Discussion Papers In Economics And Business

Estimation of High Dimensional Vector Autoregression via Sparse Precision Matrix

Benjamin Poignard
Manabu Asai

Discussion Paper 21-03

April 2021

Graduate School of Economics
Osaka University, Toyonaka, Osaka 560-0043, JAPAN

Estimation of High Dimensional Vector Autoregression via Sparse Precision Matrix*

Benjamin Poignard†
Graduate School of Economics, Osaka University and Riken-AIP, Japan

Manabu Asai‡
Faculty of Economics, Soka University, Japan

Abstract

We consider the problem of estimating sparse structural vector autoregression (SVAR) processes via penalized precision matrix. Such matrix is the output of the underlying directed acyclic graph of the SVAR process, whose zero components correspond to zero SVAR coefficients. The precision matrix estimators are deduced from the class of Bregman divergences and regularized by the SCAD, MCP and LASSO penalties. Under suitable regularity conditions, we derive error bounds for the regularized precision matrix for each Bregman divergence. Moreover, we establish the support recovery property, including the case when the penalty is non-convex. These theoretical results are supported by empirical studies.

Keywords: sparse structural vector autoregression; statistical consistency; support recovery.

*The authors are most grateful to Yoshihisa Baba for very helpful comments and suggestions. The authors acknowledge the financial support of the Japan Society for the Promotion of Science.

†E-mail address: bpoignard@econ.osaka-u.ac.jp (corresponding author)

‡E-mail address: m-asai@soka.ac.jp

1 Introduction

The Vector AutoRegressive (VAR) process is a standard model for multivariate time series. From an inferential viewpoint, the ordinary least squares method is inapplicable once fitted to high-dimensional data due to a $O(pN^2)$ complexity order, where N is the number of variables and p the number of lags. Numerous studies modelled sparsity among the VAR parameters to tackle the over-fitting issue, which justified the use of penalized OLS losses: see, e.g., Basu and Michailidis (2015), who considered a penalized OLS loss for sparse stable Gaussian VAR processes; in the same manner, Wong, Li and Tewari (2020) derived some consistency results of penalized OLS based estimators for α-mixing Gaussian VAR processes. Rather than specifying sparse VAR matrix parameters, Alquier, Bertin, Doukhan and Garnier (2020) specified a low-rank constraint on the VAR transition matrix in the presence of factors, where the motivation is to improve the prediction accuracy.

Furthermore, the Structural VAR (SVAR) model has been used to accommodate economic theory within the VAR framework: see, e.g., Blanchard and Quah (1989) and Waggoner and Zha (2003) among others. Apart from economic theory, Tunnicliffe-Wilson and Reale (2008), Oxley, Reale, Tunnicliffe-Wilson (2009), and Ahelegbey, Billio and Casarin (2016) discussed the interpretation and identification of SVAR models in terms of graphical modelling. In this paper, we consider the structure of SVAR models and their identification based on a graphical representation, extending Sims (1980) and the previous papers as an approach which is free from the constraints usually assumed in economic theory. We propose a different procedure for the analysis of the issues related to SVAR inference in high dimension: our analysis focuses on the relationships between the SVAR parameters and the precision matrix - i.e., the inverse variance-covariance - of the SVAR process, which allows for a sparse estimation of the SVAR parameters.

More precisely, in this study, we consider the following problem: given T observations of a N-dimensional SVAR(p) process $(\boldsymbol{Y}_t)$, estimate the sparse precision matrix Θ of $\boldsymbol{X}_t = (\boldsymbol{Y}_{t-p}^\top, \cdots, \boldsymbol{Y}_t^\top)^\top$, whose components provide sparse consistent estimators of the SVAR parameters by the one-to-one matching between the

zero coefficients of the SVAR parameters and the precision matrix. Such matching between the precision matrix Θ and the SVAR parameters relies on the so-called directed acyclic graph, where the elements of $\boldsymbol{X}_t$ form the nodes and past dependence is captured by the directed edges linking the nodes. By such relationship, we can show that the sparsity on Θ gives a one-to-one correspondence to the zero coefficients of the SVAR(p) parameters. Moreover, our method ensures the stationarity of the resulting process while fostering sparsity on the SVAR coefficients. There exist several methods fostering sparsity for the precision matrix: for example, Bickel and Levina (2008a) considered a thresholding operator for the inverse, which constrains the parameters toward zero should they exceed a certain threshold. Rothman, Levina, and Zhu (2009) showed that this estimator satisfies the "sparsistency" property in the sense of Lam and Fan (2007), that is the true zero parameters are correctly identified with probability tending to one and the sparse estimator is sign consistent for nonzero elements. Motivated by the structure of some time series data, Furrer and Bengtsson (2012) focused on banded precision matrices and fostered sparsity by shrinking the off-diagonal entries based on their distances with respect to the diagonal. Wu and Pourahmadi (2003) and Huang, Liu, Pourahmadi and Liu (2006) considered a Cholesky decomposition of the precision matrix with a thresholding procedure and LASSO penalization of the Cholesky factors respectively. For the aspect of "regression" in the VAR, we may carry out the latter approach based on the relationship between regression coefficients and the precision matrix for vector of relevant variables. The Cholesky decomposition automatically ensures the positive definiteness of the estimated covariance matrix. Bickel and Levina (2008b) provided the conditions for the statistical consistency of such estimators while highlighting that a key feature of such decomposition is that the latter depend on the variable ordering and thus the sparse structure highly depends on such order. Another parsimonious approach consists in assuming a particular sparse structure on the variance covariance or the precision matrix. In particular, Bickel and Levina (2008a) applied a thresholding procedure under the sparse row assumption - the ℓ_q-sparse assumption -. Finally, a significant literature is dedicated to the parsimonious analysis of precision matrices under an element-wise sparsity assumption. Yuan and Lin (2007) or Ravikumar, Wainwright, Raskutti and Yu (2011) assumed an

element-wise sparse precision matrix and proposed a LASSO Gaussian likelihood estimator. Zhang and Zou (2014) proposed an alternative LASSO penalized loss, namely the D-trace loss, to estimate such matrix. In the same vein, Loh and Wainwright (2017) extended the penalization of the Gaussian precision matrix to non-convex penalties, which allows for relaxing the incoherence - or irrepresentability - condition.

Our setting lies within such sparse element-wise precision matrix for estimating sparse SVAR parameters. Our main contributions are as follows: first, we provide a novel estimation method for SVAR parameters under the sparsity assumption; we provide error bounds for a broard range of sparse estimators of Θ in the ℓ_1, ℓ_2 and ℓ_∞ senses for specific scaling behaviours of (T, d, k_0), where d and k_0 respectively are the dimension problem and the cardinality of the true unknown sparse support; finally, we provide the conditions to satisfy the support recovery property. The estimators of Θ are deduced from the class of penalized Bregman divergence losses, including the Gaussian, least squares and von Neumann losses, thus providing new estimators for sparse precision matrices. To the best of our knowledge, this paper is the first attempt to link general penalized - potentially non-convex - M-estimators and the Bregman divergence-based inference for sparse precision matrix. Our study shares a similar spirit to that of Ravikumar et al. (2011) and Loh and Wainwright (2017), who derived the conditions for statistical consistency of component-wise sparse Gaussian based precision matrix. But our work differs from these studies in two main respects: we provide bounds on ℓ_1-, ℓ_2- and ℓ_∞- errors and the conditions to satisfy the support recovery property within the framework of Bregman divergence, thus providing new sparse estimators for precision matrix.

The framework we use to derive such error bounds is related to the study of Poignard and Fermanian (2021), which covers a broad range of non-convex objective functions for sparse M-estimation. Assuming the restricted strong convexity (see e.g. Negahban, Ravikumar, Wainwright and Yu, 2012) of the non-penalized loss function and for suitable regularity conditions of the penalty, they derived some error bounds for the penalized estimators. To establish the conditions for support recovery, our proof techniques are inspired from the primal dual witness method of Loh and Wainwright (2017). In our study, we extend these results to a broad range of sparse precision matrix estimators. We quantify the statistical accuracy and discuss the

relevance of these theoretical bounds for each M-criterion. The scaling behaviours with respect to (T, d, k_0) that we derive for support recovery highly depend on the regularity of the Bregman divergence.

The organization of the paper is as follows. In Section 2, we describe our approach for sparse SVAR modelling based on sparse precision matrix. In Section 3, we provide our sparse estimation framework based on Bregman divergences together with the theoretical properties - error bounds and support recovery - of the corresponding estimators. Section 4 illustrates these theoretical properties through simulation and real data experiments. All intermediary results, proofs and figures are contained in Section 5.

Notations. Throughout this paper, we denote the cardinality of a set E by $|E|$. For a vector $\boldsymbol{v} \in \mathbb{R}^d$, the ℓ_p norm is $\|\boldsymbol{v}\|_p = \left(\sum_{k=1}^{\mathrm{d}} |\boldsymbol{v}_k|^p\right)^{1/p}$ for $p > 0$, and $\|\boldsymbol{v}\|_\infty = \max_i |\boldsymbol{v}_i|$. Let the subset $\mathcal{A} \subseteq \{1, \cdots, d\}$, then $\boldsymbol{v}_\mathcal{A} \in \mathbb{R}^{|\mathcal{A}|}$ is the vector $\boldsymbol{v}$ restricted to $\mathcal{A}$. We write $\mathcal{M}_{d_1 \times d_2}(\mathbb{R})$ the set of $d_1 \times d_2$-dimensional matrices with real coefficients. For a matrix A, $\|A\|_s$, $\|A\|_\infty$ and $\|A\|_F$ are the spectral, infinity and Frobenius norms, respectively, and $\|A\|_{\max} = \max_{i,j} |A_{i,j}|$ is the coordinate-wise maximum (in absolute value). We write $A^\top$ (resp. $\boldsymbol{v}^\top$) to denote the transpose of the matrix A (resp. the vector $\boldsymbol{v}$). We write $\mathrm{vec}(A)$ to denote the vectorization operator that stacks the columns of A on top of one another into a vector. We denote by $A \succ 0$ (resp. $A \succeq 0$) the positive definiteness (resp. semi-definiteness) of A and $\mathrm{vech}(A)$ the $d(d+1)/2$ vector that stacks the columns of the lower triangular part of $A \in \mathcal{M}_{d \times d}(\mathbb{R})$. $\lambda_{\min}(A)$ (resp. $\lambda_{\max}(A)$) denotes the minimum (resp. maximum) eigenvalue of A. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we denote by ∇f the gradient or subgradient of f and $\nabla^2 f$ the Hessian of f. We denote by $(\nabla^2 f)_{\mathcal{A}\mathcal{A}}$ the Hessian of f restricted to the block $\mathcal{A}$. We write $\mathcal{A}^c$ to denote the complement of the set $\mathcal{A}$. The expression *with high probability* refers to event occurring with probability approaching one when (T, d, k_0) tend to infinity. The scaling results for (T, d, k_0) are expressed as $f(T) \geq M g(k_0, d)$ for $0 < M < \infty$ some universal constant and continuous functions $f(.), g(.)$.

2 High Dimensional VAR

Let $(\boldsymbol{Y}_t)$ be an N-dimensional random vector of time series and consider the following VAR model:

$$\boldsymbol{Y}_t = A_1\boldsymbol{Y}_{t-1} + \cdots + A_p\boldsymbol{Y}_{t-p} + \boldsymbol{u}_t, \quad t = 1, \ldots, T, \tag{1}$$

where $\boldsymbol{u}_t \in \mathbb{R}^N$ with $\mathbb{E}[\boldsymbol{u}_t] = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{u}_t) = \Sigma_u \in \mathcal{M}_{N\times N}(\mathbb{R})$, where $\Sigma_u \succ 0$ and for each $i = 1, \cdots, p$, $A_i \in \mathcal{M}_{N\times N}(\mathbb{R})$. The driving parameters are A_i, Σ_u. Although we assume that $\boldsymbol{Y}_t$ is a mean zero vector, our setting can straightforwardly include a constant term in (1). Now consider an invertible matrix $B_0 \in \mathcal{M}_{N\times N}(\mathbb{R})$ with diagonal elements being equal to one. Multiplying B_0 from the left of both sides of (1), we obtain the following form:

$$B_0\boldsymbol{Y}_t = B_1\boldsymbol{Y}_{t-1} + \cdots + B_p\boldsymbol{Y}_{t-p} + \boldsymbol{e}_t, \quad t = 1, \ldots, T, \tag{2}$$

where $\boldsymbol{e}_t = B_0\boldsymbol{u}_t$ and $B_i = B_0A_i$. By construction, $\mathbb{E}[\boldsymbol{e}_t] = \mathbf{0}$, $\Sigma_e = \mathrm{Var}(\boldsymbol{e}_t) = B_0\Sigma_uB_0'$, and Σ_e is positive definite. Equation (2) is known as the 'A-model' in the literature on structural VAR (SVAR) models: see, e.g., Lütkepohl (2006, 2017). A challenging task concerns SVAR identification. One way is to assume (i) Σ_e is diagonal, (ii) the diagonal elements of B_0 are one, and (iii) the number of zeros in B_0 is $N(N-1)/2$: these identification restrictions are provided in Proposition 9.1 of Lütkepohl (2006). As pointed out in Subsection 9.1.4 of Lütkepohl (2006), assuming such restrictions is not necessary, especially due to the lack of meaningful economic justifications to impose contemporaneous restrictions. For instance, Blanchard and Quah (1989) developed a framework for imposing long-run restrictions for structural VAR models: such restrictions are based on long-run neutrality properties.

Rather than imposing the above restrictions a priori on B_0 or on long-run representations, we pursue the direction of Sims (1980) for a data-oriented identification based on a graphical representation, which will be explained below. We highlight the existence of an alternative SVAR identification method based on sign restrictions: it consists in dropping doubtful restrictions one after one to identify the most likely admissible model within the set of structural VAR models that satisfy the assumed sign restrictions. For instance,

within the 'B-model' described in Subsection 9.1.2 of Lütkepohl (2006), Inoue and Killian (2013) developed an approach based on all admissible models which satisfy signs of parameters derived by economic theory.

Since our analysis focuses on potentially large dimensional VAR model, we assume sparsity either on Σ_u and A_i $(i = 0, 1, \ldots, p)$ or on B_i $(i = 0, 1, \ldots, p)$. The matrices A_i's are coefficients for regressing $\boldsymbol{Y}_t$ on $\boldsymbol{Y}_{t-1}, ..., \boldsymbol{Y}_{t-p}$, while the matrices B_i $(i = 1, \ldots, p)$ and $I_N - B_0$ are obtained by regressing recursively an element of $\boldsymbol{Y}_t$ on $\boldsymbol{Y}_{t-1}, ..., \boldsymbol{Y}_{t-p}$ and $\boldsymbol{Y}_t$ except for the own variable. In other words, the sparsity on the parameters B_0 and B_i $(i = 1, \ldots, p)$ corresponds to the absence of contemporaneous effects and dependence with respect to past observations, respectively. As investigated by Tunnicliffe-Wilson and Reale (2008), Oxley, Reale, Tunnicliffe-Wilson (2009), and Ahelegbey, Billio and Casarin (2016), Σ_u^{-1} (or B_0) and B_i $(i = 1, \ldots, p)$ in the SVAR model (2) can be interpreted in terms of directed acyclic graph in the graphical modeling. For this reason, we assume the sparsity on Σ_u^{-1} (or B_0) and B_i $(i = 1, \ldots, p)$. However, rather than fostering sparsity on the B_i's directly and thus apply a standard regularized estimation on a OLS/Gaussian MLE loss, we consider sparsity on the precision matrix of the vector $\boldsymbol{X}_t = (\boldsymbol{Y}_{t-p}^\top, \ldots, \boldsymbol{Y}_{t-1}^\top, \boldsymbol{Y}_t^\top)^\top$, whose partial correlation coefficients characterise the coefficients among the B_i's and Σ_u: see, e.g., Section 5.3 of Johnston (1972). Denote $\Sigma_x = \mathrm{Var}(\boldsymbol{X}_t)$, where $\boldsymbol{X}_t$ is the $N(p+1)$ vector defined above. Then, denoting $\Gamma_i = \mathbb{E}[\boldsymbol{Y}_t \boldsymbol{Y}'_{t-i}]$ the autocovariance matrix, we obtain

$$\Sigma_x = \begin{pmatrix} \Gamma_0 & \Gamma_1' & \cdots & \Gamma_p' \\ \Gamma_1 & \Gamma_0 & \cdots & \Gamma_{p-1}' \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_p & \Gamma_{p-1} & \cdots & \Gamma_0 \end{pmatrix}$$

Applying Example 2.2 of Dahlhaus and Eichler (2000) under the non-Gaussian assumption, equation (5) of Ahelegbey, Billio and Casarin (2016) indicates the equivalence

$$\sigma_u^{rs} = 0 \Longleftrightarrow \mathrm{Corr}(\boldsymbol{Y}_{r,t}, \boldsymbol{Y}_{s,t} | \{\boldsymbol{X}_t \backslash \{\boldsymbol{Y}_{r,t}, \boldsymbol{Y}_{s,t}\}\}) = 0,$$

$$B_{i,rs} = 0 \Longleftrightarrow \mathrm{Corr}(\boldsymbol{Y}_{r,t}, \boldsymbol{Y}_{s,t-i} | \{\boldsymbol{X}_t \backslash \{\boldsymbol{Y}_{r,t}, \boldsymbol{Y}_{s,t-i}\}\}) = 0,$$

for $i = 0, 1, \ldots, p$, where σ_u^{rs} is the (r, s)th element of Σ_u^{-1}. To derive the partial correlation coefficient between two variables in $\boldsymbol{X}_t$, we need to use the inverse of Σ_x. Denoting the (j, k)th element of Σ_x^{-1} as σ_x^{jk},

the partial correlation coefficient between the jth and the kth elements of $\boldsymbol{X}_t$ is then $\rho_x^{jk} = -\sigma_x^{jk} \Big/ \sqrt{\sigma_x^{jj}\sigma_x^{kk}}$ $(j \neq k)$. It is straightforward to show

$$\sigma_u^{rs} = \sigma_x^{Np+r,Np+s}$$

for $r, s = 1, \ldots, N$. By equation (5.31) of Johnston (1972), we obtain

$$B_{i,rs} = \frac{\Sigma_{x,kk}}{\Sigma_{x,jj}} \rho_x^{jk}, \quad j = Np + r, \quad k = N(p-i) + s,$$

for $i = 0, 1, \ldots, p$ and $r, s = 1, \ldots, N$. In view of such structure, we can impose sparsity on B_i's and Σ_u^{-1} through the sparsity of Σ_x^{-1}, the precision matrix of the random vector $\boldsymbol{X}_t$: if we are in a position to get a consistent and a positive definite estimator of Σ_x^{-1}, say $\widehat{\Sigma}_x^{-1}$, then we would obtain consistent estimators $\widehat{B}_i, \widehat{A}_i$ $(i = 1, \ldots, p)$ and $\widehat{\Sigma}_u$. There is an additional merit for imposing sparsity on Σ_x^{-1}: as we consider positive definite Σ_x or Σ_x^{-1}, the parameters satisfy the stationary condition automatically by, e.g., equation (2.1.43) of Lütkepohl (2006). Such matching between the coefficients of Σ_x^{-1} and the VAR coefficients allows for a component-wise sparse structure, in the same spirit as in Ravikumar et al. (2011) or Zhang and Zou (2014). Such element-wise sparsity assumption is a key difference with the Cholesky based sparse assumption of Huang et al. (2006), which imposes a particular variable ordering as emphasised by Bickel and Levina (2008b). The sparsity assumption is thus stated for the B_i's and Σ_u^{-1} parameters, which is equivalent to assuming a component-wise sparse structure on Σ_x^{-1}. The main contribution of our study is to propose novel sparse estimators for $\widehat{\Theta} := \widehat{\Sigma}_x^{-1}$ based on Bregman divergences and provide error bounds together with the conditions for support recovery. In particular, the von Neuman and least squares based estimators of such precision matrix have not been studied. In the rest of the paper, we denote $\Theta_0 := \Sigma_x^{-1}$ the true precision matrix of the $N(p+1)$ random vector $\boldsymbol{X}_t$.

