Supplementary File to Asymptotic Theory of the Adaptive Sparse Group Lasso

Benjamin Poignard

Abstract We provide the proofs of Theorems 6, 7 and 8. An additional simulated experiment is provided to support the performance of the adaptive Sparse Group Lasso regularization method.

1 Proof of Theorem 6

We proceed as in the proof of Theorem 2. We denote $\nu_T = (d_T/T)^{1/2}$ and we would like to prove that, for any $\boldsymbol{\epsilon} > 0$, there exists $C_{\boldsymbol{\epsilon}} > 0$ such that

$$\mathbb{P}(\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2/\tilde{\nu}_T > C_{\boldsymbol{\epsilon}}) < \boldsymbol{\epsilon}. \tag{1}$$

To prove (1), it is sufficient to show that for any $\boldsymbol{\epsilon} > 0$, there exists $C_{\boldsymbol{\epsilon}} > 0$ such that

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

by convexity. By a Taylor expansion of $\mathbb{G}_T l(\boldsymbol{\theta}_0 + \nu_T \boldsymbol{u})$, we obtain

$$\mathbb{G}_T l(\boldsymbol{\theta}_0 + \nu_T \boldsymbol{u}) = \mathbb{G}_T l(\boldsymbol{\theta}_0) + \nu_T \dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} + \frac{\nu_T^2}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} + \frac{\nu_T^3}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u},$$

where $\bar{\boldsymbol{\theta}} \in \Theta$ such that $\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 \leq C_{\boldsymbol{\epsilon}}\nu_T$. We would like to prove

$$\begin{aligned}\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : \quad & \nu_T \dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} + \frac{\nu_T^2}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} \\ & + \frac{\nu_T^3}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u} \leq 0) < \boldsymbol{\epsilon}\end{aligned} \tag{2}$$

To do so, we focus on each quantity of the Taylor expansion to extract the dominant term. First, for $a > 0$ and the Markov inequality, we have for the

Benjamin Poignard
Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-0043, Japan
E-mail: poignard@sigmath.es.osaka-u.ac.jp

score term

$$\begin{aligned}\mathbb{P}(\sup_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=C_{\boldsymbol{\epsilon}}}|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}|>a) &\leq \mathbb{P}(\sup_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=C_{\boldsymbol{\epsilon}}}\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\|_2\|\boldsymbol{u}\|_2>a)\\ &\leq \mathbb{P}(\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\|_2>\tfrac{a}{C_{\boldsymbol{\epsilon}}})\\ &\leq (\tfrac{C_{\boldsymbol{\epsilon}}}{a})^2\mathbb{E}[\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\|_2^2]\\ &\leq (\tfrac{C_{\boldsymbol{\epsilon}}}{a})^2\sum_{k=1}^{d_T}\mathbb{E}[(\partial_{\theta_k}\mathbb{G}_T l(\boldsymbol{\theta}_0))^2]\\ &= (\tfrac{C_{\boldsymbol{\epsilon}}}{a})^2\tfrac{1}{T^2}\sum_{t,t'=1}^{T}\sum_{k=1}^{d_T}\mathbb{E}[\partial_{\theta_k}l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\partial_{\theta_k}l(\boldsymbol{\epsilon}_{t'};\boldsymbol{\theta}_0)]\\ &\leq (\tfrac{C_{\boldsymbol{\epsilon}}}{a})^2\{\tfrac{1}{T^2}\sum_{t,t'=1}^{T}\Psi(|t-t'|)\}.d_T.\end{aligned}$$

By assumption 9, $\sup_{k=1,\cdots,d_T}\mathbb{E}[\partial_{\theta_k}l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\partial_{\theta_k}l(\boldsymbol{\epsilon}_{t'};\boldsymbol{\theta}_0)]\leq\Psi(|t-t'|)$ and $\frac{1}{T}\sum_{t,t'=1}^{T}\Psi(|t-t'|)<\infty$. This implies

$$\mathbb{P}(\sup_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=C_{\boldsymbol{\epsilon}}}|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}|>a)\leq\frac{C_{\boldsymbol{\epsilon}}^2 d_T}{Ta^2}K_1,$$

for some constant $K_1>0$.

We now focus on the hessian quantity that can be rewritten as

$$\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}=\boldsymbol{u}'\mathbb{E}[\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)]\boldsymbol{u}+\mathcal{R}_T(\boldsymbol{\theta}_0),$$

where $\mathcal{R}_T(\boldsymbol{\theta}_0)=\sum_{k,l=1}^{d_T}\boldsymbol{u}_k\boldsymbol{u}_l\{\partial^2_{\theta_k\theta_l}\mathbb{G}_T l(\boldsymbol{\theta}_0)-\mathbb{E}[\partial^2_{\theta_k\theta_l}\mathbb{G}_T l(\boldsymbol{\theta}_0)]\}$. We have

$$\mathbb{E}[\mathcal{R}_T(\boldsymbol{\theta}_0)]=0,\ \mathrm{Var}(\mathcal{R}_T(\boldsymbol{\theta}_0))=\frac{1}{T^2}\sum_{t,t'=1}^{T}\sum_{k,k',l,l'=1}^{d_T}\boldsymbol{u}_k\boldsymbol{u}_{k'}\boldsymbol{u}_l\boldsymbol{u}_{l'}\mathbb{E}[\zeta_{kl,t}.\zeta_{k'l',t'}],$$

where $\zeta_{kl,t}=\partial^2_{\theta_k\theta_l}l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)-\mathbb{E}[\partial^2_{\theta_k\theta_l}l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)]$. Let $b>0$, we deduce by the Markov inequality and assumption 10,

$$\mathbb{P}(|\mathcal{R}_T(\boldsymbol{\theta}_0)|>b)\leq\frac{1}{b^2}\mathbb{E}[\mathcal{R}_T^2(\boldsymbol{\theta}_0)]\leq\frac{K_2\|\boldsymbol{u}\|_2^4 d_T^2}{T}\leq\frac{K_2 C_{\boldsymbol{\epsilon}}^4 d_T^2}{b^2T},$$

where $K_2>0$. Furthermore, by assumption 7,

$$\boldsymbol{u}'\mathbb{E}[\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)]\boldsymbol{u}\geq\lambda_{\min}(\mathbb{H}_T)\boldsymbol{u}'\boldsymbol{u}.$$

As for the third order term, we have

$$\begin{aligned}&|\nabla\{\boldsymbol{u}'\dddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u}|^2\\ &\leq\frac{1}{T^2}\sum_{t,t'=1}^{T}\sum_{k_1,k_2,k_3}\sum_{l_1,l_2,l_3}|u_{k_1}u_{k_2}u_{k_3}u_{l_1}u_{l_2}u_{l_3}||\partial^3_{\theta_{k_1}\theta_{k_2}\theta_{k_3}}l(\boldsymbol{\epsilon}_t;\bar{\boldsymbol{\theta}}).\partial^3_{\theta_{l_1}\theta_{l_2}\theta_{l_3}}l(\boldsymbol{\epsilon}_{t'};\bar{\boldsymbol{\theta}})|\\ &\leq\|\boldsymbol{u}\|_2^6 d_T^3\frac{1}{T^2}\sum_{t,t'=1}^{T}\upsilon_t(C_{\boldsymbol{\epsilon}})\upsilon_{t'}(C_{\boldsymbol{\epsilon}}),\end{aligned}$$

where

$$\upsilon_t(C_0) = \sup_{k_1k_2k_3} \{ \sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2\leq \nu_T C_0} |\partial^3_{\theta_{k_1}\theta_{k_2}\theta_{k_3}} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta})|\}.$$

Note that $\upsilon_t(C_0)$ depends on d_T and C_0. By assumption 11, we have

$$\eta(C_0) := \frac{1}{T^2}\sum_{t,t'=1}^{T} \mathbb{E}[\upsilon_t(C_0)\upsilon_{t'}(C_0)] < \infty.$$

By the Markov inequality, for $c > 0$, we conclude that

$$\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : \frac{\nu_T^2}{6} \sup_{\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2\leq \nu_T C_{\boldsymbol{\epsilon}}} |\nabla\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u}| > c) \leq \frac{\nu_T^4 d_T^3 C_{\boldsymbol{\epsilon}}^6}{36c^2}\eta(C_{\boldsymbol{\epsilon}}).$$

We can now bound (2) thanks to proper choices of a, b, c and $C_{\boldsymbol{\epsilon}}$. We denote by $\delta_T = \lambda_{\min}(\mathbb{H}_T)C_{\boldsymbol{\epsilon}}^2\nu_T/2$, and using $\frac{\nu_T}{2}\mathbb{E}[\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}] \geq \delta_T$, we have

$$\begin{aligned}
&\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : \dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} + \frac{\nu_T}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} + \frac{\nu_T^2}{6}\nabla\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u} \leq 0)\\
&\leq \mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : |\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}| > \delta_T/4)\\
&+\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : \tfrac{\nu_T}{2}|\mathcal{R}_T(\boldsymbol{\theta}_0)| > \delta_T/4)\\
&+\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : \tfrac{\nu_T^2}{6} \sup_{\bar{\boldsymbol{\theta}}:\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2<\nu_T C_{\boldsymbol{\epsilon}}} |\nabla\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u}| > \delta_T/4)\\
&\leq \frac{16C_{\boldsymbol{\epsilon}}^2 d_T K_1}{T\delta_T^2} + \frac{4\nu_T^2 d_T^2 C_{\boldsymbol{\epsilon}}^4}{T\delta_T^2} + \frac{16\nu_T^4 d_T^3 C_{\boldsymbol{\epsilon}}^6}{36\delta_T^2}\eta(C_{\boldsymbol{\epsilon}})\\
&\leq C_1\frac{d_T}{TC_{\boldsymbol{\epsilon}}^2\nu_T^2} + C_2\frac{d_T^2}{T} + C_3\nu_T^2 d_T^3 C_{\boldsymbol{\epsilon}}^2\eta(C_{\boldsymbol{\epsilon}}),
\end{aligned}$$

where C_1, C_2, C_3 are strictly positive constants. We chose $\nu_T = (\frac{d_T}{T})^{\frac{1}{2}}$, we then deduce

$$\begin{aligned}
&\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : \nu_T\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} + \frac{\nu_T^2}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} + \frac{\nu_T^3}{6}\nabla\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u} \leq 0)\\
&\leq \frac{C_1}{C_{\boldsymbol{\epsilon}}^2} + C_2\frac{d_T^2}{T} + \frac{d_T^4 C_{\boldsymbol{\epsilon}}^2}{T}\eta(C_{\boldsymbol{\epsilon}}).
\end{aligned}$$

