{\boldsymbol{\theta}} \xrightarrow{\mathbb{P}} \arg\min_{\boldsymbol{x}\in\boldsymbol{B}}\{\mathbb{G}_\infty\varphi(\boldsymbol{x})\} = \boldsymbol{\theta}_0^*,$$

with

$$\mathbb{G}_\infty\varphi(\boldsymbol{x}) = \mathbb{G}_\infty l(\boldsymbol{x}) + \lambda_0\sum_{k=1}^{m}\alpha_k\|\boldsymbol{x}^{(k)}\|_1 + \gamma_0\sum_{l=1}^{m}\xi_l\|\boldsymbol{x}^{(l)}\|_2,$$

where $\mathbb{G}_\infty l(.)$ is the limit in probability of $\mathbb{G}_T l(.)$. Hence if $\lambda_T = o(T)$ and $\gamma_T = o(T)$, then $\hat{\boldsymbol{\theta}}$ is consistent.

The penalized estimator does not converge to $\boldsymbol{\theta}_0$ under the convergence rate $\lambda_T = O(T)$ and $\gamma_T = O(T)$. We assumed that $\mathbb{G}_\infty l(\boldsymbol{x})$ admits unique minimum in Sect. 2.1 so that the solution $\boldsymbol{\theta}_0^*$ is unique.

The next result provides an explicit convergence for the SGL estimator.

Theorem 2 *Under Assumptions 1–3 and 6, the sequence of penalized estimators $\hat{\boldsymbol{\theta}}$ satisfies*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p\left(T^{-1/2} + \lambda_T T^{-1}a + \gamma_T T^{-1}b\right),$$

when $\lambda_T = o(T)$ and $\gamma_T = o(T)$, and $a := card(\mathcal{A})(\max_k \alpha_k)$, $b := card(\mathcal{A})(\max_l \xi_l)$ satisfy $\lambda_T T^{-1}a_T \to 0$ and $\gamma_T T^{-1}b_T \to 0$.

This result highlights that if $\lambda_T T^{-1} = O(T^{-1/2})$ and $\gamma_T T^{-1} = O(T^{-1/2})$, then we would obtain a $\sqrt{T}$-consistent $\hat{\boldsymbol{\theta}}$.

To derive the asymptotic distribution, we rely on the convexity property of $\varphi(.)$, and hence of $\mathbb{G}_T\varphi(.)$. The intuition is as follows: Let $\mathbb{F}_T(\boldsymbol{u})$ and $\mathbb{F}_\infty(\boldsymbol{u})$, $\boldsymbol{u} \in \mathbb{R}^d$, be random convex functions such that their minimum are, respectively, $\boldsymbol{u}_T$ and $\boldsymbol{u}_\infty$. Then if $\mathbb{F}_T(.)$ converges in finite distribution to $\mathbb{F}_\infty(.)$, and $\boldsymbol{u}_\infty$ is the unique minimum of $\mathbb{F}_\infty$ with probability one, then $\boldsymbol{u}_T$ converges weakly to $\boldsymbol{u}_\infty$. This method to prove the convergence of arg min processes is called the *convexity argument*. It was developed by Pollard (1991), Davis et al. (1992), Geyer (1996) or Kato (2009). The convexity argument is stated as a lemma in the proof of Theorem 4.1 in Chernozhukov (2005). It is reported in Appendix.

Theorem 3 *Under Assumptions 1–6, if* $\lambda_T T^{-1/2} \to \lambda_0$ *and* $\gamma_T T^{-1/2} \to \gamma_0$, *then*

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \underset{\boldsymbol{u}\in\mathbb{R}^d}{\arg\ \min}\{\mathbb{F}_\infty(\boldsymbol{u})\},$$

provided $\mathbb{F}_\infty$ *is the random function in* $\mathbb{R}^d$, *where*

$$\mathbb{F}_\infty(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}'\mathbb{H}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{Z} + \lambda_0 \sum_{k=1}^{m}\alpha_k \sum_{i=1}^{\boldsymbol{c}_k}\left\{|\boldsymbol{u}_i^{(k)}|\mathbf{1}_{\theta_{0,i}^{(k)}=0} + \boldsymbol{u}_i^{(k)} sgn(\theta_{0,i}^{(k)})\mathbf{1}_{\theta_{0,i}^{(k)}\neq 0}\right\}$$
$$+ \gamma_0 \sum_{l=1}^{m}\xi_l\left\{\|\boldsymbol{u}^{(l)}\|_2\mathbf{1}_{\boldsymbol{\theta}_0^{(l)}=\mathbf{0}} + \frac{\boldsymbol{u}^{(l)'}\boldsymbol{\theta}_0^{(l)}}{\|\boldsymbol{\theta}_0^{(l)}\|_2}\mathbf{1}_{\boldsymbol{\theta}_0^{(l)}\neq\mathbf{0}}\right\},$$

with $\mathbb{H} = \mathbb{H}(\boldsymbol{\theta}_0) := \mathbb{E}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}'} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0)]$ *and some random vector* $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbb{M})$, $\mathbb{M} = \mathbb{M}(\boldsymbol{\theta}_0) := \mathbb{E}[\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}'} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0)]$.

The previous result establishes the $\sqrt{T}$-consistency of the SGL estimator. However, for $\lambda_T = O(\sqrt{T})$ and $\gamma_T = O(\sqrt{T})$, the true active set $\mathcal{A}$ can not be recovered with high probability as stated in the next proposition.

Proposition 1 *Under Assumption 1–6, if* $\lambda_T T^{-1/2} \to \lambda_0$ *and* $\gamma_T T^{-1/2} \to \gamma_0$, *then*

$$\limsup_{T\to\infty} \mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) \leq c < 1,$$

where c is a constant depending on the true model.

Proposition 1 shows that the SGL estimator as specified in (2) does not satisfy the oracle property. To fix this issue, Zou (2006) proposed the adaptive Lasso and Nardi and Rinaldo (2008) the adaptive Group Lasso for OLS models. The idea is to alter the convergence rate of the regularization parameters by considering random weights to penalize the coefficients differently. We propose to use this methodology in our SGL regularization procedure.

4.2 Adaptive SGL-regularized loss function

The adaptive specification of the proposed estimator is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\,\{\mathbb{G}_T\psi(\boldsymbol{\theta})\}, \tag{4}$$

with $\mathbb{G}_T\psi(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}) + \boldsymbol{p}_1(\lambda_T,\tilde{\boldsymbol{\theta}},\boldsymbol{\theta}) + \boldsymbol{p}_2(\gamma_T,\tilde{\boldsymbol{\theta}},\boldsymbol{\theta})$, both penalties are

$$\boldsymbol{p}_1(\lambda_T,\tilde{\boldsymbol{\theta}},\boldsymbol{\theta}) = \lambda_T T^{-1}\sum_{k=1}^{m}\sum_{i=1}^{\boldsymbol{c}_k}\alpha(\tilde{\theta}_i^{(k)})|\theta_i^{(k)}|,\ \boldsymbol{p}_2(\gamma_T,\tilde{\boldsymbol{\theta}},\boldsymbol{\theta}) = \gamma_T T^{-1}\sum_{l=1}^{m}\xi(\tilde{\boldsymbol{\theta}}^{(l)})\|\boldsymbol{\theta}^{(l)}\|_2.$$

These penalties are now randomized through the $\tilde{\boldsymbol{\theta}}$ argument in the weights α's and ξ's. This first step estimator $\tilde{\boldsymbol{\theta}}$ is supposed to be a $T^{1/2}$-consistent estimator of $\boldsymbol{\theta}_0$. For instance, it can be defined as an M-estimator of the unpenalized empirical criterion $\mathbb{G}_T l(.)$, that is, $\tilde{\boldsymbol{\theta}} = \arg\min\,\mathbb{G}_T l(\boldsymbol{\theta})$ with $\boldsymbol{\theta}\in\Theta$. The weights are now random, and for any group k or l, $\boldsymbol{\alpha}(\tilde{\boldsymbol{\theta}}^{(\boldsymbol{k})}) \in \mathbb{R}_+^{\boldsymbol{c_k}}$, $\xi(\tilde{\boldsymbol{\theta}}^{(l)}) \in \mathbb{R}_+$ are specified as

$$\overset{(k)}{T} := (\tilde{\boldsymbol{\theta}}^{(k)}) = (|\tilde{\theta}_i^{(k)}|^{-\eta}, i = 1,\ldots,\boldsymbol{c}_k),\ \xi_{T,l} := \xi(\tilde{\boldsymbol{\theta}}^{(l)}) = \|\tilde{\boldsymbol{\theta}}^{(l)}\|_2^{-\mu},$$

for some constants $\eta > 0$ and $\mu > 0$ (to be specified).

Theorem 4 *Under Assumptions 1–3 and 6, the sequence of penalized estimators $\hat{\boldsymbol{\theta}}$ satisfies*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p\left(T^{-1/2} + \lambda_T T^{-1} a_T + \gamma_T T^{-1} b_T\right),$$

with $a_T = card(\mathcal{A}).(\max_{k\in\mathcal{S}}(\max_{i\in\mathcal{A}_k}\alpha_{T,i}^{(k)}))$, $b_T = card(\mathcal{A}).(\max_{l\in\mathcal{S}}\xi_{T,l})$ stochastic quantities, such that $\lambda_T T^{-1} a_T \xrightarrow{\mathbb{P}} 0$ and $\gamma_T T^{-1} b_T \xrightarrow{\mathbb{P}} 0$.

This result is similar to Theorem 2, the difference being that a_T and b_T are stochastic. The following theorem shows that the adaptive SGL satisfies the oracle property under proper convergence rates of λ_T and γ_T and provides the trade-off between the l^1/l^2 regularizer and the l^1 regularizer.

Theorem 5 *Under Assumptions 1–6, if $\lambda_T T^{-1/2} \to 0$, $\gamma_T T^{-1/2} \to 0$, $T^{(\eta-1)/2}\lambda_T \to \infty$, $T^{(\mu-1)/2}\gamma_T \to \infty$ and $T^{(\mu-\eta)/2}\gamma_T\lambda_T^{-1} \to \infty$, then $\hat{\boldsymbol{\theta}}$ obtained in (4) satisfies*

$$\lim_{T\to\infty}\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) = 1,\ \textit{and}$$
$$\sqrt{T}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{H}_{\mathcal{A}\mathcal{A}}^{-1}\mathbb{M}_{\mathcal{A}\mathcal{A}}\mathbb{H}_{\mathcal{A}\mathcal{A}}^{-1}\right).$$

5 Asymptotic properties: diverging *d*

5.1 Framework, assumptions and properties

From now on, we consider the case where $d = d_T$, so that $d_T \to \infty$ as $T \to \infty$. We have $\text{card}(\mathcal{S}) = O(\text{card}(\mathcal{A})) = O(d_T)$. The dimension is supposed to be $d_T = O(T^c)$ for some $q_2 < c < q_1$. In this section, we prove that the adaptive SGL satisfies the oracle property for proper choices of $0 \leq q_2 < q_1 < 1$. This work has not been performed so far for the SGL estimator. All the proofs of this section are reported in the Supplementary File.

The quantities depend on d_T, hence on T and should be indexed by T. We denote $\mathbb{H}_T := \mathbb{E}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0)]$ and $\mathbb{M}_T := \mathbb{E}[\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0) \nabla_{\boldsymbol{\theta}'} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0)]$ in the rest of the paper. To make the reading easier, we do not index other quantities by T, which will be implicit. The criterion is

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \{\mathbb{G}_T l(\boldsymbol{\theta}) + \frac{\lambda_T}{T} \sum_{k=1}^{m} \sum_{i=1}^{\boldsymbol{c}_k} \alpha_{T,i}^{(k)} |\theta_i^{(k)}| + \frac{\gamma_T}{T} \sum_{l=1}^{m} \xi_{T,l} \|\boldsymbol{\theta}^{(l)}\|_2\}, \tag{5}$$

with $\alpha_{T,i}^{(k)} = |\tilde{\theta}_i^{(k)}|^{-\eta}$ and $\xi_{T,l} = \|\tilde{\boldsymbol{\theta}}^{(l)}\|_2^{-\mu}$, where $\eta > 0$, $\mu > 0$, and $\tilde{\boldsymbol{\theta}}$ is a first step estimator satisfying $\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \{\mathbb{G}_T l(\boldsymbol{\theta})\}$.

The two next assumptions are similar to condition (F) of Fan and Peng (2004) and allow for controlling the minimum and maximum eigenvalues of the limits of the empirical Hessian and the score cross-product. We denote by $\lambda_{\min}(\boldsymbol{M})$ and $\lambda_{\max}(\boldsymbol{M})$ the minimum and maximum eigenvalues of any positive definite square matrix $\boldsymbol{M}$.