It is worth mentioning the competing estimation techniques for SVAR models. The inference for the 'A-model' can be perfomed by maximum likelihood under suitable restrictions as described in Subsection 9.3.1 of Lütkepohl (2005). Lütkepohl (2017) considered the method of moments for estimating B_0 using the OLS residuals from the VAR model. Should one consider a graphical approach for SVAR identification,

a Bayesian inference procedure can be carried out by Markov chain Monte Carlo method as in Ahelegbey, Billio and Casarin (2016). Unlike the latter work, the matching we highlighted between the SVAR parameters and the precision matrix enables us to carry out a novel sparse inference.

3 SVAR inference through sparse precision matrix

3.1 Framework

Under the sparsity assumption of $B_i, i = 0, \cdots, p$ and Σ_u and thus of Σ_x^{-1}, we aim at recovering the true sparse support $\mathcal{A}$ of the inverse of the variance covariance matrix parameter $\Theta_0 = \Sigma_x^{-1} \in \mathcal{M}_{d\times d}(\mathbb{R})$ with $d = N(p+1)$. The sparsity assumption is specified as follows.

Assumption 1. *The true parameter $\theta_0 = vec(\Theta_0)$ is sparse so that $k_0 = \text{card}(\mathcal{A})$, where $\mathcal{A} = \{1 \leq i \leq d^2 : \theta_{i,0} \neq 0\}$ with $k_0 < d^2$ the total number of parameters.*

To estimate such sparse Θ_0, we rely on a regularized M-estimation problem given by:

$$\widehat{\Theta} = \underset{\Theta\in\Omega}{\arg\min} \Big\{\mathbb{L}_T(\Theta) + \boldsymbol{p}(\lambda_T, \theta)\Big\}, \tag{3}$$

where $\theta = \text{vec}(\Theta)$, $\boldsymbol{p}(\lambda_T, .) : \mathbb{R}^{d^2} \to \mathbb{R}$ is the penalty function, with λ_T the regularization parameter, which depends on the sample size, and enforce a particular type of sparse structure in the solution $\widehat{\Theta}$. Here $\mathbb{L}_T : \mathbb{R}^{d^2} \times \mathbb{R}^{TN} \to \mathbb{R}$ is the non-penalized loss function, which evaluates the precision of the fit with the sample $(\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_T)$. As stated in Subsections 3.2 and 5.2, the scaling behaviour of (T, d, k_0) together with statistical consistency highly depend on the choice of $\mathbb{L}_T(.)$ and its regularity. Ω denotes a $d \times d$-variance convex covariance matrix subset defined as $\Omega = \Big\{\Theta : \ \Theta \succ 0, \ \|\theta\|_1 \leq R\Big\}$. The constraint through R may be somewhat arbitrary but it enforces the estimated optimum to be close enough to the theoretical optimum and ensures that $\mathbb{L}_T(.)$ is lower bounded. Indeed, due to the potential non-convexity of the criterion, we include the side condition $\|\theta\|_1 \leq R$, where R is a supplementary regularization parameter to ensure the existence of local/global optima: more details on this constraint can be found in Section 3 of Loh and Wainwright (2017) or in Poignard and Fermanian (2021). Alternative presentations of the sparse

precision matrix setting are possible, where only the off-diagonal entries of Θ are penalized. Similar results for statistical consistency actually hold in this case. In our framework, we assume that all components are equally penalized to clarify our arguments. Similar settings are also considered by Loh and Wainwright (2015) and Fan, Feng and Wu (2009).

As for the loss $\mathbb{L}_T(.)$ in (3), we consider the general framework of the Bregman divergence criterion $D_\phi(\Theta, \Theta_0)$, defined as a dissimilarity measure between two symmetric positive definite matrices, the candidate Θ and the true inverse Θ_0, defined as

$$D_\phi(\Theta, \Theta_0) = \phi(\Theta) - \phi(\Theta_0) - \text{tr}(\big(\nabla\phi(\Theta_0)\big)^\top\big(\Theta - \Theta_0\big)),$$

where ϕ is a differentiable and strictly convex function over the space of real and symmetric positive definite matrices. Obviously, such discrepancy can not be optimized with respect to Θ unless Θ_0 is replaced by some known quantity. We propose to replace Θ_0 by its empirical version $\widehat{S}^{-1}$, the inverse of the sample variance covariance matrix of $\mathbf{X}_t$, where we assume $\widehat{S} \succ 0$. Thus, replacing Θ_0 by $\widehat{S}^{-1}$, we consider the loss function in (3) as $\mathbb{L}_T(.) = \mathbb{L}_{T,\eta}(.)$ with

$$L_{T,\eta}(\Theta) = \eta D_\phi(\Theta, S^{-1}) + (1-\eta) D_\phi(S^{-1}, \Theta),$$

where $0 \leq \eta \leq 1$ is a known scalar value. This loss function is an extension of the standard Bregman divergence setting. Indeed, the loss function is a balance between the Bregman divergence and its switched argument version. As it will be emphasised in our theoretical analysis, suitable choices of η will enable us to provide the probability for which statistical accuracy holds. We propose the following specifications of ϕ and hence of $\mathbb{L}_T(.)$:

(i) $\phi(\Theta) = -\log(|\Theta|)$, so that the corresponding Bregman divergence can be written as

$$\mathbb{L}_{T,\eta}(\Theta) = (1-2\eta)\log(|\Theta|) + \text{tr}\Big(\eta\widehat{S}\Theta + (1-\eta)\Theta^{-1}\widehat{S}^{-1}\Big),$$

where the terms independent of Θ are discarded. This function is known as Stein's loss, and is closely related to the standard Gaussian QML criterion up to some constants. When $\eta = 0$, one obtain the

standard Gaussian QML criterion for inverse variance covariance estimation.

(ii) $\phi(\Theta) = \text{tr}(\Theta^2)$, then the Bregman divergence is a least squares loss defined as $\mathbb{L}_{T,\eta}(\Theta) = \|\Theta - \widehat{S}^{-1}\|_F^2$, which is η independent: in that case, the dependence of $\mathbb{L}_{T,\eta}(.)$ on η is simply skipped. We use $\mathbb{L}_T(.)$ as the notation for the least squares case.

(iii) $\phi(\Theta) = \text{tr}(\Theta\log(\Theta) - \Theta)$, then the derivative becomes $\nabla_\Theta\phi(\Theta) = \text{tr}(\log(\Theta))$ (see exercise 13.31 of Abadir and Magnus, 2005, for the matrix logarithm derivative) and the Bregman divergence becomes the von Neumann. Then, our loss is defined as

$$\begin{aligned}\mathbb{L}_{T,\eta}(\Theta) &= (2\eta-1)\text{tr}\Big(\widehat{S}^{-1} - \Theta\Big) + \eta\text{tr}\Big(\Theta\log(\Theta)\Big) + (1-\eta)\text{tr}\Big(\widehat{S}^{-1}\log(\widehat{S}^{-1})\Big) \\ &\quad -\eta\text{tr}\Big(\log(\widehat{S}^{-1})\Theta\Big) - (1-\eta)\text{tr}\Big(\log(\Theta)\widehat{S}^{-1}\Big).\end{aligned}$$

It is interesting to compare the above losses with the D-trace loss function $\mathbb{L}_T(\Theta) = \frac{1}{2}\text{tr}\Big(\Theta^2\widehat{S} - 2\Theta\Big)$ that was proposed by Zhang and Zou (2014), who performed a LASSO regularization to obtain a sparse precision matrix. We will extend their framework to non-convex penalty functions, which will in particular allow for relaxing their so-called incoherence condition when analysing the support recovery property.

To summarize, the loss function $\mathbb{L}_T(.)$ in (3) will be taken as the Stein's and von Neumann losses with switched arguments $\mathbb{L}_{T,\eta}(.)$, the least squares loss and the D-trace loss. As for the penalty function $\boldsymbol{p}(\lambda_T, .)$ in (3), we rely on the following assumption.

Assumption 2. *We consider penalty functions that are assumed to be amenable regularizers defined as follows. We denote $\boldsymbol{p}(.,.) : \mathbb{R}_+ \times \mathbb{R}^q$, with q denoting the dimension problem, the penalty function - or regularizer -, which is assumed to be coordinate-separable with respect to $\theta \in \mathbb{R}^q$, idest $\boldsymbol{p}(\lambda_T, \theta) = \sum_{k=1}^{q} p(\lambda_T, \theta_i)$. Furthermore, let $\mu \geq 0$, and $\boldsymbol{p}(\lambda_T, .)$ is μ-amenable if*

(i) $x \mapsto p(\lambda_T, x)$ is symmetric around zero and $\boldsymbol{p}(\lambda_T, 0) = 0$.

(ii) $x \mapsto p(\lambda_T, x)$ is non-decreasing on $\mathbb{R}^+$.

(iii) $x \mapsto \frac{p(\lambda_T, x)}{x}$ is non-increasing on $\mathbb{R}_\star^+$.

(iv) $x \mapsto p(\lambda_T, x)$ *is differentiable for any* $x \neq 0$.

(v) $\lim_{x \to 0^+} \partial_x p(\lambda_T, x) = \lambda_T$.

(vi) $x \mapsto p(\lambda_T, x) + \frac{\mu}{2} x^2$ *is convex for some* $\mu \geq 0$.

The regularizer $\boldsymbol{p}(\lambda_T, .)$ *is* (μ, ζ)*-amenable if in addition*

(vii) There exists $\zeta \in (0, \infty)$ *such that* $\partial_x p(\lambda_T, x) = 0$ *for* $x \geq \lambda_T \zeta$.

Let $\boldsymbol{q} : \mathbb{R}^+ \times \mathbb{R}^q \to \mathbb{R}$ *be* $\boldsymbol{q}(\lambda_T, x) = \lambda_T \|x\|_1 - \boldsymbol{p}(\lambda_T, x)$ *so that the function* $\frac{\mu}{2}\|x\|_2^2 - \boldsymbol{q}(\lambda_T, x)$ *is convex.*

Assumption 1 implies that the true support (unknown) is sparse, that is the matrix Θ_0 contains zero components. The regularization - or penalization - procedure provides an estimator of $\mathcal{A}$. To derive our theoretical properties, assumption 2 provides regularity conditions that potentially encompass non-convex functions. These regularity conditions are the same than Loh and Wainwright (2015, 2017) or Poignard and Fermanian (2021). In this paper, we focus on the LASSO, the SCAD due to Fan and Li (2001) and the MCP due to Zhang (2010), respectively defined as

$$\begin{aligned}
\textbf{LASSO}: \ & \boldsymbol{p}(\lambda_T, \rho) = \lambda_T |\rho|, \\
\textbf{MCP}: \ & \boldsymbol{p}(\lambda_T, \rho) = \operatorname{sign}(\rho) \lambda_T \int_0^{|\rho|} (1 - z/(\lambda_T b_{\text{mcp}}))_+ dz, \\
\textbf{SCAD}: \ & \boldsymbol{p}(\lambda_T, \rho) = \begin{cases} \lambda_T |\rho|, & \text{for } |\rho| \leq \lambda_T, \\ -(\rho^2 - 2 b_{\text{scad}} \lambda_T |\rho| + \lambda_T^2)/(2(b_{\text{scad}} - 1)), & \text{for } \lambda_T \leq |\rho| \leq b_{\text{scad}} \lambda_T, \\ (b_{\text{scad}} + 1)\lambda_T^2/2, & \text{for } |\rho| > b_{\text{scad}} \lambda_T, \end{cases}
\end{aligned}$$

where $b_{\text{scad}} > 2$ and $b_{\text{mcp}} > 0$ are fixed parameters for the SCAD and MCP respectively. The LASSO is a μ-amenable regularizer, whereas the SCAD and the MCP are (μ, ζ)-amenable. More precisely, $\mu = 0$ (resp. $\mu = 1/(b_{\text{scad}} - 1)$, resp. $\mu = 1/b_{\text{mcp}}$) for the Lasso (resp. SCAD, resp. MCP). The parameter μ can be interpreted as a coefficient of non-convexity level: the larger, the more non-convex the penalty becomes.

The penalized problem (3) may not be convex depending on the choice of the penalty - SCAD or MCP - and/or for a specific Bregman divergence. Therefore, we would like to weaken the convexity assumption so that we could evaluate the accuracy of $\widehat{\Theta}$. To do so, the restricted strong convexity is a key ingredient

to handle non-convex loss functions. Intuitively, we would like to handle a loss function that locally admits some curvature. To ensure this property, we rely on the strong convexity (local) of the loss function. The strong convexity of a differentiable loss function corresponds to a strictly positive lower bound on the eigenvalues of the Hessian matrix uniformly valid over a local region around the true parameter. This amounts to a curvature condition. More precisely, we are interested in a particular direction, that is the difference $\Delta = \widehat{\theta} - \theta_0$. Hence the notion of restricted strong convexity weakens the (local) strong convexity by adding a tolerance term. A detailed explanation is provided in Negahban et al. (2012).

Slightly extending the definition of Loh and Wainwright (2017), we say that an empirical loss function $\mathbb{G}_T(.)$ satisfies the restricted strong convexity condition (RSC) at $\theta \in \mathbb{R}^q$ if there exist two positive functions α_1, α_2 and two nonnegative functions τ_1, τ_2 of (θ, T, q) such that, for any $\Delta \in \mathbb{R}^q$,

$$\begin{aligned}
\langle \nabla_\theta \mathbb{G}_T(\theta + \Delta) - \nabla_\theta \mathbb{G}_T(\theta), \Delta \rangle &\geq \alpha_1 \|\Delta\|_2^2 - \tau_1 \frac{\log(q)}{T} \|\Delta\|_1^2, \text{ if } \|\Delta\|_2 \leq 1, \\
\langle \nabla_\theta \mathbb{G}_T(\theta + \Delta) - \nabla_\theta \mathbb{G}_T(\theta), \Delta \rangle &\geq \alpha_2 \|\Delta\|_2 - \tau_2 \sqrt{\frac{\log(q)}{T}} \|\Delta\|_1, \text{ if } \|\Delta\|_2 \geq 1.
\end{aligned}$$

The RSC property is fundamentally local and α_k, τ_k, $k = 1, 2$ depend on the chosen θ. The RSC condition of Loh and Wainwright (2015) is similar but uniform with respect to (T, q). Moreover, to weaken notations, we simply write α_k and τ_k, $k = 1, 2$, by skipping their implicit arguments (θ, T, q). The threshold for $\|\Delta\|_2$ has been set for convenience and one can reparameterize the model with $\bar{\theta} := r\theta$ for some $r > 0$.

3.2 Error bounds

We first provide some error bounds for the estimator (3) assuming that $\mathbb{L}_T(.)$ satisfies the RSC condition and the penalty is μ-amenable. We assume that the population risk function $\mathbb{L}(\Theta) = \mathbb{E}[\mathbb{L}_T(\Theta)]$ is assumed to be uniquely minimized at $\theta_0 = \text{vec}(\Theta_0) \in \mathbb{R}^{d^2}$. Then we have the following Theorem.

Theorem 3.1. *Assume $\theta \in \mathbb{R}^{d^2}$ and the objective function $\mathbb{L}_T(.) : \mathbb{R}^{d^2} \mapsto \mathbb{R}$ satisfies the RSC condition and $\boldsymbol{p}(\lambda_T, .)$ is μ-amenable, with $\frac{3}{4}\mu \leq \alpha_1$. Choose*

$$4 \max \left\{ \|\nabla_\theta \mathbb{L}_T(\theta_0)\|_\infty, \alpha_2 \sqrt{\frac{\log(d^2)}{T}} \right\} \leq \lambda_T \leq \frac{\alpha_2}{6R}, \tag{4}$$

and assume $T \geq \dfrac{16R^2 \max\{\tau_1^2, \tau_2^2\}}{\alpha_2^2} \log(d^2)$. *Let* $\widehat{\theta}$ *be a stationary point of (3). Then* $\widehat{\theta}$ *satisfies*

$$\|\widehat{\theta} - \theta_0\|_2 \leq \frac{6\lambda_T \sqrt{k_0}}{4\alpha_1 - 3\mu}, \ \|\widehat{\theta} - \theta_0\|_1 \leq \frac{6(16\alpha_1 - 9\mu)}{(4\alpha_1 - 3\mu)^2} \lambda_T k_0.$$

Remark. (i) This result is based on an optimization reasoning only and is obtained in a deterministic way; the proof can be found in Poignard and Fermanian (2021), Theorem 1. As will be clarified in the following Corollaries, to apply Theorem 3.1, we will need to check the conditions for which the loss function $\mathbb{L}_T(.)$ satisfies the RSC condition. Moreover, we will show that suitable choices of λ_T and R provide the probability to satisfy the conditions of Theorem 3.1 with high probability.

(ii) About (α_1, μ): the tightness of the error bounds are sensitive to the difference $4\alpha_1 - 3\mu$, assuming λ_T, k_0 fixed. Here, α_1 should be thought as the curvature of $\mathbb{L}_T$: the bigger α_1 is, the larger the curvature becomes. On the other hand, μ measures the non-convexity of the penalty function: the larger μ is, the more non-convex $\boldsymbol{p}(\lambda_T, .)$ becomes. Thus, there is a trade-off between α_1 and μ when satisfying the constraint $4\alpha_1 > 3\mu$.

One of the purposes of the paper is to answer the following points: given $\mathbb{L}_T(.)$ - Stein's and von Neumann loss with $\mathbb{L}_T := \mathbb{L}_{T,\eta}$, least squares and D-trace loss -, is the RSC condition satisfied? Can we apply Theorem 3.1 and evaluate the probability of (4)? To tackle the latter issue, we derive an exponential-type inequality. To do so, we rely on a Bernstein-type based inequality applied to the difference $\|\widehat{S} - \Sigma_x\|_{\max}$. Such quantity is key when bounding the gradient of $\mathbb{L}_T(.)$. The sample variance covariance matrix is defined as $\widehat{S} := \frac{1}{T} \sum_{i=1}^{T} \mathbf{X}_t \mathbf{X}_t^\top$. Suitable assumptions on the variables are required to adapt the exponential bound to the data dependent setting. This is the motivation of the next assumptions. To do so, we consider standard assumptions on the reduced form VAR equation (1) that can be written as $\sum_{i=0}^{p} L(i) \boldsymbol{Y}_{t-i} = \boldsymbol{u}_t$, with $L(0) = I_N$ and $L(i) = -A_i$. We assume:

Assumption 3. *The absolute values of the zeros of the polynomial* $det(P(z))$ *with* $P(z) = \sum_{i=1}^{p} L(i) z^i, z \in \mathbb{C}$ *are strictly greater than one.*

Assumption 4. *The probability distribution of* $\boldsymbol{u}_t$ *is absolutely continuous with respect to the Lebesgue measure on* $\mathbb{R}^N$.

Equipped with assumptions 3 and 4, the process $(\boldsymbol{Y}_t)$ is a geometrically completely regular process by Theorem 1 of Mokkadem (1988): thus $(\boldsymbol{Y}_t)$ is a strongly mixing process. Hence, setting p finite, $(\mathbf{X}_t)$ is also strongly mixing by, e.g., Theorem 3.49 of White (2001). The mixing condition will be crucial in our analysis to evaluate the probability of the discrepancy $\|\widehat{S} - \Sigma_x\|_{\max}$ exceeding a certain threshold.

Assumption 5. $(\mathbf{X}_t)$ *is strongly mixing with mixing coefficient* $\alpha(\tau) \leq \exp(-c\tau^{\gamma_1})$ *with* γ_1 *and* c *positive constants. Moreover,* $\exists \gamma_2 > 0, b > 0$ *such that* $\forall \delta > 0$ *and* $\forall i \leq d$, $\mathbb{P}(|\mathbf{X}_{i,t}| \geq \delta) \leq \exp(-\big(\frac{\delta}{b}\big)^{\gamma_2})$.