Now we fix $C_{\boldsymbol{\epsilon}}$ sufficiently large enough, such that $C_1/C_{\boldsymbol{\epsilon}}^2 < \boldsymbol{\epsilon}/3$. Once this constant is fixed, there exists a T_0 such that for $T > T_0$ we have $C_2\frac{d_T^2}{T} < \boldsymbol{\epsilon}/3$ and $C_3\frac{d_T^4 C_{\boldsymbol{\epsilon}}^2}{T}\eta(C_{\boldsymbol{\epsilon}}) < \boldsymbol{\epsilon}/3$ under the assumption that $d_T^4 = o(T)$. Consequently, we obtain

$$\begin{aligned}
\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : \quad &\mathbb{G}_T l(\boldsymbol{\theta}_0) + \nu_T\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} + \frac{\nu_T^2}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}\\
&+ \frac{\nu_T^3}{6}\nabla\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u} \leq 0) < \boldsymbol{\epsilon}.
\end{aligned}$$

This proves (1), that is $\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = O_p((\frac{d_T}{T})^{\frac{1}{2}})$. □

2 Proof of Theorem 7

We proceed as we did for proving Theorem 6. Let $\nu_T = (d_T/T)^{1/2}$. We would like to prove that for any $\boldsymbol{\epsilon} > 0$, there exists $C_{\boldsymbol{\epsilon}} > 0$ such that

$$\mathbb{P}(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2/\nu_T > C_{\boldsymbol{\epsilon}}) < \boldsymbol{\epsilon}. \tag{3}$$

To prove (3), we show

$$\begin{aligned}\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : \dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} + \frac{\nu_T}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} + \frac{\nu_T^2}{6}\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u} \\ +\nu_T^{-1}\{\boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T\boldsymbol{u}) - \boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0) + \boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T\boldsymbol{u}) - \boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0)\} \leq 0) \\ < \boldsymbol{\epsilon},\end{aligned} \tag{4}$$

a relationship obtained by convexity and a Taylor expansion.

The score quantity can be upper bounded as

$$|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}| \leq \|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\|_2\|\boldsymbol{u}\|_2 = O_p((\frac{d_T}{T})^{\frac{1}{2}})\|\boldsymbol{u}\|_2 = O_p(\nu_T)\|\boldsymbol{u}\|_2,$$

where we used assumption 9 to obtain the bound in probability of the score.

As for the third order term, we have by the Cauchy-Schwartz inequality

$$\begin{aligned}&|\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u}|^2 \\ &\leq \|\boldsymbol{u}\|_2^6 d_T^3 \frac{1}{T^2}\sum_{t,t'=1}^{T^2}\{\sum_{k_1,l_1,m_1=1}^{d_T}\sum_{k_2,l_2,m_2=1}^{d_T}\partial^3_{\theta_{k_1}\theta_{l_1}\theta_{m_1}} l(\boldsymbol{\epsilon}_t;\bar{\boldsymbol{\theta}})\partial^3_{\theta_{k_2}\theta_{l_2}\theta_{m_2}} l(\boldsymbol{\epsilon}_{t'};\bar{\boldsymbol{\theta}})\} \\ &= \|\boldsymbol{u}\|_2^6 d_T^3 \eta(C_{\boldsymbol{\epsilon}}).\end{aligned}$$

This implies

$$\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u} = O_p(d_T^{3/2}\|\boldsymbol{u}\|_2^3).$$

Hence by the Markov inequality

$$\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : |\nu_T^2\nabla'\{\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u}| > a) \leq \frac{\nu_T^4 C_{\boldsymbol{\epsilon}}^6 d_T^3}{a^2}\eta(C_{\boldsymbol{\epsilon}}),$$

where we used assumption 11.

Finally, the hessian quantity can be treated as in the proof of Theorem 6. We denote by $\mathcal{R}_T(\boldsymbol{\theta}_0) = \sum_{k,l=1}^{d_T}\boldsymbol{u}_k\boldsymbol{u}_l\{\partial^2_{\theta_k\theta_l}\mathbb{G}_T l(\boldsymbol{\theta}_0) - \mathbb{E}[\partial^2_{\theta_k\theta_l}\mathbb{G}_T l(\boldsymbol{\theta}_0)]\}$. We have

$$\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} = \boldsymbol{u}'\mathbb{E}[\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)]\boldsymbol{u} + \mathcal{R}_T(\boldsymbol{\theta}_0).$$

By assumption 10 and the Markov inequality, for any $\kappa > 0$, we obtain

$$\mathbb{P}(|\mathcal{R}_T(\boldsymbol{\theta}_0)| > \kappa) \leq \frac{1}{\kappa^2}\mathbb{E}[\mathcal{R}_T^2(\boldsymbol{\theta}_0)] \leq \frac{K_2}{\kappa^2}\frac{\|\boldsymbol{u}\|_2^4 d_T^2}{T} \leq \frac{K_2 C_{\boldsymbol{\epsilon}}^4 d_T^2}{\kappa^2 T},$$

with $K_2 > 0$. This relationship holds for any $\kappa > 0$. Then for T large enough, we deduce that $|\mathcal{R}_T(\boldsymbol{\theta}_0)| = o_p(1)$. Consequently

$$\frac{\nu_T^2}{2}\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} \geq \frac{\nu_T^2}{2}\lambda_{\min}(\mathbb{H}_T)\|\boldsymbol{u}\|_2^2 + o_p(1)\nu_T^2\|\boldsymbol{u}\|_2^2.$$

We focus on the penalty terms. We have

$$\begin{aligned}\boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T\boldsymbol{u}) - \boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0) &= \tfrac{\lambda_T}{T}\sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k}\alpha_{T,i}^{(k)}\{|\theta_{0,i}^{(k)} + \nu_T\boldsymbol{u}_i^{(k)}| - |\theta_{0,i}^{(k)}|\},\\ \text{and } |\boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T\boldsymbol{u}) - \boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0)| &\leq \tfrac{\lambda_T}{T}\sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k}\alpha_{T,i}^{(k)}\nu_T|\boldsymbol{u}_i^{(k)}|.\end{aligned}$$

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

$$\begin{aligned}\boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T\boldsymbol{u}) - \boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0) &= \tfrac{\gamma_T}{T}\sum_{l\in\mathcal{S}1}\xi_{T,l}\{\|\boldsymbol{\theta}_0^{(l)} + \nu_T\boldsymbol{u}\|_2 - \|\boldsymbol{\theta}_0^{(l)}\|_2\}\\ \text{and } |\boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T\boldsymbol{u}) - \boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0)| &\leq \tfrac{\gamma_T}{T}\sum_{l\in\mathcal{S}}\xi_{T,l}\nu_T\|\boldsymbol{u}^{(l)}\|_2.\end{aligned}$$

For the l^1 norm penalty, using $\{\min_{k\in\mathcal{S}, i\in\mathcal{A}_k}|\tilde{\tilde{\theta}}_i^{(k)}|\}^{-\eta} \leq T^{\kappa\eta}$, then

$$\begin{aligned}\tfrac{\lambda_T}{T}\sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k}\alpha_{T,i}^{(k)}\nu_T|\boldsymbol{u}_i^{(k)}| &\leq \tfrac{\lambda_T}{T}\nu_T\{\sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k}|\tilde{\tilde{\theta}}_i^{(k)}|^{-2\eta}\}^{1/2}\|\boldsymbol{u}\|_2\\ &\leq \tfrac{\lambda_T}{T}\nu_T\frac{\sqrt{d_T}}{\{\min_{k\in\mathcal{S}, i\in\mathcal{A}_k}|\tilde{\tilde{\theta}}_i^{(k)}|\}^{\eta}}\|\boldsymbol{u}\|_2\\ &\leq \tfrac{\lambda_T}{T}\nu_T\sqrt{d_T}T^{\kappa\eta}\|\boldsymbol{u}\|_2,\end{aligned}$$

by the Cauchy-Schwartz inequality. Then if $\lambda_T T^{\frac{c}{2}-1+\kappa\eta}$ is bounded, we obtain

$$\boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T\boldsymbol{u}) - \boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0) = O(\nu_T^2)\|\boldsymbol{u}\|_2.$$

As for the l^1/l^2 term, using $\{\min_{l\in\mathcal{S}}\|\tilde{\tilde{\boldsymbol{\theta}}}^{(l)}\|_2\}^{-\mu} \leq T^{\kappa\mu}$, we obtain

$$\begin{aligned}\tfrac{\gamma_T}{T}\sum_{l=1}^{m}\xi_{T,l}\nu_T\|\boldsymbol{u}^{(l)}\|_2 &\leq \tfrac{\gamma_T}{T}\nu_T\{\sum_{l\in\mathcal{S}}\|\tilde{\tilde{\boldsymbol{\theta}}}^{(l)}\|_2^{-2\mu}\}^{1/2}\|\boldsymbol{u}\|_2\\ &\leq \tfrac{\gamma_T}{T}\nu_T\frac{\sqrt{d_T}}{\{\min_{l\in\mathcal{S}}\|\tilde{\tilde{\boldsymbol{\theta}}}^{(l)}\|_2\}^{\mu}}\|\boldsymbol{u}\|_2\\ &\leq \tfrac{\gamma_T}{T}\nu_T\sqrt{d_T}T^{\kappa\mu}\|\boldsymbol{u}\|_2,\end{aligned}$$

by the Cauchy-Schwartz inequality. Then if $\gamma_T T^{\frac{c}{2}-1+\kappa\mu}$ is bounded, we obtain

$$\boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T\boldsymbol{u}) - \boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0) = O(\nu_T^2)\|\boldsymbol{u}\|_2.$$

We now can prove (4). Let $\delta_T = \lambda_{\min}(\mathbb{H}_T)C_{\boldsymbol{\epsilon}}^2\nu_T/2$ and using $\frac{\nu_T}{2}\mathbb{E}[\boldsymbol{u}'\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}] \geq$