Assumption 7 $\mathbb{H}_T$ and $\mathbb{M}_T$ exist. $\mathbb{H}_T$ is non-singular, and there exist b_1, b_2 with $0 < b_1 < b_2 < \infty$ and c_1, c_2 with $0 < c_1 < c_2 < \infty$ such that, for all T,

$$b_1 < \lambda_{\min}(\mathbb{M}_T) < \lambda_{\max}(\mathbb{M}_T) < b_2, \; c_1 < \lambda_{\min}(\mathbb{H}_T) < \lambda_{\max}(\mathbb{H}_T) < c_2.$$

Let $\mathbb{V}_T = \mathbb{H}_T^{-1} \mathbb{M}_T \mathbb{H}_T^{-1}$, we deduce there exist a_1, a_2 with $0 < a_1 < a_2 < \infty$ such that, for all T, $a_1 < \lambda_{\min}(\mathbb{V}_T) < \lambda_{\max}(\mathbb{V}_T) < a_2$.

Assumption 8 $\mathbb{E}[\{\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0) \nabla_{\boldsymbol{\theta}'} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0)\}^2] < \infty$, for every d_T (and then of T).

Assumption 9 There exist some functions $\Psi(.)$ such that, for all T,

$$\sup_{k=1,\ldots,d_T} \mathbb{E}[\partial_{\boldsymbol{\theta}_k} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}) \partial_{\boldsymbol{\theta}_k} l(\boldsymbol{\epsilon}_{t'}; \boldsymbol{\theta})] \leq \Psi(|t - t'|), \quad \text{and} \quad \sup_T \frac{1}{T} \sum_{t,t'=1}^{T} \Psi(|t - t'|) < \infty.$$

Assumption 10 Let $\zeta_{kl,t} := \partial^2_{\theta_k \theta_l} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0) - \mathbb{E}[\partial^2_{\theta_k \theta_l} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0)]$. There exist some functions $\chi(.)$ such that

$$|\mathbb{E}[\zeta_{kl,t} \zeta_{k'l',t'}]| \leq \chi(|t - t'|), \quad \text{and} \quad \sup_T \frac{1}{T} \sum_{t,t'=1}^{T} \chi(|t - t'|) < \infty.$$

Assumption 11 Let $\upsilon_t(C) := \sup_{k,l,m=1,\ldots,d_T} \{ \sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2 \leq \nu_T C} |\partial^3_{\theta_k\theta_l\theta_m} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta})|\}$, where $C > 0$ is a fixed constant and $\nu_T = (d_T/T)^{1/2}$. Then

$$\eta(C) := \frac{1}{T^2}\sum_{t,t'=1}^{T} \mathbb{E}[\upsilon_t(C)\upsilon_{t'}(C)] < \infty.$$

Theorem 6 *Under Assumptions 1–3, 7–11 and if $d_T^4 = o(T)$, the sequence of unpenalized M-estimators solving $\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta}\{\mathbb{G}_T l(\boldsymbol{\theta})\}$ satisfies*

$$\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = O_p\left(\left(\frac{d_T}{T}\right)^{\frac{1}{2}}\right).$$

$\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ depend on T such that $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_T$ and $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_{0,T} := \boldsymbol{\theta}_{0,\infty}.e_T$.

$d_T^4 = o(T)$ is also used in Theorem 1 of Fan and Peng (2004). The convergence rate of $\hat{\boldsymbol{\theta}}$ is the same as the analysis of the M-estimator by Huber (1973).

The first step estimator used for the adaptive weights is $(T/d_T)^{1/2}$-consistent. However, the estimated quantities on $\mathcal{A}^c$ converge to zero by consistency. We then propose a slight modification of the first step estimator, denoted $\tilde{\tilde{\boldsymbol{\theta}}}$, which disappears asymptotically as follows: $\tilde{\tilde{\boldsymbol{\theta}}} = \tilde{\boldsymbol{\theta}} + e_T$ so that $e_T \to 0$ is a strictly positive quantity. We choose $e_T = T^{-\kappa}$ with $\kappa > 0$. This means we add in the adaptive weights a power of T to the first step estimator, that is

$$\alpha_{T,i}^{(k)} = |\tilde{\tilde{\theta}}_i^{(k)}|^{-\eta} = |\tilde{\boldsymbol{\theta}} + T^{-\kappa}|^{-\eta},\ \xi_{T,l} = \|\tilde{\tilde{\boldsymbol{\theta}}}^{(l)}\|_2^{-\mu} = \|\tilde{\boldsymbol{\theta}}^{(l)} + T^{-\kappa}\|_2^{-\mu}.$$

Theorem 7 *Under Assumptions 1–3, 7–11, if $d_T^4 = o(T)$, and if $\frac{\gamma_T}{\sqrt{T}} T^{\frac{c}{2}+\kappa\mu} \underset{T\to\infty}{\longrightarrow} 0$, $\frac{\lambda_T}{\sqrt{T}} T^{\kappa\eta} \underset{T\to\infty}{\longrightarrow} 0$, then the sequence of penalized estimators $\hat{\boldsymbol{\theta}}$ solving 5 satisfies*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = O_p\left(\left(\frac{d_T}{T}\right)^{\frac{1}{2}}\right).$$

We make additional assumptions regarding the adaptive penalty components so that the adaptive SGL satisfies the oracle property.

Assumption 12 For any T, there exists β such that $0 < \beta < \min_{i\in\mathcal{A}_k} \theta_{0,i,\mathcal{A}_k}, k \in \mathcal{S}$. Moreover,

$$\beta^{-1}T^{-1}\left\{\lambda_T d_T^{1/2}\mathbb{E}\left[\max_{k\in\mathcal{S}, i\in\mathcal{A}_k} \alpha_{T,\mathcal{A}_k,i}\right] + \gamma_T \mathbb{E}\left[\max_{k\in\mathcal{S}} \xi_{T,k}\right]\right\} \underset{T\to\infty}{\longrightarrow} 0.$$

Assumption 13 The model complexity is assumed to behave as $d_T^5 = o(T)$, which implies that $0 < c < \frac{1}{5}$. The regularization parameters are chosen such that they satisfy

$$\frac{\gamma_T}{\sqrt{T}} T^{\frac{c}{2}+\kappa\mu} \underset{T\to\infty}{\longrightarrow} 0, \quad \frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(1-c)-1]} \underset{T\to\infty}{\longrightarrow} \infty, \quad \frac{\lambda_T}{\sqrt{T}} T^{\kappa\eta} \underset{T\to\infty}{\longrightarrow} 0,$$
$$\frac{\lambda_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\eta)(1-c)-1]} \underset{T\to\infty}{\longrightarrow} \infty, \quad \frac{\gamma_T}{\lambda_T^{1+\mu}} T^{(1+\mu)(1-\frac{c}{2}-\kappa\eta)-1} \underset{T\to\infty}{\longrightarrow} \infty.$$

The rate $d_T^5 = o(T)$ is also assumed in Theorem 2 of Fan and Peng (2004). Moreover, the convergence rates of the regularization parameters are closely related to condition (A5) of Zou and Zhang (2009). In Sect. 6, we provide further details about the choice of the adaptive weights and μ, η, κ.

Assumption 14 Let $X_{T,t} = \sqrt{T} Q_T \mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2} \mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1} \dot{\mathbb{G}}_T l_t(\boldsymbol{\theta}_0)_{\mathcal{A}}$ with (Q_T) a sequence of $r \times \text{card}(\mathcal{A})$ matrices such that $Q_T \times Q_T' \xrightarrow{\mathbb{P}} \mathbb{C}$, for some $r \times r$ nonnegative symmetric matrix $\mathbb{C}$, $\mathbb{V}_{T,\mathcal{A}\mathcal{A}} = (\mathbb{H}_T^{-1}\mathbb{M}_T\mathbb{H}_T^{-1})_{\mathcal{A}\mathcal{A}}$ and $\dot{\mathbb{G}}_T l_t(\boldsymbol{\theta}_0)_{\mathcal{A}} = \frac{1}{T}\nabla_{\mathcal{A}} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0)$. Let $\mathcal{F}_t^T = \sigma(X_{T,s}, s \leq t)$, then $X_{T,t}$ is a martingale difference and we have

$$\mathbb{E}[\sup_{i,j=1,\ldots,d_T} \mathbb{E}\left[\left\{\partial_{\theta_i} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0)\partial_{\theta_j} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0)\right\}^2 |\mathcal{F}_{t-1}^T\right] \lambda_{\max,t-1}(\mathbb{H}_{t-1}^T)] \leq \bar{B} < \infty,$$

with $\mathbb{H}_{t-1}^T := \mathbb{E}[\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}'} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0)|\mathcal{F}_{t-1}^T]$ and $\lambda_{\max,t-1}(\mathbb{H}_{t-1}^T) < \infty$.

Theorem 8 *Under Assumptions 1–3, and Assumptions 7–14, the sequence of adaptive estimator $\hat{\boldsymbol{\theta}}$ solving (5) satisfies*

$$\lim_{T\to\infty} \mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) = 1, \text{ and}$$
$$\sqrt{T} Q_T \mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2} (\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}) \xrightarrow{d} \mathcal{N}(0, \mathbb{C}),$$

where (Q_T) is a sequence of $r \times \text{card}(\mathcal{A})$ matrices such that $Q_T \times Q_T' \xrightarrow{\mathbb{P}} \mathbb{C}$, for some $r \times r$ nonnegative symmetric matrix $\mathbb{C}$ and $\mathbb{V}_{T,\mathcal{A}\mathcal{A}} = (\mathbb{H}_T^{-1}\mathbb{M}_T\mathbb{H}_T^{-1})_{\mathcal{A}\mathcal{A}}$.

5.2 From fixed to double asymptotic: discussion

The sample size indexing on $d := d_T$ alters significantly the theoretical analysis. In Sect. 5, the regularity conditions on $(l(_t; \boldsymbol{\theta}))$ have been strengthened to keep uniform properties for the double-asymptotic analysis: Hence Assumptions 9, 10 and 11 differ from Assumption 6. Assumptions 7 and 8 are stronger than Assumption 5, but they facilitate the theoretical analysis. Assumption 12 might be artificial, but it is key to obtain the oracle property and is in line with assumption (H) of Fan and Peng (2004). This assumption shows the rate at which the penalized criterion distinguishes nonzero parameters from zero parameters.

Furthermore, the convergence rates on d_T are different from Sect. 4 due to the necessary control on the third-order term of the Taylor expansions. As a consequence, the conditions $d_T^4 = O(T)$ for the consistency result and $d_T^5 = O(T)$ for the oracle property must be assumed. This issue was encountered by Fan and Peng (2004) in an i.i.d. and non-adaptive framework. This problem is moved aside when considering the linear model, where the third-order derivative vanishes. For instance, Zou and Zhang (2009) proved the oracle property of the adaptive elastic-net in a double-asymptotic framework for linear models where $0 \leq c < 1$. Nardi and Rinaldo (2008) in Theorem 4.2 provide a model selection consistency result for the adaptive Group Lasso when $\log(d_T)/T \to 0$, a rate also obtained in Theorem 1 of Wainwright (2009). This scaling between d_T and T is obtained for the linear regression model by applying standard results on the maximum of a Gaussian vector. Their assumption (S4) also allows the dimension d_T to grow at a faster rate than T for a suitable choice of the adaptive weights. Since we consider a general penalized likelihood setting, we rely on Assumption 11 to control for the third-order term, which in turn implies the convergence $d_T^5 = o(T)$ stated in Assumption 13. The latter also controls the convergence rates of the regularization parameters and provides a trade-off between the group and the within-group regularizations.

Moreover, the double asymptotic requires the use of explicit vector/matrix norms, especially for Theorems 6 and 7: Because of the norm equivalences, some constants may appear so that these constant may depend on the size d_T and thus on T. Hence the norms of the latter theorems are explicit, contrary to Theorems 2 and 4.

Finally, $|\mathcal{A}|$ is allowed to diverge, which implies that the vector size of $(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})$ also diverges. To derive the distribution in Theorem 8, we thus multiplied the discrepancy $\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)_{\mathcal{A}}$ by a matrix sequence (Q_T) of size $r \times \text{card}(\mathcal{A})$, r being arbitrary but finite. This method was also used by Fan and Peng (2004) or Zou and Zhang (2009). The derivation of the asymptotic distribution heavily relies on Assumption 14.

6 Empirical applications

Our empirical experiments are based on linear dynamic systems with fixed dimension so that the statistical problem corresponds to a penalized OLS criterion. We consider 6 penalization methods: the Lasso (L), the adaptive Lasso (AL), the Group Lasso (GL), the adaptive Group Lasso (AGL), the Sparse Group Lasso (SGL) and the adaptive Sparse Group Lasso (ASGL). Table 1 reports for simulated data the variable selection performance through the number of zero coefficients correctly estimated, denoted as C and the number of nonzero coefficients incorrectly estimated, denoted IC. Besides, the mean squared error is reported as an estimation accuracy measure. Table 2 reports the regularization performances based on real data sets.