The latter assumption requires the exponential-type tails for the distribution of $(\mathbf{X}_{1,t}, \cdots, \mathbf{X}_{d,t})$. This allows us to derive an exponential bound on $\frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_{i,t}\mathbf{X}_{j,t} - \Sigma_{ij,x}$. We have the following lemma.

Lemma 3.2. *Let* $\gamma < 1$ *with* $1/\gamma = 1/\gamma_1 + 3/\gamma_2$. *Under assumptions 3, 4 and 5, assume* $T \geq 4$, *there exist positive constants* C_1, C_2, C_3, C_4, C_5 *depending only on* b, γ_1, γ_2 *such that* $\forall \epsilon > 0$,

$$\mathbb{P}(\|\widehat{S} - \Sigma_x\|_{\max} \geq \epsilon) \leq d^2\Big\{T\exp\big(-\frac{(T\epsilon)^{\gamma}}{C_1}\big) + \exp\big(-\frac{(T\epsilon)^2}{C_2(1+TC_3)}\big) + \exp\big(-\frac{(T\epsilon)^2}{C_4 T}\exp\big(\frac{(T\epsilon)^{\gamma(1-\gamma)}}{C_5(\log(T\epsilon))^{\gamma}}\big)\big)\Big\}.$$

In particular, let $L > 0$, *then* $\mathbb{P}(\|\widehat{S} - \Sigma_x\|_{\max} \geq L\sqrt{\frac{\log(d^2)}{T}}) = O(\frac{1}{d^2})$.

Remark. (i) This concentration inequality will be applied when bounding the random quantity $\|\mathbb{L}_{T,\eta}(\Sigma)\|_{\infty}$, where the difference $\widehat{S} - \Sigma_x$ will typically be bounded in this score function.

(ii) The choice of ϵ proportional to $\sqrt{\frac{\log(d^2)}{T}}$ is motivated by condition (4) in Theorem 3.1, where we aim at evaluating the probability of satisfying such condition.

(iii) Alternatively, we can use exponential bounds for separately Lipschitz functions such as Dedecker and Fan (2015) or Alquier et al. (2020). To do so, some contraction property on the data generating process of the reduced form VAR would be necessary.

Indeed, Theorem 3.1 is stated in a deterministic manner. We show that for suitable parameter choices (λ_T, R), the conditions of Theorem 3.1 hold with high probability. To do so, this requires bounding the random quantity $\|\nabla_\theta \mathbb{L}_{T,\eta}(\Theta_0)\|_\infty$ and verifying the RSC conditions. This motivates the use of Lemma 3.2.

We are in a position to provide the conditions for consistency of the Bregman divergence based sparse estimators. First, let us consider the Stein's loss case. For $\phi(\Theta) = -\log(|\Theta|)$, the statistical criterion is

$$\left\{ \begin{array}{rcl} \widehat{\Theta}^{\mathrm{g}} & = & \underset{\Theta \in \Omega}{\arg\min} \Big\{ \mathbb{L}_{T,\eta}(\Theta) + \boldsymbol{p}(\lambda_T, \theta) \Big\}, \ \text{with} \\ \mathbb{L}_{T,\eta}(\Theta) & = & (1-2\eta)\log(|\Theta|) + \mathrm{tr}\Big(\eta \widehat{S}\Theta + (1-\eta)\Theta^{-1}\widehat{S}^{-1}\Big). \end{array} \right. \tag{5}$$

Here, the loss function is optimized over the convex set $\Omega = \Omega = \Big\{\Theta : \ \Theta \succ 0, \ \|\theta\|_1 \leq R\Big\}$. The side constraint in Ω is in the same spirit as in Loh and Wainwright (2015).

Corollary 3.3. *Assume the regularizer is μ-amenable, under the sample size $T \geq CR^2\alpha_2^{-2}\log(d^2)$, with $C > 0$ a sufficiently large constant, with $\alpha_2 = (2\eta-1)\big\{\lambda_{\max}(\Theta_0)+1\big\}^{-2} + 2(1-\eta)\big\{\lambda_{\max}(\Theta_0)+1\big\}^{-3}\lambda_{\min}(\widehat{S}^{-1})$, if the regularization parameter satisfies*

$$4\max\Big\{\|(1-2\eta)\Theta_0^{-1} + \eta\widehat{S} - (1-\eta)\Theta_0^{-1}\widehat{S}^{-1}\Theta_0^{-1}\|_{\max}, \alpha_2\sqrt{\frac{\log\left(d^2\right)}{T}}\Big\} \leq \lambda_T \leq \frac{\alpha_2}{6R}, \tag{6}$$

where $\Theta_0 \in \Omega$, suppose $\frac{3}{4}\mu < \alpha_1$ with $\alpha_1 = \alpha_2$. Then any local optimum $\widehat{\Theta}^{\mathrm{g}}$ of program (5) satisfies

$$\|\widehat{\Theta}^{\mathrm{g}} - \Theta_0\|_F \leq \frac{6\lambda_T\sqrt{k_0}}{4\alpha_1 - 3\mu}, \quad \|\mathrm{vec}(\widehat{\Theta}^{\mathrm{g}}) - \mathrm{vec}(\Theta_0)\|_1 \leq \frac{6(16\alpha_1 - 9\mu)\lambda_T k_0}{(4\alpha_1 - 3\mu)^2}, \tag{7}$$

with $\alpha_1 = \alpha_2$, $k_0 = |\mathcal{A}|$ and $\mathcal{A} = \{1 \leq i \leq d^2 : \theta_{i,0} := \mathrm{vec}(\Theta_0)_i \neq 0\}$.

Furthermore, for $\eta = 1$, under assumptions 3, 4 and 5 so that the sample variance-covariance estimator satisfies the bound in Lemma 3.2, if (λ_T, R) are chosen so that $C_1\sqrt{\log(d^2)/T} \leq \lambda_T \leq C_2/R$ and for a sample size $T \geq L\Big\{\log(d^2)\max\left(R^2, k_0\right) \vee \log(d^2)^{2/\gamma-1}\Big\}$, for C_1, C_2, L large constants, then (7) hold with probability at least $1 - O(\exp\big(-\log(d^2)\big)) - o(\exp\big(-\log(d^2)\big))$.

Remark. (i) The proof relies on the following two steps: first, we verify the RSC condition for $\mathbb{L}_T(\Theta)$, derive the quantities $\alpha_1, \alpha_2, \tau_1, \tau_2$ by lower bounding $\lambda_{\min}(\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta))$ and obtain the error bounds (7) applying Theorem 3.1; second, we bound $\|\nabla_\theta \mathbb{L}_T(\Theta_0)\|_\infty$ using Lemma 3.2 for a fixed ϵ proportional to $\sqrt{\log(d^2)/T}$. In that case, the required rate becomes $\max\left(R^2, k_0\right)\log(d^2) = O(T)$.

(ii) Should we consider the Stein's loss for estimating the variance-covariance $\Sigma_x = \Theta_0^{-1}$, the regularity of the loss is significantly altered. The RSC parameters would involve the minimum eigenvalue $\lambda_{\min}(2\widehat{S} - \Sigma_x)$, which thus must be assumed positive to ensure positive RSC parameters α_1, α_2: see, e.g., Poignard and Terada (2020), who considered the RSC property for the Gaussian QML estimator of the factor model based variance covariance matrix.

(iii) The RSC parameters are sample dependent through $\widehat{S}^{-1}$. Actually, using the exponential bound on $\|\widehat{S} - \Sigma_x\|_{\max}$, we could express α_1, α_2 with respect to $\lambda_{\max}(\Sigma_x) = \lambda_{\max}(\Theta_0^{-1})$ as:

$$\lambda_{\min}(\widehat{S}^{-1}) \geq \lambda_{\max}(\widehat{S})^{-1} \geq \Big(\|\widehat{S} - \Sigma_x\|_s + \|\Sigma_x\|_s\Big)^{-1} \geq \Big(L\sqrt{d^2\frac{\log(d^2)}{T}} + \|\Sigma_x\|_s\Big)^{-1},$$

with high probability using Lemma 3.2 with $L > 0$ and for a suitable sample size.

(iv) When $\boldsymbol{p}(\lambda_T, \theta) = \lambda_T\|\theta\|_1$, then setting $\lambda_T \geq L\sqrt{\log(d^2)/T}$ and $R = m_0\sqrt{k_0}$ with a constant $m_0 \geq \|\theta_0\|_2$, we have the scaling $T \geq Mk_0\log(d^2)$.

We now consider the consistency of the least squares estimator $\widehat{\Theta}^{\text{ls}}$ (case $\phi(\Theta) = \text{tr}(\Theta^2)$) that satisfies:

$$\left\{\begin{array}{rcl} \widehat{\Theta}^{\text{ls}} & = & \underset{\Theta\in\Omega}{\arg\min}\Big\{\mathbb{L}_T(\Theta) + \boldsymbol{p}(\lambda_T, \theta)\Big\}, \text{ with} \\ \mathbb{L}_T(\Theta) & = & \|\Theta - \widehat{S}^{-1}\|_F^2, \;\; \Omega = \Big\{\Theta : \; \Theta \succ 0, \; \|\theta\|_1 \leq R\Big\}. \end{array}\right. \tag{8}$$

Corollary 3.4. *Assume the regularizer is μ-amenable, under the sample size $T \geq C\max(R^2, k_0^2)\alpha_2^{-2}\log(d^2)$, with $C > 0$ a sufficiently large constant, with $\alpha_2 = 2$, if the regularization parameter satisfies*

$$4\max\Big\{\|2\Big(\Theta_0 - \widehat{S}^{-1}\Big)\|_{\max}, \alpha_2\sqrt{\frac{\log\left(d^2\right)}{T}}\Big\} \leq \lambda_T \leq \frac{\alpha_2}{6R}, \tag{9}$$

where $\Theta_0 \in \Omega$, suppose $\frac{3}{4}\mu < \alpha_1$ with $\alpha_1 = \alpha_2$. Then any local optimum $\widehat{\Theta}^{\text{ls}}$ of program (8) satisfies

$$\|\widehat{\Theta}^{\text{ls}} - \Theta_0\|_F \leq \frac{6\lambda_T\sqrt{k_0}}{8 - 3\mu}, \quad \|\text{vec}(\widehat{\Theta}^{\text{ls}}) - \text{vec}(\Theta_0)\|_1 \leq \frac{6(32 - 9\mu)\lambda_T k_0}{(8 - 3\mu)^2}, \tag{10}$$

with $k_0 = |\mathcal{A}|$ and $\mathcal{A} = \{1 \leq i \leq d^2 : \theta_{i,0} := \text{vec}(\Theta_0)_i \neq 0\}$.

Furthermore, under assumptions 3, 4 and 5 so that the sample variance-covariance estimator satisfies the bound in Lemma 3.2, if (λ_T, R) are chosen so that $C_1\sqrt{d^2\log(d^2)/T} \leq \lambda_T \leq C_2/R$ and for a sample

size $T \geq L\Big\{d^2\log(d^2)\max(R^2,k_0)\vee\log(d^2)^{2/\gamma-1}\Big\}$, *for* $C_1,C_2,L>0$ *large constants, then (10) hold with probability at least* $1-O(\exp\big(-\log(d^2)\big))-o(\exp\big(-\log(d^2)\big))$.

Remark. Evaluating the probability of (9) requires upper bounding $\|\Sigma_x^{-1}-\widehat{S}^{-1}\|_{\max}$. Using the exponential bound on $\|\Sigma_x-\widehat{S}\|_{\max}$ of Lemma 3.2, one can get a bound on $\|\Sigma_x^{-1}-\widehat{S}^{-1}\|_{\max}$ at a cost d.

Let us focus on the von Neumann estimator $\widehat{\Theta}^{\text{vn}}$. To analyse its properties, we rely on the power series expansion $\log(A)=-\sum_{k=1}^{\infty}\frac{1}{k}(I_d-A)^k$, where A is a d-square symmetric positive definite matrix, to handle the matrix logarithm part. This expression is valid when $\|I_d-A\|_s<1$. In our case, instead of assuming $\|I_d-\Theta\|_s<1$ for expanding $\log(\Theta)$, which might be restrictive, we consider Θ/ν, where ν is the constant $\nu>\lambda_{\max}(\Theta)$ and integrate such constraint in the parameter space. Thus, we consider the criterion

$$\left\{\begin{array}{rcl}\widehat{\Theta}^{\text{vn}} & = & \underset{\Theta\in\Omega^{\text{vn}}}{\arg\min}\Big\{\mathbb{L}_{T,\eta}(\Theta)+\boldsymbol{p}(\lambda_T,\theta)\Big\},\ \text{with}\\ \mathbb{L}_{T,\eta}(\Theta) & = & (2\eta-1)\text{tr}\Big(\widehat{S}^{-1}-\Theta\Big)/\nu+\eta\text{tr}\Big(\Theta\log(\Theta/\nu)\Big)/\nu+(1-\eta)\text{tr}\Big(\widehat{S}^{-1}\log(\widehat{S}^{-1}/\nu)\Big)/\nu\\ & & -\eta\text{tr}\Big(\log(\widehat{S}^{-1}/\nu)\Theta\Big)/\nu-(1-\eta)\text{tr}\Big(\log(\Theta/\nu)\widehat{S}^{-1}\Big)/\nu.\end{array}\right. \tag{11}$$

where Ω^{vn} is defined as $\Omega^{\text{vn}}=\Big\{\Theta:\Theta\succ 0,\|\Theta\|_s<\nu,\|\theta\|_1\leq R\Big\}$.

Corollary 3.5. *Assume the regularizer is* μ*-amenable, under the sample size* $T\geq CR^2\alpha_2^{-2}\log(d^2)$*, with* $C>0$ *large enough, let* $k_0=|\mathcal{A}|$ *and* $\mathcal{A}=\{1\leq i\leq d^2:\theta_{i,0}:=\text{vec}(\Theta_0)_i\neq 0\}$*, then*

(i) $\eta=1$*:* $\alpha_2=1/(\nu d)$*, if the regularization parameter satisfies*

$$4\max\Big\{\|\frac{1}{\nu}\Big(\log(\Theta_0/\nu)-\log(\widehat{S}^{-1}/\nu)\Big)\|_{\max},\alpha_2\sqrt{\frac{\log\left(d^2\right)}{T}}\Big\}\leq\lambda_T\leq\frac{\alpha_2}{6R}, \tag{12}$$

with $\Theta_0\in\Omega^{vn}$*, suppose* $\frac{3}{4}\mu<\alpha_1$ *with* $\alpha_1=\alpha_2$*. Then any local optimum* $\widehat{\Theta}^{\text{vn}}$ *of (11) satisfies*

$$\|\widehat{\Theta}^{\text{vn}}-\Theta_0\|_F\leq\frac{6\lambda_T\sqrt{k_0}}{4/(\nu d)-3\mu},\quad\|\text{vec}(\widehat{\Theta}^{\text{vn}})-\text{vec}(\Theta_0)\|_1\leq\frac{6(16/(\nu d)-9\mu)\lambda_T k_0}{(4/(\nu d)-3\mu)^2}. \tag{13}$$

(ii) $\eta=0$*:* $\alpha_2=\lambda_{\min}(\widehat{S}^{-1}/\nu)\{\lambda_{\max}(\Theta_0)+1\}^{-2}$*, if the regularization parameter satisfies*

$$4\max\Big\{\|\frac{1}{\nu}\Theta_0^{-1}\Big(\Theta_0-\widehat{S}^{-1}\Big)\|_{\max},\alpha_2\sqrt{\frac{\log\left(d^2\right)}{T}}\Big\}\leq\lambda_T\leq\frac{\alpha_2}{6R}, \tag{14}$$

with $\Theta_0 \in \Omega$, suppose $\frac{3}{4}\mu < \alpha_1$ with $\alpha_1 = \alpha_2$. Then any local optimum $\widehat{\Theta}^{\mathrm{vn}}$ of (11) satisfies

$$\begin{aligned} \|\widehat{\Theta}^{\mathrm{vn}} - \Theta_0\|_F &\leq \frac{6\lambda_T\sqrt{k_0}}{4\Big(\lambda_{\min}(\widehat{S}^{-1}/\nu)\{\lambda_{\max}(\Theta_0)+1\}^{-2}\Big) - 3\mu}, \\ \|\mathrm{vec}(\widehat{\Theta}^{\mathrm{vn}}) - \mathrm{vec}(\Theta_0)\|_1 &\leq \frac{6(16\Big(\lambda_{\min}(\widehat{S}^{-1}/\nu)\{\lambda_{\max}(\Theta_0)+1\}^{-2}\Big) - 9\mu)\lambda_T k_0}{(4\Big(\lambda_{\min}(\widehat{S}^{-1}/\nu)\{\lambda_{\max}(\Theta_0)+1\}^{-2}\Big) - 3\mu)^2}. \end{aligned} \tag{15}$$

Furthermore, under assumptions 3, 4 and 5 so that the sample variance-covariance estimator satisfies the bound in Lemma 3.2, if (λ_T, R) are chosen so that $C_1\sqrt{d^2\log(d^2)/T} \leq \lambda_T \leq C_2/R$ and for a sample size $T \geq L\Big\{\|\Theta_0^{-1}/\nu\|_s^2 d^2\log(d^2)\max(R^2,k_0)\vee\log(d^2)^{2/\gamma-1}\Big\}$, for $C_1, C_2, L > 0$ large constants, then (15) hold with probability at least $1 - O(\exp\big(-\log(d^2)\big)) - o(\exp\big(-\log(d^2)\big))$.

Remark. (i) As in Corollary 3.3, the RSC parameters are sample dependent through $\widehat{S}^{-1}$ and can be expressed with respect to $\lambda_{\max}(\Sigma_x)$.

(ii) For $\eta = 1$, $\nabla_\Theta \mathbb{L}_{T,1}(\Theta_0) = \frac{1}{\nu}\Big(\log(\Theta_0/\nu) - \log(\widehat{S}^{-1}/\nu)\Big)$. Then upper bounding $\log(\Theta_0/\nu) - \log(\widehat{S}^{-1}/\nu)$ is challenging. For instance, should we use power series expansion, we would need to control for

$$\|\sum_{k=1}^{\infty}\frac{1}{k}\Big\{(I_d - \widehat{S}^{-1}/\nu)^k - (I_d - \Theta_0/\nu)^k\Big\}\|_{\max} \leq \sum_{k=1}^{\infty}\frac{1}{k}\|(I_d - \widehat{S}^{-1}/\nu)^k - (I_d - \Theta_0/\nu)^k\|_s.$$

Thus, unless we assume further conditions such as Θ_0 and $\widehat{S}^{-1}$ commuting, deriving an exponential bound is a challenging task.

(iii) When $\eta = 0$, the scaling behaviour (T, d, k_0) in the von Neumann case is similar to the least squares case. Indeed, evaluating (14) requires controlling for $\frac{1}{\nu}\big(\Theta_0 - \widehat{S}^{-1}\big) = \frac{1}{\nu}\big(\Sigma_x^{-1} - \widehat{S}^{-1}\big)$. For this reason, we are able to provide the probability for which inequalities in (15) hold.

We finally investigate the D-trace loss based sparse estimator $\widehat{\Theta}^{\mathrm{dt}}$, which satisfies the criterion

$$\begin{cases} \widehat{\Theta}^{\mathrm{dt}} &= \underset{\Theta\in\Omega}{\arg\min}\Big\{\mathbb{L}_T(\Theta) + \boldsymbol{p}(\lambda_T,\theta)\Big\}, \\ \mathbb{L}_T(\Theta) &= \frac{1}{2}\mathrm{tr}\Big(\Theta^2\widehat{S}\Big) - \mathrm{tr}\Big(\Theta\Big), \;\; \Omega = \Omega = \Big\{\Theta :\; \Theta \succ 0,\; \|\theta\|_1 \leq R\Big\}. \end{cases} \tag{16}$$

This loss function was proposed by Zhang and Zou (2014). We then have the following error bounds.