δ_T, we have

$$\begin{aligned}
&\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : \dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u} + \nu_T \boldsymbol{u}' \ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}/2 + \nu_T^2 \nabla\{\boldsymbol{u}' \ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u}/6 \\
&+\nu_T^{-1}\{\boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T \boldsymbol{u}) - \boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0) + \boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T \boldsymbol{u}) - \boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0)\} \leq 0) \\
&\leq \mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : |\nu_T \boldsymbol{u}' \ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}/2| \leq |\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}| + |\nu_T^2 \nabla\{\boldsymbol{u}' \ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\boldsymbol{u}\}\boldsymbol{u}/6| \\
&+\nu_T^{-1}\{|\boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0) - \boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T \boldsymbol{u})| + |\boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0) - \boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T \boldsymbol{u})|\}) \\
&\leq \mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : |\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\boldsymbol{u}| > \delta_T/8) \\
&+\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : \tfrac{\nu_T}{2}|\mathcal{R}_T(\boldsymbol{\theta}_0)| > \delta_T/8) \\
&+\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : |\tfrac{\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} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : |\boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0) - \boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T \boldsymbol{u})| > \nu_T \delta_T/8) \\
&+\mathbb{P}((\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : |\boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0) - \boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T \boldsymbol{u})| > \nu_T \delta_T/8) \\
&\leq \frac{C_{st}}{C_{\boldsymbol{\epsilon}}^2} + C_{st}\{\nu_T^2 C_{\boldsymbol{\epsilon}}^6 d_T^3 \eta(C_{\boldsymbol{\epsilon}})\} + \frac{C_{st}\nu_T^2 d_T^2 C_{\boldsymbol{\epsilon}}^4}{T\delta_T^2} + \boldsymbol{\epsilon}/5 + \boldsymbol{\epsilon}/5 \\
&< \boldsymbol{\epsilon},
\end{aligned}$$

with $C_{st} > 0$ a generic constant. We used $d_T^4 = o(T)$ and for $C_{\boldsymbol{\epsilon}}$ large enough

$$\begin{aligned}
&\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : |\boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0) - \boldsymbol{p}_1(\lambda_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T \boldsymbol{u})| > \nu_T \delta_T/8) \leq \boldsymbol{\epsilon}/5, \\
&\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_T}, \|\boldsymbol{u}\|_2 = C_{\boldsymbol{\epsilon}} : |\boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0) - \boldsymbol{p}_2(\gamma_T, \tilde{\tilde{\boldsymbol{\theta}}}, \boldsymbol{\theta}_0 + \nu_T \boldsymbol{u})| > \nu_T \delta_T/8) \leq \boldsymbol{\epsilon}/5.
\end{aligned}$$

Thus we obtain for $C_{\boldsymbol{\epsilon}}$ and T large enough, with the conditions $\gamma_T T^{\frac{c}{2}-1+\kappa\mu} \to 0$ and $\lambda_T T^{\frac{c}{2}-1+\kappa\eta} \to 0$ that

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

□

3 Proof of Theorem 8

Model selection consistency consists of proving that the probability of the event $\{\hat{\mathcal{A}} = \mathcal{A}\}$ tends to one asymptotically. This event is

$$\{\hat{\mathcal{A}} = \mathcal{A}\} = \{\forall k \in \mathcal{S}, \forall i \in \mathcal{A}_k, |\hat{\theta}_i^{(k)}| > 0\} \cap \{\forall k = 1, \cdots, m, \forall i \in \mathcal{A}_k^c, \hat{\theta}_i^{(k)} = 0\}.$$

Hence we prove

$$\mathbb{P}(\{\forall k \in \mathcal{S}, \forall i \in \mathcal{A}_k, |\hat{\theta}_i^{(k)}| > 0\} \cap \{\forall k = 1, \cdots, m, \forall i \in \mathcal{A}_k^c, \hat{\theta}_i^{(k)} = 0\}) \underset{T\to\infty}{\longrightarrow} 1. \tag{5}$$

Model selection consistency can be decomposed into two parts: recovering the active indices by estimating nonzero coefficients; discarding the inactive indices by shrinking to zero the related coefficients. Now (5) can be proved by first showing that for any T, there exists β such that $0 < \beta < \min_{i \in \mathcal{A}_k} \theta_{0,i,\mathcal{A}_k}$, with $k \in \mathcal{S}$ and

$$\mathbb{P}(\|\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}\|_2 < \beta) \underset{T\to\infty}{\longrightarrow} 1. \tag{6}$$

The second part regarding nonactive indices can be proved as

$$\begin{cases} \mathbb{P}(\bigcap_{k\in\mathcal{S}^c}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2<1\}) \underset{T\to\infty}{\longrightarrow} 1, \\ \mathbb{P}(\bigcap_{k\in\mathcal{S}}\bigcap_{i\in\mathcal{A}_k^c}\{|\hat{\boldsymbol{w}}_i^{(k)}|<1\}) \underset{T\to\infty}{\longrightarrow} 1, \end{cases} \tag{7}$$

where $\hat{\boldsymbol{z}}^{(k)}$ (resp. $\hat{\boldsymbol{w}}^{(k)}$) is the subgradient of $\|\hat{\boldsymbol{\theta}}^{(k)}\|_2$ (resp. $\|\hat{\boldsymbol{\theta}}^{(k)}\|_1$) given in Section 3. Hence (6) and (7) prove (5).

We first focus on (6), which is equivalent to $\mathbb{P}(\|\hat{\boldsymbol{\theta}}_{\mathcal{A}_k}-\boldsymbol{\theta}_{0,\mathcal{A}_k}\|_2>\beta) \underset{T\to\infty}{\longrightarrow} 0$. By the Karush-Kuhn-Tucker optimality conditions, we have

$$\dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}})_{\mathcal{A}}+\frac{\lambda_T}{T}\alpha_{T,\mathcal{A}}\odot\operatorname{sgn}(\hat{\boldsymbol{\theta}}_{\mathcal{A}})+\frac{\gamma_T}{T}\varsigma_T=0,$$

where $\varsigma_T=\operatorname{vec}(\xi_{T,k}\dfrac{\hat{\boldsymbol{\theta}}_{\mathcal{A}_k}}{\|\hat{\boldsymbol{\theta}}_{\mathcal{A}_k}\|_2}, k\in\mathcal{S})$. We denote by $\alpha_{T,\mathcal{A}_k}=(\alpha_{T,i}, i\in\mathcal{A}_k)$, a vector of size $\mathbb{R}^{\boldsymbol{c}_{\mathcal{A}_k}}$. By a Taylor expansion of the gradient component around $\boldsymbol{\theta}_{0,\mathcal{A}}$, we have

$$\begin{aligned}
&\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{\mathcal{A}}+\mathbb{H}_{T,\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}}-\boldsymbol{\theta}_{0,\mathcal{A}})+\mathcal{P}_T(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_{\mathcal{A}}-\boldsymbol{\theta}_{0,\mathcal{A}}) \\
&+\frac{1}{2}\nabla'_{\mathcal{A}}\{(\hat{\boldsymbol{\theta}}_{\mathcal{A}}-\boldsymbol{\theta}_{0,\mathcal{A}})'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{\mathcal{A}}-\boldsymbol{\theta}_{0,\mathcal{A}})\}+\tfrac{\lambda_T}{T}\alpha_{T,\mathcal{A}}\odot\operatorname{sgn}(\hat{\boldsymbol{\theta}}_{\mathcal{A}})+\tfrac{\gamma_T}{T}\varsigma_T=0 \\
&\Leftrightarrow \hat{\boldsymbol{\theta}}_{\mathcal{A}}=\boldsymbol{\theta}_{0,\mathcal{A}}-\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}(\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{\mathcal{A}}+\tfrac{\lambda_T}{T}\alpha_{T,\mathcal{A}}\odot\operatorname{sgn}(\hat{\boldsymbol{\theta}}_{\mathcal{A}})+\tfrac{\gamma_T}{T}\varsigma_T \\
&-\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\frac{1}{2}\nabla'_{\mathcal{A}}\{(\hat{\boldsymbol{\theta}}_{\mathcal{A}}-\boldsymbol{\theta}_{0,\mathcal{A}})'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}}-\boldsymbol{\theta}_{0,\mathcal{A}})\}-\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\mathcal{P}_T(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_{\mathcal{A}}-\boldsymbol{\theta}_{0,\mathcal{A}}),
\end{aligned}$$

where $\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2\le\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2$, $\mathcal{P}_T(\boldsymbol{\theta}_0)=\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{\mathcal{A}\mathcal{A}}-\mathbb{H}_{T,\mathcal{A}\mathcal{A}}$ and $\mathbb{H}_{T,\mathcal{A}\mathcal{A}}=\mathbb{E}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}'}l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)]_{\mathcal{A}\mathcal{A}}$. Then using $\|\hat{\boldsymbol{\theta}}_{\mathcal{A}}-\boldsymbol{\theta}_{0,\mathcal{A}}\|_2=O_p((\dfrac{d_T}{T})^{\frac{1}{2}})$, we obtain

$$\begin{aligned}
&\mathbb{P}(\|\hat{\boldsymbol{\theta}}_{\mathcal{A}}-\boldsymbol{\theta}_{0,\mathcal{A}}\|_2>\beta)\le\mathbb{P}(\|\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{\mathcal{A}}\|_2+\|\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\|_2\|\frac{\lambda_T}{T}\alpha_{T,\mathcal{A}}\odot\operatorname{sgn}(\hat{\boldsymbol{\theta}}_{\mathcal{A}})\|_2 \\
&+\|\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\|_2\|\frac{\gamma_T}{T}\varsigma_T\|_2+\|\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\|_2\|\nabla'_{\mathcal{A}}\{(\hat{\boldsymbol{\theta}}_{\mathcal{A}}-\boldsymbol{\theta}_{0,\mathcal{A}})'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}}-\boldsymbol{\theta}_{0,\mathcal{A}})\}/2\|_2 \\
&+\|\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\|_2\|\mathcal{P}_T(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_{\mathcal{A}}-\boldsymbol{\theta}_{0,\mathcal{A}})\|_2>\beta) \\
&\le\mathbb{P}(\lambda^{-1}_{\min}(\mathbb{H}_T)\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{\mathcal{A}}\|_2+\lambda^{-1}_{\min}(\mathbb{H}_T)\frac{\lambda_T}{T}\|\alpha_{T,\mathcal{A}}\|_2 \\
&+\lambda^{-1}_{\min}(\mathbb{H}_T)\frac{\gamma_T}{T}\|\varsigma_T\|_2+\lambda^{-1}_{\min}(\mathbb{H}_T)C_0^2(d_T/2T)\|\nabla'_{\mathcal{A}}\{\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{\mathcal{A}\mathcal{A}}\}\|_2 \\
&+\lambda^{-1}_{\min}(\mathbb{H}_T)C_0(d_T/T)^{1/2}\|\mathcal{P}_T(\boldsymbol{\theta}_0)\|_2>\beta)+\mathbb{P}(\|\hat{\boldsymbol{\theta}}_{\mathcal{A}}-\boldsymbol{\theta}_{0,\mathcal{A}}\|_2>(d_T/T)^{1/2}C_0),
\end{aligned}$$

for $C_0>0$ large enough, and we used $\|\mathbb{H}_T^{-1}\boldsymbol{x}\|_2\le\lambda^{-1}_{\min}(\mathbb{H}_T)\|\boldsymbol{x}\|_2$ for any vector $\boldsymbol{x}\in\mathbb{R}^{d_T}$. Let us proceed element-by-element. We have by the Markov inequality

$$\mathbb{P}(\lambda^{-1}_{\min}(\mathbb{H}_T)C_0\sqrt{\frac{d_T}{T}}\|\mathcal{P}_T(\boldsymbol{\theta}_0)\|_2>\frac{\beta}{6})\le\frac{36\lambda^{-2}_{\min}(\mathbb{H}_T)C_0^2d_T}{T\beta^2}\mathbb{E}[\|\mathcal{P}_T(\boldsymbol{\theta}_0)\|_2^2].$$