6.1 Tuning of the regularization parameters and the adaptive weights

They both must satisfy some convergence rate provided in Theorem 5 for fixed dimensions and Theorem 8 for diverging dimensions to satisfy the oracle prop-

Table 1 Model selection and precision accuracy based on 100 replications

Model	MSE	C	IC
Truth		47	0
Lasso	0.1558	22.34	5.44
aLasso	0.1270	24.56	5.48
GLasso	0.1153	28.59	2.50
AGLasso	0.0414	40.91	2.40
SGL	0.0407	39.42	2.49
ASGL	0.0392	43.34	1.90

Table 2 Mean square error based on 100 test sets

Data set	Lasso	aLasso	GLasso	AGLasso	SGL	ASGL
MPG	0.592	0.581	0.985	0.909	0.274	0.257
Automobile	0.311	0.263	0.889	0.424	0.283	0.238

erty. More precisely, we suppose $\lambda_T = T^{\beta}$ and $\gamma_T = T^{\nu}$, where β and ν are both strictly positive constant. Within the double-asymptotic framework, regarding Assumption 13 and Theorem 8, taking condition $\gamma_T T^{\frac{1}{2}[(1+\mu)(1-c)-1]-\frac{1}{2}} \to \infty$ means $T^{\nu-\frac{1}{2}+\frac{1}{2}[(1+\mu)(1-c)-1]} \to \infty$, which implies $\nu - \frac{1}{2} + \frac{1}{2}[(1+\mu)(1-c)-1] > 0$. Thus the set of conditions is

$$\begin{cases} \nu + \frac{c}{2} + \kappa\mu - \frac{1}{2} < 0, \\ \nu - \frac{1}{2} + \frac{1}{2}[(1+\mu)(1-c)-1] > 0, \\ \beta + \kappa\eta - \frac{1}{2} < 0, \\ \beta - \frac{1}{2} + \frac{1}{2}[(1+\eta)(1-c)-1] > 0, \\ (1+\mu)[1 - \frac{c}{2} - \kappa\eta - \beta] + \nu - 1 > 0. \end{cases}$$

This system allows for flexibility when choosing μ and η once κ, c, ν and β are fixed. For instance, for $c = 1/6$, $\kappa = 0.05$, $\nu = 1/10$ and $\beta = 1/10$, then $\mu \in [0.4, 6.3]$ and $\eta \in [0.6, 7.9]$. If $\nu = \beta = 1/5$ and for $c = 1/6$ and $\kappa = 0.05$, then $\mu \in [0.4, 4.3]$ and $\eta \in [0.3, 5.9]$.

As for the *fixed* dimensional case, the conditions in Theorem 5 are

$$\begin{cases} \beta - \frac{1}{2} < 0, \nu - \frac{1}{2} < 0, \\ \beta + \frac{\eta}{2} - \frac{1}{2} > 0, \nu + \frac{\mu}{2} - \frac{1}{2} > 0, \\ \nu - \beta + \frac{\mu-\eta}{2} > 0, \end{cases} \tag{6}$$

For instance, let $\nu = 1/3$ and $\beta = 1/5$, then we would have $1 \leq \eta \leq \mu$.

We used a cross-validation (CV) procedure to select both parameters λ_T and γ_T such that both terms are defined by $\lambda_T = T^{\beta}$ and $\gamma_T = T^{\nu}$, and $\beta = \nu = 1/8$. The adaptive weights are computed as follows: We first compute an OLS estimator $\tilde{\boldsymbol{\theta}}$ such

that the adaptive weights entering the penalties correspond to $\tilde{\tilde{\boldsymbol{\theta}}} = \tilde{\boldsymbol{\theta}} + T^{-\kappa}$, with $\kappa = 0.2$. As for the adaptive weights, they are chosen such that the above system is satisfied: We set $\eta = 2.5$ and $\mu = 1.5$. The standard CV developed for i.i.d. data can not be used in dependent framework. To fix this issue, we used the hv-CV procedure devised by Racine (2000), which consists in leaving a gap between the test sample and the training sample, on both sides of the test sample.

6.2 Numerical procedure

There are several methods to numerically solve the non-differentiable statistical problem (4) or (5). Fan and Li (2001) proposed a local quadratic approximation (LQA) of the first-order derivative of the penalty function and a Newton–Raphson-type algorithm. To circumvent numerical instability, they suggest to shrunk to zero coefficients that are close to zero, that is, a coefficient $|\theta_j| < \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} > 0$ to be calibrated. The drawback is that once it is set to zero, it will be excluded at any step of the LQA algorithm. Hunter and Li (2005) proposed a more sophisticated version of the LQA algorithm to avoid the drawback of the stepwise selection and numerical instability. When one consider the OLS loss function, closed form algorithms can be applied to our problem. Bühlmann and van de Geer (2011) compiled these methodologies for solving the Lasso and the Group Lasso using gradient descent methods for general penalized convex empirical function. We used these algorithms in our study for solving the group Lasso. As for the Lasso, we applied the shooting algorithm developed by Fu (1998), which is a particular case of the gradient descent method. Finally, we used the alternative direction method of multipliers provided by Li et al. (2014) for solving the SGL penalization.

6.3 Simulated experiment

We consider a data generating process

$$\begin{aligned} y_t &= \sigma_t \eta_t, \\ x_{1,t} &= \beta_1 x_{1,t-1} + \nu_{1,t}, x_{2,t} = \beta_2 x_{2,t-1} + \nu_{2,t}, \\ \sigma_t^2 &= \sum_{k=1}^{p} a_k y_{t-k}^2 + \sum_{l=1}^{q} \{b_l |x_{1,t-l}| + c_l |x_{2,t-1}|\}, \end{aligned}$$

where the exogenous variable $(|x_{1,t}|)$ and $(|x_{2,t}|)$ are positive and stationary. They are simulated as $\beta_1 \sim \mathcal{U}([0.85, 0.95])$ and $\beta_2 \sim \mathcal{U}([0.6, 0.75])$ with $\mathcal{U}(.)$ the uniform distribution. Moreover, (η_t) is uncorrelated with $(\nu_t) = (\nu_{1,t}, \nu_{2,t}) \sim \mathcal{N}(0, \Gamma)$ where

$$\Gamma = \begin{pmatrix} 0.05 & 0.035 \\ 0.035 & 0.04 \end{pmatrix}.$$

(σ_t^2) corresponds to an ARCH(p) model with exogenous variables (q lags). We set $T = 5000$ $p = 5$, $q = 2$ and $a_k \sim \mathcal{U}([0.01, 0.2])$, $b_l, c_l \sim \mathcal{U}([0.01, 0.2])$ so that

the stationarity conditions derived by Francq and Thieu (2015) for GARCH models with exogenous variables are satisfied. They are also supposed to satisfy an ordering constraint, id est $\forall k \geq 2, a_k \leq a_{k-1}$, $\forall l \geq 2, b_l \leq b_{l-1}$ and $\forall l \geq 2, c_l \leq c_{l-1}$.

This model can be estimated by an ordinary least squares procedure as (y_t) and (x_t) are observed variables. We would have the linear model

$$y_t^2 = \sum_{k=1}^{p} a_k y_{t-k}^2 + \sum_{l=1}^{q} \{b_l |x_{1,t-l}| + c_l |x_{2,t-1}|\} + u_t,$$

where $u_t = y_t^2 - \sigma_t^2$ is the error term. For instance, (y_t) can be a stock index (e.g., Apple), $(x_{1,t})$ the S&P 500 return index and $x_{2,t}$ the NASDAQ return index. We specify an initial number of lags $m = 20$ and estimate

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \Theta} \left\{ \frac{1}{T} \sum_{t=m+1}^{T} \sum_{k=1}^{m} (y_t^2 - Z_{t-k}' \boldsymbol{\theta}^{(k)})^2 + \frac{\lambda_T}{T} \sum_{i=1}^{m} \sum_{j=1}^{\boldsymbol{c}_i} \alpha_{T,j}^{(i)} |\theta_i^{(k)}| + \sum_{l=1}^{m} \xi_{T,l} \|\boldsymbol{\theta}^{(l)}\|_2 \right\},$$

with $Z_{t-k} = (y_{t-k}^2, |x_{1,t-k}|, |x_{2,t-k}|)'$, $\boldsymbol{\theta}^{(k)} = (a_k, b_k, c_k)'$. The weights are constructed as $\alpha_{T,j}^{(i)} = |\tilde{\boldsymbol{\theta}}_j^{(i)}|^{-\eta}$ the jth component of group i, $\xi_{T,l} = \|\tilde{\boldsymbol{\theta}}^{(l)}\|_2^{-\mu}$ with $\tilde{\beta}$ defined as an unpenalized OLS version of the previous criterion. In this setting, $\boldsymbol{c}_i = 3$. The parameters are subject to nonnegative constraints. The regularization procedures aim at correctly selecting the variables, that is, we would like to discard the y_{t-k}^2 and $|x_{i,t-l}|$ for $k > 5$ and $l > 2$, for any $i = 1, 2$. That means the total number of zeros to be identified is equal to 47.

We discuss how this OLS objective function satisfies the assumptions of Theorem 5. By construction, the vector of observation $\boldsymbol{\epsilon}_t = (y_t, x_{1,t}, x_{2,t})'$ is a strictly stationary and ergodic process. The loss function is quadratic with respect to $\boldsymbol{\theta}$ so it is convex (Assumption 3). The score of the unpenalized part would be

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \left(y_t^2 - Z_{t-k}' \boldsymbol{\theta}^{(k)} \right)^2 = Z_{t-k} \left(y_t^2 - Z_{t-k}' \boldsymbol{\theta}^{(k)} \right) = Z_{t-k} u_t,$$

so that $\mathbb{E}[\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}) | \mathcal{F}_{t-1}] = 0$ as the error (u_t) is uncorrelated with past observations. Hence Assumption 4 is satisfied, that is, $(u_t, \mathcal{F}_t)$ is a martingale difference when $\mathbb{E}[y_t^2] = \sigma_t^2 < \infty$. Assumption 5 is satisfied using step (ii) of the proof of Theorem 6.1 of Francq and Zakoïan (2010), where they show the invertibility of $\mathbb{E}[Z_t Z_t']$ by contradiction. As for Assumption 6, the third-order term vanishes in the OLS model. Finally, the convergence rates provided in 6 must be satisfied to satisfy the oracle property.

Table 1 reports the performances of the regularization methods. The adaptive versions of the Lasso, the Group Lasso or the SGL outperform their non-adaptive versions. The difference is significant for the adaptive Lasso and the adaptive SGL. This is in line with the asymptotic theory. The adaptive SGL performs well as it can discard inactive groups and inactive indices among active groups and outperform other adaptive penalization methods.

An additional simulated experiment for VAR models is reported in the Supplementary file, Sect. 4.

6.4 Real data experiment

We carry out a performance analysis of the regularization methods on two data sets from UCI Machine Learning Repository: the auto MPG and the Automobile data. Both contain real and categorical data, which are dummy encoded. For the Automobile data set, the car's price is predicted from 3 categorical variables (car's style, engine type, fuel system) and 9 real-valued variables, which are grouped as follows: a group for the car's dimension (height, width, length, curb-weight), a group for the engine's properties (size, bore and stroke, compression ratio, horsepower, the peak of power band), one group for the miles per gallon (city and highway); each set of indicator variables corresponding to a given categorical covariate are grouped together. There are 195 observations and 30 parameters to estimate. As for the MPG data, the city-cycle fuel consumption is predicted by four real-valued predictors (horsepower, weight, displacement, acceleration), each of them corresponding to one group, and 3 categorical variables (cylinders, model year and origin) so that the dummy variables from one covariate are grouped. There are 392 observations and 25 parameters to estimate. The grouping structure is arbitrary.

The OLS problem—after centering and standardizing the variables, no intercept is included—was considered for prediction purposes with the Lasso, the Group Lasso and the Sparse Group Lasso regularization procedures together with their adaptive versions. For the Automobile (resp. MPG) data, 145 (resp. 312) observations were randomly chosen to fit the penalized OLS models and the 50 (resp. 80) remaining observations were used as a test set. The procedure was repeated 100 times so that an average mean square error for prediction is reported in Table 2. The adaptive SGL still outperforms the other methods. The procedure is well adapted in the presence of both categorical factors and continuous covariates. The prediction performances also emphasize the gain to consider adaptive weights.

Acknowledgements I would like to thank Alexandre Tsybakov, Arnak Dalalyan, Jean-Michel Zakoïan and Christian Francq for all the theoretical references they provided. And I thank warmly Jean-David Fermanian for his significant help and helpful comments. I gratefully acknowledge the Ecodec Laboratory for its support and the Japan Society for the Promotion of Science.

Appendix

We first introduce some preleminary results. The dependent setting requires the use of more sophisticated probabilistic tools to derive asymptotic results than the i.i.d. case. Assumptions 1 and 4 allow for using the central limit theorem of Billingsley (1961). We remind this result stated as a corollary in Billingsley (1961).

Corollary 1 (Billingsley 1961) *If* $(x_t, \mathcal{F}_t)$ *is a stationary and ergodic sequence of square integrable martingale increments such that* $\sigma_x^2 = Var(x_t) \neq 0$*, then* $T^{-1/2} \sum_{t=1}^{T} x_t \xrightarrow{d} \mathcal{N}(0, \sigma_x^2)$.

Note that the square martingale difference condition can be relaxed by α-mixing and moment conditions. For instance, Rio (2013) provides a central limit theorem for strongly mixing and stationary sequences.

To prove Theorem 1, we remind of Theorem II.1 of Anderson and Gill (1982) which proves that pointwise convergence in probability of random concave functions implies uniform convergence on compact subspaces.