Corollary 3.6. *Assume the regularizer is μ-amenable, under the sample size $T \geq C \max(R^2, k_0^2)\alpha_2^{-2}\log(d^2)$, with $C > 0$ a sufficiently large constant, with $\alpha_2 = \lambda_{\min}(\widehat{S})$, if the regularization parameter satisfies*

$$4\max\Big\{\|\frac{1}{2}\Theta_0\Big(\widehat{S} - \Theta_0^{-1}\Big) + \frac{1}{2}\Big(\widehat{S} - \Theta_0^{-1}\Big)\Theta_0\|_{\max}, \alpha_2\sqrt{\frac{\log(d^2)}{T}}\Big\} \leq \lambda_T \leq \frac{\alpha_2}{6R}, \tag{17}$$

where $\Theta_0 \in \Omega$, suppose $\frac{3}{4}\mu < \alpha_1$ with $\alpha_1 = \alpha_2$, then any local optimum $\widehat{\Theta}^{\mathrm{dt}}$ of program (16) satisfies

$$\|\widehat{\Theta}^{\mathrm{dt}} - \Theta_0\|_F \leq \frac{6\lambda_T\sqrt{k_0}}{4\lambda_{\min}(\widehat{S}) - 3\mu}, \quad \|\mathrm{vec}(\widehat{\Theta}^{\mathrm{dt}}) - \mathrm{vec}(\Theta_0)\|_1 \leq \frac{6(16\lambda_{\min}(\widehat{S}) - 9\mu)\lambda_T k_0}{(4\lambda_{\min}(\widehat{S}) - 3\mu)^2}, \tag{18}$$

with $k_0 = |\mathcal{A}|$ and $\mathcal{A} = \{1 \leq i \leq d^2 : \theta_{i,0} := \mathrm{vec}(\Theta_0)_i \neq 0\}$.

Furthermore, under assumptions 3, 4 and 5 so that the sample variance-covariance estimator satisfies the bound in Lemma 3.2, if (λ_T, R) are chosen so that $C_1\sqrt{d^2\log(d^2)/T} \leq \lambda_T \leq C_2/R$ and for a sample size $T \geq L\Big\{\|\Theta_0\|_s^2 d^2\log(d^2)\max\big(R^2, k_0\big) \vee \log(d^2)^{2/\gamma - 1}\Big\}$, for $C_1, C_2, L > 0$ large constants, then (18) hold with probability at least $1 - O(\exp\big(-\log(d^2)\big)) - o(\exp\big(-\log(d^2)\big))$.

Remark. (i) As in remark (i) following Corollary 3.5, the RSC parameters can be expressed with respect to $\lambda_{\min}(\Sigma_x)$ using the exponential bound of Lemma 3.2.

(ii) Zhang and Zou (2014), in their Theorem 2, obtained a similar rate for the $\|.\|_F$ consistency of the D-trace estimator; and our scaling behaviour (T, d, k_0) shares the same sample size order as theirs. However, contrary to their result, our upper bound explicitly links the regularity of the D-trace loss with the non-convexity degree of the penalty function through the RSC property. In another context, using the D-Trace loss function, Wu and Li (2020) focused on the consistency of the LASSO penalized precision matrix difference $\widehat{\Delta} = \widehat{\Theta}_Y - \widehat{\Theta}_X$, where $\Theta_Y = \Sigma_Y^{-1}, \Theta_X = \Sigma_X^{-1}$ are the precision matrices of the two state vectors Y, X assumed sub-Gaussian. In their Theorem 2, they also obtained a similar $\|.\|_F$ consistency but with much larger dimension dependent constants.

All else being equal, the curvature of the loss $\mathbb{L}_T(.)$ in (3) significantly influences how informative the theoretical upper bounds are. For the Stein's loss, the curvature parameter α_1 involves the spectral norm of the true parameter Θ_0: $\lambda_{\max}(\Theta_0)$ determines the convexity degree of $\mathbb{L}_T(.)$. As it will be highlighted in

Section 4, a small α_1 value requires a low degree of non-convexity in the penalty to satisfy $4\alpha_1 > 3\mu$: the smaller μ is, the less non-convex $\boldsymbol{p}(\lambda_T, .)$ should be. The dimension impacts also the von Neumann and D-trace losses. Interestingly, the least squares loss involves a dimension free expression for α_1.

3.3 Support recovery

Based on the Karush-Kuhn-Tucker optimality conditions, Wainwright (2009) developed the primal dual witness (PDW) approach to derive selection consistency for convex problems. There exist similar approaches in Candès and Plan (2009) or Zhao and Yu (2006). The PDW approach consists in plugging the true subset model $\mathcal{A}$ in the KKT optimality conditions, which are necessary and sufficient if the problem is convex, and checking if they can be satisfied. It means that any solution of the non restricted problem (the original problem providing $\mathcal{A}$) is also a solution to the restricted problem (the regularized one). Loh and Wainwright (2017) showed that this approach can be extended to a nonconvex problem and thus to any stationary point, which is their key contribution. They prove that all stationary points are consistent for variable selection via a strict dual feasibility condition and second-order conditions. To obtain the support recovery property, the RSC condition of the loss function with parameters $(\alpha_k, \tau_k)_{k=1,2}$ and the μ-amenability of the penalty are key assumptions. More details can be found in Subsection 5.2: there, in Theorem 5.1, we provide the conditions of Loh and Wainwright (2017) to ensure the success of the PDW construction - corresponding to **Step 3.** -, that is the scaling of (λ_T, R) and the so-called strict feasibility condition, which characterize the solution of the PDW construction; Theorem 5.2 establishes the support recovery property together with consistency in the $\|.\|_\infty$-sense under the RSC condition, μ-amenable penalties and strict dual feasibility; finally, two sufficient conditions in Proposition 5.3 ensure that strict dual feasibility holds for (μ, ζ)-amenable penalties. Within this setting, we provide ℓ_∞-guarantees for the regularized $\widehat{\Theta}$ together with the conditions to satisfy the support recovery property: for the Stein's loss, we restrict our analysis to $\eta = 1$, the most commonly used loss for sparse precision matrix estimation; $\eta = 0$ for the von Neumann case. Rather than stating the support recovery property in a deterministic manner, we directly evaluate

the probability of satisfying the latter property. For all Bregman losses, we show that any local/global optimum of (3) corresponds to the oracle estimator with high probability. The latter is given as

$$\widehat{\Theta}^{\mathcal{O}} := \underset{\theta\in\mathbb{R}^{|\mathcal{A}|}:\Theta\in\Omega}{\arg\min} \Big\{\mathbb{L}_T(\Theta)\Big\}, \tag{19}$$

with $\mathrm{vec}(\widehat{\Theta}^{\mathcal{O}}) = (\mathrm{vec}(\widehat{\Theta}^{\mathcal{O}}_{\mathcal{A}}), \mathbf{0}_{\mathcal{A}^c})$. We denote the Fisher information matrix $\boldsymbol{K}_0 = \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]$. The conditions we derive hold for all stationary points of (3), idest for local/global optimum. $\mathbb{L}_T(\Theta)$ is the non-penalized loss of problem (3). Thus, we denote $\widehat{\Theta}^{\mathrm{g},\mathcal{O}}, \widehat{\Theta}^{\mathrm{ls},\mathcal{O}}, \widehat{\Theta}^{\mathrm{vn},\mathcal{O}}, \widehat{\Theta}^{\mathrm{dt},\mathcal{O}}$ the oracle estimators of (19) respectively for the Stein, least squares, von Neumann and D-trace loss.

For the Stein's loss, we consider $\eta = 1$ so that the Fisher information matrix becomes $\boldsymbol{K}_0 = \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)] = \Big(\Theta_0^{-1}\otimes\Theta_0^{-1}\Big)$. The conditions for support recovery for the Stein case are given as follows:

Corollary 3.7. *Assume $T \geq L\Big\{\|\Sigma_x\|^4_\infty k_0^2 \log(d^2) \vee \log(d^2)^{2/\gamma-1}\Big\}$ with $L > 0$ large enough, the regularization parameters (λ_T, R) are chosen so that $\|\mathrm{vec}(\Theta_0)\|_1 \leq \frac{R}{2}$ and $C_1\sqrt{\frac{\log(d^2)}{T}} \leq \lambda_T \leq \frac{C_2}{R}$, for $C_1, C_2 > 0$, assume $\|\boldsymbol{K}_0^{-1}\|_\infty \leq \beta_\infty$ and assumptions 3, 4 and 5 hold. Then:*

(i) Assume $\boldsymbol{p}(\lambda_T, .)$ is μ-amenable penalty and $\|\Big(\Theta_0^{-1}\otimes\Theta_0^{-1}\Big)_{\mathcal{A}^c\mathcal{A}}\Big(\Theta_0^{-1}\otimes\Theta_0^{-1}\Big)^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty \leq \omega < 1$ (incoherence condition), then with probability at least $1 - O(\exp\big(-\log(d^2)\big) - o(\exp\big(-\log(d^2)\big)$, the objective function (5) admits a unique optimum so that $\widehat{\mathcal{A}} \subseteq \mathcal{A}$ and for a sufficiently large $\tilde{L} > 0$

$$\|\widehat{\Theta}^{\mathrm{g}} - \Theta_0\|_{\max} \leq \tilde{L}\sqrt{\frac{\log(d^2)}{T}} + \lambda_T\beta_\infty.$$

(ii) Assume $\boldsymbol{p}(\lambda_T, .)$ is (μ, ζ)-amenable and $\min_{i\in\mathcal{A}}|\mathrm{vech}(\Theta_0)_i| \geq \lambda_T(\zeta + 2\beta_\infty) + \tilde{L}\sqrt{\frac{\log(d^2)}{T}}$ for a sufficiently large $\tilde{L} > 0$, then with probability $1 - O(\exp\big(-\log(d^2)\big) - o(\exp\big(-\log(d^2)\big)$, (5) admits a unique optimum $\widehat{\Theta}^{\mathrm{g}}$, which agrees with the oracle estimator $\widehat{\Theta}^{\mathrm{g},\mathcal{O}}$ so that

$$\|\widehat{\Theta}^{\mathrm{g}} - \Theta_0\|_{\max} \leq \tilde{L}\sqrt{\frac{\log(d^2)}{T}}.$$

Remark. (i) The proof relies on the use of Theorem 5.2. To do so, strict dual feasibility must be proved (since Theorem 5.2 relies on the conditions of Theorem 5.1 and strict dual feasibility). To establish

strict dual feasibility, we use Theorem 5.1 for μ-amenable penalty functions and Proposition 5.3 for (μ, ζ)-amenable penalty functions. The main proof steps can be summarized as:

(a) Establishing strict dual feasibility by upper bounding the quantities $\|\nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty$ and $\|\hat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\hat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,1}(\Theta_0)\|_\infty$ by $(1-\delta)/2\lambda_T$ for $\delta \in [0,1]$ defined in Theorem 5.1 - with $\tau_1 = 0$ since the RSC condition for the Gaussian loss is satisfied with $\tau_1 = \tau_2 = 0$; here $\hat{\boldsymbol{K}}$ is defined as in Theorem 5.2 for the Gaussian loss. These bounds correspond to inequalities (27) and (28) in Proposition 5.3. Note that for the μ-amenable penalty case, the additional quantity $\|\hat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\hat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty$ must be upper-bounded. Once strict dual feasibility is established, we compute the upper bound of $\|\widehat{\Theta}^{\mathrm{g}} - \Theta_0\|_{\max}$ in point (i) of Theorem 5.2.

(b) Establishing point (ii) of Corollary 3.7 uses the exact same steps as in point (i), except that the (μ, ζ)-amenability allows for a simplification in the upper bound of $\|\widehat{\Theta}^{\mathrm{g}} - \Theta_0\|_{\max}$: the term involving $\|\hat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\hat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty$ can be discarded as well as the incoherence condition.

(ii) Corollary 3.7 is not expressed in a deterministic manner since we evaluate the probability of satisfying the inequalities (24) and (25) in Theorem 5.1. This implies controlling, among others, for the infimum norm of the score function evaluated at Θ_0. We thus obtain λ_T proportional to $\sqrt{\log(d^2)/T}$.

(iii) The scaling behaviour we obtained is in the same vein as in Loh and Wainwright (2017) or Ravikumar et al. (2011). Loh and Wainwright (2017) require $n > L\|\Gamma_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty^2 \|\Sigma_x\|_\infty^6 r_0^2 \log(d^2)$ in their Corollary 4 for i.i.d. data, for $L > 0$ a large constant, where $\|.\|_\infty$ denotes the ℓ_∞ operator norm, r_0 the number of nonzero entries per row and $\Gamma = \nabla^2_{\theta\theta}\mathbb{L}_{T,1}(\Theta_0)$. Such scaling is obtained for the side constraint $\|\Theta\|_s \leq \kappa$ rather than $\|\mathrm{vec}(\Theta)\|_1 \leq R$, which allows their sparsity assumption to be stated at a row/column level rather than over the whole matrix, which is the main difference with our setting.

(iv) We emphasize that there is an alternative method for constructing $\mathrm{vec}(\widehat{\Theta}^{\mathrm{g}})_{\mathcal{A}}$ such that $\mathrm{supp}(\widehat{\Theta}^{\mathrm{g}}_{\mathcal{A}}) \subseteq \mathcal{A}$ and $\mathrm{vec}(\widehat{\Theta}^{\mathrm{g}})_{\mathcal{A}}$ is a zero-subgradient point of the program (22) in **Step 1** of the Primal Dual Witness method: this method is based on the Brouwer's fixed point Theorem. Intuitively, the idea is to prove

that if there is a zero sub-gradient vector of the penalized estimator $\mathbb{L}_{T,1}(\Theta) + \boldsymbol{p}(\lambda_T; \text{vec}(\Theta))$ within the set $\{\Theta \in \Omega^{\text{g}}, \text{supp}(\Theta) \subseteq \text{supp}(\Theta_0)\}$, then this vector is the unique optimum. Then Brouwer's fixed point Theorem is used to show that such optimum lies in a neighbourhood of the true value $\text{vech}(\Theta_0)$ in the $\|.\|_\infty$-sense. Such method was developed by Ravikumar et al. (2011) for LASSO penalized Θ or Loh and Wainwright (2017) for LASSO/SCAD/MCP penalized Θ.

We now consider the least squares estimator $\widehat{\Theta}^{\text{ls}}$. The oracle $\widehat{\Theta}^{\text{ls},\mathcal{O}}$ satisfies (19) with $\mathbb{L}_T(\Theta) = \|\Theta - \widehat{S}^{-1}\|_F^2$.

Corollary 3.8. *Under assumptions 3, 4 and 5, assume* $T \geq L \max\Big\{[(d^2 - k_0) \vee k_0] \log(d^2), \log(d^2)^{2/\gamma - 1}\Big\}$ *with* $L > 0$ *large enough, assume the regularization parameters* (λ_T, R) *are chosen so that* $\|\text{vec}(\Theta_0)\|_1 \leq \frac{R}{2}$ *and* $C_1\sqrt{\frac{\log(d^2)}{T}} \leq \lambda_T \leq \frac{C_2}{R}$ *with* $C_1, C_2 > 0$. *Assume* $\boldsymbol{p}(\lambda_T, .)$ *is a* (μ, ζ)*-amenable penalty and* $\min_{i \in \mathcal{A}} |\text{vech}(\Theta_0)_i| \geq \lambda_T(\zeta + 2) + L\sqrt{\frac{\log(d^2)}{T}}$ *for* $L > 0$ *large enough, then with probability* $1 - O(\exp\big(-\log(d^2)\big) - o(\exp\big(-\log(d^2)\big)$, *(8) admits a unique optimum* $\widehat{\Theta}^{\text{ls}}$, *which agrees with the oracle estimator* $\widehat{\Theta}^{\text{ls},\mathcal{O}}$ *so that*

$$\|\widehat{\Theta}^{\text{ls}} - \Theta_0\|_{\max} \leq L\sqrt{\frac{\log(d^2)}{T}}.$$

Remark. (i) The proof follows the same steps as in Corollary 3.7: establishing inequalities (27) and (28) in Proposition 5.3 to use Theorem 5.2. Note that upper bounding $\|\nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}^c}\|_\infty$ and $\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty$ by $(1-\delta)/2\lambda_T$, for $\delta \in [0,1]$ defined in Theorem 5.1 - with $\tau_1 = 0$ since the RSC condition for the least squares loss is satisfied with $\tau_1 = \tau_2 = 0$, is more straightforward due to the linearity of the least squares loss, contrary to the case $\phi(\Theta) = -\log(|\Theta|)$.

(ii) If the Lasso is considered, which is a μ-amenable regularizer, then the mutual incoherence condition does not hold since $\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty = 1$. Consequently, strict dual feasibility can not be established for μ-amenable penalties when the least squares type loss function is considered.

Let us consider the conditions for support recovery for the von Neumann case, where we restrict our analysis to $\eta = 0$ to clarify our arguments. The matrix $\boldsymbol{K}_0 = \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]$ is given by (31) in the Appendix.

Corollary 3.9. *Assume* $T \geq L \max\left\{\frac{1}{\nu^2}\left[(d^2-k_0)\vee\frac{k_0^9}{\nu^2}\|\Sigma_x\|_s^4\left(1-\|\Theta_0/\nu\|_s\right)^{-2}\right]\|\Sigma_x\|_s^2\log(d^2), \log(d^2)^{2/\gamma-1}\right\}$ *with* $L>0$ *large enough, assume* $\mathbb{E}[\widehat{S}^{-1}]<\infty$, *choose the regularization parameters* (λ_T, R) *as* $\|\mathrm{vec}(\Theta_0)\|_1 \leq \frac{R}{2}$ *and* $C_1\sqrt{\frac{\log(d^2)}{T}} \leq \lambda_T \leq \frac{C_2}{R}$, *suppose* $\|\boldsymbol{K}_0^{-1}\|_\infty \leq \beta_\infty$ *and assumptions 3, 4 and 5 hold. Then:*

(i) Assume $\boldsymbol{p}(\lambda_T,.)$ *is a* μ*-amenable penalty and* $\|\boldsymbol{K}_{0,\mathcal{A}^c\mathcal{A}}\boldsymbol{K}_{0,\mathcal{A}\mathcal{A}}^{-1}\|_\infty \leq \omega<1$ *(incoherence condition), then with probability at least* $1-O(\exp\big(-\log(d^2)\big)-o(\exp\big(-\log(d^2)\big)$, *the objective function (11) admits a unique optimum so that* $\widehat{\mathcal{A}}\subseteq\mathcal{A}$ *and for a sufficiently large* $\tilde{L}>0$

$$\|\widehat{\Theta}^{\mathrm{vn}}-\Theta_0\|_{\max} \leq \tilde{L}\sqrt{\frac{\log(d^2)}{T}}+\lambda_T\beta_\infty.$$

(ii) Assume $\boldsymbol{p}(\lambda_T,.)$ *is* (μ,ζ)*-amenable and* $\min_{i\in\mathcal{A}}|\mathrm{vech}(\Theta_0)_i| \geq \lambda_T(\zeta+2\beta_\infty)+\tilde{L}\sqrt{d^2\frac{\log(d^2)}{T}}$ *for a sufficiently large* $\tilde{L}>0$, *then with probability* $1-O(\exp\big(-\log(d^2)\big)-o(\exp\big(-\log(d^2)\big)$, *(11) admits a unique optimum* $\widehat{\Theta}^{\mathrm{vn}}$, *which agrees with the oracle estimator* $\widehat{\Theta}^{\mathrm{vn},\mathcal{O}}$ *so that*

$$\|\widehat{\Theta}^{\mathrm{vn}}-\Theta_0\|_{\max} \leq \tilde{L}\sqrt{\frac{\log(d^2)}{T}}.$$

Remark. The scaling behaviour (T,d,k_0) is less favorable compared to the Stein's loss: this is because the Hessian of the Von Neumann loss requires the control of $\|\widehat{S}^{-1}/\nu\|$, a quantity, which does not appear in the Hessian of the Stein's loss for $\eta=1$.