We have

$$\mathbb{E}[\|\mathcal{P}_T\|_2^2] = \frac{1}{T^2}\sum_{t,t'=1}^{T}\sum_{k,k'\in\mathcal{A}}\sum_{l,l'\in\mathcal{A}}\mathbb{E}[\zeta_{kl,t}\zeta_{k'l',t'}],$$

where $\zeta_{kl,t} = \partial^2_{\theta_k\theta_l} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0) - \mathbb{E}[\partial^2_{\theta_k\theta_l} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)]$. By assumption 10, we obtain

$$\mathbb{P}(\lambda^{-1}_{\min}(\mathbb{H}_T)C_0\sqrt{\frac{d_T}{T}}\|\mathcal{P}_T(\boldsymbol{\theta}_0)\|_2 > \frac{\beta}{6}) \leq \frac{36\lambda^{-2}_{\min}(\mathbb{H}_T)C_0^2}{\beta^2}\frac{d_T^3}{T^2}.$$

As for the third order term, by the Markov inequality

$$\begin{aligned}&\mathbb{P}(\frac{1}{2}\lambda^{-1}_{\min}(\mathbb{H}_T)C_0^2\frac{d_T}{T}\|\nabla'_{\mathcal{A}}\{\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{\mathcal{A}\mathcal{A}}\}\|_2 > \frac{\beta}{6})\\ &\leq \frac{9\lambda^{-2}_{\min}(\mathbb{H}_T)C_0^4 d_T^2}{T^2}\mathbb{E}[\|\nabla'_{\mathcal{A}}\{\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{\mathcal{A}\mathcal{A}}\}\|_2^2].\end{aligned}$$

We obtain

$$\begin{aligned}&\mathbb{E}[\|\nabla'_{\mathcal{A}}\{\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{\mathcal{A}\mathcal{A}}\}\|_2^2]\\ &\leq \frac{1}{T^2}\sum_{t,t'=1}^{T}\sum_{k_1,k_2,k_3\in\mathcal{A}}\sum_{l_1,l_2,l_3\in\mathcal{A}}\mathbb{E}[|\partial^3_{\theta_{k_1}\theta_{k_2}\theta_{k_3}} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0).\partial^3_{\theta_{l_1}\theta_{l_2}\theta_{l_3}} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0) l(\boldsymbol{\epsilon}_{t'};\boldsymbol{\theta}_0)|]\\ &\leq \frac{1}{T^2}d_T^3\sum_{t,t'=1}^{T}\mathbb{E}[\upsilon_t(C_0)\upsilon_{t'}(C_0)] = \eta(C_0)d_T^3,\end{aligned}$$

by assumption 11, where $\upsilon_t(C_0) = \sup_{k_1k_2k_3}\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2\leq\sqrt{\frac{d_T}{T}}C_0}|\partial^3_{\theta_{k_1}\theta_{k_2}\theta_{k_3}} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)|$.
We deduce that

$$\mathbb{P}(\frac{1}{2}\lambda^{-1}_{\min}(\mathbb{H}_T)C_0^2\frac{d_T}{T}\|\nabla'_{\mathcal{A}}\{\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{\mathcal{A}\mathcal{A}}\}\|_2 > \frac{\beta}{6}) \leq \frac{9\lambda^{-2}_{\min}(\mathbb{H}_T)C_0^4}{4}\frac{d_T^5}{T^2}\eta(C_0).$$

As for the score, by the Markov inequality and assumption 9

$$\begin{aligned}&\mathbb{P}(\lambda^{-1}_{\min}(\mathbb{H}_T)\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{\mathcal{A}}\|_2 > \beta/6) \leq \frac{\lambda^{-2}_{\min}(\mathbb{H}_T)36}{\beta^2}\mathbb{E}[\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{\mathcal{A}}\|_2]\\ &\leq \frac{\lambda^{-2}_{\min}(\mathbb{H}_T)36}{\beta^2}\frac{1}{T^2}\sum_{t,t'=1}^{T}\sum_{k\in\mathcal{A}}\mathbb{E}[\partial_{\theta_k} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\partial_{\theta_k} l(\boldsymbol{\epsilon}_{t'};\boldsymbol{\theta}_0)]\\ &\leq \frac{\lambda^{-2}_{\min}(\mathbb{H}_T)36}{\beta^2}\frac{1}{T}\{\frac{1}{T}\sum_{t,t'=1}^{T}\Psi(|t-t'|)\}d_T \leq \frac{\lambda^{-2}_{\min}(\mathbb{H}_T)36Kd_T}{T\beta^2},\end{aligned}$$

with $K > 0$. Hence we deduce

$$\begin{aligned}
&\mathbb{P}(\|\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}\|_2 > \beta) \leq \mathbb{P}(\lambda_{\min}^{-1}(\mathbb{H}_T)\frac{\lambda_T}{T}\|\alpha_{T,\mathcal{A}}\|_2 + \lambda_{\min}^{-1}(\mathbb{H}_T)\frac{\gamma_T}{T}\|\varsigma_T\|_2 > \beta/2) \\
&+ \mathbb{P}(\|\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}\|_2 > (d_T/T)^{1/2}C_0) + \mathbb{P}(\lambda_{\min}^{-1}(\mathbb{H}_T)C_0(d_T/T)^{1/2}\|\mathcal{P}_T(\boldsymbol{\theta}_0)\|_2 > \beta/6) \\
&+ \mathbb{P}(\lambda_{\min}^{-1}(\mathbb{H}_T)C_0^2(d_T/2T)\|\nabla'_{\mathcal{A}}\{\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{\mathcal{A}\mathcal{A}}\}\|_2 > \beta/6) \\
&+ \mathbb{P}(\lambda_{\min}^{-1}(\mathbb{H}_T)\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{\mathcal{A}}\|_2 > \beta/6) \\
&\leq \frac{2\lambda_{\min}^{-1}(\mathbb{H}_T)}{\beta}\{\frac{\lambda_T}{T}d_T^{1/2}\mathbb{E}[\max_{k\in\mathcal{S},i\in\mathcal{A}_k}\alpha_{T,\mathcal{A}_k,i}] + \frac{\gamma_T}{T}\mathbb{E}[\max_{k\in\mathcal{S}}\xi_{T,k}]\} \\
&+ \frac{36\lambda_{\min}^{-2}(\mathbb{H}_T)Kd_T}{T\beta^2} + \frac{9\lambda_{\min}^{-2}(\mathbb{H}_T)C_0^4}{4}\frac{d_T^5}{T^2}\eta(C_0) + \frac{36\lambda_{\min}^{-2}(\mathbb{H}_T)C_0^2}{\beta^2}\frac{d_T^3}{T^2} + \boldsymbol{\epsilon}.
\end{aligned}$$

For T and C_0 large enough, if $d_T^5 = o(T)$, by assumption 12, that is if

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

then $\mathbb{P}(\|\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}\|_2 > \beta) \underset{T\to\infty}{\longrightarrow} 0$.

We now turn to the second step of model selection consistency. First we prove

$$\mathbb{P}(\bigcap_{k\in\mathcal{S}^c}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 < 1\}) \underset{T\to\infty}{\longrightarrow} 1 \Leftrightarrow \mathbb{P}(\bigcup_{k\in\mathcal{S}^c}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 \geq 1\}) \underset{T\to\infty}{\longrightarrow} 0. \tag{8}$$

This is equivalent to proving

$$\mathbb{P}(\bigcup_{k\in\mathcal{S}^c}\{\|\dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}})_{(k)} + \frac{\lambda_T}{T}\alpha_T^{(k)} \odot \hat{\boldsymbol{w}}^{(k)}\|_2 \geq \frac{\gamma_T}{T}\xi_{T,k}\}) \underset{T\to\infty}{\longrightarrow} 0.$$

We have for $k \in \mathcal{S}^c$ that $\|\hat{\boldsymbol{w}}^{(k)}\|_\infty \leq 1$, which implies by the optimality conditions of Karush-Kuhn-Tucker that

$$\begin{aligned}
&\mathbb{P}(\bigcup_{k\in\mathcal{S}^c}\{\|\dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}})_{(k)} + \tfrac{\lambda_T}{T}\alpha_T^{(k)} \odot \hat{\boldsymbol{w}}^{(k)}\|_2 \geq \tfrac{\gamma_T}{T}\xi_{T,k}\}) \\
&\leq \mathbb{P}(\bigcup_{k\in\mathcal{S}^c}\{\|\dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}})_{(k)}\|_2 \geq \tfrac{\gamma_T}{T}\xi_{T,k} - \tfrac{\lambda_T}{T}\|\alpha_T^{(k)}\|_2\}).
\end{aligned}$$

By a Taylor expansion around $\boldsymbol{\theta}_0$, let $\bar{\boldsymbol{\theta}}$ such that $\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \leq \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|$, we have

$$\begin{aligned}
&\mathbb{P}(\bigcup_{k\in\mathcal{S}^c}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 \geq 1\}) \leq \mathbb{P}(\bigcup_{k\in\mathcal{S}^c}\{\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k)}\|_2 \geq \frac{\gamma_T}{T}\xi_{T,k} - \frac{\lambda_T}{T}\|\alpha_T^{(k)}\|_2 \\
&- \|\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k)(k)}\|_2\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 - \|\nabla'\{\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{(k)(k)}\}_{(k)}\|_2\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2^2\}) \\
&\leq \mathbb{P}(\bigcup_{k\in\mathcal{S}^c}\{\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k)}\|_2 \geq \frac{\gamma_T}{T}\|\tilde{\bar{\boldsymbol{\theta}}}^{(k)}\|_2^{-\mu} - \frac{\lambda_T}{T}d_T^{1/2}\max_{k\in\mathcal{S}^c,i\in\mathcal{G}_k}(|\tilde{\bar{\theta}}_i^{(k)}|^{-\eta}) \\
&- \|\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k)(k)}\|_2\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 - \|\nabla'\{\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{(k)(k)}\}_{(k)}\|_2\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2^2\}),
\end{aligned}$$

where we used $\|\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k)(k)}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|_2 \leq \|\ddot{\mathbb{G}}_T l(\boldsymbol{\theta})_{(k)(k)}\|_2\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2$. Let $\boldsymbol{\epsilon} > 0$, and $K_{\boldsymbol{\epsilon}}$ strictly positive constants, we proved for T large enough that $\mathbb{P}(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 > K_{\boldsymbol{\epsilon}}(d_T/T)^{1/2}) < \boldsymbol{\epsilon}/6$. We deduce that

$$\mathbb{P}(\underset{k\in\mathcal{S}^c}{\cup}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 \geq 1\}) \leq \mathbb{P}(\underset{k\in\mathcal{S}^c}{\cup}\{\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k)}\|_2 \geq \frac{\gamma_T}{T}\|\tilde{\tilde{\boldsymbol{\theta}}}^{(k)}\|_2^{-\mu}$$
$$- \frac{\lambda_T}{T}d_T^{1/2}\max_{k\in\mathcal{S}^c, i\in\mathcal{G}_k}(|\tilde{\tilde{\theta}}_i^{(k)}|^{-\eta}) - \|\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k)(k)}\|_2(d_T/T)^{1/2}K_{\boldsymbol{\epsilon}}$$
$$- \|\nabla'\{\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{(k)(k)}\}_{(k)}\|_2(\frac{d_T}{T})^2 K_{\boldsymbol{\epsilon}}^2\}) + \boldsymbol{\epsilon}/6.$$