Theorem 9 (Anderson and Gill 1982) *Let E be an open convex subset of $\mathbb{R}^p$, and let $F_1, F_2, \ldots,$ be a sequence of random concave functions on E such that $F_n(x) \xrightarrow[n\to\infty]{\mathbb{P}} f(x)$ for every $x \in E$ where f is some real function on E. Then f is also concave, and for all compact $A \subset E$,*

$$\sup_{x\in A} |F_n(x) - f(x)| \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

The proof of this theorem is based on a diagonal argument and Theorem 10.8 of Rockafeller (1970), that is, the pointwise convergence of concave random functions on a dense and countable subset of an open set implies uniform convergence on any compact subset of the open set. Then the following corollary is stated.

Corollary 2 (Anderson and Gill 1982) *Assume $F_n(x) \xrightarrow[n\to\infty]{\mathbb{P}} f(x)$, for every $x \in E$, an open convex subset of $\mathbb{R}^p$. Suppose f has a unique maximum at $x_0 \in E$. Let $\hat{X}_n$ maximize F_n. Then $\hat{X}_n \xrightarrow[n\to\infty]{\mathbb{P}} x_0$.*

Newey and Powell (1987) use a similar theorem to prove the consistency of asymmetric least squares estimators without any compacity assumption on Θ. We apply these results in our framework, where the parameter set Θ is supposed to be convex.

We used the *convexity argument* to derive the asymptotic distribution of the SGL estimator. Chernozhukov and Hong (2004) and Chernozhukov (2005) use this convexity argument to obtain the asymptotic distribution of quantile regression-type estimators. This argument relies on the convexity lemma, which is a key result to obtain an asymptotic distribution when the objective function is not differentiable. It only requires the lower-semicontinuity and convexity of the empirical criterion. The convexity lemma, as in Chernozhukov (2005), proof of Theorem 4.1, can be stated as follows:

Lemma 1 (Chernozhukov 2005) *Suppose*

(i) a sequence of convex lower-semicontinuous $\mathbb{F}_T : \mathbb{R}^d \to \bar{\mathbb{R}}$ marginally converges to $\mathbb{F}_\infty : \mathbb{R}^d \to \bar{\mathbb{R}}$ over a dense subset of $\mathbb{R}^d$;
(ii) $\mathbb{F}_\infty$ is finite over a non-empty open set $E \subset \mathbb{R}^d$;
(iii) $\mathbb{F}_\infty$ is uniquely minimized at a random vector $\boldsymbol{u}_\infty$.

Then

$$\arg\min_{z\in\mathbb{R}^d} \mathbb{F}_T(z) \xrightarrow{d} \arg\min_{z\in\mathbb{R}^d} \mathbb{F}_\infty(z), \text{ that is } \boldsymbol{u}_T \xrightarrow{d} \boldsymbol{u}_\infty.$$

This is a key argument used in Theorem 3, Proposition 1 and Theorem 5.

When we consider a diverging number of parameters, the empirical criterion can be viewed as a sequence of dependent arrays for which we need refined asymptotic results. Shiryaev (1991) proposed a version of the central limit theorem for dependent sequence of arrays, provided this sequence is a square integrable martingale difference satisfying the so-called Lindeberg condition. A similar theorem can be found in Billingsley (1995, Theorem 35.12, p.476). We provide here the theorem of Shiryaev (see Theorem 4, p.543 of Shiryaev 1991) that we will use to derive the asymptotic distribution of the adaptive SGL estimator.

Theorem 10 (Shiryaev 1991) *Let a sequence of square integrable martingale differences* $\xi^n = (\xi_{nk}, \mathcal{F}_k^n), n \geq 1$, *with* $\mathcal{F}_k^n = \sigma(\xi_{ns}, s \leq k)$, *satisfy the Lindeberg condition for any* $0 < t \leq 1$, *for* $\epsilon > 0$, *given by*

$$\sum_{k=0}^{\lfloor nt \rfloor} \mathbb{E}\left[\xi_{nk}^2 \mathbf{1}_{|\xi_{nk}|>\epsilon} | \mathcal{F}_{k-1}^n\right] \xrightarrow[n\to\infty]{\mathbb{P}} 0,$$

then if $\sum_{k=0}^{\lfloor nt \rfloor} \mathbb{E}[\xi_{nk}^2 | \mathcal{F}_{k-1}^n] \xrightarrow[n\to\infty]{\mathbb{P}} \sigma_t^2$, *or* $\sum_{k=0}^{\lfloor nt \rfloor} \xi_{nk}^2 \xrightarrow[n\to\infty]{\mathbb{P}} \sigma_t^2$, *then* $\sum_{k=0}^{\lfloor nt \rfloor} \xi_{nk} \xrightarrow{d} \mathcal{N}(0, \sigma_t^2)$.

There exist central limit results relaxing the stationarity and martingale difference assumptions for sequences of arrays. Neumann (2013) proposed such a central limit theorem for weakly dependent sequences of arrays. Such sequences should also satisfy a Lindeberg condition and conditions on covariances. Equipped with these preliminary results, we now report the proofs of Sect. 4.

Proof of Theorem 1 By definition, $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \{\mathbb{G}_T \varphi(\boldsymbol{\theta})\}$. In a first step, we prove the uniform convergence of $\mathbb{G}_T\varphi(.)$ to the limit quantity $\mathbb{G}_\infty\varphi(.)$ on any compact set $\boldsymbol{B} \subset \Theta$, idest

$$\sup_{\boldsymbol{x}\in\boldsymbol{B}} |\mathbb{G}_T\varphi(\boldsymbol{x}) - \mathbb{G}_\infty\varphi(\boldsymbol{x})| \xrightarrow[T\to\infty]{\mathbb{P}} 0. \tag{7}$$

We define $\mathcal{C} \subset \Theta$ an open convex set and pick $\boldsymbol{x} \in \mathcal{C}$. Then by Assumption 1, the law of large number implies

$$\mathbb{G}_T l(\boldsymbol{x}) \xrightarrow[T\to\infty]{\mathbb{P}} \mathbb{G}_\infty l(\boldsymbol{x}).$$

Consequently, if $\lambda_T/T \to \lambda_0 \geq 0$ and $\gamma_T/T \to \gamma_0 \geq 0$, we obtain the pointwise convergence

$$|\mathbb{G}_T\varphi(\boldsymbol{x}) - \mathbb{G}_\infty\varphi(\boldsymbol{x})| \xrightarrow[T\to\infty]{\mathbb{P}} 0.$$

By Theorem 9 of Anderson and Gill (1982), $\mathbb{G}_\infty\varphi(.)$ is a convex function and we deduce the desired uniform convergence over any compact subset of Θ, that is (7).

Now we would like that $\arg\min\{\mathbb{G}_T\varphi(.)\} \xrightarrow[T\to\infty]{\mathbb{P}} \arg\min\{\mathbb{G}_\infty\varphi(.)\}$. By Assumption 3, $\varphi(.)$ is convex, which implies

$$|\mathbb{G}_T\varphi(\boldsymbol{\theta})| \xrightarrow[\|\boldsymbol{\theta}\|\to\infty]{\mathbb{P}} \infty.$$

Consequently, $\arg\min\{\mathbb{G}_T\varphi(\boldsymbol{x})\} = O(1)$, such that $\hat{\boldsymbol{\theta}} \in \mathcal{B}_o(\boldsymbol{\theta}_0, C)$ with probability approaching one for C large enough, with $\mathcal{B}_o(\boldsymbol{\theta}_0, C)$ an open ball centered at $\boldsymbol{\theta}_0$ and of radius C. Furthermore, as $\mathbb{G}_\infty\varphi(.)$ is convex, continuous, then $\arg\min_{\boldsymbol{x}\in B}\{\mathbb{G}_\infty\varphi(\boldsymbol{x})\}$ exists and is unique. Then by Corollary 2 of Andersen and Gill, we obtain

$$\arg\min_{\boldsymbol{x}\in\boldsymbol{B}}\{\mathbb{G}_T\varphi(\boldsymbol{x})\} \xrightarrow[T\to\infty]{\mathbb{P}} \arg\min_{\boldsymbol{x}\in\boldsymbol{B}}\{\mathbb{G}_\infty\varphi(\boldsymbol{x})\}, \quad \text{that is } \hat{\boldsymbol{\theta}} \xrightarrow[T\to\infty]{\mathbb{P}} \boldsymbol{\theta}_0^*.$$

□

Proof of Theorem 2 We denote $\nu_T = T^{-1/2} + \lambda_T T^{-1} a + \gamma_T T^{-1} b$, with $a = \text{card}(\mathcal{A})(\max_k \alpha_k)$ and $b = \text{card}(\mathcal{A})(\max_l \xi_l)$. We would like to prove that for any $\boldsymbol{\epsilon} > 0$, there exists $C_{\boldsymbol{\epsilon}} > 0$ such that $\mathbb{P}(\nu_T^{-1}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| > C_{\boldsymbol{\epsilon}}) < \boldsymbol{\epsilon}$. We have

$$\mathbb{P}(\nu_T^{-1}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| > C_{\boldsymbol{\epsilon}}) \leq \mathbb{P}\left(\exists \boldsymbol{u} \in \mathbb{R}^d, \|\boldsymbol{u}\|_2 \geq C_{\boldsymbol{\epsilon}} : \mathbb{G}_T\varphi(\boldsymbol{\theta}_0 + \nu_T\boldsymbol{u}) \leq \mathbb{G}_T\varphi(\boldsymbol{\theta}_0)\right).$$

$\|\boldsymbol{u}\|_2$ can potentially be large as it represents the discrepancy $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ normalized by ν_T. Now based on the convexity of the objective function, we have

$$\begin{aligned} &\left\{\exists \boldsymbol{u}^*, \|\boldsymbol{u}^*\|_2 \geq C_{\boldsymbol{\epsilon}}, \mathbb{G}_T\varphi(\boldsymbol{\theta}_0 + \nu_T\boldsymbol{u}^*) \leq \mathbb{G}_T\varphi(\boldsymbol{\theta}_0)\right\} \\ &\subset \left\{\exists \bar{\boldsymbol{u}}, \|\bar{\boldsymbol{u}}\|_2 = C_{\boldsymbol{\epsilon}}, \mathbb{G}_T\varphi(\boldsymbol{\theta}_0 + \nu_T\bar{\boldsymbol{u}}) \leq \mathbb{G}_T\varphi(\boldsymbol{\theta}_0)\right\}, \end{aligned} \tag{8}$$

a relationship that allows us to work with a fixed $\|\boldsymbol{u}\|_2$. Let us define $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \nu_T\boldsymbol{u}^*$ such that $\mathbb{G}_T\varphi(\boldsymbol{\theta}_1) \leq \mathbb{G}_T\varphi(\boldsymbol{\theta}_0)$. Let $\alpha \in (0, 1)$ and $\boldsymbol{\theta} = \alpha\boldsymbol{\theta}_1 + (1-\alpha)\boldsymbol{\theta}_0$. Then by convexity of $\mathbb{G}_T\varphi(.)$, we obtain

$$\mathbb{G}_T\varphi(\boldsymbol{\theta}) \leq \alpha\mathbb{G}_T\varphi(\boldsymbol{\theta}_1) + (1-\alpha)\mathbb{G}_T\varphi(\boldsymbol{\theta}_0) \leq \mathbb{G}_T\varphi(\boldsymbol{\theta}_0).$$

We pick α such that $\|\bar{\boldsymbol{u}}\| = C_{\boldsymbol{\epsilon}}$ with $\bar{\boldsymbol{u}} := \alpha\boldsymbol{\theta}_1 + (1-\alpha)\boldsymbol{\theta}_0$. Hence (8) holds, which implies

$$\begin{aligned} \mathbb{P}(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| > C_{\boldsymbol{\epsilon}}\nu_T) &\leq \mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^d, \|\boldsymbol{u}\|_2 \geq C_{\boldsymbol{\epsilon}} : \mathbb{G}_T\varphi(\boldsymbol{\theta}_0 + \nu_T\boldsymbol{u}) \leq \mathbb{G}_T\varphi(\boldsymbol{\theta}_0)) \\ &\leq \mathbb{P}(\exists \bar{\boldsymbol{u}}, \|\bar{\boldsymbol{u}}\|_2 = C_{\boldsymbol{\epsilon}} : \mathbb{G}_T\varphi(\boldsymbol{\theta}_0 + \nu_T\bar{\boldsymbol{u}}) \leq \mathbb{G}_T\varphi(\boldsymbol{\theta}_0)). \end{aligned}$$