Finally, we consider the D-trace loss case. The Fisher information matrix is $\boldsymbol{K}_0=\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)] = \frac{1}{2}\Big(\Theta_0^{-1}\otimes I_d + I_d\otimes\Theta_0^{-1}\Big)$. Our analysis extends the setting of Zhang and Zou (2014) to non-convex penalties allowing for relaxing their incoherence condition in point (ii) of the following Corollary.

Corollary 3.10. *Assume* $T>L\max\Big\{k_0^2\Big[\|\Sigma_x\|_\infty^2\vee\|\Theta_0\|_s^2\Big]\log(d^2), (d^2-k_0)\|\Theta_0\|_s^2\log(d^2), \log(d^2)^{2/\gamma-1}\Big\}$, *choose the regularization parameters* (λ_T,R) *so that* $\|\mathrm{vec}(\Theta_0)\|_1 \leq \frac{R}{2}$ *and* $C_1\sqrt{\frac{\log(d^2)}{T}} \leq \lambda_T \leq \frac{C_2}{R}$ *for* $C_1,C_2>0$, *assume* $\|\boldsymbol{K}_0^{-1}\|_\infty\leq\beta_\infty$ *and assumptions 3, 4 and 5 hold. Then:*

(i) Assume $\boldsymbol{p}(\lambda_T,.)$ *is* μ*-amenable and* $\|\Big(\Theta_0^{-1}\otimes I_d+I_d\otimes\Theta_0^{-1}\Big)_{\mathcal{A}^c\mathcal{A}}\Big(\Theta_0^{-1}\otimes I_d+I_d\otimes\Theta_0^{-1}\Big)^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty\leq\omega<1$ *(incoherence condition), then with probability at least* $1-O(\exp(-\log(d^2)))-o(\exp(-\log(d^2)))$, *the*

objective function (16) admits a unique optimum with $\widehat{\mathcal{A}} \subseteq \mathcal{A}$ *and for* $L > 0$ *large enough*

$$\|\widehat{\Theta}^{\mathrm{dt}} - \Theta_0\|_{\max} \leq L\sqrt{\frac{\log(d^2)}{T}} + \lambda_T \beta_\infty.$$

(ii) Assume $\boldsymbol{p}(\lambda_T, .)$ *is* (μ, ζ)*-amenable and* $\min_{i \in \mathcal{A}} |\mathrm{vech}(\Theta_0)_i| \geq \lambda_T(\zeta + 2\beta_\infty) + L\sqrt{\frac{\log(d^2)}{T}}$ *for a sufficiently large* $\tilde{L} > 0$*, then with probability* $1 - O(\exp(-\log(d^2))) - o(\exp(-\log(d^2)))$*, (16) admits a unique optimum* $\widehat{\Theta}^{\mathrm{dt}}$*, which agrees with the oracle estimator* $\widehat{\Theta}^{\mathrm{dt},\mathcal{O}}$ *so that*

$$\|\widehat{\Theta}^{\mathrm{dt}} - \Theta_0\|_{\max} \leq \tilde{L}\sqrt{\frac{\log(d^2)}{T}}.$$

Remark. Our scaling involves k_0, which represents the total sparsity over the whole precision matrix. In their Theorem 2, Zhang and Zhou (2014) obtain a similar $\|.\|_{\max}$ consistency rate, but their scaling condition for support recovery for i.i.d. data is such that $n > Lr_0^2 \log(d)$, with r_0 the maximum number of nonzero off-diagonal entries in any row (or column), a rate similar to Corollary 4 of Loh and Wainwright (2017).

4 Empirical applications

In all our simulation experiments, we simulate the N-dimensional VAR model

$$\forall 1 \leq t \leq T,\ \boldsymbol{Y}_t = \sum_{i=1}^{p} A_i \boldsymbol{Y}_{t-i} + \boldsymbol{u}_t,\ \ \boldsymbol{u}_t \sim \mathcal{N}_{\mathbb{R}^N}(0, \Sigma_u), \tag{20}$$

where the matrix coefficients A_i's are deduced from the sparse matrices B_0 and $B_i, 1 \leq i \leq p$. The matrix coefficients are simulated so that the conditions of assumption 3 are satisfied. The object of interest is the sparse inverse variance covariance matrix Θ_0, deduced from the sparse Σ_u^{-1}, B_i's coefficients. To recover such sparse element, we consider problems (5), (8), (11) and (16) with SCAD, MCP and LASSO penalization, providing 12 estimators. To solve the penalized optimization problem under the positive-definite constraint, we apply the numerical optimization *fmincon* on Matlab with $\widehat{S}^{-1}$ for initialization. Alternatively, one can follow the composite gradient descent procedure of Loh and Wainwright (2015) - see their section 4 -, which consists in a three step updating procedure of the optimized parameter. To manage

the positive-definitness constraint of Θ, one could include an alternating direction method of multipliers step in the same spirit than the ADMM algorithm provided in Appendix 3 of Bien and Tibshirani (2011).

4.1 Sensitivity analysis

We propose a sensitivity analysis of the statistical consistency and the theoretical error bound with respect to λ_T. We consider two penalization rates: $4.5\sqrt{\frac{\log(d^2)}{T}}$ and $6.5\sqrt{\frac{\log(d^2)}{T}}$, where $d = N(p+1)$. As for the side constraint parameter R, we select $R = \frac{2}{\lambda_T}\boldsymbol{p}(\lambda_T, \text{vec}(\Theta_0))$ to ensure the feasibility of Θ_0 following Loh and Wainwright (2015, 2017). For general data sets, R cannot be computed since the true underlying model is unknown, so that a data-driven method such as cross-validation is required. We set $N = 15$ and $p = 2$ so that $d = 45$ and the total number of parameters in Θ is 2025 (idest 1035 distinct elements). The total number of zero components is 1508 so that $k_0 = 517$, and $\|\text{vec}(\Theta_0)\|_1 = 91.12$, $\|\Theta_0\|_F = 7.81$, $\|\Theta_0\|_{\max} = 1.013, \|\Theta_0\|_s = 2.101$. The Stein's (resp. von Neumann) loss is calibrated for $\eta = 1$ (resp. $\eta = 0$). As for the RSC parameters, to satisfy $4\alpha_1 > 3\mu$, we consider the setting:

(i) *Stein's loss:* $b_{\text{scad}} = 30, b_{\text{mcp}} = 20$. When $\eta = 1$, then $\alpha_1 = (\lambda_{\max}(\Theta_0) + 1)^{-2} = 0.104$. For the SCAD, $4\alpha_1 - 3\mu = 0.313$; for the MCP, $4\alpha_1 - 3\mu = 0.266$; for the Lasso, $4\alpha_1 - 3\mu = 0.4159$.

(ii) *Least squares loss:* $\alpha_1 = 2$. We choose $b_{\text{scad}} = 30, b_{\text{mcp}} = 20$. For the SCAD, $4\alpha_1 - 3\mu = 7.896$; for the MCP $4\alpha_1 - 3\mu = 7.850$; for the Lasso, $4\alpha_1 - 3\mu = 8$.

(iii) *von Neumann loss:* for $\eta = 0$, then $\alpha_1 = \lambda_{\min}(\widehat{S}^{-1}/\nu)(\lambda_{\max}(\Theta_0) + 1)^{-2} = 0.0167$, with $d = 45$. We set $\nu = 1$ for the sake of clarification to compute the theoretical bounds. To evaluate $\lambda_{\min}(\widehat{S}/\nu)$, we simulated 100 times the DGP for $T = 20000$ and took the average of these hundred minimum eigenvalues, so that $\lambda_{\min}(\widehat{S}^{-1}/\nu) = 0.079$ and $\alpha = 0.0082$. To satisfy $4\alpha_1 > 3\mu$, we chose $b_{\text{scad}} = 110, b_{\text{mcp}} = 100$ so that: for the SCAD, $4\alpha_1 - 3\mu = 0.0051$; for the MCP $4\alpha_1 - 3\mu = 0.0026$; for the Lasso, $4\alpha_1 - 3\mu = 0.0326$.

(iv) *D-trace loss:* $\alpha_1 = \lambda_{\min}(\widehat{S})$; based on 100 simulations of the DGP for $T = 20000$, we obtained 100 estimates $\widehat{S}$ and computed the average of these 100 minimum eigenvalues and obtained $\alpha_1 = 0.557$.

We set $b_{\text{scad}} = 30, b_{\text{mcp}} = 20$. For the SCAD, $4\alpha_1 - 3\mu = 2.005$; for the MCP, $4\alpha_1 - 3\mu = 1.958$; for the Lasso, $4\alpha_1 - 3\mu = 2.108$.

We consider samples with sizes $T = 500, 1000, 1500, \cdots, 20000$, and for each sample size, we simulate 100 times the process $(\boldsymbol{Y}_t)$. For each of these 100 simulated paths, we consider the sample variance covariance matrix $\widehat{S} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{X}_t \boldsymbol{X}_t^\top$. Thus, for each sample size, we obtain 100 LASSO/SCAD/MCP sparsity-based estimators $\widehat{\Theta}^{\text{g}}, \widehat{\Theta}^{\text{ls}}, \widehat{\Theta}^{\text{vn}}, \widehat{\Theta}^{\text{dt}}$. For the Stein (resp. von Neumann) case, we set $\eta = 1$ (resp. $\eta = 1$). The $\|.\|_2$ (resp. $\|.\|_1$) consistency patterns and the theoretical upper bounds for $\lambda_T = 4.5\sqrt{\log(d^2)/T}$ and $\lambda_T = 6.5\sqrt{\log(d^2)/T}$ are reported in Panels 1a-1f (resp. Panels 2a-2f). For all cases, the consistency is more favorable when using the rate $2\sqrt{\log(d^2)/T}$, which is in line with the theoretical upper bounds that depend on λ_T: using a tighter rate provides tighter bounds as depicted in our figures. In small samples, the least squares based estimator is less precise but performs well for large samples. Note that for the ℓ_1-consistency and $\lambda_T = 4.5\sqrt{\log(d^2)/T}$, only the theoretical upper bound for the LASSO is reported: for all other cases, a large sample size $T >> 20000$ would be required to reach from above $\|\text{vec}(\Theta_0)\|_1$. Panels 3a-3c report the consistency patterns for all loss based estimators for $\lambda_T = 4.5\sqrt{\log(d^2)/T}$. The von Neumann based estimators are close to each other since the $b_{\text{scad}}, b_{\text{mcp}}$ parameters are large so that the resulting penalization rate behaves as a LASSO one. A significant issue is how "informative" these error bounds are. Their rates depend on the regularization parameter λ_T, on the curvature of the loss function through the RSC parameters α_1 and the non-convexity of the penalty, where the trade-off expressed through the constraint $4\alpha_1 > 3\mu$ is key in our theoretical analysis. For the von Neumann loss, $4\alpha_1 > 3\mu$ is satisfied for significantly large values of b_{scad} and b_{mcp} so that μ is small enough compared to α_1. Hence the denominator for the von Neumann case implies that the upper bounds for both the $\|.\|_1$ and $\|.\|_2$ errors are the least informative. For the ℓ_2-consistency, it would require a dramatically large sample size T for the LASSO theoretical upper bound to cut the $\|\Theta_0\|_F$ line from above. As for the Stein's loss, it would require a sample size $T > 200000$ for the LASSO theoretical upper bound to cut the $\|\Theta_0\|_F$ line from above and

is thus not reported in Panels 1a-1f. This feature changes when considering the least squares and D-trace loss functions, where the RSC parameter α_1 is large enough so that the denominator becomes larger and the theoretical upper bounds become informative.

4.2 An illustration of the support recovery property

In this subsection, the number of zero and non-zero coefficients are expressed with respect to vech(Θ_0). We set $N = 30$ and $p = 2$ so that $d = 90$ and the total number of parameters in Θ is 8100 (idest 4095 distinct elements). The total number of zero components in vech(Θ) is 3676 so that the number of non-zero coefficients in vech(Θ) is 419. Here, $\|\text{vec}(\Theta_0)\|_1 = 141.35$, $\|\Theta_0\|_F = 10.28$, $\|\Theta_0\|_{\max} = 1.001, \|\Theta_0\|_s = 2.30$. Note that the total number of parameters in B_0 (resp. B_1, B_2; resp. vech(Σ_u)) is 788 (resp. 1664; resp. 394). Using the proposed sparse estimation to recover $\mathcal{A}$, for a sample size T, we simulated (20) a hundred times and assess the ability to correctly identify the support. We report in Table 1 the variable selection performance through the number of zero coefficients correctly estimated, denoted as C, the number of zero coefficients incorrectly estimated (i.e. an estimated zero coefficient whereas the true parameter is non-zero), denoted as IC1, the number of nonzero coefficients incorrectly estimated (i.e. an estimated non-zero coefficient whereas the true parameter is zero), denoted IC2, averaged for these hundred batches. The mean squared error is reported as an estimation accuracy measure. The Stein's (resp. von Neumann) loss is calibrated with $\eta = 1$ (resp. $\eta = 0$). The $b_{\text{scad}}, b_{\text{mcp}}$ are set so that $4\alpha_1 - 3\mu > 0$ is still satisfied: $b_{\text{scad}} = 50, b_{\text{mcp}} = 45$ (resp. $30, 20$, resp. $150, 155$, resp. $50, 45$) for the Stein's (resp. Least squares, resp. von Neumann, resp. D-Trace) loss. The regularization parameter λ_T is set as $c\sqrt{\log(d^2)/T}$, with $c = 4$ a value calibrated by a cross-validation (CV) procedure and selected as optimal for the Stein's loss and D-Trace loss for $T = 30000$. For the sake of clarification, we set the same value for all losses. We used the data-dependent 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. Our simulation results show the challenge to perfectly recover the sparse model for all sparse estimators. First, Θ_0 contains a large number

of small non-zero coefficients, where $\min_{i \in \mathcal{A}} |\text{vech}(\Theta_0)| = 0.0022$ and the number of coefficients in absolute value smaller than 0.05 is 223. This mainly justifies the IC1 figures. For all cases, the LASSO provides higher MSE, which agrees with the property that the LASSO penalizes all coefficients with the same rate, thus generating a larger bias. We note that for larger sample sizes, the recovery becomes more precise since both IC1 and IC2 diminish.

4.3 Application to real data

To assess the relevance of the new estimation technique, we consider a VAR model based on a vector of squared stock returns to obtain forecasts using the three loss functions with the three penalties. We use the stocks listed in the Dow Jones Industrial Average, excluding Dow Inc. ($N = 29$) for the period starting from February 18, 2010 to June 19, 2018, providing 2100 observations. Dow Inc. is left aside since it went public on April 1st, 2019. The last 100 observations are used to carry out the forecasting analysis, where we apply a rolling window with sample size $T = 2000$. We selected the regularization parameters $\lambda_T = \sqrt{\log((p+1)^2 N^2)/T}$ for the Gaussian and the D-Trace loss functions and $\lambda_T = 0.2\sqrt{\log((p+1)^2 N^2)/T}$ for the LS and von Neumann loss functions: to do so, we applied a cross validation over the 2000 first observations and then applied the same regularization rate over the next 100 forecasting periods. As a benchmark, we estimated the VAR model (1) with $p = 1, \ldots, 5$ by the ordinary least squares (OLS) method. Table 2 presents the number of parameters in the model and AIC, indicating that the VAR(1) model has the minimum AIC.

We obtained the one-step-ahead forecasts for the VAR(1) model in order to provide the mean squared forecast error (MSFE) and the sample mean of the number of non-zero parameter estimates in B_0 and B_1 for the 100 forecasting period. 3 shows the forecasting results. The minimum number of non-zero parameters is represents less than 15 percent of the standard non-sparse OLS case. On the other hand, the sparse modelling of SVAR models have smaller MSFEs than the traditional VAR has. Among the proposed penalized losses, the von Neumann with SCAD has the minimum MSFE. Furthermore, to examine the

significance of the forecast loss, Table 3 presents p-value for the Model Confidence Set (MCS) of Hansen, Lunde and Nason (2011). The MCS procedure selects a set of models that contains the best model with 95%. Except for OLS and the Gaussian loss function with LASSO, all approaches are included in the same confidence set. Except for these LASSO cases, the new approach provides significantly superior results than the OLS does. The results for the real data analysis support the relevance of the method over the standard OLS method.

References

Abadir, K.M. and J.R, Magnus (2005) *Matrix algebra*. Cambridge University Press.

Ahelegbey, D. F., M. Billio and R. Casarin (2016) Bayesian graphical models for structural vector autoregressive processes. *Journal of Applied Econometrics*, **31**, 357–386.

Alquier, P., K. Bertin, P. Doukhan and R. Garnier (2020) High dimensional VAR with low rank transition. *Statistics and Computing*, **30**, 1139–1153.

Basu, S. and G. Michailidis (2015) Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, **43**, No. 4, 1535–1567.

Bickel, P. J. and E. Levina (2008a) Covariance regularization by thresholding. *The Annals of Statistics*, **36**, 2577–2604.

Bickel, P. J. and E. Levina (2008b) Regularized estimation of large covariance matrices. *The Annals of Statistics*, **36**, 199–227.

Bien, J. and R. Tibshirani (2011) Sparse Estimation of a Covariance Matrix. *Biometrika*, **98**, 807–820.

Blanchard, O. J. and D. Quah (1989) The dynamic effects of aggregate demand and supply disturbances. *American Economic Review*, **79**, 655–673.

Candès, E.J. and Y. Plan (2009) Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, **37**, No. 5A, 2145-2177.

Dahlhaus, R. and M. Eichler (2000) Causality and graphical models for time series. In P. Green, N. Hjort, and S. Richardson, eds., *Highly Structured Stochastic Systems*, Oxford University Press.

Dedecker, J. and X. Fan (2015) Deviation inequalities for separately Lipschitz functionals of iterated random functions. *Stochastic Processes and their Applications*, **125**, No. 1, 60–90.

Fan, J., Y. Feng and Y. Wu (2009) Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, **3**, No. 2, 521–541.

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

Fan, J., Y. Liao and M. Mincheva (2011) High-dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, **39**, No. 6, 3320–3356

Hansen, P. R., Lunde, A., anf J. M. Nason (2011) The model confidence set. *Econometrica*, **79**, 453–497.

Huang, J., N. Liu, M. Pourahmadi, and L. Liu (2006) Covariance matrix selection and estimation via penalized normal likelihood. *Biometrika*, **93**, 85–98.

Inoue, A. and L. Killian (2013) Inference on impulse response functions in structural VAR models. *Journal of Econometrics*, **177**, 1–13.

Johnston, J. (1972) *Econometric Methods*, 2nd ed., McGraw-Hill: New York.

Lam, C. and J. Fan (2009) Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, **37**, 4254–4278.

Loh, P.L. and M.J. Wainwright (2015) regularized M-estimators with non-convexity: statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, **16**, 559-616.

Loh, P.L. and M.J. Wainwright (2017) Support recovery without incoherence: a case for non-convex regularization. *The Annals of Statistics*, **45**, No. 6, 2455-2482.

Lütkepohl, H. (2006) *New Introduction to Multiple Time Series Analyais*, New York: Springer-Verlag.

Lütkepohl, H. (2017) Estimation of structural vector autoregressive models. *Communications for Statistical Applications and Methods*, **24**, 421-441.

Merlevède F., M. Peligrad and E. Rio (2011) A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, **151**, 435-474.

Negahban, S.N, P. Ravikumar, M.J. Wainwright and B. Yu (2012) A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, **27**, No. 4, 538-557.

Oxley, L., M. Reale and G. Tunnicliffe-Wilson (2009) Constructing Structural VAR Models with Conditional Independence Graphs. *Mathematics and Computers in Simulation*, **79**, 2910–2916.