Let $M_{1,T} = (\frac{\gamma_T}{T})^{\frac{1}{1+\mu}}$, then

$$\mathbb{P}(\underset{k\in\mathcal{S}^c}{\cup}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 \geq 1\}) \leq \boldsymbol{\epsilon}/6 + \sum_{k\in\mathcal{S}^c}\{\mathbb{P}(\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k)}\|_2 \geq \frac{\gamma_T}{T}\|\tilde{\tilde{\boldsymbol{\theta}}}^{(k)}\|_2^{-\mu}$$
$$- \frac{\lambda_T}{T}d_T^{1/2}\max_{k\in\mathcal{S}^c, i\in\mathcal{G}_k}(|\tilde{\tilde{\theta}}_i^{(k)}|^{-\eta}) - \|\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k)(k)}\|_2(d_T/T)^{\frac{1}{2}}K_{\boldsymbol{\epsilon}}$$
$$- \|\nabla'\{\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{(k)(k)}\}_{(k)}\|_2(\frac{d_T}{T})^2 K_{\boldsymbol{\epsilon}}^2, \|\tilde{\tilde{\boldsymbol{\theta}}}^{(k)}\|_2 \leq M_{1,T}) + \mathbb{P}(\|\tilde{\tilde{\boldsymbol{\theta}}}^{(k)}\|_2 > M_{1,T})\}.$$

Consequently, we have the relationship

$$\mathbb{P}(\underset{k\in\mathcal{S}^c}{\cup}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 \geq 1\}) \leq \sum_{k\in\mathcal{S}^c}\{\mathbb{P}(\|\tilde{\tilde{\boldsymbol{\theta}}}^{(k)}\|_2 > M_{1,T}) + \mathbb{P}(\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k)}\|_2 \geq \frac{\gamma_T}{T}M_{1,T}^{-\mu}/4)$$
$$+ \mathbb{P}(\frac{\lambda_T}{T}d_T^{1/2}\max_{k\in\mathcal{S}^c, i\in\mathcal{G}_k}(|\tilde{\tilde{\theta}}_i^{(k)}|^{-\eta}) > \frac{\gamma_T}{T}M_{1,T}^{-\mu}/4)$$
$$+ \mathbb{P}(\|\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k)(k)}\|_2(d_T/T)^{1/2}K_{\boldsymbol{\epsilon}} > \frac{\gamma_T}{T}M_{1,T}^{-\mu}/4)$$
$$+ \mathbb{P}(\|\nabla'\{\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{(k)(k)}\}_{(k)}\|_2(\frac{d_T}{T})^2 K_{\boldsymbol{\epsilon}}^2 > \frac{\gamma_T}{T}M_{1,T}^{-\mu}/4)\} + \boldsymbol{\epsilon}/6 := \sum_{i=1}^5 T_i + \boldsymbol{\epsilon}/6.$$

We then focus on each T_i. We have by the Markov inequality

$$T_1 := \sum_{k\in\mathcal{S}^c}\mathbb{P}(\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k)}\|_2 > \frac{\gamma_T}{T}M_{1,T}^{-\mu}/4) \leq \sum_{k\in\mathcal{S}^c}\frac{16\mathbb{E}[\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k)}\|_2^2]}{\{\frac{\gamma_T}{T}M_{1,T}^{-\mu}\}^2}$$
$$\leq \frac{16\mathbb{E}[\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\|_2^2]}{\{\frac{\gamma_T}{T}M_{1,T}^{-\mu}\}^2} \leq \frac{16d_T}{T\{\frac{\gamma_T}{T}M_{1,T}^{-\mu}\}^2} = O((\frac{\gamma_T}{\sqrt{T}}T^{\frac{1}{2}[(1+\mu)(1-c)-1]})^{-\frac{2}{1+\mu}}).$$

Furthermore, using $|\tilde{\tilde{\theta}}_i^{(k)}|^{-\eta} \leq T^{\kappa\eta}$, we have for T_2 that

$$\mathbb{P}(\frac{\lambda_T}{T}d_T^{1/2}\max_{k\in\mathcal{S}^c, i\in\mathcal{G}_k}(|\tilde{\tilde{\theta}}_i^{(k)}|^{-\eta}) > \frac{\gamma_T}{T}M_{1,T}^{-\mu}/4)$$
$$\leq \mathbb{P}(\frac{\lambda_T}{T}d_T^{1/2}T^{\kappa\eta} > \frac{\gamma_T}{T}M_{1,T}^{-\mu}/4)$$
$$\leq \mathbb{P}(\frac{\gamma_T}{T}M_{1,T}^{-\mu}/4\{1 - 4\lambda_T\gamma_T^{-1}d_T^{1/2}M_{1,T}^{\mu}T^{\kappa\eta}\} < 0). \quad (9)$$

The quantity of interest is $\gamma_T \lambda_T^{-1} d_T^{-1/2} M_{1,T}^{-\mu} T^{-\kappa\eta}$ that has to converge to ∞ so that (9) converges to zero for T sufficiently large enough. We have

$$\gamma_T \lambda_T^{-1} d_T^{-1/2} M_{1,T}^{-\mu} T^{-\kappa\eta} \to \infty \Leftrightarrow \frac{\gamma_T}{\lambda_T^{1+\mu}} d_T^{-\frac{1+\mu}{2}} T^{-\kappa\eta(1+\mu)+\mu} \to \infty.$$

As for T_3, we have by the Markov inequality

$$\begin{aligned}
T_3 &:= \sum_{k \in \mathcal{S}^c} \mathbb{P}((\|\mathbb{H}_{T,(k)(k)}\|_2 + \|\mathcal{R}_{T,(k)}(\boldsymbol{\theta}_0)\|_2)(d_T/T)^{1/2} K_{\boldsymbol{\epsilon}} > \frac{\gamma_T}{T} M_{1,T}^{-\mu}/4) \\
&\leq \sum_{k \in \mathcal{S}^c} \mathbb{P}((\|\mathcal{R}_{T,(k)}(\boldsymbol{\theta}_0)\|_2 (d_T/T)^{1/2} K_{\boldsymbol{\epsilon}} > \frac{\gamma_T}{T} M_{1,T}^{-\mu}/4 - \|\mathbb{H}_{T,(k),(k)}\|_2 (d_T/T)^{1/2} K_{\boldsymbol{\epsilon}}) \\
&\leq \sum_{k \in \mathcal{S}^c} \{\mathbb{P}((\|\mathcal{R}_{T,(k)}(\boldsymbol{\theta}_0)\|_2 (d_T/T)^{1/2} K_{\boldsymbol{\epsilon}} > \frac{\gamma_T}{T} M_{1,T}^{-\mu}/8) \\
&+ \mathbb{P}(\|\mathbb{H}_{T,(k),(k)}\|_2 (d_T/T)^{1/2} K_{\boldsymbol{\epsilon}} > \frac{\gamma_T}{T} M_{1,T}^{-\mu}/8)\} \\
&\leq \sum_{k \in \mathcal{S}^c} \{\frac{64 K_{\boldsymbol{\epsilon}}^2 d_T \mathbb{E}[\|\mathcal{R}_{T,(k)}(\boldsymbol{\theta}_0)\|_2^2]}{T \gamma_T^2 T^{-2} M_{1,T}^{-2\mu}} + \frac{64 K_{\boldsymbol{\epsilon}}^2 d_T \|\mathbb{H}_{T,(k)(k)}\|_2^2}{T \gamma_T^2 T^{-2} M_{1,T}^{-2\mu}}\} \\
&\leq \frac{64 K_{\boldsymbol{\epsilon}}^2 d_T \|\mathbb{H}_T\|_2^2}{\gamma_T^2 T^{-1} M_{1,T}^{-2\mu}} + \frac{64 K_{\boldsymbol{\epsilon}}^2 \mathbb{E}[\|\mathcal{R}_T(\boldsymbol{\theta}_0)\|_2^2]}{\gamma_T^2 M_{1,T}^{-2\mu}} \leq \frac{64 K_{\boldsymbol{\epsilon}}^2 d_T \lambda_{\max}^2(\mathbb{H}_T)}{\gamma_T^2 T^{-1} M_{1,T}^{-2\mu}} + \frac{64 K_{\boldsymbol{\epsilon}}^2 d_T^3}{\gamma_T^2 M_{1,T}^{-2\mu}} \\
&\leq \frac{64 K_{\boldsymbol{\epsilon}}^2 \lambda_{\max}^2(\mathbb{H}_T)}{\{\gamma_T T^{-1/2} d_T^{-1/2} M_{1,T}^{-\mu}\}^2} + \frac{64 K_{\boldsymbol{\epsilon}}^2}{\{\gamma_T d_T^{-3/2} M_{1,T}^{-\mu}\}^2} \\
&= O((\frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(1-c)-1]})^{-\frac{2}{1+\mu}}) + O((\frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(2-3c)-1]})^{-\frac{2}{1+\mu}}).
\end{aligned}$$

We obtain for T_4 by the Markov inequality

$$\begin{aligned}
T_4 &:= \sum_{k \in \mathcal{S}^c} \mathbb{P}(\|\nabla' \{\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{(k)(k)}\}_{(k)}\|_2 (\frac{d_T}{T})^2 K_{\boldsymbol{\epsilon}}^2 > \frac{\gamma_T}{T} M_{1,T}^{-\mu}/4) \\
&\leq \sum_{k \in \mathcal{S}^c} \frac{16 K_{\boldsymbol{\epsilon}}^4 d_T^2 \mathbb{E}[\|\nabla' \{\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{(k)(k)}\}_{(k)}\|_2^2]}{T^2 \gamma_T T^{-2} M_{1,T}^{-2\mu}} \leq \frac{16 K_{\boldsymbol{\epsilon}}^4 d_T^5 \mathbb{E}[\|\nabla' \{\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})\}\|_2^2]}{\gamma_T^2 M_{1,T}^{-2\mu}} \\
&\leq \frac{16 K_{\boldsymbol{\epsilon}}^4 d_T^5 \eta(K_{\boldsymbol{\epsilon}})}{\gamma_T^2 M_{1,T}^{-2\mu}} = \frac{16 K_{\boldsymbol{\epsilon}}^4 \eta(K_{\boldsymbol{\epsilon}})}{\{\gamma_T d_T^{-5/2} M_{1,T}^{-\mu}\}^2} = O((\frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(2-5c)-1]})^{-\frac{2}{1+\mu}}).
\end{aligned}$$