Hence, we pick a $\boldsymbol{u}$ such that $\|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}}$. Using $\boldsymbol{p}_1(\lambda_T, \alpha, 0) = 0$ and $\boldsymbol{p}_2(\gamma_T, \xi, 0) = 0$, by a Taylor expansion to $\mathbb{G}_T l(\boldsymbol{\theta}_0 + \nu_T\boldsymbol{u})$, we obtain

$$\mathbb{G}_T\varphi(\boldsymbol{\theta}_0+\nu_T\boldsymbol{u})-\mathbb{G}_T\varphi(\boldsymbol{\theta}_0)=\nu_T\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}+\frac{\nu_T^2}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}$$
$$+\frac{\nu_T^3}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u}+\boldsymbol{p}_1(\lambda_T,\alpha,\boldsymbol{\theta}_T)$$
$$-\boldsymbol{p}_1(\lambda_T,\alpha,\boldsymbol{\theta}_0)+\boldsymbol{p}_2(\gamma_T,\xi,\boldsymbol{\theta}_T)-\boldsymbol{p}_2(\gamma_T,\xi,\boldsymbol{\theta}_0),$$

where $\bar{\boldsymbol{\theta}}$ is defined as $\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|\leq\|\boldsymbol{\theta}_T-\boldsymbol{\theta}_0\|$. We want to prove

$$\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_{\boldsymbol{\epsilon}}:\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}+\frac{\nu_T}{2}\mathbb{E}[\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}]+\frac{\nu_T}{2}\mathcal{R}_T(\boldsymbol{\theta}_0)$$
$$+\frac{\nu_T^2}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u}+\nu_T^{-1}\{\boldsymbol{p}_1(\lambda_T,\alpha,\boldsymbol{\theta}_T)-\boldsymbol{p}_1(\lambda_T,\alpha,\boldsymbol{\theta}_0)$$
$$+\boldsymbol{p}_2(\gamma_T,\xi,\boldsymbol{\theta}_T)-\boldsymbol{p}_2(\gamma_T,\xi,\boldsymbol{\theta}_0)\}\leq 0)<\boldsymbol{\epsilon}, \tag{9}$$

where $\mathcal{R}_T(\boldsymbol{\theta}_0)=\sum_{k,l=1}^{d}\boldsymbol{u}_k\boldsymbol{u}_l\{\partial^2_{\theta_k\theta_l}\mathbb{G}_T l(\boldsymbol{\theta}_0)-\mathbb{E}[\partial^2_{\theta_k\theta_l}\mathbb{G}_T l(\boldsymbol{\theta}_0)]\}$. By Assumption 1, $(\boldsymbol{\epsilon}_t)$ is a non-anticipative stationary solution and is ergodic. As a square integrable martingale difference by Assumption 4,

$$\sqrt{T}\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}\xrightarrow{d}\mathcal{N}(0,\boldsymbol{u}'\mathbb{M}\boldsymbol{u}),$$

by the central limit theorem of Billingsley (1961), which implies $\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}=O_p(T^{-1/2}\boldsymbol{u}'\mathbb{M}\boldsymbol{u})$. By the ergodic theorem of Billingsley (1995), we have

$$\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\xrightarrow[T\to\infty]{\mathbb{P}}\mathbb{H}.$$

This implies $\mathcal{R}_T(\boldsymbol{\theta}_0)=o_p(1)$. Furthermore, by the Markov inequality, for $b>0$

$$\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_{\boldsymbol{\epsilon}}:\sup_{\bar{\boldsymbol{\theta}}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2\leq\nu_T C_{\boldsymbol{\epsilon}}}|\frac{\nu_T^2}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u}|>b)\leq\frac{\nu_T^4 C_{\boldsymbol{\epsilon}}^6}{36b^2}\eta(C_{\boldsymbol{\epsilon}}),$$

where $\eta(C_{\boldsymbol{\epsilon}})$ is defined in Assumption 6. We now focus on the penalty terms. As $\boldsymbol{p}_1(\lambda_T,\alpha,0)=0$, for the l^1 norm penalty, we have

$$\boldsymbol{p}_1(\lambda_T,\alpha,\boldsymbol{\theta}_T)-\boldsymbol{p}_1(\lambda_T,\alpha,\boldsymbol{\theta}_0)=\lambda_T T^{-1}\sum_{k\in\mathcal{S}}\alpha_k\left\{\|\boldsymbol{\theta}_0^{(k)}+\nu_T\boldsymbol{u}^{(k)}\|_1-\|\boldsymbol{\theta}_0^{(k)}\|_1\right\},$$

and $|\boldsymbol{p}_1(\lambda_T,\alpha,\boldsymbol{\theta}_T)-\boldsymbol{p}_1(\lambda_T,\alpha,\boldsymbol{\theta}_0)|\leq\mathrm{card}(\mathcal{S})\{\max_{k\in\mathcal{S}}\alpha_k\}\lambda_T T^{-1}\nu_T\|\boldsymbol{u}\|_1$.

As for the l^1/l^2 norm, we obtain

$$\boldsymbol{p}_2(\gamma_T,\xi,\boldsymbol{\theta}_T)-\boldsymbol{p}_2(\gamma_T,\xi,\boldsymbol{\theta}_0)=\gamma_T T^{-1}\sum_{l\in\mathcal{S}}\xi_l\left\{\|\boldsymbol{\theta}_T^{(l)}\|_2-\|\boldsymbol{\theta}_0^{(l)}\|_2\right\},$$

$$\begin{aligned}\text{and } |\boldsymbol{p}_2(\gamma_T,\xi,\boldsymbol{\theta}_T)-\boldsymbol{p}_2(\gamma_T,\xi,\boldsymbol{\theta}_0)| &\leq \gamma_T T^{-1}\sum_{l\in\mathcal{S}}\xi_l\nu_T\|\boldsymbol{u}^{(l)}\|_2\\ &\leq \text{card}(\mathcal{S})\left\{\max_{l\in\mathcal{S}}\xi_l\right\}\gamma_T T^{-1}\nu_T\|\boldsymbol{u}\|_2.\end{aligned}$$

Then denoting by $\delta_T=\lambda_{\min}(\mathbb{H})C_\epsilon^2\nu_T/2$, and using $\frac{\nu_T}{2}\mathbb{E}[\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}]\geq\delta_T$, we deduce that (9) can be bounded as

$$\begin{aligned}\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon:\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}+\frac{\nu_T}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}+\frac{\nu_T^2}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u}\\ +\nu_T^{-1}\{\boldsymbol{p}_1(\lambda_T,\alpha,\boldsymbol{\theta}_T)-\boldsymbol{p}_1(\lambda_T,\alpha,\boldsymbol{\theta}_0)+\boldsymbol{p}_2(\gamma_T,\xi,\boldsymbol{\theta}_T)\\ -\boldsymbol{p}_2(\gamma_T,\xi,\boldsymbol{\theta}_0)\}\leq 0)\\ \leq\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon:|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}|>\delta_T/8)+\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2\\ =C_\epsilon:\frac{\nu_T}{2}|\mathcal{R}_T(\boldsymbol{\theta}_0)|>\delta_T/8)\\ +\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon:|\frac{\nu_T^2}{6}\nabla'\left\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\right\}\boldsymbol{u}|>\delta_T/8)\\ +\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon:|\boldsymbol{p}_1(\lambda_T,\alpha,\boldsymbol{\theta}_T)-\boldsymbol{p}_1(\lambda_T,\alpha,\boldsymbol{\theta}_0)|>\nu_T\delta_T/8)\\ +\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon:|\boldsymbol{p}_2(\gamma_T,\xi,\boldsymbol{\theta}_T)-\boldsymbol{p}_2(\gamma_T,\xi,\boldsymbol{\theta}_0)|>\nu_T\delta_T/8).\end{aligned}$$

We also have for C_ϵ and T large enough, and using norm equivalences that

$$\begin{aligned}\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon:|\boldsymbol{p}_1(\lambda_T,\alpha,\boldsymbol{\theta}_T)-\boldsymbol{p}_1(\lambda_T,\alpha,\boldsymbol{\theta}_0)|>\nu_T\delta_T/8)\\ \leq\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon:\text{card}(\mathcal{S})\{\max_{k\in\mathcal{S}}\alpha_k\}\lambda_T T^{-1}\nu_T\|\boldsymbol{u}\|_1>\nu_T\delta_T/8)<\epsilon/5,\\ \mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon:|\boldsymbol{p}_2(\gamma_T,\xi,\boldsymbol{\theta}_T)-\boldsymbol{p}_2(\gamma_T,\xi,\boldsymbol{\theta}_0)|>\nu_T\delta_T/8)\\ \leq\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon:\text{card}(\mathcal{S})\{\max_{l\in\mathcal{S}}\xi_l\}\gamma_T T^{-1}\nu_T\|\boldsymbol{u}\|_2>\nu_T\delta_T/8)<\epsilon/5.\end{aligned}$$

Moreover, if $\nu_T=T^{-1/2}+\lambda_T T^{-1}a+\gamma_T T^{-1}b$, then for C_ϵ large enough

$$\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon:|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}|>\delta_T/8)\leq\frac{C_\epsilon^2 C_{st}}{T\delta_T^2}\leq\frac{C_{st}}{C_\epsilon^4}<\epsilon/5.$$

Moreover

$$\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon:\sup_{\bar{\boldsymbol{\theta}}:\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2<\nu_T C_\epsilon}|\frac{\nu_T^2}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u}|>\delta_T/8)$$

$$\leq \frac{C_{st}\nu_T^4\eta(C_\epsilon)}{\delta_T^2} \leq C_{st}\nu_T^2 C_\epsilon^2\eta(C_\epsilon)$$

where $C_{st} > 0$ is a generic constant. We obtain

$$\begin{aligned}\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}| > \delta_T/8) + \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \frac{\nu_T}{2}|\mathcal{R}_T(\boldsymbol{\theta}_0)| > \delta_T/8)\\ + \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\frac{\nu_T^2}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u}| > \delta_T/8)\\ + \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\boldsymbol{p}_1(\lambda_T, \alpha, \boldsymbol{\theta}_0) - \boldsymbol{p}_1(\lambda_T, \alpha, \boldsymbol{\theta}_T)| > \nu_T\delta_T/8)\\ + \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\boldsymbol{p}_2(\gamma_T, \xi, \boldsymbol{\theta}_0) - \boldsymbol{p}_2(\gamma_T, \xi, \boldsymbol{\theta}_T)| > \nu_T\delta_T/8)\\ \leq \frac{C_{st}}{C_\epsilon^4} + \nu_T^2 C_\epsilon^2\eta(C_\epsilon)C_{st} + 3\epsilon/5 \leq \epsilon,\end{aligned}$$

for C_ϵ and T large enough. We then deduce $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(\nu_T)$. □

Proof of Theorem 3 Let $\boldsymbol{u} \in \mathbb{R}^d$ such that $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \boldsymbol{u}/T^{1/2}$ and we define the empirical criterion $\mathbb{F}_T(\boldsymbol{u}) = T\mathbb{G}_T(\varphi(\boldsymbol{\theta}_0 + \boldsymbol{u}/T^{1/2}) - \varphi(\boldsymbol{\theta}_0))$. First, we are going to prove the finite distributional convergence of $\mathbb{F}_T$ to $\mathbb{F}_\infty$. Then we use the convexity of $\mathbb{F}_T(.)$ to obtain the convergence in distribution of the arg min empirical criterion to the arg min process limit. To do so, let $\boldsymbol{u} = \sqrt{T}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$. We have

$$\begin{aligned}\mathbb{F}_T(\boldsymbol{u}) &= T\left\{\mathbb{G}_T(l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0)) + \boldsymbol{p}_1(\lambda_T, \alpha, \boldsymbol{\theta}) - \boldsymbol{p}_1(\lambda_T, \alpha, \boldsymbol{\theta}_0) + \boldsymbol{p}_2(\gamma_T, \xi, \boldsymbol{\theta})\right.\\ &\quad \left. - \boldsymbol{p}_2(\gamma_T, \xi, \boldsymbol{\theta}_0)\right\}\\ &= T\mathbb{G}_T(l(\boldsymbol{\theta}_0 + \boldsymbol{u}/T^{1/2}) - l(\boldsymbol{\theta}_0)) + \lambda_T\sum_{k=1}^m \alpha_k\left[\|\boldsymbol{\theta}_0^{(k)} + \boldsymbol{u}^{(k)}/\sqrt{T}\|_1 - \|\boldsymbol{\theta}_0^{(k)}\|_1\right]\\ &\quad + \gamma_T\sum_{l=1}^m \xi_l\left[\|\boldsymbol{\theta}_0^{(l)} + \boldsymbol{u}^{(l)}/\sqrt{T}\|_2 - \|\boldsymbol{\theta}_0^{(l)}\|_2\right],\end{aligned}$$

where $\mathbb{F}_T(.)$ is convex and $C^0(\mathbb{R}^d)$. We now prove the finite dimensional distribution of $\mathbb{F}_T$ to $\mathbb{F}_\infty$ to apply Lemma 1. For the l^1 penalty, for any group k, we have for T sufficiently large

$$\|\boldsymbol{\theta}_0^{(k)} + \boldsymbol{u}^{(k)}/\sqrt{T}\|_1 - \|\boldsymbol{\theta}_0^{(k)}\|_1 = T^{-1/2}\sum_{i=1}^{c_k}\left\{|\boldsymbol{u}_i^{(k)}|\mathbf{1}_{\theta_{0,i}^{(k)}=0} + \boldsymbol{u}_i^{(k)}\mathrm{sgn}(\theta_{0,i}^{(k)})\mathbf{1}_{\theta_{0,i}^{(k)}\neq 0}\right\},$$

which implies that

$$\begin{aligned}\lambda_T\sum_{k=1}^m \alpha_k\left[\|\boldsymbol{\theta}_0^{(k)} + \boldsymbol{u}^{(k)}/\sqrt{T}\|_1 - \|\boldsymbol{\theta}_0^{(k)}\|_1\right] &\underset{T\to\infty}{\longrightarrow} \lambda_0\sum_{k=1}^m \alpha_k\sum_{i=1}^{c_k}\left\{|\boldsymbol{u}_i^{(k)}|\mathbf{1}_{\theta_{0,i}^{(k)}=0}\right.\\ &\quad \left. + \boldsymbol{u}_i^{(k)}\mathrm{sgn}(\theta_{0,i}^{(k)})\mathbf{1}_{\theta_{0,i}^{(k)}\neq 0}\right\},\end{aligned}$$

under the condition that $\lambda_T/\sqrt{T} \to \lambda_0$. As for the l^1/l^2 quantity, for any group l, we have

$$\|\boldsymbol{\theta}_0^{(l)} + u^{(l)}/\sqrt{T}\|_2 - \|\boldsymbol{\theta}_0^{(l)}\|_2 = T^{-1/2}\left\{\|u^{(l)}\|_2 \mathbf{1}_{\boldsymbol{\theta}_0^{(l)}=\mathbf{0}} + \frac{u^{(l)'}\boldsymbol{\theta}_0^{(l)}}{\|\boldsymbol{\theta}_0^{(l)}\|_2}\mathbf{1}_{\boldsymbol{\theta}_0^{(l)}\neq\mathbf{0}}\right\} + o(T^{-1}).$$