Poignard, B. and J.D. Fermanian (2021) Finite sample properties of Sparse M-estimators with Pseudo-Observations. To appear in *Annals of the Institute of Statistical Mathematics*.

Poignard, B. and Y. Terada (2020) Statistical Analysis of Sparse Approximate Factor Models. *Electronic Journal of Statistics*, **14**, 3315–3365.

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

Ravikumar, P., M.J. Wainwright, G. Raskutti and B. Yu (2011) High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, **5**, 935–980.

Rothman, A. J., P. J. Bickel, E. Levina, and J. Zhu (2008) Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, **2**, 494–515.

Rothman, A. J., E. Levina, and J. Zhu (2009) Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, **104**, 177–186.

Rothman, A. J., E. Levina, and J. Zhu (2010) A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika*, **97**, 539–550.

Sims, C. A. (1980) Macroeconomics and Reality. *Econometrica*, **48**, 1–48.

Tunnicliffe-Wilson, G. and M. Reale (2008) The sampling properties of conditional independence graphs for I(1) structural VAR models. *Journal of Time Series Analysis*, **29**, 802–810.

Waggoner, D. F. and T. Zha (2003) Likelihood preserving normalization in multiple equation models. *Journal of Econometrics*, **114**, 329–347.

Wainwright, M.J. (2009) Sharpe thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, **55**, No. 5, 2183-2202.

White, H. (2001) *Asymptotic Theory for Econometricians*, 2nd edition, Emerald, UK

Wong, K.C., Z. Li and A. Tewari (2020) Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, **48**, No. 2, 1124–1142.

Wu, Y. and T. Li (2020) Differential network inference via the fused D-trace loss with cross variables. *Electronic Journal of Statistics*, **14**, 1269–1301.

Wu, W. B. and M. Pourahmadi (2003) Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, **90** 831–844.

Yuan, M. and Y. Lin (2007) Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19–35.

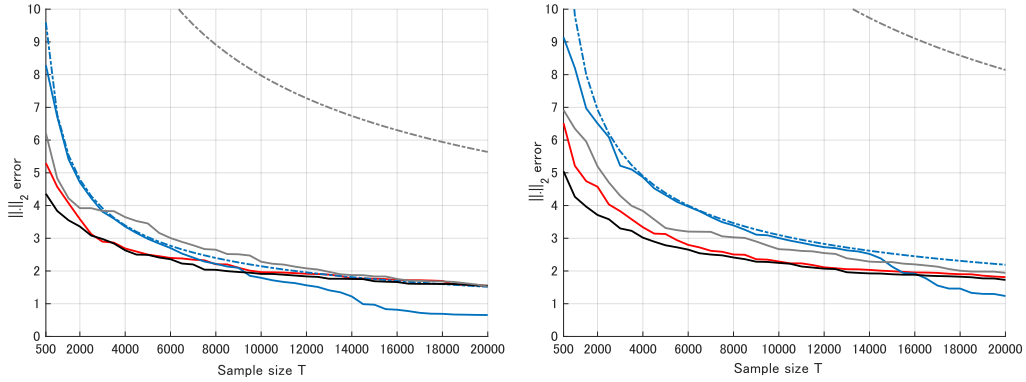
Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**, 894-942.

Zhang, T. and H. Zou (2014) Sparse precision matrix estimation via lasso penalized D-trace loss. *Biometrika*, **101**, 103–120.

Zhao, P. and B. Yu (2006) On model selection consistency of Lasso. *Journal of Machine Learning Research*, **7**, 2541-2567.

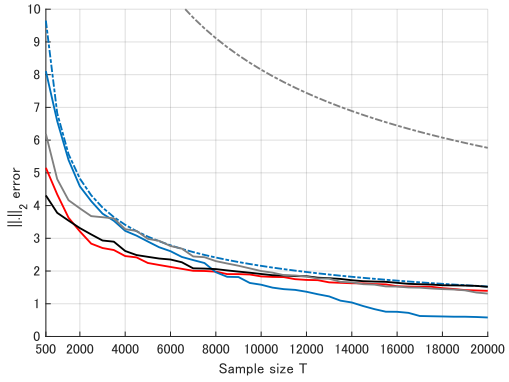
5 Technical appendix

5.1 Figures and Tables

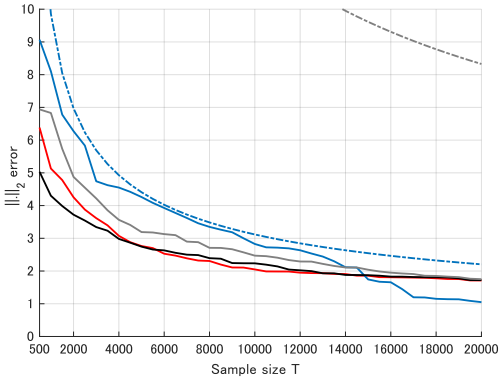


(a) SCAD - $\lambda_T = 4.5\sqrt{\log(d^2/T)}$

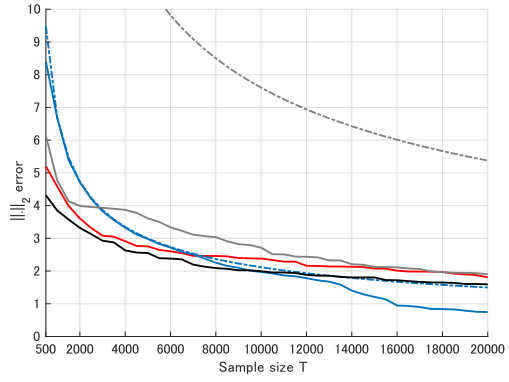
(b) SCAD - $\lambda_T = 6.5\sqrt{\log(d^2/T)}$



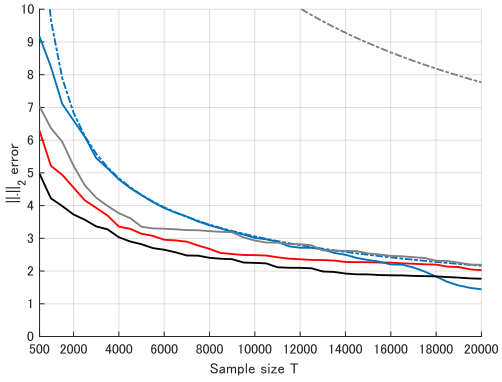
(c) MCP - $\lambda_T = 4.5\sqrt{\log(d^2/T)}$



(d) MCP - $\lambda_T = 6.5\sqrt{\log(d^2/T)}$

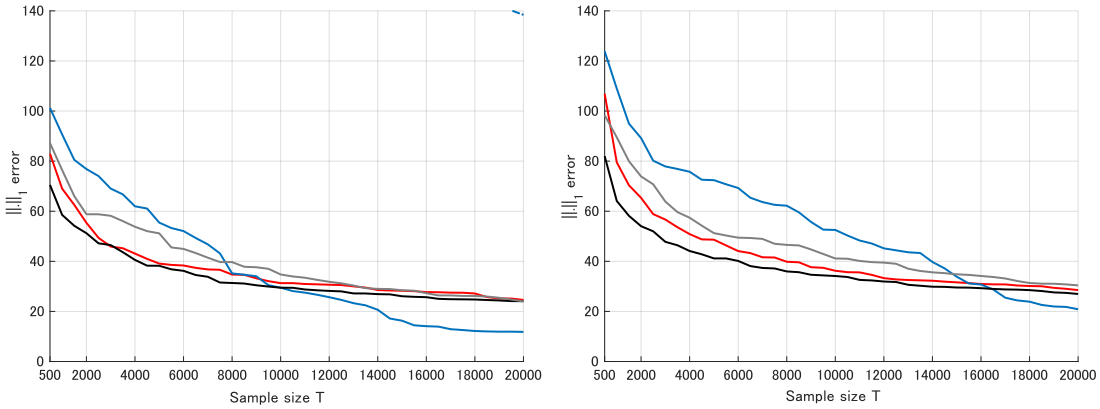


(e) LASSO - $\lambda_T = 4.5\sqrt{\log(d^2/T)}$



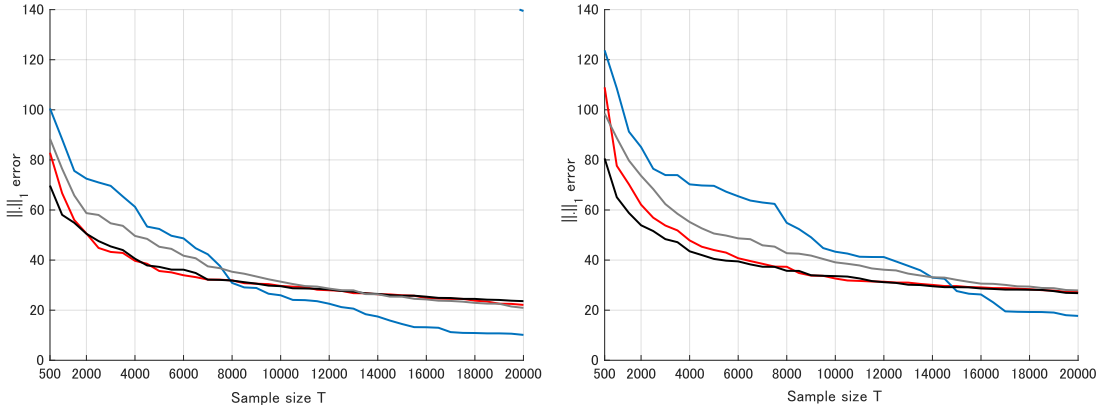
(f) LASSO - $\lambda_T = 6.5\sqrt{\log(d^2/T)}$

Figure 1: $\|.\|_2$ consistency for the setting of Subsection 4.1 for $\lambda_T = 4.5\sqrt{\log(d^2)/T}$ and $\lambda_T = 6.5\sqrt{\log(d^2)/T}$. The Stein, least squares, von Neumann and D-trace cases are represented in solid red, blue, black and gray respectively. Each point represents an average of 200 trials for each sample size. The corresponding theoretical upper bounds are in dashed-dotted lines.



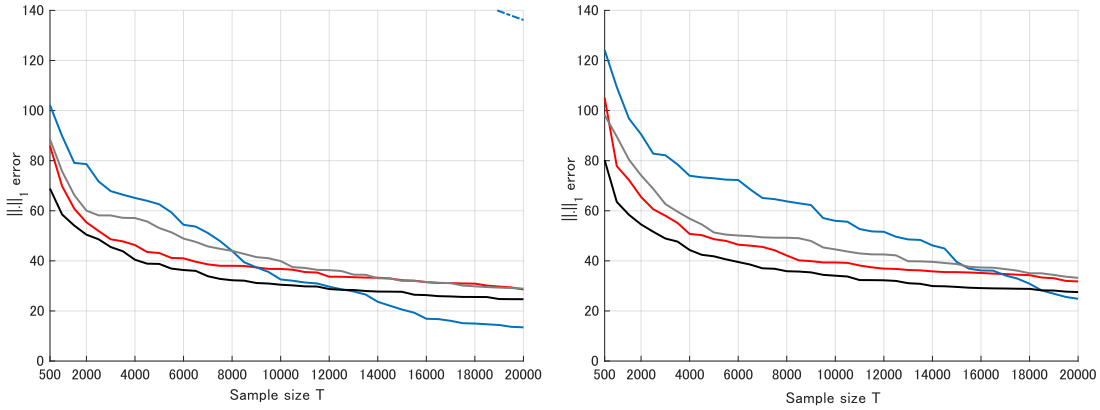
(a) SCAD - $\lambda_T = 4.5\sqrt{\log(d^2/T)}$

(b) SCAD - $\lambda_T = 6.5\sqrt{\log(d^2/T)}$



(c) MCP - $\lambda_T = 4.5\sqrt{\log(d^2/T)}$

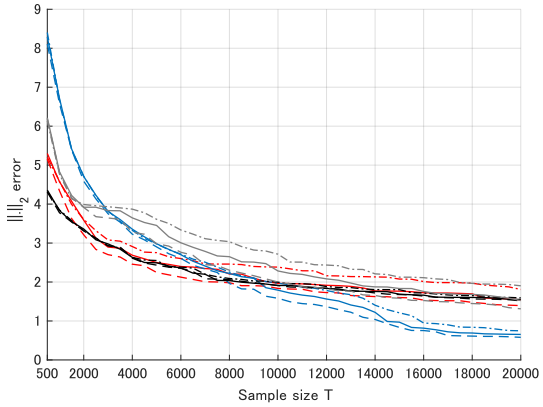
(d) MCP - $\lambda_T = 6.5\sqrt{\log(d^2/T)}$



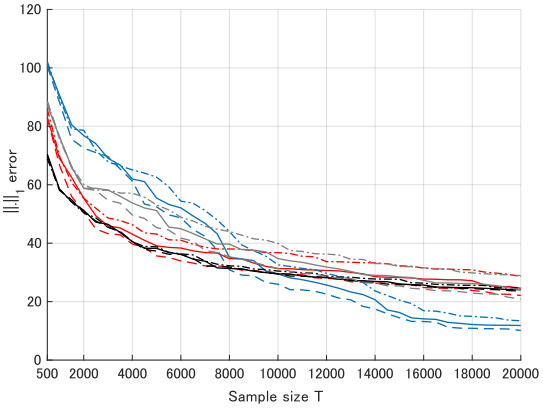
(e) LASSO - $\lambda_T = 4.5\sqrt{\log(d^2/T)}$

(f) LASSO - $\lambda_T = 6.5\sqrt{\log(d^2/T)}$

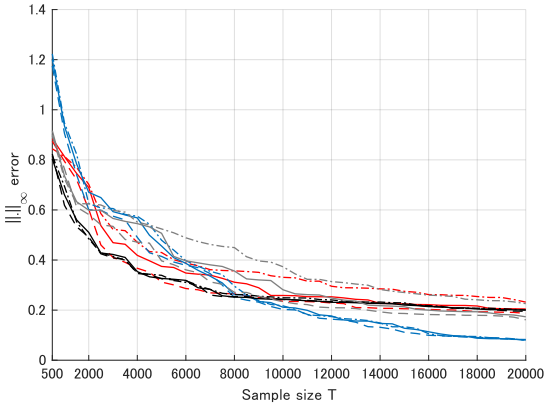
Figure 2: $\|.\|_1$ consistency for the setting of Subsection 4.1 for $\lambda_T = 4.5\sqrt{\log(d^2)/T}$ and $\lambda_T = 6.5\sqrt{\log(d^2)/T}$. The Stein, least squares, von Neumann and D-trace cases are represented in solid red, blue, black and gray respectively. Each point represents an average of 200 trials for each sample size. The corresponding theoretical upper bounds are in dashed-dotted lines.



(a) $\|.\|_2$-consistency



(b) $\|.\|_1$-consistency



(c) $\|.\|_\infty$-consistency

Figure 3: $\|.\|_2, \|\|_1, \|.\|_\infty$ consistencies for the setting of Subsection 4.1 for $\lambda_T = 4.5\sqrt{\log(d^2)/T}$. The Stein, least squares, von Neumann and D-trace cases are represented in red, blue, black and gray color respectively. For each loss, the SCAD (resp. MCP, resp. LASSO) is in solid (resp. dashed, resp. dashed-dotted) line. Each point represents an average of 200 trials for each sample size.

Table 1: *Model selection and precision accuracy for $N = 30$ and $p = 2$ based on 100 replications. For each penalized Stein, Least squares, Von Neumann and D-trace, the penalty cases are reported according to the order: SCAD, MCP and LASSO*

	Truth	Stein	Least-squares	von Neumann	D-trace
$T = 5000$					
C	3676	3469.1 − 3458.8 − 3464.8	3535.4 − 3523.2 − 3535.1	3432.9 − 3422.5 − 3426.6	3462.6 − 3463.6 − 3462.3
IC1	0	222.5 − 221.3 − 222.5	217.8 − 218.6 − 218.5	222.6 − 220.9 − 221.8	223.7 − 222.2 − 224.9
IC2	0	206.9 − 217.1 − 211.2	140.6 − 152.8 − 140.9	243.1 − 253.5 − 249.4	213.4 − 212.7 − 213.7
MSE		3.291 − 3.342 − 3.549	7.085 − 7.522 − 7.058	1.462 − 1.459 − 1.418	3.267 − 3.347 − 3.598
$T = 10000$					
C	3676	3603.4 − 3610.2 − 3609.9	3602.1 − 3601.9 − 3607.7	3583.5 − 3574.2 − 3579.8	3612.6 − 3601.8 − 3602.1
IC1	0	216.1 − 214.7 − 217.2	210.3 − 206.5 − 219.6	215.9 − 211.7 − 220.8	217.7 − 208.1 − 221.9
IC2	0	72.5 − 65.8 − 66.0	73.9 − 74.1 − 68.3	92.5 − 101.7 − 96.2	63.4 − 74.2 − 73.9
MSE		2.332 − 2.341 − 2.787	5.663 − 6.014 − 6.405	0.970 − 0.967 − 1.211	1.951 − 1.955 − 2.329
$T = 20000$					
C	3676	3643.5 − 3643.6 − 3650.3	3639.1 − 3637.4 − 3644.3	3649.9 − 3650.9 − 3648.7	3653.5 − 3654.2 − 3657.3
IC1	0	192.4 − 183.8 − 203.7	193.6 − 180.2 − 196.1	189.3 − 179.8 − 196.2	184.2 − 171.8 − 193.6
IC2	0	32.4 − 32.3 − 35.7	36.9 − 38.6 − 41.6	26.1 − 25.0 − 27.3	22.5 − 21.8 − 28.6
MSE		1.446 − 1.491 − 1.964	2.554 − 2.514 − 3.242	0.643 − 0.637 − 0.758	0.961 − 1.012 − 1.391
$T = 30000$					
C	3676	3665.1 − 3670.2 − 3674.9	3668.6 − 3673.4 − 3674.4	3675.7 − 3675.5 − 3675.6	3665.1 − 3665.9 − 3675.8
IC1	0	172.8 − 161.8 − 188.2	187.7 − 170.2 − 188.9	182.3 − 171.1 − 191.3	178.1 − 163.4 − 186.7
IC2	0	16.9 − 17.7 − 19.1	7.4 − 2.5 − 1.6	0.3 − 0.5 − 0.8	10.9 − 7.3 − 0.2
MSE		0.185 − 0.226 − 0.396	0.254 − 0.156 − 0.293	0.113 − 0.110 − 0.163	0.102 − 0.117 − 0.202

Table 2: OLS Results

VAR(p)	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$
# of parameters	1276	2117	2958	3799	4640
AIC	16673.7	16841.8	17009.9	17178.1	17346.2

Table 3: Sparse VAR Forecasts

Loss	Penalty	# of Non-Zero	MSFE	MCS
Gaussian	SCAD	180.0	448.01	0.080
Gaussian	MCP	181.8	443.88	0.080
Gaussian	LASSO	180.5	451.76	0.035
LS	SCAD	479.0	442.37	0.080
LS	MCP	477.3	441.44	0.080
LS	LASSO	488.1	441.77	0.080
D-Trace	SCAD	359.7	435.42	0.454
D-Trace	MCP	356.7	436.31	0.080
D-Trace	LASSO	359.7	435.27	0.454
von Neumann	SCAD	296.6	431.94	1.000
von Neumann	MCP	289.9	432.63	0.454
von Neumann	LASSO	291.9	432.02	0.911
OLS	—	1276	457.65	0.035

Note: The results are based on the one-step-ahead forecasts for the last 100 observations using the rolling window with the sample size $T = 2000$. '# of Non-Zero' indicates the sample mean of the number of non-zero estimates for B_0 and B_1 in the (sparse) VAR(1) model. 'MSFE' denotes the mean squared forecasts error. The column labeled MCS presents the p-values associated with the Model Confidence Set of Hansen et al. (2011).