Finally, we have for T_5 that

$$\begin{aligned}
T_5 &:= \sum_{k \in \mathcal{S}^c} \mathbb{P}(\|\tilde{\tilde{\boldsymbol{\theta}}}^{(k)}\|_2 > M_{1,T}) \leq \sum_{k \in \mathcal{S}^c} \frac{\mathbb{E}[\|\tilde{\tilde{\boldsymbol{\theta}}}^{(k)}\|_2^2]}{M_{1,T}^2} \\
&\leq \frac{\mathbb{E}[\|\tilde{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}_0\|_2^2]}{M_{1,T}^2} = O((\frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(1-c)-1]})^{-\frac{2}{1+\mu}}).
\end{aligned}$$

Hence we obtain from these relationships and using assumption 13

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

so that the latter implies

$$\frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(2-3c)-1]} \underset{T\to\infty}{\longrightarrow} \infty, \quad \frac{\gamma_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\mu)(2-5c)-1]} \underset{T\to\infty}{\longrightarrow} \infty.$$

Consequently each T_i converges to zero for T large enough. Hence

$$\mathbb{P}(\underset{k\in\mathcal{S}^c}{\cup}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 \geq 1\}) \leq \sum_{i=1}^{5} T_i + \boldsymbol{\epsilon}/6 \underset{T\to\infty}{\longrightarrow} \boldsymbol{\epsilon}.$$

For $\boldsymbol{\epsilon} \to 0$, we proved $\mathbb{P}(\underset{k\in\mathcal{S}^c}{\cup}\{\|\hat{\boldsymbol{z}}^{(k)}\|_2 \geq 1\}) \to 0$ for T large enough.

As for the second part of the model selection procedure, we prove that

$$\mathbb{P}(\underset{k\in\mathcal{S}}{\cap}\underset{i\in\mathcal{A}_k^c}{\cap}\{|\hat{\boldsymbol{w}}_i^{(k)}| < 1\}) \underset{T\to\infty}{\longrightarrow} 1 \Leftrightarrow \mathbb{P}(\underset{k\in\mathcal{S}}{\cup}\underset{i\in\mathcal{A}_k^c}{\cup}\{|\hat{\boldsymbol{w}}_i^{(k)}| \geq 1\}) \underset{T\to\infty}{\longrightarrow} 0.$$

By the optimality conditions, we have

$$\mathbb{P}(\underset{k\in\mathcal{S}}{\cup}\underset{i\in\mathcal{A}_k^c}{\cup}\{|\hat{\boldsymbol{w}}_i^{(k)}| \geq 1\}) = \mathbb{P}(\underset{k\in\mathcal{S}}{\cup}\underset{i\in\mathcal{A}_k^c}{\cup}\{|\dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}})_{(k),i}| \geq \frac{\lambda_T}{T}\alpha_{T,i}^{(k)}\}).$$

Then by a Taylor expansion around $\boldsymbol{\theta}_0$, with $\bar{\boldsymbol{\theta}}$ between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, we have

$$\begin{aligned}
&\mathbb{P}(\underset{k\in\mathcal{S}}{\cup}\underset{i\in\mathcal{A}_k^c}{\cup}\{|\hat{\boldsymbol{w}}_i^{(k)}| \geq 1\}) = \mathbb{P}(\underset{k\in\mathcal{S}}{\cup}\underset{i\in\mathcal{A}_k^c}{\cup}\{|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k),i} + [\sum_j \partial_{ij}^2 \mathbb{G}_T l(\boldsymbol{\theta}_0)(\hat{\theta}_j - \theta_{0,j})]_i \\
&+ [\sum_{j,k}\frac{1}{T}\sum_{t=1}^{T}\partial_{ijk}^3 l(\boldsymbol{\epsilon}_t;\bar{\boldsymbol{\theta}})(\hat{\theta}_j - \theta_{0,j})^2/2]_i| \geq \frac{\lambda_T}{T}\alpha_{T,i}^{(k)}\}) \\
&\leq \mathbb{P}(\underset{k\in\mathcal{S}}{\cup}\underset{i\in\mathcal{A}_k^c}{\cup}\{|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k),i}| \geq \frac{\lambda_T}{T}\alpha_{T,i}^{(k)} - [\sum_j \partial_{ij}^2 \mathbb{G}_T l(\boldsymbol{\theta}_0)(\hat{\theta}_j - \theta_{0,j})]_i \\
&- [\sum_{j,k} T^{-1}\sum_{t=1}^{T}\partial_{ijk}^3 l(\boldsymbol{\epsilon}_t;\bar{\boldsymbol{\theta}})(\hat{\theta}_j - \theta_{0,j})^2/2]_i|\}).
\end{aligned}$$

Let $M_{2,T} = (\frac{\lambda_T}{T})^{\frac{1}{1+\eta}}$. By $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = O_p((\frac{d_T}{T})^{\frac{1}{2}})$ and the Cauchy-Schwartz inequality

$$
\begin{aligned}
&\mathbb{P}(\underset{k\in\mathcal{S}}{\cup}\underset{i\in\mathcal{A}_k^c}{\cup}\{|\hat{\boldsymbol{w}}_i^{(k)}| \geq 1\}) \leq \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\{\mathbb{P}(|\tilde{\tilde{\theta}}_i^{(k)}| > M_{2,T}) + \mathbb{P}(|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k),i}| \geq \frac{\lambda_T}{T}\alpha_{T,i}^{(k)} \\
&- [\sum_j \partial_{ij}^2 \mathbb{G}_T l(\boldsymbol{\theta}_0)(\hat{\theta}_j - \theta_{0,j})]_i - [\sum_{j,k} T^{-1}\sum_{t=1}^T \partial_{ijk}^3 l(\boldsymbol{\epsilon}_t;\bar{\boldsymbol{\theta}})(\hat{\theta}_j - \theta_{0,j})^2/2]_i|, |\tilde{\tilde{\theta}}_i^{(k)}| \leq M_{2,T})\} \\
&\leq \boldsymbol{\epsilon}/5 + \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\{\mathbb{P}(|\tilde{\tilde{\theta}}_i^{(k)}| > M_{2,T}) + \mathbb{P}(|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k),i}| \geq \frac{\lambda_T}{T} M_{2,T}^{-\eta} \\
&- \{\sum_j (\partial_{ij}^2 \mathbb{G}_T l(\boldsymbol{\theta}_0))^2\}^{1/2} K_{\boldsymbol{\epsilon}} (d_T/T)^{1/2} \\
&- \{\sum_{j,k,l,m} T^{-2}\sum_{t,t'=1}^T \partial_{ijk}^3 l(\boldsymbol{\epsilon}_t;\bar{\boldsymbol{\theta}})\partial_{ilm}^3 l(\boldsymbol{\epsilon}_{t'};\bar{\boldsymbol{\theta}})\}^{1/2} K_{\boldsymbol{\epsilon}}^2 (d_T/T))\} \\
&\leq \boldsymbol{\epsilon}/5 + \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\{\mathbb{P}(|\tilde{\tilde{\theta}}_i^{(k)}| > M_{2,T}) + \mathbb{P}(|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k),i}| \geq \frac{\lambda_T}{T} M_{2,T}^{-\eta}/3) \\
&+ \mathbb{P}(\{\sum_j (\partial_{ij}^2 \mathbb{G}_T l(\boldsymbol{\theta}_0))^2\}^{1/2} K_{\boldsymbol{\epsilon}} (d_T/T)^{1/2} > \frac{\lambda_T}{T} M_{2,T}^{-\eta}/3) \\
&+ \mathbb{P}(\{\sum_{j,k,l,m} T^{-2}\sum_{t,t'=1}^T \partial_{ijk}^3 l(\boldsymbol{\epsilon}_t;\bar{\boldsymbol{\theta}})\partial_{ilm}^3 l(\boldsymbol{\epsilon}_{t'};\bar{\boldsymbol{\theta}})\}^{1/2} K_{\boldsymbol{\epsilon}}^2 (d_T/T) > \frac{\lambda_T}{T} M_{2,T}^{-\eta}/3)\} \\
&:= \sum_{i=1}^4 T_i + \boldsymbol{\epsilon}/5.
\end{aligned}
$$

We proceed as for inactive groups. For T_1, we have by the Markov inequality

$$
\begin{aligned}
T_1 &:= \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\mathbb{P}(|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k),i}| \geq \frac{\lambda_T}{T} M_{2,T}^{-\eta}/3) \\
&\leq \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c} \frac{9\mathbb{E}[|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{(k),i}|^2]}{\{\frac{\lambda_T}{T} M_{2,T}^{-\eta}\}^2} \leq \frac{9\mathbb{E}[\|\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)\|_2^2]}{\{\frac{\lambda_T}{T} M_{2,T}^{-\eta}\}^2} \\
&= O((\frac{\lambda_T}{\sqrt{T}} T^{\frac{1}{2}[(1+\eta)(1-c)-1]})^{-\frac{2}{1+\eta}}).
\end{aligned}
$$

As for T_2, we have

$$
\begin{aligned}
T_2 &:= \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\mathbb{P}(\{\sum_j(\partial^2_{ij}\mathbb{G}_T l(\boldsymbol{\theta}_0))^2\}^{1/2}K_{\boldsymbol{\epsilon}}(d_T/T)^{1/2} > \tfrac{\lambda_T}{T}M_{2,T}^{-\eta}/3)\\
&= \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\mathbb{P}((\{\sum_j\mathcal{P}_{T,(k),j}(\boldsymbol{\theta}_0)\}^{1/2} + \{\sum_j\mathbb{H}^2_{T,(k),j}\}^{1/2})K_{\boldsymbol{\epsilon}}(d_T/T)^{1/2} > \tfrac{\lambda_T}{T}M_{2,T}^{-\eta}/3)\\
&\leq \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\sum_j\{\frac{36d_T\mathbb{E}[\mathcal{P}^2_{T,(k),j}(\boldsymbol{\theta}_0)]}{T\{\frac{\lambda_T}{T}M_{2,T}^{-\eta}\}^2}\} + \frac{36d_T\|\mathbb{H}_T\|_2^2}{T\{\frac{\lambda_T}{T}M_{2,T}^{-\eta}\}^2}\\
&\leq \frac{36d_T\lambda^2_{\max}(\mathbb{H}_T)}{T\{\frac{\lambda_T}{T}M_{2,T}^{-\eta}\}^2} + \frac{36d_T\mathbb{E}[\|\mathcal{P}_T(\boldsymbol{\theta}_0)\|_2^2]}{T\{\frac{\lambda_T}{T}M_{2,T}^{-\eta}\}^2}\\
&= O((\frac{\lambda_T}{\sqrt{T}}T^{\frac{1}{2}[(1+\eta)(1-c)-1]})^{-\frac{2}{1+\eta}}) + O((\frac{\lambda_T}{\sqrt{T}}T^{\frac{1}{2}[(1+\eta)(2-3c)-1]})^{-\frac{2}{1+\eta}}).
\end{aligned}
$$