Consequently, if $\gamma_T T^{-1/2} \to \gamma_0 \geq 0$, we obtain

$$\gamma_T \sum_{l=1}^{m} \xi_l \left[\|\boldsymbol{\theta}_0^{(l)} + u^{(l)}/\sqrt{T}\|_2 - \|\boldsymbol{\theta}_0^{(l)}\|_2\right] = \gamma_0 \sum_{l=1}^{m} \xi_l \left\{\|u^{(l)}\|_2 \mathbf{1}_{\theta_{0,k}^{(l)}=\mathbf{0}} + \frac{u^{(l)'}\boldsymbol{\theta}_0^{(l)}}{\|\boldsymbol{\theta}_0^{(l)}\|_2}\mathbf{1}_{\boldsymbol{\theta}_0^{(l)}\neq\mathbf{0}}\right\} + o(T^{-1})\gamma_T.$$

Now for the unpenalized criterion $\mathbb{G}_T l(.)$, by a Taylor expansion, we have

$$T\mathbb{G}_T(l(\boldsymbol{\theta}_0 + \boldsymbol{u}/T^{1/2}) - l(\boldsymbol{\theta}_0)) = \boldsymbol{u}' T^{1/2}\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0) + \frac{1}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} + \frac{1}{6T^{1/3}}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u},$$

where $\bar{\boldsymbol{\theta}}$ is defined as $\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \leq \|\boldsymbol{u}\|/\sqrt{T}$. Then by Assumption 4, we have the central limit theorem of Billingsley (1961)

$\sqrt{T}\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbb{M})$, and by the ergodic theorem $\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0) \xrightarrow[T\to\infty]{\mathbb{P}} \mathbb{H}$. Furthermore, we have by Assumption 6

$$\begin{aligned}
&|\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u}|^2 \\
&\leq \frac{1}{T^2}\sum_{t,t'=1}^{T}\sum_{k_1,l_1,m_1}^{d}\sum_{k_2,l_2,m_2}^{d} \boldsymbol{u}_{k_1}\boldsymbol{u}_{l_1}\boldsymbol{u}_{m_1}\boldsymbol{u}_{k_2}\boldsymbol{u}_{l_2}\boldsymbol{u}_{m_2}|\partial^3_{\theta_{k_1}\theta_{l_1}\theta_{m_1}} l(\boldsymbol{\epsilon}_t;\bar{\boldsymbol{\theta}}).\partial^3_{\theta_{k_2}\theta_{l_2}\theta_{m_2}} l(\boldsymbol{\epsilon}_{t'};\bar{\boldsymbol{\theta}})| \\
&\leq \frac{1}{T^2}\sum_{t,t'=1}^{T}\sum_{k_1,l_1,m_1}^{d}\sum_{k_2,l_2,m_2}^{d} \boldsymbol{u}_{k_1}\boldsymbol{u}_{l_1}\boldsymbol{u}_{m_1}\boldsymbol{u}_{k_2}\boldsymbol{u}_{l_2}\boldsymbol{u}_{m_2}\upsilon_t(C)\upsilon_{t'}(C),
\end{aligned}$$

for C large enough, such that $\upsilon_t(C) = \sup_{k,l,m=1,\ldots,d}\{\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2\leq\nu_T C}|\partial^3_{\theta_k\theta_l\theta_m} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta})|\}$ with $\nu_T = T^{-1/2} + \lambda_T T^{-1}a_T + \gamma_T T^{-1}b_T$. We deduce $\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u} = O_p(\|\boldsymbol{u}\|_2^3\eta(C))$. We obtain

$$\frac{1}{6T^{1/3}}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u} \xrightarrow[T\to\infty]{\mathbb{P}} 0.$$

Then we proved that $\mathbb{F}_T(\boldsymbol{u}) \xrightarrow{d} \mathbb{F}_\infty(\boldsymbol{u})$, for a fixed $\boldsymbol{u}$. Let us observe that

$$\boldsymbol{u}_T^* = \arg\min_{\boldsymbol{u}} \{\mathbb{F}_T(\boldsymbol{u})\},$$

and $\mathbb{F}_T(.)$ admits as a minimizer $\boldsymbol{u}_T^* = \sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. As $\mathbb{F}_T$ is convex and $\mathbb{F}_\infty$ is continuous, convex and has a unique minimum by Assumption 5, then by convexity Lemma 1, we obtain

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \arg\min_{\boldsymbol{u}}\{\mathbb{F}_T\} \xrightarrow{d} \arg\min_{\boldsymbol{u}}\{\mathbb{F}_\infty\}.$$

□

Proof of Proposition 1 In Theorem 3, we proved $\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) := \arg\min_{\boldsymbol{u}\in\mathbb{R}^d}\{\mathbb{F}_T\} \xrightarrow{d} \arg\min_{\boldsymbol{u}\in\mathbb{R}^d}\{\mathbb{F}_\infty\}$ for $\lambda_T/\sqrt{T} \to \lambda_0$ and $\gamma_T/\sqrt{T} \to \gamma_0$. The limit random function is

$$\begin{aligned}\mathbb{F}_\infty(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}'\mathbb{H}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{Z} + \lambda_0\sum_{k=1}^{m}\alpha_k\sum_{i=1}^{\boldsymbol{c}_k}\left\{|\boldsymbol{u}_i^{(k)}|\mathbf{1}_{\theta_{0,i}^{(k)}=0} + \boldsymbol{u}_i^{(k)}\operatorname{sgn}(\theta_{0,i}^{(k)})\mathbf{1}_{\theta_{0,i}^{(k)}\neq 0}\right\}\\ + \gamma_0\sum_{l=1}^{m}\xi_l\{\|\boldsymbol{u}^{(l)}\|_2\mathbf{1}_{\boldsymbol{\theta}_0^{(l)}=\mathbf{0}} + \frac{\boldsymbol{u}^{(l)'}\boldsymbol{\theta}_0^{(l)}}{\|\boldsymbol{\theta}_0^{(l)}\|_2}\mathbf{1}_{\boldsymbol{\theta}_0^{(l)}\neq\mathbf{0}}\}.\end{aligned}$$

First, let us observe that

$$\{\hat{\mathcal{A}}=\mathcal{A}\}=\left\{\forall k=1,\ldots,m,\ i\in\mathcal{A}_k^c, \hat{\theta}_i^{(k)}=0\right\}\cap\left\{\forall k=1,\ldots,m,\ i\in\hat{\mathcal{A}}_k^c, \theta_{0,i}^{(k)}=0\right\}.$$

Both sets describing $\{\hat{\mathcal{A}} = \mathcal{A}\}$ are symmetric, and thus we can focus on

$$\{\hat{\mathcal{A}} = \mathcal{A}\} \Rightarrow \left\{\forall k = 1,\ldots,m,\ i \in \mathcal{A}_k^c, T^{1/2}\hat{\theta}_i^{(k)} = 0\right\}.$$

Hence

$$\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) \leq \mathbb{P}\left(\forall k = 1,\ldots,m, \forall i \in \mathcal{A}_k^c, T^{1/2}\hat{\theta}_i^{(k)} = 0\right).$$

Denoting by $\boldsymbol{u}^* := \arg\min_{\boldsymbol{u}\in\mathbb{R}^d}\{\mathbb{F}_\infty(\boldsymbol{u})\}$, Theorem 3 corresponds to $\sqrt{T}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}) \xrightarrow{d} \boldsymbol{u}_{\mathcal{A}}^*$. By the Portmanteau theorem (see Wellner and van der Vaart 1996), we have

$$\limsup_{T\to\infty}\mathbb{P}(\forall k = 1,\ldots,m, \forall i \in \mathcal{A}_k^c, T^{1/2}\hat{\theta}_i^{(k)} = 0) \leq \mathbb{P}(\forall k = 1,\ldots,m, \forall i \in \mathcal{A}_k^c, \boldsymbol{u}_i^{(k)*} = 0),$$

as $\boldsymbol{\theta}_{0,\mathcal{A}^c} = \mathbf{0}$. Consequently, we need to prove that the probability of the right-hand side is strictly inferior to 1, which is upper-bounded by

$$\begin{aligned}&\mathbb{P}(\forall k = 1, \dots, m, \forall i \in \mathcal{A}_k^c, \boldsymbol{u}_i^{(k)*} = 0) \leq \\ &\min(\mathbb{P}(k \notin \mathcal{S}, \boldsymbol{u}^{(k)*} = 0), \mathbb{P}(k \in \mathcal{S}, \forall i \in \mathcal{A}_k^c, \boldsymbol{u}_i^{(k)*} = 0)).\end{aligned} \tag{10}$$

If $\lambda_0 = \gamma_0 = 0$, then $\boldsymbol{u}^* = -\mathbb{H}^{-1}\boldsymbol{Z}$ so that $\mathbb{P}_{\boldsymbol{u}^*} = \mathcal{N}(0, \mathbb{H}^{-1}\mathbb{M}\mathbb{H}^{-1})$. Hence $c = 0$.

If $\lambda_0 \neq 0$ or $\gamma_0 \neq 0$, the necessary and sufficient optimality conditions for a group k tell us that $\boldsymbol{u}^*$ satisfies

$$\begin{cases}(\mathbb{H}\boldsymbol{u}^* + \boldsymbol{Z})_{(k)} + \lambda_0\alpha_k \boldsymbol{p}^{(k)} + \gamma_0\xi_k \dfrac{\boldsymbol{\theta}_0^{(k)}}{\|\boldsymbol{\theta}_0^{(k)}\|_2} = 0, & k \in \mathcal{S}, \\ (\mathbb{H}\boldsymbol{u}^* + \boldsymbol{Z})_{(k)} + \lambda_0\alpha_k \boldsymbol{w}^{(k)} + \gamma_0\xi_k \boldsymbol{z}^{(k)} = 0, & \text{otherwise},\end{cases} \tag{11}$$

where $\boldsymbol{w}^{(k)}$ and $\boldsymbol{z}^{(k)}$ are the subgradients of $\|\boldsymbol{u}^{(k)}\|_1$ and $\|\boldsymbol{u}^{(k)}\|_2$ given by

$$\boldsymbol{w}_i^{(k)} \begin{cases} = \operatorname{sgn}(\boldsymbol{u}_i^{(k)}) \text{ if } \boldsymbol{u}_i^{(k)} \neq 0, \\ \in \{\boldsymbol{w}_i^{(k)} : |\boldsymbol{w}_i^{(k)}| \leq 1\} \text{ if } \boldsymbol{u}_i^{(k)} = 0, \end{cases} \qquad \boldsymbol{z}^{(k)} \begin{cases} = \dfrac{\boldsymbol{u}^{(k)}}{\|\boldsymbol{u}^{(k)}\|_2} \text{ if } \boldsymbol{u}^{(k)} \neq 0, \\ \in \{\boldsymbol{z}^{(k)} : \|\boldsymbol{z}^{(k)}\|_2 \leq 1\} \text{ if } \boldsymbol{u}^{(k)} = 0, \end{cases}$$

and $\boldsymbol{p}_i^{(k)} = \partial_{\boldsymbol{u}_i}\{|\boldsymbol{u}_i^{(k)}|\mathbf{1}_{\theta_{0,i}^{(k)}=0} + \boldsymbol{u}_i^{(k)}\operatorname{sgn}(\theta_{0,i}^{(k)})\mathbf{1}_{\theta_{0,i}^{(k)}\neq 0}\}$.