5.2 Intermediary results

We provide the primal dual witness method as in Loh and Wainwright (2017), an approach that relies on the following steps. Here, the parameter of interest is $\theta \in \mathbb{R}^q$ and q denotes the dimension. The problem of interest is a regularized M-estimation one, where a generic loss $\mathbb{G}_T(.)$ is penalized by $\boldsymbol{p}(\lambda_T, .)$:

$$\widehat{\theta} = \arg\min_{\theta \in \Omega} \Big\{ \mathbb{G}_T(\theta) + \boldsymbol{p}(\lambda_T, \theta) \Big\}, \quad \Omega = \Big\{ \theta \in \Theta \subseteq \mathbb{R}^q, \|\theta\|_1 \leq R \Big\}. \tag{21}$$

The loss function $\mathbb{G}_T(.)$ satisfies the RSC condition and $\boldsymbol{p}(\lambda_T, .)$ assumption 2.

Step 1. We define the estimator

$$\widehat{\theta}_{\mathcal{A}} = \underset{\theta \in \mathbb{R}^{|\mathcal{A}|} : g(\theta) \leq R, \theta \in \Omega}{\arg\min} \Big\{ \mathbb{G}_T(\theta) + \boldsymbol{p}(\lambda_T, \theta) \Big\}. \tag{22}$$

We solve problem (22), under the constraint $\widehat{\mathcal{A}} \subseteq \mathcal{A}$ and prove $\|\widehat{\theta}_{\mathcal{A}}\|_1 < R$.

Step 2. Defining $\widehat{\boldsymbol{z}}_{\mathcal{A}} \in \partial \|\widehat{\theta}_{\mathcal{A}}\|$, we choose $\widehat{\boldsymbol{z}}_{\mathcal{A}^c}$ satisfying the orthogonality condition

$$\nabla_\theta \mathbb{G}_T(\widehat{\theta}) - \nabla_\theta \boldsymbol{q}(\lambda_T, \widehat{\theta}) + \lambda_T \widehat{\boldsymbol{z}} = 0, \tag{23}$$

with $\widehat{\boldsymbol{z}} = (\widehat{\boldsymbol{z}}_{\mathcal{A}}, \widehat{\boldsymbol{z}}_{\mathcal{A}^c})$, $\widehat{\theta} = (\widehat{\theta}_{\mathcal{A}}, 0_{\mathcal{A}^c})$, $\boldsymbol{q}(\lambda_T, \rho) = \lambda_T \rho - \boldsymbol{p}(\lambda_T, \rho)$. We prove strict dual feasibility $\|\widehat{\boldsymbol{z}}_{\mathcal{A}^c}\|_\infty < 1$.

Step 3. We prove that $\widehat{\theta}$ is a local optimum of (21) and that any stationary point of (21) satisfies $\text{supp}(\widehat{\theta}) \subseteq \mathcal{A}$.

The PDW procedure does not allow for *practically* solving the regularization problem (21) as step 1 requires to know the true subset model $\mathcal{A}$. However, this approach is useful as a proof method to characterize the optimal solution $\widehat{\Theta}$. In **Step 1**, the criterion (22) is striclty convex under the RSC condition. This implies that for $\|\widehat{\theta}_{\mathcal{A}}\|_1 < 1$, the subgradient condition (23) must hold at $\widehat{\theta}_{\mathcal{A}}$ for the restricted problem (22). Loh and Wainwright (2017) proves that, although problem (22) may be non-convex, the RSC condition and regularity conditions on the penalty function allow them to prove that the optimum obtained in **Step 3** is a local optimum: see in particular their Lemma 10.

Using optimization reasoning, Loh and Wainwright (2017) provide conditions on λ_T, R to ensure the success of the PDW technique, which depends on **Step 3**, under the assumption that $\mathbb{G}_T(.)$ satisfies the RSC condition with parameters $(\alpha_k, \tau_k)_{k=1,2}$ and $4\alpha_1 > 3\mu$. Indeed, these conditions guarantee that the support of $\widehat{\theta}$ satisfying (23) in **Step 2** is the unique stationary point of the criterion (21): to be precise, the first condition concerns the suitable scaling of λ_T and R; the second condition ensures strict dual feasibility - that is $\|\widehat{\boldsymbol{z}}_{\mathcal{A}^c}\|_\infty < 1$ in **Step 2**. This is the object of the following Theorem.

Theorem 5.1. Loh and Wainwright (Theorem 1, 2017)

Suppose $\mathbb{G}_T(.)$ satisfies the RSC condition with $(\alpha_k, \tau_k)_{k=1,2}$ parameters and $\boldsymbol{p}(\lambda_T, .)$ is a μ-amenable penalty, with $0 \leq \mu < \alpha_1$. Suppose

(i) The parameters (λ_T, R) satisfy

$$4\max\Big\{\|\nabla_\theta \mathbb{G}_T(\theta_0)\|_\infty, \alpha_2\sqrt{\frac{\log(k_0)}{T}}\Big\} \leq \quad \lambda_T \quad \leq \sqrt{\frac{(4\alpha_1 - 3\mu)\alpha_2}{384 k_0}}, \tag{24}$$

$$\max\Big\{2\|\theta_0\|_1, \frac{48 k_0 \lambda_T}{4\alpha_1 - 3\mu}\Big\} \leq \quad R \quad \leq \min\{\frac{\alpha_2}{8\lambda_T}, \frac{\alpha_2}{\tau_2}\sqrt{\frac{T}{\log(q)}}\}. \tag{25}$$

(ii) For some $\delta \in [\frac{4R\tau_1 \log(q)}{T\lambda_T}, 1]$, the vector $\widehat{\boldsymbol{z}}$ from the PDW construction satisfies the strict dual feasibility condition

$$\|\widehat{\boldsymbol{z}}_{\mathcal{A}^c}\|_\infty \leq 1 - \delta. \tag{26}$$

Then for any k_0-sparse vector θ_0, the program (21) with a sample size $T \geq \frac{2\tau_1}{2\alpha_1 - \mu} k_0 \log q$ has a unique stationary point given by the primal output $\widehat{\theta}$ of the PDW construction.

Suitable calibrations of λ_T and R, and thus proper scaling behaviours (T, q, k_0), are necessary. Using exponential bounds, it is possible to evaluate the probability of satisfying (24) and (25) and thus the probability of the PDW success. In all their applications of interest - linear model, generalized linear model, Gaussian graphical Lasso - Loh and Wainwright (2017) obtain the upper bound $\|\nabla_\theta \mathbb{G}_T(\theta_0)\|_\infty \leq C\sqrt{\log(q)/T}$ with high probability. This motivates the choice λ_T proportional to $\sqrt{\log(q)/T}$ to satisfy

(24). Finally, it is worth noting that the trade-off between the curvature of the loss function through α_1 and the non-convexity degree of the penalty function through μ appears. As our simulations emphasize this trade-off for the Stein or von Neumann loss in particular, significantly large values for $b_{\text{scad}}, b_{\text{mcp}}$ are necessary to ensure $4\alpha_1 > 3\mu$.

In their Theorem 2, Loh and Wainwright (2017) provide an additional error bound under the conditions of Theorem 5.1. It also provides the guarantees that the unique optimum - local or global - (21) is the oracle estimator. The latter is defined as the non-penalized estimator obtained from minimizing the criterion $\mathbb{G}_T(\theta)$ over the true support $\mathcal{A}$. This is the object of the following Theorem.

Theorem 5.2. Loh and Wainwright (Theorem 2, 2017)

Under the conditions of Theorem 5.1, suppose strict dual feasibility (26) holds, suppose $\boldsymbol{p}(\lambda_T, .)$ *is* μ*-amenable with* $\mu \in [0, \alpha_1)$*. Then the unique stationary solution of (21) satisfies*

(i)

$$\|\widehat{\theta} - \theta_0\|_\infty \leq \|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} \nabla_\theta \mathbb{G}_T(\theta_0)_\mathcal{A}\|_\infty + \lambda_T \|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty,$$

with $\widehat{\boldsymbol{K}} = \int_0^1 \nabla^2_{\theta\theta^\top} \mathbb{G}_T(\theta_0 + u(\widehat{\theta} - \theta_0))du$.

(ii) If $\boldsymbol{p}(\lambda_T, .)$ *is* (μ, ζ)*-amenable and if the lower bound*

$$\min_{i \in \mathcal{A}} |\theta_{0,i}| \geq \lambda_T \big(\zeta + \|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty\big) + \|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} \nabla_\theta \mathbb{G}_T(\theta_0)_\mathcal{A}\|_\infty,$$

holds, then $\widehat{\theta}$ *agrees with the oracle estimator* $\widehat{\theta}^{\mathcal{O}}$ *and we have the bound*

$$\|\widehat{\theta} - \theta_0\|_\infty \leq \|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} \nabla_\theta \mathbb{G}_T(\theta_0)_\mathcal{A}\|_\infty.$$

These inequalities are expressed in a deterministic manner. As in Theorem 5.1, exponential bounds allow for upper bounding $\|\nabla_\theta \mathbb{G}_T(\theta_0)_\mathcal{A}\|_\infty$, which will provide explicit convergence rates over the $\|.\|_\infty$-error. The application of Theorem 5.2 requires that strict dual feasibility holds under the RSC condition. In their Proposition 1, Loh and Wainwright (2017) provide sufficient conditions to satisfy strict dual feasibility

for (μ, ζ)-amenable penalties, which thus allows for using Theorem 5.2. These conditions are given in the following Proposition.

Proposition 5.3. Loh and Wainwright (Proposition 1, 2017)

Under the conditions of Theorem 5.1, suppose $\boldsymbol{p}(\lambda_T, .)$ *is* (μ, ζ)*-amenable. Suppose*

$$\theta_{0,\min} \geq \lambda_T(\zeta + \|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty) + \|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{G}_T(\theta_0)_{\mathcal{A}}\|_\infty,$$

with $\theta_{0,\min} = \min_{i\in\mathcal{A}}|\theta_i|$ *and* $\widehat{\boldsymbol{K}} = \int_0^1 \nabla^2_{\theta\theta^\top}\mathbb{G}_T(\theta_0 + u(\widehat{\theta} - \theta_0))du$*. Then strict dual feasibility holds provided*

$$\|\nabla_\theta \mathbb{G}_T(\theta_0)\|_\infty \leq \frac{1-\delta}{2}\lambda_T, \qquad \textit{and} \tag{27}$$

$$\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{G}_T(\theta_0)_{\mathcal{A}}\|_\infty \quad \leq \quad \frac{1-\delta}{2}\lambda_T. \tag{28}$$

6 Proofs

Proof of Lemma 3.2. Under assumptions 3, 4 and 5, by Lemma A.2 of Fan, Liao and Mincheva (2013), $\mathbf{X}_{i,t}\mathbf{X}_{j,t}$ satisfies the exponential tail condition with parameter $\gamma_2/3$. Thus, by Theorem 1 of Merlevède, Peligrad and Rio (2011), there exist constants C_1, C_2, C_3, C_4, C_5 depending only on c, γ and γ_1 such that, for any positive ϵ and any $i, j \leq d^2$,

$$\begin{aligned}
&\mathbb{P}(|\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_{i,t}\mathbf{X}_{j,t} - \Sigma_{x,ij}| > \epsilon) \\
&\quad \leq \quad T\exp\big(-\frac{(T\epsilon)^\gamma}{C_1}\big) + \exp\big(-\frac{(T\epsilon)^2}{C_2(1+TC_3)}\big) + \exp\big(-\frac{(T\epsilon)^2}{C_4T}\exp\big(\frac{(T\epsilon)^{\gamma(1-\gamma)}}{C_5(\log(T\epsilon))^\gamma}\big)\big),
\end{aligned}$$

so that $\mathbb{P}(\|\widehat{S} - \Sigma_x\|_{\max} > \epsilon) = \mathbb{P}(\max_{1\leq i,j\leq d}|\widehat{S}_{ij} - \Sigma_{x,ij}| > \epsilon) \leq d^2 \max_{1\leq i,j\leq d}\mathbb{P}(|\widehat{S}_{ij} - \Sigma_{x,ij}| > \epsilon)$. We then deduce

$$\begin{aligned}
&\mathbb{P}(|\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_{i,t}\mathbf{X}_{j,t} - \Sigma_{x,ij}| > \epsilon) \\
&\quad \leq \quad d^2\Big\{T\exp\big(-\frac{(T\epsilon)^\gamma}{C_1}\big) + \exp\big(-\frac{(T\epsilon)^2}{C_2(1+TC_3)}\big) + \exp\big(-\frac{(T\epsilon)^2}{C_4T}\exp\big(\frac{(T\epsilon)^{\gamma(1-\gamma)}}{C_5(\log(T\epsilon))^\gamma}\big)\big)\Big\}.
\end{aligned}$$

Let us now fix $\epsilon = L\sqrt{\log(d^2)/T}$ with $L > 0$ a constant large enough. For the scaling behaviour $T \geq K \log(d^2)^{2/\gamma - 1}$ for a sufficiently large and positive constant K, then

$$\begin{aligned} & d^2 T \exp\big(-\frac{(T\epsilon)^\gamma}{C_1}\big) + d^2 \exp\Big(-\frac{(T\epsilon)^2}{C_4 T}\exp\big(\frac{(T\epsilon)^{\gamma(1-\gamma)}}{C_5(\log(T\epsilon))^\gamma}\big)\Big) \\ &= \exp\big(-\frac{T^\gamma L^\gamma(\sqrt{\log(d^2)/T})^\gamma}{C_1} + \log(d^2) + \log(T)\big) + \exp\Big(-\frac{T^2\log(d^2)/T}{C_4 T}\exp\big(\frac{T^{\gamma(1-\gamma)}(L\sqrt{\log(d^2)/T})^{\gamma(1-\gamma)}}{C_5(\log(TL\sqrt{\log(d^2)/T}))^\gamma}\big)\Big) \\ &= o(\exp\big(-\log(d^2)\big)), \end{aligned}$$

and

$$d^2 \exp\big(-\frac{(T\epsilon)^2}{C_2(1+TC_3)}\big) = \exp\big(-\frac{T^2(L\sqrt{\log(d^2)/T})^2}{C_2(1+TC_3)} + \log(d^2)\big) = O(\exp\big(-\log(d^2)\big)).$$

As a consequence, if we fix $\epsilon = L\sqrt{\log(d^2)/T}$, then $\|\widehat{S} - \Sigma_x\|_{\max} \leq L\sqrt{\log(d^2)/T}$ with probability at least $1 - O(\exp\big(-\log(d^2)\big)) - o(\exp\big(-\log(d^2)\big))$. □

Proof of Corollary 3.3. We first establish the RSC property. To do so, we derive the first and second order derivatives of the Stein's loss criterion $\mathbb{L}_{T,\eta}(\Theta)$ defined in (5). Using the differential operator applied with respect to Θ, we obtain

$$\boldsymbol{d}\mathbb{L}_{T,\eta}(\Theta) = (1-2\eta)\Theta^{-1}(\boldsymbol{d}\Theta) + \mathrm{tr}\Big(\eta\widehat{S}(\boldsymbol{d}\Theta) - (1-\eta)\Theta^{-1}(\boldsymbol{d}\Theta)\Theta^{-1}\widehat{S}^{-1}\Big).$$

Hence in vector and matrix forms, the derivatives become

$$\nabla_\theta \mathbb{L}_{T,\eta}(\Theta) = \mathrm{vec}\Big((1-2\eta)\Theta^{-1} + \eta\widehat{S} - (1-\eta)\Theta^{-1}\widehat{S}^{-1}\Theta^{-1}\Big), \quad \nabla_\Theta \mathbb{L}_{T,\eta}(\Theta) = (1-2\eta)\Theta^{-1} + \eta\widehat{S} - (1-\eta)\Theta^{-1}\widehat{S}^{-1}\Theta^{-1}.$$

Note that alternatively, the vech(.) operator could be applied to treat the redundant terms. Taking the $\|.\|_\infty$ norm, over Θ_0, the gradient becomes

$$\|\nabla_\theta \mathbb{L}_T(\Theta_0)\|_\infty = \|\nabla_\Theta \mathbb{L}_T(\Theta_0)\|_{\max} \leq \|(1-2\eta)\Theta^{-1} + \eta\widehat{S} - (1-\eta)\Theta^{-1}\widehat{S}^{-1}\Theta^{-1}\|_{\max}.$$

We now focus on the Hessian matrix. The second order differential is given by

$$\boldsymbol{d}^2\mathbb{L}_{T,\eta}(\Theta) = (2\eta-1)\mathrm{tr}\Big(\Theta^{-1}(\boldsymbol{d}\Theta)\Theta^{-1}(\boldsymbol{d}\Theta)\Big) + (1-\eta)\mathrm{tr}\Big(\Theta^{-1}(\boldsymbol{d}\Theta)\Theta^{-1}\widehat{S}^{-1}\Theta^{-1}(\boldsymbol{d}\Theta) + \Theta^{-1}\widehat{S}^{-1}\Theta^{-1}(\boldsymbol{d}\Theta)\Theta^{-1}(\boldsymbol{d}\Theta)\Big).$$

We aim at extracting the form $\mathrm{tr}(L(\boldsymbol{d}\Lambda)^\top M(\boldsymbol{d}\Lambda))$ for L (resp. M) any square $m\times m$ matrix (resp. $p\times p$). We have

$$\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,\eta}(\Theta)=(2\eta-1)\Big(\Theta^{-1}\otimes\Theta^{-1}\Big)+(1-\eta)\Big\{\Theta^{-1}\otimes\Theta^{-1}\widehat{S}^{-1}\Theta^{-1}+\Theta^{-1}\widehat{S}^{-1}\Theta^{-1}\otimes\Theta^{-1}\Big\}.$$

Now for some $\Theta_1\in\Omega$ and $u\in[0,1]$, let us define $\Theta=\Theta_0+u\Gamma$ with $\Gamma=\Theta_1-\Theta_0$. Then $\Theta\in\Omega$ and

$$\begin{aligned}
f_T(\Theta) &:= \mathrm{vec}(\Gamma)^\top\nabla^2_{\mathrm{vec}(\Theta)\mathrm{vec}(\Theta)^\top}\mathbb{L}_{T,\eta}(\Theta)\mathrm{vec}(\Gamma)\\
&\geq \mathrm{vec}(\Gamma)^\top\Big[(2\eta-1)\Big(\Theta^{-1}\otimes\Theta^{-1}\Big)+(1-\eta)\Big\{\Theta^{-1}\otimes\Theta^{-1}\widehat{S}^{-1}\Theta^{-1}+\Theta^{-1}\widehat{S}^{-1}\Theta^{-1}\otimes\Theta^{-1}\Big\}\Big]\mathrm{vec}(\Gamma)\\
&\geq \|\Gamma\|_F^2\Big[(2\eta-1)\lambda_{\min}(\Theta^{-1}\otimes\Theta^{-1})+(1-\eta)\lambda_{\min}(\Big\{\Theta^{-1}\otimes\Theta^{-1}\widehat{S}^{-1}\Theta^{-1}+\Theta^{-1}\widehat{S}^{-1}\Theta^{-1}\otimes\Theta^{-1}\Big\})\Big]\\
&\geq \|\Gamma\|_F^2\Big[(2\eta-1)\lambda_{\min}(\Theta^{-1})^2+2(1-\eta)\lambda_{\min}(\Theta^{-1})^3\lambda_{\min}(\widehat{S}^{-1})\Big]
\end{aligned}$$

because the spectrum of $A\otimes B$ is the cross product of the spectrums of A and B (see, e.g., Lütkepohl, 1996, Subsection 5.2.1), and $\lambda_{\min}(\Theta)=\inf_{\boldsymbol{x}}\boldsymbol{x}^\top\Theta\boldsymbol{x}/\|\boldsymbol{x}\|_2^2$. We now focus on $\lambda_{\min}(\Theta^{-1})^2$. We have $\lambda_{\max}(\Theta)\leq\lambda_{\max}(\Theta_0)+\lambda_{\max}(\Theta_1-\Theta_0)$, which implies $\lambda_{\min}(\Theta^{-1})^2\geq\{\lambda_{\max}(\Theta_0)+1\}^{-2}$. Therefore

$$f_T(\Theta)\geq\|\Gamma\|_F^2\Big[(2\eta-1)\big\{\lambda_{\max}(\Theta_0)+1\big\}^{-2}+2(1-\eta)\big\{\lambda_{\max}(\Theta_0)+1\big\}^{-3}\lambda_{\min}(\widehat{S}^{-1})\Big].$$

The true vector of parameters is $\theta_0=\mathrm{vec}(\Theta_0)$ and we have $\|\Gamma\|_F^2=\|\Theta-\Theta_0\|_2^2$. As a consequence,

$$(\theta-\theta_0)^\top\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\overline{\Theta})(\theta-\theta_0)\geq\|\theta-\theta_0\|_2^2\Big[(2\eta-1)\big\{\lambda_{\max}(\Theta_0)+1\big\}^{-2}+2(1-\eta)\big\{\lambda_{\max}(\Theta_0)+1\big\}^{-3}\lambda_{\min}(\widehat{S}^{-1})\Big],$$

for any $\overline{\Theta}$ that lies between Θ and Θ_0. Thus, at Θ_0, the RSC condition is satisfied with coefficients $\alpha_1=(2\eta-1)\big\{\lambda_{\max}(\Theta_0)+1\big\}^{-2}+2(1-\eta)\big\{\lambda_{\max}(\Theta_0)+1\big\}^{-3}\lambda_{\min}(\widehat{S}^{-1})$ and $\alpha_2=\alpha_1$, $\tau_1=\tau_2=0$. Should we take $\eta=1$, then $\alpha_1=\big\{\lambda_{\min}(\Theta_0)+1\big\}^{-2}$ since the Hessian simplifies as $\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,\eta}(\Theta)=\Theta^{-1}\otimes\Theta^{-1}$.