Furthermore, for the third order term in T_3, we have

$$
\begin{aligned}
T_3 &:= \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\mathbb{P}(\{\sum_{j,k,l,m}T^{-2}\sum_{t,t'=1}^{T}\partial^3_{ijk}l(\boldsymbol{\epsilon}_t;\bar{\boldsymbol{\theta}})\partial^3_{ilm}l(\boldsymbol{\epsilon}_{t'};\bar{\boldsymbol{\theta}})\}^{1/2}K^2_{\boldsymbol{\epsilon}}(d_T/T) > \tfrac{\lambda_T}{T}M_{2,T}^{-\eta}/3)\\
&\leq \frac{9d_T^2\mathbb{E}[\|\nabla'\{\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}}\}\|_2^2]}{T^2\{\frac{\lambda_T}{T}M_{2,T}^{-\eta}\}^2} = O((\frac{\lambda_T}{\sqrt{T}}T^{\frac{1}{2}[(1+\eta)(2-5c)-1]})^{-\frac{2}{1+\eta}}).
\end{aligned}
$$

Finally, we have for T_4 that

$$
\begin{aligned}
T_4 &:= \sum_{i\in\mathcal{A}_k^c}\mathbb{P}(|\tilde{\tilde{\theta}}_i^{(k)}| > M_{2,T}) \leq \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{A}_k^c}\frac{\mathbb{E}[|\tilde{\tilde{\theta}}_i^{(k)}|^2]}{M^2_{2,T}}\\
&\leq \frac{\mathbb{E}[\|\tilde{\tilde{\boldsymbol{\theta}}}-\boldsymbol{\theta}_0\|_2^2]}{M^2_{2,T}} = O((\frac{\lambda_T}{\sqrt{T}}T^{\frac{1}{2}[(1+\eta)(1-c)-1]})^{-\frac{2}{1+\eta}}).
\end{aligned}
$$

From these relationships and assumption 13, $\frac{\lambda_T}{\sqrt{T}}T^{\frac{1}{2}[(1+\eta)(1-c)-1]} \underset{T\to\infty}{\longrightarrow} \infty$ implying

$$
\frac{\lambda_T}{\sqrt{T}}T^{\frac{1}{2}[(1+\eta)(2-3c)-1]} \underset{T\to\infty}{\longrightarrow} \infty, \quad \frac{\lambda_T}{\sqrt{T}}T^{\frac{1}{2}[(1+\eta)(2-5c)-1]} \underset{T\to\infty}{\longrightarrow} \infty.
$$

We deduce $\mathbb{P}(\underset{k\in\mathcal{S}}{\cup}\underset{i\in\mathcal{A}_k^c}{\cup}\{|\hat{\boldsymbol{w}}_i^{(k)}|\geq 1\}) \underset{T\to\infty}{\longrightarrow} \boldsymbol{\epsilon}$. We have then concluded the model selection consistency.

We now focus on the asymptotic normality. Model selection implies that

$$
\mathbb{P}(\{k\in\mathcal{S}, i\in\mathcal{A}_k, : \hat{\theta}_i^{(k)} \neq 0\} = \mathcal{A}) \underset{T\to\infty}{\longrightarrow} 1.
$$

As a consequence, the next relationship holds

$$
\mathbb{P}(\forall k\in\mathcal{S}, \dot{\mathbb{G}}_T l(\hat{\boldsymbol{\theta}})_{\mathcal{A}_k} + \frac{\lambda_T}{T}\alpha_{T,\mathcal{A}_k}\odot\operatorname{sgn}(\hat{\boldsymbol{\theta}}_{\mathcal{A}_k}) + \frac{\gamma_T}{T}\xi_{T,k}\frac{\hat{\boldsymbol{\theta}}_{\mathcal{A}_k}}{\|\hat{\boldsymbol{\theta}}_{\mathcal{A}_k}\|_2} = 0) \underset{T\to\infty}{\longrightarrow} 1.
$$

By a Taylor expansion of the gradient term around $\boldsymbol{\theta}_{0,\mathcal{A}}$, we obtain

$$\mathbb{P}(\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{\mathcal{A}} + \ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}) + \frac{1}{2}\nabla'\{(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})\} + \tfrac{\lambda_T}{T}\alpha_{T,\mathcal{A}} \odot \operatorname{sgn}(\hat{\boldsymbol{\theta}}_{\mathcal{A}}) + \tfrac{\gamma_T}{T}\eta_T = 0) \underset{T\to\infty}{\longrightarrow} 1,$$

where $\eta_T = \operatorname{vec}(\xi_{T,k}\frac{\hat{\boldsymbol{\theta}}_{\mathcal{A}_k}}{\|\hat{\boldsymbol{\theta}}_{\mathcal{A}_k}\|_2}, k \in \mathcal{S})$ and $\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 \leq \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2$. As a consequence, we have

$$\mathcal{P}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}) + \mathbb{H}_{T,\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}) = -\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{\mathcal{A}}$$
$$- \frac{1}{2}\nabla'\{(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})\} - \frac{\lambda_T}{T}\alpha_{T,\mathcal{A}} \odot \operatorname{sgn}(\hat{\boldsymbol{\theta}}_{\mathcal{A}}) - \frac{\gamma_T}{T}\eta_T + o_p(1),$$

where $\mathcal{P}(\boldsymbol{\theta}_0) = \ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{\mathcal{A}\mathcal{A}} - \mathbb{H}_{T,\mathcal{A}\mathcal{A}}$ and $\mathbb{H}_{T,\mathcal{A}\mathcal{A}} = \mathbb{E}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}'} l(\boldsymbol{\epsilon}_t; \boldsymbol{\theta}_0)]_{\mathcal{A}\mathcal{A}}$. Then multiplying by $\sqrt{T}Q_T\mathbb{V}^{-1/2}_{T,\mathcal{A}\mathcal{A}}$, we obtain

$$\sqrt{T}Q_T\mathbb{V}^{-1/2}_{T,\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}) = -\sqrt{T}Q_T\mathbb{V}^{-1/2}_{T,\mathcal{A}\mathcal{A}}\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}(\frac{\lambda_T}{T}\alpha_{T,\mathcal{A}} \odot \operatorname{sgn}(\hat{\boldsymbol{\theta}}_{\mathcal{A}}) + \frac{\gamma_T}{T}\eta_T)$$
$$- \sqrt{T}Q_T\mathbb{V}^{-1/2}_{T,\mathcal{A}\mathcal{A}}\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{\mathcal{A}} - \sqrt{T}Q_T\mathbb{V}^{-1/2}_{T,\mathcal{A}\mathcal{A}}\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\mathcal{P}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})$$
$$- \sqrt{T}/2Q_T\mathbb{V}^{-1/2}_{T,\mathcal{A}\mathcal{A}}\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\nabla'\{(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})\} + o_p(1).$$

We focus on the l^1 penalty term, which can be upper bounded as

$$\begin{aligned}
&N_{1,T} := |\sqrt{T}Q_T\mathbb{V}^{-1/2}_{T,\mathcal{A}\mathcal{A}}\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}(\frac{\lambda_T}{T}\alpha_{T,\mathcal{A}} \odot \operatorname{sgn}(\hat{\boldsymbol{\theta}}_{\mathcal{A}}))| \\
&\leq |Q_T\mathbb{V}^{-1/2}_{T,\mathcal{A}\mathcal{A}}||\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}|\lambda_T T^{-1/2}\max_{k\in\mathcal{S}, i\in\mathcal{A}_k}\alpha_{T,i,\mathcal{A}} \\
&\leq |Q_T\mathbb{V}^{-1/2}_{T,\mathcal{A}\mathcal{A}}|\lambda^{-1}_{\min}(\mathbb{H}_{T,\mathcal{A}\mathcal{A}})\lambda_T T^{-1/2}\{\min_{k\in\mathcal{S}, i\in\mathcal{A}_k}|\tilde{\tilde{\theta}}^{(k)}_i|\}^{-\eta} \\
&\leq |Q_T\mathbb{V}^{-1/2}_{T,\mathcal{A}\mathcal{A}}|\lambda^{-1}_{\min}(\mathbb{H}_{T,\mathcal{A}\mathcal{A}})\lambda_T T^{\kappa\eta-\frac{1}{2}}.
\end{aligned}$$

If $\lambda_T T^{\kappa\eta} \to 0$, then $N_{1,T} = o_p(1)$.

As for the l^1/l^2 penalty, it can be upper bounded as

$$\begin{aligned}
&N_{2,T} := |\sqrt{T}Q_T\mathbb{V}^{-1/2}_{T,\mathcal{A}\mathcal{A}}\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}\frac{\gamma_T}{T}\eta_T| \\
&\leq |Q_T\mathbb{V}^{-1/2}_{T,\mathcal{A}\mathcal{A}}||\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}|\gamma_T T^{-1/2}\|\eta_T\|_2 \\
&\leq |Q_T\mathbb{V}^{-1/2}_{T,\mathcal{A}\mathcal{A}}||\mathbb{H}^{-1}_{T,\mathcal{A}\mathcal{A}}|\gamma_T T^{-1/2}\sqrt{\sum_{k\in\mathcal{S}}\|\tilde{\tilde{\boldsymbol{\theta}}}^{(k)}\|_2^{-2\mu}} \\
&\leq |Q_T\mathbb{V}^{-1/2}_{T,\mathcal{A}\mathcal{A}}|\lambda^{-1}_{\min}(\mathbb{H}_{T,\mathcal{A}\mathcal{A}})\gamma_T T^{-1/2}d_T^{1/2}\{\min_{k\in\mathcal{S}}\|\tilde{\tilde{\boldsymbol{\theta}}}^{(k)}\|_2\}^{-\mu} \\
&\leq |Q_T\mathbb{V}^{-1/2}_{T,\mathcal{A}\mathcal{A}}|\lambda^{-1}_{\min}(\mathbb{H}_{T,\mathcal{A}\mathcal{A}})\gamma_T T^{-1/2}d_T^{1/2}T^{\kappa\mu}.
\end{aligned}$$

Using $d_T = O(T^c)$, if $\gamma_T T^{\frac{c-1}{2}+\kappa\mu} \to 0$, then $N_{2,T} = o_p(1)$. Consequently, we have $N_{1,T} + N_{2,T} = o_p(1)$.