If $\boldsymbol{u}^{(m)*} = 0, \forall m \notin \mathcal{S}$, then the optimality conditions (11) become

$$\begin{cases}\mathbb{H}_{\mathcal{S}\mathcal{S}}\boldsymbol{u}_{\mathcal{S}}^* + \boldsymbol{Z}_{\mathcal{S}} + \lambda_0\tau_{\mathcal{S}} + \gamma_0\zeta_{\mathcal{S}} = 0, \\ \| - \mathbb{H}_{(l)\mathcal{S}}\boldsymbol{u}_{\mathcal{S}}^* - \boldsymbol{Z}_{(l)} - \lambda_0\alpha_l \boldsymbol{w}^{(l)}\|_2 \leq \gamma_0\xi_l, \text{ as } \|\boldsymbol{z}^{(l)}\|_2 \leq 1, \ l \in \mathcal{S}^c,\end{cases} \tag{12}$$

with $\tau_{\mathcal{S}} = \operatorname{vec}(k \in \mathcal{S}, \alpha_k \boldsymbol{p}^{(k)})$ and $\zeta_{\mathcal{S}} = \operatorname{vec}(k \in \mathcal{S}, \xi_k \dfrac{\boldsymbol{\theta}_0^{(k)}}{\|\boldsymbol{\theta}_0^{(k)}\|_2})$, which are vectors of $\mathbb{R}^{\operatorname{card}(\mathcal{S})}$.

For $k \in \mathcal{S}$, that is, the vector $\boldsymbol{\theta}_0^{(k)}$ is at least nonzero, then

$$\begin{cases}(\mathbb{H}\boldsymbol{u}^* + \boldsymbol{Z})_i + \lambda_0\alpha_k \operatorname{sgn}(\theta_{0,i}^{(k)}) + \gamma_0\xi_k \dfrac{\theta_{0,i}^{(k)}}{\|\boldsymbol{\theta}_0^{(k)}\|_2} = 0, \text{ if } k \in \mathcal{S}, i \in \mathcal{A}_k, \\ (\mathbb{H}\boldsymbol{u}^* + \boldsymbol{Z})_i + \lambda_0\alpha_k \boldsymbol{w}_i^{(k)} = 0, \ i \in \mathcal{A}_k^c.\end{cases} \tag{13}$$

Consequently, if $\boldsymbol{u}_i^{(k)*} = 0, \forall i \in \mathcal{A}_k^c$, with $k \in \mathcal{S}$, then the conditions (13) become

$$\begin{cases}\mathbb{H}_{\mathcal{A}_k\mathcal{A}_k}\boldsymbol{u}_{\mathcal{A}_k}^* + \boldsymbol{Z}_{\mathcal{A}_k} + \lambda_0\alpha_k \operatorname{sgn}(\boldsymbol{\theta}_{0,\mathcal{A}_k}) + \gamma_0\xi_k \dfrac{\boldsymbol{\theta}_{0,\mathcal{A}_k}}{\|\boldsymbol{\theta}_{0,\mathcal{A}_k}\|_2} = 0, \\ |-(\mathbb{H}_{\mathcal{A}_k^c\mathcal{A}_k}\boldsymbol{u}_{\mathcal{A}_k}^* + \boldsymbol{Z}_{\mathcal{A}_k^c})_i| \leq \lambda_0\alpha_k.\end{cases}$$

Combining relationships in (12), we obtain

$$\|\mathbb{H}_{(l)\mathcal{S}}\mathbb{H}_{\mathcal{S}\mathcal{S}}^{-1}(\boldsymbol{Z}_{\mathcal{S}}+\lambda_0\tau_{\mathcal{S}}+\gamma_0\zeta_{\mathcal{S}})-\boldsymbol{Z}_{(l)}-\lambda_0\alpha_l\boldsymbol{w}^{(l)}\|_2\leq\gamma_0\xi_l, l\in\mathcal{S}^c.$$

The same reasoning applies for active groups with inactive components, so that combining relationships in (13), we obtain

$$|\left(\mathbb{H}_{\mathcal{A}_k^c\mathcal{A}_k}\mathbb{H}_{\mathcal{A}_k\mathcal{A}_k}^{-1}\left(\boldsymbol{Z}_{\mathcal{A}_k}+\lambda_0\alpha_k\text{sgn}(\boldsymbol{\theta}_{0,\mathcal{A}_k})+\gamma_0\xi_k\frac{\boldsymbol{\theta}_{0,\mathcal{A}_k}}{\|\boldsymbol{\theta}_{0,\mathcal{A}_k}\|_2}\right)-\boldsymbol{Z}_{\mathcal{A}_k^c}\right)_i|\leq\lambda_0\alpha_k.$$

Hence we deduce

$$\begin{aligned}&\mathbb{P}(\forall k=1,\ldots,m,\forall\in\mathcal{A}_k^c,\boldsymbol{u}_i^{(k)*}=0)\leq\\&\min(\mathbb{P}(k\notin\mathcal{S},\boldsymbol{u}^{(k)*}=0),\mathbb{P}(k\in\mathcal{S},\forall i\in\mathcal{A}_k^c,\boldsymbol{u}_i^{(k)*}=0))\coloneqq\min(a_1,a_2).\end{aligned}$$

Under the assumption that $\lambda_0<\infty$ and $\gamma_0<\infty$, we obtain

$$\begin{aligned}a_1&=\mathbb{P}(l\in\mathcal{S}^c,\|\mathbb{H}_{(l)\mathcal{S}}\mathbb{H}_{\mathcal{S}\mathcal{S}}^{-1}(\boldsymbol{Z}_{\mathcal{S}}+\lambda_0\tau_{\mathcal{S}}+\gamma_0\zeta_{\mathcal{S}})-\boldsymbol{Z}_{(l)}-\lambda_0\alpha_l\boldsymbol{w}^{(l)}\|_2\leq\gamma_0\xi_l)<1,\\a_2&=\mathbb{P}(k\in\mathcal{S},i\in\mathcal{A}_k^c,|(\mathbb{H}_{\mathcal{A}_k^c\mathcal{A}_k}\mathbb{H}_{\mathcal{A}_k\mathcal{A}_k}^{-1}(\boldsymbol{Z}_{\mathcal{A}_k}+\lambda_0\alpha_k\text{sgn}(\boldsymbol{\theta}_{0,\mathcal{A}_k})\\&\quad+\gamma_0\xi_k\frac{\boldsymbol{\theta}_{0,\mathcal{A}_k}}{\|\boldsymbol{\theta}_{0,\mathcal{A}_k}\|_2})-\boldsymbol{Z}_{\mathcal{A}_k^c})_i|\leq\lambda_0\alpha_k)<1.\end{aligned}$$

Thus $c<1$, which proves (10), that is proposition 1. □

Proof of Theorem 4 The proof relies on the same steps as in the proof of Theorem 2. □

Proof of Theorem 5 We start with the asymptotic distribution and proceed as in the proof of Theorem 3, where we used Lemma 1. To do so, we prove the finite dimensional convergence in distribution of the empirical criterion $\mathbb{F}_T(\boldsymbol{u})$ to $\mathbb{F}_\infty(\boldsymbol{u})$ with $\boldsymbol{u}\in\mathbb{R}^d$, where these quantities are, respectively, defined as

$$\begin{aligned}\mathbb{F}_T(\boldsymbol{u})&=T\mathbb{G}_T(\psi(\boldsymbol{\theta}_0+\boldsymbol{u}/\sqrt{T})-\psi(\boldsymbol{\theta}_0))\\&=T\mathbb{G}_T(l(\boldsymbol{\theta}_0+\boldsymbol{u}/\sqrt{T})-l(\boldsymbol{\theta}_0))+\lambda_T\sum_{k=1}^m\sum_{i=1}^{\boldsymbol{c}_k}\alpha_{T,i}^{(k)}\left[|\theta_{0,i}^{(k)}+\boldsymbol{u}_i^{(k)}/\sqrt{T}|-|\theta_{0,i}^{(k)}|\right]\\&\quad+\gamma_T\sum_{l=1}^m\xi_{T,l}\left[\|\boldsymbol{\theta}_0^{(l)}+\boldsymbol{u}^{(l)}/\sqrt{T}\|_2-\|\boldsymbol{\theta}_0^{(l)}\|_2\right],\end{aligned}$$

and

$$\mathbb{F}_\infty(\boldsymbol{u})=\begin{cases}\frac{1}{2}\boldsymbol{u}'_{\mathcal{A}}\mathbb{H}_{\mathcal{A}\mathcal{A}}\boldsymbol{u}_{\mathcal{A}}+\boldsymbol{u}'_{\mathcal{A}}\boldsymbol{Z}_{\mathcal{A}} & \text{if } \boldsymbol{u}_i=0,\ \text{when } i\notin\mathcal{A},\ \text{and}\\ \infty & \text{otherwise,}\end{cases}\tag{14}$$

with $\boldsymbol{Z}_{\mathcal{A}} \sim \mathcal{N}(0, \mathbb{M}_{\mathcal{A}\mathcal{A}})$. By Lemma 1, the finite dimensional convergence in distribution implies $\arg\min_{\boldsymbol{u}\in\mathbb{R}^d}\{\mathbb{F}_T(\boldsymbol{u})\} \xrightarrow{d} \arg\min_{\boldsymbol{u}\in\mathbb{R}^d}\{\mathbb{F}_\infty(\boldsymbol{u})\}$. We first consider the unpenalized empirical criterion of $\mathbb{F}_T(.)$, which can be expanded as

$$T\mathbb{G}_T(\psi(\boldsymbol{\theta}_0 + \boldsymbol{u}/\sqrt{T}) - \psi(\boldsymbol{\theta}_0)) = T^{1/2}\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} + \frac{\boldsymbol{u}'}{2}\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} + \frac{1}{6T^{1/3}}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\}\boldsymbol{u},$$

where $\bar{\boldsymbol{\theta}}$ lies between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0 + \boldsymbol{u}/\sqrt{T}$. First, using the same reasoning on the third-order term, we obtain $\frac{1}{6T^{1/3}}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\}\boldsymbol{u} \xrightarrow[T\to\infty]{\mathbb{P}} 0$. By the ergodic theorem, we deduce $\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0) \xrightarrow[T\to\infty]{\mathbb{P}} \mathbb{H}$ and by Assumption 4, $\sqrt{T}\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbb{M})$.

We now focus on the penalty terms of (4), we remind that $\alpha_{T,i}^{(k)} = |\tilde{\theta}_i^{(k)}|^{-\eta}$, so that for $i \in \mathcal{A}_k, k \in \mathcal{S}$, $\tilde{\theta}_i^{(k)} \xrightarrow[T\to\infty]{\mathbb{P}} \theta_{0,i}^{(k)} \neq 0$. Note that

$$\sqrt{T}(|\boldsymbol{\theta}_{0,i}^{(k)} + \boldsymbol{u}_i^{(k)}/\sqrt{T}| - |\boldsymbol{\theta}_0^{(k)}|] \xrightarrow[T\to\infty]{\mathbb{P}} \boldsymbol{u}_i^{(k)}\mathrm{sgn}(\theta_{0,i}^{(k)})\mathbf{1}_{\theta_{0,i}^{(k)}\neq 0}.$$

This implies that, for $i \in \mathcal{A}_k, k \in \mathcal{S}$, we have

$$\lambda_T T^{-1/2}\sum_{i=1}^{\boldsymbol{c}_k}\alpha_{T,i}^{(k)}\sqrt{T}(|\theta_{0,i}^{(k)} + \boldsymbol{u}_i^{(k)}/\sqrt{T}| - |\theta_{0,i}^{(k)}|) \xrightarrow[T\to\infty]{\mathbb{P}} 0,$$

under the condition $\lambda_T T^{-1/2} \to 0$. For $i \in \mathcal{A}_k^c, \theta_{0,i}^{(k)} = 0$, then $T^{\eta/2}(|\tilde{\theta}_i^{(k)}|)^\eta = O_p(1)$. Hence under the assumption $\lambda_T T^{(\eta-1)/2} \to \infty$, we obtain

$$\lambda_T T^{-1/2}\alpha_{T,i}^{(k)}\sqrt{T}\left(|\theta_{0,i}^{(k)} + \boldsymbol{u}_i^{(k)}/\sqrt{T}| - |\theta_{0,i}^{(k)}|\right) = \lambda_T T^{-1/2}|\boldsymbol{u}_i^{(k)}|\frac{T^{\eta/2}}{(T^{1/2}|\tilde{\theta}_i^{(k)}|)^\eta} \xrightarrow[T\to\infty]{\mathbb{P}} \infty. \tag{15}$$

As for the l^1/l^2 quantity, we remind that $\xi_{T,l} = \|\tilde{\boldsymbol{\theta}}^{(l)}\|_2^{-\mu}$, so that for $l \in \mathcal{S}$, $\tilde{\boldsymbol{\theta}}^{(l)} \xrightarrow[T\to\infty]{\mathbb{P}} \boldsymbol{\theta}_0^{(l)}$, and in this case

$$\sqrt{T}\left\{\|\boldsymbol{\theta}_0^{(l)} + \boldsymbol{u}^{(l)}/\sqrt{T}\|_2 - \|\boldsymbol{\theta}_0^{(l)}\|_2\right\} = \frac{\boldsymbol{u}^{(l)\prime}\boldsymbol{\theta}_0^{(l)}}{\|\boldsymbol{\theta}_0^{(l)}\|_2} + o\left(T^{-1/2}\right).$$

Consequently, using $\gamma_T T^{-1/2} \to 0$, and for $l \in \mathcal{S}$, we obtain

$$\gamma_T T^{-1/2}\sqrt{T}\xi_{T,l}\left(\|\boldsymbol{\theta}_0^{(l)} + \boldsymbol{u}^{(l)}/\sqrt{T}\|_2 - \|\boldsymbol{\theta}_0^{(l)}\|_2\right) \xrightarrow[T\to\infty]{\mathbb{P}} 0.$$

Combining the fact $k \in \mathcal{S}$ and $\boldsymbol{\theta}_0^{(k)}$ is partially zero, that is $i \in \mathcal{A}_k^c$, we obtain the divergence given in (15). Furthermore, if $l \notin \mathcal{S}$, that is $\boldsymbol{\theta}_0^{(l)} = 0$, then

$$\sqrt{T}\left\{\|\boldsymbol{\theta}_0^{(l)} + \boldsymbol{u}^{(l)}/\sqrt{T}\|_2 - \|\boldsymbol{\theta}_0^{(l)}\|_2\right\} = \|\boldsymbol{u}^{(l)}\|_2,$$

and $T^{\mu/2}(\|\tilde{\boldsymbol{\theta}}^{(l)}\|_2)^{\mu} = O_p(1)$. Then by $\gamma_T T^{(\mu-1)/2} \to \infty$ we have

$$\begin{aligned}&\gamma_T T^{-1/2}\xi_{T,l}\sqrt{T}\left[\|\boldsymbol{\theta}_0^{(l)} + \boldsymbol{u}^{(l)}/\sqrt{T}\|_2 - \|\boldsymbol{\theta}_0^{(l)}\|_2\right]\\ &= \gamma_T T^{-1/2}\|\boldsymbol{u}^{(l)}\|_2 \frac{T^{\mu/2}}{(T^{1/2}\|\tilde{\boldsymbol{\theta}}^{(l)}\|_2)^{\mu}} \xrightarrow[T\to\infty]{\mathbb{P}} \infty.\end{aligned}$$

We deduce the pointwise convergence $\mathbb{F}_T(\boldsymbol{u}) \xrightarrow{d} \mathbb{F}_\infty(\boldsymbol{u})$, where $\mathbb{F}_\infty(.)$ is given in (14). As $\mathbb{F}_T(.)$ is convex and $\mathbb{F}_\infty(.)$ is convex and has a unique minimum $(\mathbb{H}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{Z}_{\mathcal{A}}, \boldsymbol{0}_{\mathcal{A}^c})$ since $\mathbb{H}$ is positive definite, by Lemma 1, we obtain

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \underset{\boldsymbol{u}\in\mathbb{R}^d}{\arg\min}\{\mathbb{F}_T(\boldsymbol{u})\} \xrightarrow{d} \underset{\boldsymbol{u}\in\mathbb{R}^d}{\arg\min}\{\mathbb{F}_\infty(\boldsymbol{u})\},$$

that is to say $\sqrt{T}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}) \xrightarrow{d} \mathbb{H}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{Z}_{\mathcal{A}}$, and $\sqrt{T}(\hat{\boldsymbol{\theta}}_{\mathcal{A}^c} - \boldsymbol{\theta}_{0,\mathcal{A}^c}) \xrightarrow{d} \boldsymbol{0}_{\mathcal{A}^c}$.

We now prove the model selection consistency. Let $i \in \mathcal{A}_k$, then by the asymptotic normality result, $\hat{\theta}_i^{(k)} \xrightarrow[T\to\infty]{\mathbb{P}} \boldsymbol{\theta}_0^{(k)}$, which implies $\mathbb{P}(i \in \hat{\mathcal{A}}_k) \to 1$. Thus the proof consists of proving

$$\forall k = 1, \ldots, m, \forall i \in \mathcal{A}_k^c, \mathbb{P}(i \in \hat{\mathcal{A}}_k) \to 0.$$

This problem can be split into two parts as

$$\forall k \notin \mathcal{S}, \mathbb{P}(k \in \hat{\mathcal{S}}) \to 0, \text{ and } \forall k \in \mathcal{S}, \forall i \in \mathcal{A}_k^c, \mathbb{P}(i \in \hat{A}_k) \to 0. \tag{16}$$

Let us start with the case $k \notin \mathcal{S}$. If $k \in \hat{\mathcal{S}}$, by the optimality conditions given by the Karush–Kuhn–Tucker theorem applied on $\mathbb{G}_T\psi(\hat{\boldsymbol{\theta}})$, we have

$$\dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}})_{(k)} + \frac{\lambda_T}{T}\alpha_T^{(k)} \odot \hat{\boldsymbol{w}}^{(k)} + \frac{\gamma_T}{T}\xi_{T,k}\frac{\hat{\boldsymbol{\theta}}^{(k)}}{\|\hat{\boldsymbol{\theta}}^{(k)}\|_2} = 0,$$

$\odot$ is the element-by-element vector product, and

$$\hat{\boldsymbol{w}}_i^{(k)} \begin{cases} = \text{sgn}(\hat{\theta}_i^{(k)}) \text{ if } \hat{\theta}_i^{(k)} \neq 0, \\ \in \{\hat{\boldsymbol{w}}_i^{(k)} : |\hat{\boldsymbol{w}}_i^{(k)}| \leq 1\} \text{ if } \hat{\theta}_i^{(k)} = 0. \end{cases}$$

Multiplying the unpenalized part by $T^{1/2}$, we have the expansion

$$\begin{aligned} T^{1/2}\dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}})_{(k)} = T^{1/2}\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k)} + T^{1/2}\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k)(k)}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)_{(k)} \\ + T^{1/2}\nabla'\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'_{(k)}\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{(k)(k)}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)_{(k)}\}, \end{aligned}$$

which is asymptotically normal by consistency, Assumption 6 regarding the bound on the third-order term, the Slutsky theorem and the central limit theorem of Billingsley (1961). Furthermore, we have

$$\gamma_T T^{-1/2}\xi_{T,k}\frac{\hat{\boldsymbol{\theta}}^{(k)}}{\|\hat{\boldsymbol{\theta}}^{(k)}\|_2} = \gamma_T T^{(\mu-1)/2}(T^{1/2}\|\tilde{\boldsymbol{\theta}}^{(k)}\|_2)^{-\mu}\frac{\hat{\boldsymbol{\theta}}^{(k)}}{\|\hat{\boldsymbol{\theta}}^{(k)}\|_2} \xrightarrow[T\to\infty]{\mathbb{P}} \infty.$$

Then using $T^{(\mu-\eta)/2}\gamma_T\lambda_T^{-1} \to \infty$, we have

$$\forall k \notin \mathcal{S}, \mathbb{P}(k \in \hat{\mathcal{S}}) \leq \mathbb{P}\left(-\dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}})_{(k)} = \frac{\lambda_T}{T}\alpha_T^{(k)} \odot \hat{\boldsymbol{w}}_i^{(k)} + \frac{\gamma_T}{T}\xi_{T,k}\frac{\hat{\boldsymbol{\theta}}^{(k)}}{\|\hat{\boldsymbol{\theta}}^{(k)}\|_2}\right) \to 0.$$

We now pick $k \in \mathcal{S}$ and consider the event $\{i \in \hat{\mathcal{A}}_k\}$. Then the Karush–Kuhn–Tucker conditions for $\mathbb{G}_T\psi(\hat{\boldsymbol{\theta}})$ are given by

$$(\dot{\mathbb{G}}_T l(\hat{\theta}))_{(k),i} + \frac{\lambda_T}{T}\alpha_{T,i}^{(k)}\text{sgn}(\hat{\theta}_{T,i}^{(k)}) + \frac{\gamma_T}{T}\xi_{T,k}\frac{\hat{\theta}_i^{(k)}}{\|\hat{\boldsymbol{\theta}}^{(k)}\|_2} = 0.$$

Using the same reasoning as previously, $T^{1/2}(\dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}}))_{(k),i}$ is also asymptotically normal, and $\tilde{\boldsymbol{\theta}}^{(k)} \xrightarrow[T\to\infty]{\mathbb{P}} \boldsymbol{\theta}_0^{(k)}$ for $k \in \mathcal{S}$, and besides

$$\lambda_T T^{-1/2}\alpha_{T,i}^{(k)}\text{sgn}(\hat{\theta}_i^{(k)}) = \lambda_T\frac{T^{(\eta-1)/2}}{(T^{1/2}|\tilde{\theta}_i^{(k)}|)^\eta} \xrightarrow[T\to\infty]{\mathbb{P}} \infty,$$

so that we obtain the same when adding $\gamma_T T^{-1/2}\xi_{T,k}\frac{\hat{\theta}_i^{(k)}}{\|\hat{\boldsymbol{\theta}}^{(k)}\|_2}$. Therefore, we have for any $k \in \mathcal{S}$ and $i \notin \mathcal{A}_k$

$$\mathbb{P}(i \in \hat{\mathcal{A}}_k) \leq \mathbb{P}\left(-(\dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}}))_{(k),i} = \frac{\lambda_T}{T}\alpha_{T,i}^{(k)}\mathrm{sgn}(\hat{\theta}_i^{(k)}) + \frac{\gamma_T}{T}\xi_{T,k}\frac{\hat{\theta}_i^{(k)}}{\|\hat{\boldsymbol{\theta}}^{(k)}\|_2}\right) \to 0.$$

We have proved (16). □

References

Anderson, P. K., Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, *10*(4), 1100–1120.

Bertsekas, D. (1995). *Nonlinear programming*. Belmont, MA: Athena Scientific.

Billingsley, P. (1961). The Lindeberg–Levy theorem for martingales. *Proceedings of the American Mathematical Society*, *12*, 788792.

Billingsley, P. (1995). *Probability and measure*. New York: Wiley.

Bühlmann, P., van de Geer, S. (2011). *Statistics for high-dimensional data*: *Methods, theory and applications*. Springer series in statistics Berlin: Springer.

Chernozhukov, V. (2005). Extremal quantile regression. *The Annals of Statistics*, *33*(2), 806–839.

Chernozhukov, V., Hong, H. (2004). Likelihood estimation and inference in a class of nonregular econometric models. *Econometrica*, *72*(5), 1445–1480.

Davis, R. A., Knight, K., Liu, J. (1992). M-estimation for autoregressions with infinite variance. *Stochastic Processes and Their Applications*, *40*, 145–180.

Fan, J. (1997). Comments on wavelets in statistics: A review by A. Antoniadis. *Journal of the Italian Statistical Association*, *6*, 131138.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fan, J., Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, *32*(3), 928–961.

Francq, C., Thieu, L. Q. (2015). *QML inference for volatility models with covariates*. MPRA paper no. 63198.

Francq, C., Zakoïan, J. M. (2010). *GARCH models*. Chichester: Wiley.

Fu, W. J. (1998). Penalized regression: the Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, *7*, 397–416.

Geyer, C. J. (1996). *On the asymptotics of convex stochastic optimization*. Unpublished manuscript.

Hjort, N. L., Pollard, D. (1993). *Asymptotics for minimisers of convex processes*. Unpublished manuscript.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, *1*(5), 799821.

Hunter, D. R., Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics*, *33*(4), 1617–1642.

Kato, K. (2009). Asymptotics for argmin processes: Convexity arguments. *Journal of Multivariate Analysis*, *100*, 1816–1829.

Knight, K., Fu, W. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics*, *28*(5), 1356–1378.

Li, X., Mo, L., Yuan, X., Zhang, J. (2014). Linearized alternating direction method of multipliers for Sparse Group and Fused Lasso models. *Computational Statistics and Data Analysis*, *79*, 203–221.

Nardi, Y., Rinaldo, A. (2008). On the asymptotic properties of the Group Lasso estimator for linear models. *Electronic Journal of Statistics*, *2*, 605–633.

Neumann, M. H. (2013). A central limit theorem for triangular arrays of weakly dependent random variables, with applications in statistics. *Probability and Statistics*, *17*, 120–134.

Newey, W. K., Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, *55*(4), 819–847.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, *7*(2), 186–199.

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, *99*, 39–61.

Rio, E. (2013). Inequalities and limit theorems for weakly dependent sequences. *3 ème Cycle*, *cel–00867106*, 170.

Rockafeller, R. T. (1970). *Convex analysis*. Princeton: Princeton University Press.
Shiryaev, A. N. (1991). *Probability*. Berlin: Springer.
Simon, N., Friedman, J., Hastie, T., Tibshirani, R. (2013). A Sparse Group Lasso. *Journal of Computational and Graphical Statistics*, *22*(2), 231–245.
Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, *58*(1), 267–288.
Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l^1-constrained quadratic programming. *IEEE Transactions on Information Theory*, *55*(5), 2183–2202.
Wellner, J. A., van der Vaart, A. W. (1996). *Weak convergence and empirical processes. With applications to statistics*. New York, NY: Springer.
Yuan, M., Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B*, *68*(1), 49–67.
Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.
Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*: *Series B*, *67*(2), 301–320.
Zou, H., Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, *37*(4), 1733–1751.