We now turn to the hessian quantity of the Taylor expansion and prove that the discrepancy $\mathcal{P}(\boldsymbol{\theta}_0)$ converges uniformly to zero in probability. For any $\boldsymbol{\epsilon} > 0$, by the Markov's inequality, we have

$$\begin{aligned}
&\mathbb{P}(\|\ddot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{\mathcal{A}\mathcal{A}} - \mathbb{H}_{T,\mathcal{A}\mathcal{A}}\|_2^2 > (\boldsymbol{\epsilon}/d_T)^2) \\
&\leq \frac{d_T^2}{\boldsymbol{\epsilon}^2 T^2}\mathbb{E}[\sum_{(k,l)\in\mathcal{A}} \{\partial^2_{\theta_k\theta_l} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0) - \mathbb{E}[\nabla^2_{\theta_k\theta_l} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)]\}^2] \\
&\leq \frac{d_T^4}{\boldsymbol{\epsilon}^2 T^2}\lambda^2_{\max}(\mathbb{H}_{T,\mathcal{A}\mathcal{A}}).
\end{aligned}$$

As for the third order term, by the Cauchy-Schwartz inequality

$$\begin{aligned}
&\|\nabla'\{(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})'\ddot{\mathbb{G}}_T l(\bar{\boldsymbol{\theta}})_{\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})\}\|_2^2 \\
&\leq \frac{1}{T^2}\sum_{t=1}^{T}\{\sum_{(k,l,m)\in\mathcal{A}} \partial^3_{\theta_k\theta_l\theta_m} l_T^2(\boldsymbol{\epsilon}_t;\bar{\boldsymbol{\theta}})\}\|\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}\|_2^4 \\
&\leq \frac{1}{T^2}\sum_{t=1}^{T}\{\sum_{(k,l,m)\in\mathcal{A}} \psi_T^2(\boldsymbol{\epsilon}_t)\}\|\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}\|_2^4 = O_p(\frac{d_T^5}{T^2}) = o_p(\frac{1}{T}).
\end{aligned}$$

We now prove $X_{T,t} = \sqrt{T}Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}\mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1}\dot{\mathbb{G}}_T l_t(\boldsymbol{\theta}_0)_{\mathcal{A}}, t = 1,\cdots,T$, is asymptotically normal by checking the Lindeberg-Feller's condition for applying Theorem 10 of Shiryaev. We remind that $\dot{\mathbb{G}}_T l_{T,t}(\boldsymbol{\theta}_0)$ is the t-th point of the score of the empirical criterion. Let $\beta > 0$, and to the Theorem of Shiryaev, we need to prove that for any $\boldsymbol{\epsilon} > 0$, we have

$$\mathbb{P}(\sum_{t=0}^{T}\mathbb{E}[\|X_{T,t}\|_2^2\mathbf{1}_{\|X_{T,t}\|_2>\beta}|\mathcal{F}_{t-1}^T] > \boldsymbol{\epsilon}) \underset{T\to\infty}{\longrightarrow} 0.$$

By the Markov inequality, we obtain

$$\begin{aligned}
&\mathbb{P}(\sum_{t=0}^{T}\mathbb{E}[\|X_{T,t}\|_2^2\mathbf{1}_{\|X_{T,t}\|_2>\beta}|\mathcal{F}_{t-1}^T] > \boldsymbol{\epsilon}) \leq \frac{1}{\boldsymbol{\epsilon}}\sum_{t=0}^{T}\mathbb{E}[\|X_{T,t}\|_2^2\mathbf{1}_{\|X_{T,t}\|_2>\beta}|\mathcal{F}_{t-1}^T] \\
&\leq \frac{1}{\boldsymbol{\epsilon}}\sum_{t=0}^{T}\mathbb{E}[\mathbb{E}[\|X_{T,t}\|_2^4|\mathcal{F}_{t-1}^T]^{1/2}\mathbb{P}(\|X_{T,t}\|_2 > \beta|\mathcal{F}_{t-1}^T)^{1/2}] \\
&\leq \frac{1}{\boldsymbol{\epsilon}}\sum_{t=0}^{T}\mathbb{E}[\{\frac{C_{st}}{T^2}\mathbb{E}[\|\nabla l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\nabla' l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\|_2^2|\mathcal{F}_{t-1}^T]\}^{1/2} \\
&\times \frac{1}{\beta}\mathbb{E}[\|\sqrt{T}Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}\mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1}\dot{\mathbb{G}}_T l_t(\boldsymbol{\theta}_0)_{\mathcal{A}}\|_2^2|\mathcal{F}_{t-1}^T]^{1/2}],
\end{aligned}$$

with $C_{st} > 0$. First, let $\mathbb{K}_T = Q_T \mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2} \mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1}$, we have

$$\begin{aligned}
&\mathbb{E}[\|\sqrt{T}\mathbb{K}_T \dot{\mathbb{G}}_T l_t(\boldsymbol{\theta}_0)_{\mathcal{A}}\|_2^2 | \mathcal{F}_{t-1}^T] = \frac{1}{T}\mathbb{E}[\nabla' l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\mathbb{K}_T'\mathbb{K}_T \nabla l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)|\mathcal{F}_{t-1}^T] \\
&= \frac{1}{T}\mathbb{E}[\text{Trace}(\nabla' l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\mathbb{K}_T'\mathbb{K}_T \nabla l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0))|\mathcal{F}_{t-1}^T] \\
&= \frac{1}{T}\text{Trace}(\mathbb{E}[\nabla l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\nabla' l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)|\mathcal{F}_{t-1}^T]\mathbb{K}_T'\mathbb{K}_T) \leq \frac{1}{T}\lambda_{\max}(\mathbb{H}_{t-1}^T)\tilde{C}_{st},
\end{aligned}$$

where $\tilde{C}_{st} > 0$. Furthermore, we have

$$\begin{aligned}
\mathbb{E}[\|\nabla l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\nabla' l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\|_2^2|\mathcal{F}_{t-1}^T] &= \mathbb{E}[\sum_{i,j=0}^{d_T}\{\partial_{\theta_i} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\partial_{\theta_j} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\}^2|\mathcal{F}_{t-1}^T] \\
&\leq d_T^2 \sup_{i,j=1,\cdots,d_T}\mathbb{E}[\{\partial_{\theta_i} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\partial_{\theta_j} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\}^2|\mathcal{F}_{t-1}^T].
\end{aligned}$$

By assumption 14, we have

$$\begin{aligned}
&\mathbb{P}(\sum_{t=0}^{T}\mathbb{E}[\|X_{T,t}\|_2^2 \mathbf{1}_{\|X_{T,t}\|_2>\beta}|\mathcal{F}_{t-1}^T] > \boldsymbol{\epsilon}) \\
&\leq \frac{C_{st}^{\frac{1}{2}}\tilde{C}_{st}^{\frac{1}{2}} d_T}{T^{\frac{3}{2}}}\sum_{t=0}^{T}\mathbb{E}[\sup_{i,j=1,\cdots,d_T}\mathbb{E}[\{\partial_{\theta_i} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\partial_{\theta_j} l(\boldsymbol{\epsilon}_t;\boldsymbol{\theta}_0)\}^2|\mathcal{F}_{t-1}^T]\lambda_{\max}(\mathbb{H}_{t-1}^T)] \\
&\leq \frac{C_{st}^{\frac{1}{2}}\tilde{C}_{st}^{\frac{1}{2}}\bar{B}Td_T}{T^{\frac{3}{2}}}.
\end{aligned}$$

Consequently, we obtain $\sum_{t=0}^{T}\mathbb{E}[\|X_{T,t}\|_2^2 \mathbf{1}_{\|X_{T,t}\|_2>\beta}|\mathcal{F}_{t-1}^T] = o_p(1)$. We deduce that $X_{T,t}$ satisfies the Lindeberg-Feller condition, and by Theorem 10, we obtain that $\sqrt{T}Q_T\mathbb{V}_{T,\mathcal{A}\mathcal{A}}^{-1/2}\mathbb{H}_{T,\mathcal{A}\mathcal{A}}^{-1}\dot{\mathbb{G}}_T l(\boldsymbol{\theta}_0)_{\mathcal{A}}$ is asymptotically normally distributed. The asymptotic distribution of Theorem 8 follows. □

4 Additional simulated experiment

This section provides a further insight in the support recovery of the adaptive Sparse Group Lasso through an additional simulated experiment on VAR models. We consider a data generating process

$$y_t = \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + u_t,$$

with $y_t = (y_{1,t}, \cdots, y_{d,t})'$, $u_t \sim \mathcal{N}_{\mathbb{R}^d}(0, \Sigma)$ such that $\Sigma = D^{\frac{1}{2}} R D^{\frac{1}{2}}$, with $R_{ij} = \rho^{|i-j|}, 1 \leq j, i \leq d$, $D = \text{diag}(\sigma_1^2, \cdots, \sigma_d^2)$, $\forall i, \sigma_i \in \mathcal{U}([0.03; 0.1])$ and $\rho \in \mathcal{U}([0.4, 0.9])$. It corresponds to a VAR(p) dynamic, with $p = 2$ such that we generate Φ_1 and Φ_2 under the usual stationarity constraints together with an ordering constraint, idest $\forall i, j$, $\Phi_{2,ij} \leq \Phi_{1,ij}$. We also set zero coefficients among Φ_1 and Φ_2. We consider $d = 5$ (resp. $d = 10$) such that the number of

zeros is 30 (resp. 105) and the number of nonzero coefficients is 20 (resp. 95). Each of these active coefficients is simulated in $\mathcal{U}([0.05, 0.9])$.

Then we estimate a VAR(p) model, with $p = 4$ and using a similar criterion as in the simulated experiment 1. The total number of estimated parameters would be $d = p \times N^2$ and the total number of zero $|\mathcal{A}|$ to recover is 80 for $d = 5$ and 305 for $d = 10$. In this setting d is not indexed by T. We define the group as the lags for the Group Lasso and the SGL procedures, which implies there are 4 groups in total, with 2 active groups.

The results reported in Table 1 clearly illustrate the ability of the adaptive SGL procedure to properly perform for variable selection. The number of zero coefficients correctly estimated is denoted as C and the number of nonzero coefficients incorrectly estimated is denoted IC. MSE stands for the mean square error. The adaptive Lasso also provide proper performance results regarding both estimation precision and variable selection.

Table 1 *Model selection and precision accuracy based on 100 replications*

d	$\|\mathcal{A}\|$	Model	MSE	C	IC
5	80	Truth		80	0
		Lasso	0.0308	44.25	0.31
		aLasso	0.0283	50.45	0.24
		GLasso	0.0170	63.03	0.08
		AGLasso	0.0175	65.34	0.07
		SGL	0.0149	53.93	0.06
		ASGL	0.0094	77.97	0.06
10	305	Truth		305	0
		Lasso	0.2827	110.30	13.45
		aLasso	0.2071	129.55	9.12
		GLasso	0.0683	236.94	3.64
		AGLasso	0.0591	256.73	3.12
		SGL	0.0529	220.52	3.05
		ASGL	0.0526	295.36	2.96