We now evaluate the probability of satisfying (4) when $\eta=1$, a situation most commonly used when dealing with inverse variance covariance estimation. Then:

$$\|\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)\|_\infty=\|\nabla_\Theta\mathbb{L}_{T,1}(\Theta_0)\|_{\max}=\|\widehat{S}-\Sigma_x\|_{\max},$$

which provides (6). By Lemma 3.2, we conclude $\|\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)\|_\infty\leq K\sqrt{\frac{\log(d)}{T}}$, with probability at least $1-O(\exp\big(-\log(d^2)\big))-o(\exp\big(-\log(d^2)\big))$ for a sample size $T\geq L\log(d^2)^{2/\gamma-1}$. □

Proof of Corollary 3.4. We first establish the RSC condition for the least squares loss function. Using the differential operator with respect to Θ, we have $\boldsymbol{d}\mathbb{L}_T(\Theta) = 2\mathrm{tr}\Big(\Theta - \widehat{S}^{-1}\Big)(\boldsymbol{d}\Theta)$. Hence

$$\nabla_\theta \mathbb{L}_T(\Theta) = 2\mathrm{vec}(\Theta - \widehat{S}^{-1}), \;\; \nabla_\Theta \mathbb{L}_T(\Theta) = 2\Big(\Theta - \widehat{S}^{-1}\Big).$$

As for the Hessian, by identification, we deduce

$$\nabla^2_{\theta\theta^\top} \mathbb{L}_T(\Theta) = 2\Big(I_d \otimes I_d\Big).$$

We thus deduce $(\theta - \theta_0)^\top \nabla^2_{\theta\theta^\top} \mathbb{L}_T(\overline{\Theta})(\theta - \theta_0) \geq 2\|\theta - \theta_0\|_2^2$, for any $\overline{\Theta}$ that lies between Θ and Θ_0. Thus, at Θ_0, the RSC condition is satisfied with coefficients $\alpha_1 = 2$ and $\alpha_2 = \alpha_1$, $\tau_1 = \tau_2 = 0$.

We now evaluate the probability of satisfying (4). We have

$$\|\nabla_\theta \mathbb{L}_T(\Theta_0)\|_\infty = \|\nabla_\Theta \mathbb{L}_T(\Theta_0)\|_{\max} = 2\|\Theta_0 - \widehat{S}^{-1}\|_{\max} \leq 2\|\Theta_0 - \widehat{S}^{-1}\|_s.$$

Now, with high probability, for a sufficiently large constant K, $\|\widehat{S} - \Sigma_x\|_{\max} \leq K\sqrt{\frac{\log(d^2)}{T}}$, and we aim at bounding $\|\widehat{S}^{-1} - \Sigma_x^{-1}\|_{\max}$. To do so, we use series expansion for the inverse of matrices. By the Woodbury matrix identities, we have for two symmetric positive definite matrices A and B:

$$(A - B)^{-1} = \sum_{k=0}^{\infty} (A^{-1}B)^k A^{-1} = A^{-1} + \sum_{k=1}^{\infty} (A^{-1}B)^k A^{-1}.$$

Thus

$$(A + B)^{-1} = A^{-1} + \sum_{k=1}^{\infty} (-1)^k (A^{-1}B)^k A^{-1}.$$

Now, taking $B = \widehat{S} - \Sigma_x$ and $A = \Sigma_x$, then we obtain

$$\widehat{S}^{-1} - \Sigma_x^{-1} = \sum_{k=1}^{\infty} (-1)^k \Big(\Sigma_x^{-1}(\widehat{S} - \Sigma_x)\Big)^k \Sigma_x^{-1}.$$

Hence, using this relationship, we can upper bound the norm of the difference $\widehat{S}^{-1} - \Sigma_x^{-1}$ by the norm of $\widehat{S} - \Sigma_x^{-1}$ for which we have an exponential bound. For any sub-multiplicative matrix norm $\|.\|_.$, we deduce

$$\|\widehat{S}^{-1} - \Sigma_x^{-1}\|_. \leq \|\Sigma_x^{-1}\|_. \sum_{k=1}^{\infty} \Big(\|\Sigma_x^{-1}\|_. \|\widehat{S} - \Sigma_x\|_.\Big)^k = \|\Sigma_x^{-1}\|_.^2 \|\widehat{S} - \Sigma_x\|_. \sum_{k=0}^{\infty} \Big(\|\Sigma_x^{-1}\|_. \|\widehat{S} - \Sigma_x\|_.\Big)^k.$$

For a square symmetric matrix M, under the condition $\|M\|_. < 1$, then $\sum_{k=0}^{\infty} \|M\|_.^k = (1 - \|M\|_.)^{-1}$. Now, with high probability, for a sufficiently large sample size, we thus have

$$\|\widehat{S}^{-1} - \Sigma_x^{-1}\|_. \leq \frac{\|\Sigma_x^{-1}\|_.^2 \|\widehat{S} - \Sigma_x\|_.}{1 - \|\Sigma_x^{-1}\|_. \|\widehat{S} - \Sigma_x\|_.},$$

so that bounding $\|\widehat{S} - \Sigma_x\|_.$ implies bounding $\|\widehat{S}^{-1} - \Sigma_x^{-1}\|_.$. Now, we have

$$\|\Sigma_x^{-1} - \widehat{S}^{-1}\|_{\max} \leq \|\Sigma_x^{-1} - \widehat{S}^{-1}\|_s \leq \frac{\|\widehat{S}^{-1}\|_s^2 \|\Sigma_x - \widehat{S}\|_s}{1 - \|\widehat{S}^{-1}\|_s \|\Sigma_x - \widehat{S}\|_s}.$$

Moreover, since $\|\Sigma_x - \widehat{S}\|_s \leq d\|\Sigma_x - \widehat{S}\|_{\max}$, we obtain $\|\nabla_\Theta \mathbb{L}_T(\Theta_0)\|_{\max} \leq Ld\sqrt{\frac{\log(d^2)}{T}}$, for $L > 0$ sufficiently large. Hence, for a sample size $T \geq Md^2 \log(d^2) \max(R^2, k_0) \vee L \log(d^2)^{2/\gamma - 1}$, then (10) holds with probability at least $1 - O(\exp\big(-\log(d^2)\big)) - o(\exp\big(-\log(d^2)\big))$. □

Proof of Corollary 3.5. First, let us consider the series expansion of the logarithm of a p-square positive definite matrix A, given as $\log(A) = -\sum_{k=1}^{\infty} \frac{1}{k}(I_p - A)^k$. Obviously, when the spectral radius of $I_p - A$ is strictly inferior to one, this expansion does exist. Moreover, the inverse of A can be developed as a Neumann series $A^{-1} = \sum_{k=0}^{\infty} (I_p - A)^k$. Let us consider the first order differential of $\mathbb{L}_{T,\eta}(\Theta)$:

$$\begin{aligned}
\boldsymbol{d}\mathbb{L}_{T,\eta}(\Theta) &= \eta \mathrm{tr}\Big((\boldsymbol{d}\Theta/\nu)\log(\Theta/\nu) + (\Theta/\nu)(\boldsymbol{d}\log(\Theta/\nu))\Big) \\
&+(2\eta - 1)\mathrm{tr}\Big(-(\boldsymbol{d}\Theta/\nu)\Big) - \eta \mathrm{tr}\Big(\log(\widehat{S}^{-1}/\nu)(\boldsymbol{d}\Theta/\nu)\Big) - (1-\eta)\mathrm{tr}\Big((\boldsymbol{d}\log(\Theta/\nu))(\widehat{S}^{-1}/\nu)\Big)
\end{aligned}$$

Now, let us treat the trace of $(\Theta/\nu)\boldsymbol{d}\log(\Theta/\nu)$. Following Abadir and Magnus (2006) (see exercise 13.31), we have

$$\begin{aligned}
\mathrm{tr}\Big((\Theta/\nu)(\boldsymbol{d}\log(\Theta/\nu))\Big) &= \mathrm{tr}\Big(\sum_{k=1}^{\infty} \tfrac{1}{k}\Big\{\sum_{l=1}^{k}(I_p - \Theta/\nu)^{l-1}(\boldsymbol{d}\Theta/\nu)(I_p - \Theta/\nu)^{k-l}\Big\}(\Theta/\nu)\Big) \\
&= \sum_{k=1}^{\infty} \tfrac{1}{k}\sum_{l=1}^{k} \mathrm{tr}\Big((I_p - \Theta/\nu)^{l-1}(\boldsymbol{d}\Theta/\nu)(I_p - \Theta/\nu)^{k-l}(\Theta/\nu)\Big) \\
&= \sum_{k=1}^{\infty} \tfrac{1}{k}\sum_{l=1}^{k} \mathrm{tr}\Big((I_p - \Theta/\nu)^{k-l}(\Theta/\nu)(I_p - \Theta/\nu)^{l-1}(\boldsymbol{d}\Theta/\nu)\Big).
\end{aligned}$$

In the same spirit,

$$\mathrm{tr}\Big((\widehat{S}^{-1}/\nu)(\boldsymbol{d}\log(\Theta/\nu))\Big) = \sum_{k=1}^{\infty} \frac{1}{k}\sum_{l=1}^{k} \mathrm{tr}\Big((I_p - \Theta/\nu)^{k-l}(\widehat{S}^{-1}/\nu)(I_p - \Theta/\nu)^{l-1}(\boldsymbol{d}\Theta/\nu)\Big).$$

Now let us consider the cases $\eta = 0, 1$ separately.

(i) $\eta = 1$: the differential becomes

$$\boldsymbol{d}\mathbb{L}_{T,1}(\Theta) = \mathrm{tr}\Big((\boldsymbol{d}\Theta/\nu)\log(\Theta/\nu) + (\Theta/\nu)(\boldsymbol{d}\log(\Theta/\nu)) - (\boldsymbol{d}\Theta/\nu) - \log(\widehat{S}^{-1}/\nu)(\boldsymbol{d}\Theta/\nu)\Big).$$

Thus, the differential becomes

$$\boldsymbol{d}\mathbb{L}_{T,1}(\Theta) = \mathrm{tr}\Big(\big[\log(\Theta/\nu) - \log(\widehat{S}^{-1}/\nu)\big](\boldsymbol{d}\Theta/\nu) + \sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=1}^{k}(I_p - \Theta/\nu)^{k-l}(\Theta/\nu)(I_p - \Theta/\nu)^{l-1}(\boldsymbol{d}\Theta/\nu) - (\boldsymbol{d}\Theta/\nu)\Big).$$

Hence, by the trace properties and the von Neumann based series expansion of the inverse matrix, the gradient in vector and matrix forms are, respectively,

$$\nabla_\theta \mathbb{L}_{T,1}(\Theta) = \mathrm{vec}(\log(\Theta/\nu) - \log(\widehat{S}^{-1}/\nu))/\nu, \quad \nabla_\Theta \mathbb{L}_{T,1}(\Theta) = \frac{1}{\nu}\Big(\log(\Theta/\nu) - \log(\widehat{S}^{-1}/\nu)\Big).$$

Let us now focus on the Hessian matrix, where we only need to focus on Starting from $\mathrm{tr}\Big((\boldsymbol{d}\Theta/\nu)\log(\Theta/\nu)\Big)/\nu$. The second order differential becomes

$$\boldsymbol{d}^2\mathbb{L}_{T,1}(\Theta) = \mathrm{tr}\Big((\boldsymbol{d}\Theta/\nu)(\boldsymbol{d}\log(\Theta/\nu))\Big)/\nu = \mathrm{tr}\Big((\boldsymbol{d}\Theta/\nu)\Big[\sum_{k=1}^{\infty}\frac{1}{k}\Big\{\sum_{l=1}^{k}(I_p - \Theta/\nu)^{l-1}(\boldsymbol{d}\Theta/\nu)(I_p - \Theta/\nu)^{k-l}\Big\}\Big]\Big).$$

Using the Hessian identification, we obtain

$$\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,\eta}(\Theta) = \sum_{k=1}^{\infty}\frac{1}{k}\Big\{\sum_{l=1}^{k}(I_p - \Theta/\nu)^{k-l} \otimes (I_p - \Theta/\nu)^{l-1}\Big\}/\nu^2$$

By 10.20 (b) of Abadir and Magnus (2006), we have

$$f_T(\Theta) = \mathrm{vec}(\Gamma)^\top \nabla^2_{\theta\theta^\top}\mathbb{L}_{T,\eta}(\Theta)\mathrm{vec}(\Gamma) = \mathrm{tr}\Big(\sum_{k=1}^{\infty}\frac{1}{k}\Big\{\sum_{l=1}^{k}(I_p - \Theta/\nu)^{k-l}\Gamma(I_p - \Theta/\nu)^{l-1}\Gamma\Big\}\Big)/\nu^2.$$

where $\Theta = \Theta_0 + s\Gamma$, $\Gamma = \Theta_1 - \Theta_0$ with $\Theta_1 \in \Omega^{\mathrm{vn}}$. Since Θ_0 and Θ_1 can be diagonalized in the same basis, idest $\Theta_0 = P\Lambda_0 P^{-1}$ and $\Theta_1 = P\Lambda_1 P^{-1}$, with P invertible containing the eigenvectors and $\Lambda_j = \mathrm{diag}(\lambda_{1,j}, \cdots, \lambda_{d,j}), j = 0, 1$, a diagonal matrix containing the eigenvalues of $\Lambda_j, j = 0, 1$. Hence, Θ can also be diagonalized in the same basis with eigenvalues $\tilde{\lambda}_k, k = 1, \cdots, d$, so that we can manipulate the trace of diagonal matrices. Hence $\mathrm{tr}\Big((I_p - \Theta/\nu)^{k-l}\Gamma(I_p - \Theta/\nu)^{l-1}\Gamma\Big) = \mathrm{tr}\Big((I_p - \Theta/\nu)^{k-1}\Gamma^2\Big)$. As a consequence,

$$\begin{aligned} f_T(\Theta) &\geq \sum_{k=1}^{\infty}\frac{1}{\nu^2}\sum_{n=1}^{p}(1 - \tilde{\lambda}_n/\nu)^{k-1}(\lambda_{n,1} - \lambda_{n,0})^2 \\ &\geq \sum_{n=1}^{p}\frac{(\lambda_{n,1} - \lambda_{n,0})^2}{\tilde{\lambda}_n \nu} \geq \sum_{n=1}^{p}\frac{(\lambda_{n,1} - \lambda_{n,0})^2}{\nu(\lambda_{\max}(\Theta_0) \vee \lambda_{\max}(\Theta_1))} \geq \|\Gamma\|_F^2 \frac{1}{\nu d}. \end{aligned}$$

We deduce

$$(\theta-\theta_0)^\top \nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\bar\Theta)(\theta-\theta_0) \geq \|\mathrm{vec}(\Theta)-\mathrm{vec}(\Theta_0)\|_2^2 \frac{1}{\nu d}.$$

The RSC parameters are thus given by $\alpha_1=\alpha_2=1/(\nu d)$, $\tau_1=\tau_2=0$.

(ii) $\eta=0$: the differential becomes

$$\begin{aligned} \boldsymbol{d}\mathbb{L}_{T,0}(\Theta) &= \mathrm{tr}\Big((\boldsymbol{d}\Theta/\nu)\Big) - \mathrm{tr}\Big((\boldsymbol{d}\log(\Theta/\nu))(\widehat{S}^{-1}/\nu)\Big) \\ &= \mathrm{tr}\Big((\boldsymbol{d}\Theta/\nu)\Big) - \sum_{k=1}^{\infty}\tfrac{1}{k}\sum_{l=1}^{k}\mathrm{tr}\Big((I_p-\Theta/\nu)^{k-l}(\widehat{S}^{-1}/\nu)(I_p-\Theta/\nu)^{l-1}(\boldsymbol{d}\Theta/\nu)\Big). \end{aligned} \tag{29}$$

The gradient in matrix form becomes

$$\begin{aligned} \nabla_\Theta \mathbb{L}_{T,0}(\Theta) &= \mathrm{tr}\Big(I_p\Big)/\nu - \sum_{k=1}^{\infty}\tfrac{1}{k}\sum_{l=1}^{k}\mathrm{tr}\Big((I_p-\Theta/\nu)^{k-l}(\widehat{S}^{-1}/\nu)(I_p-\Theta/\nu)^{l-1}\Big)/\nu \\ &= \mathrm{tr}\Big(I_p\Big)/\nu - \sum_{k=1}^{\infty}\mathrm{tr}\Big((I_p-\Theta/\nu)^{k-1}(\widehat{S}^{-1}/\nu)\Big)/\nu \\ &= \mathrm{tr}\Big(I_p\Big)/\nu - \mathrm{tr}\Big((\Theta/\nu)^{-1}\widehat{S}^{-1}/\nu\Big)/\nu. \end{aligned}$$

As for the Hessian, we aim at extracting the form $\mathrm{tr}(L(\boldsymbol{d}\Theta)M(\boldsymbol{d}\Theta))$ for L (resp. M) any square $m\times m$ matrix (resp. $n\times n$). Applying the differential operator on (29), we obtain

$$\begin{aligned} &\boldsymbol{d}^2\mathbb{L}_{T,0}(\Theta) \\ &= \sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=1}^{k}\Big[\mathrm{tr}\Big(\sum_{j=1}^{k-l}(I_p-\Theta/\nu)^{j-1}(\boldsymbol{d}\Theta/\nu)(I_p-\Theta/\nu)^{k-l-j}(\widehat{S}^{-1}/\nu)(I_p-\Theta/\nu)^{l-1}(\boldsymbol{d}\Theta/\nu)\Big)\Big] \\ &+ \sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=1}^{k}\Big[\mathrm{tr}\Big((I_p-\Theta/\nu)^{k-l}(\widehat{S}^{-1}/\nu)\sum_{i=1}^{l-1}(I_p-\Theta/\nu)^{i-1}(\boldsymbol{d}\Theta/\nu)(I_p-\Theta/\nu)^{l-1-i}(\boldsymbol{d}\Theta/\nu)\Big)\Big] \\ &= \sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=1}^{k}\sum_{j=1}^{k-l}\Big[\mathrm{tr}\Big((I_p-\Theta/\nu)^{j-1}(\boldsymbol{d}\Theta/\nu)(I_p-\Theta/\nu)^{k-l-j}(\widehat{S}^{-1}/\nu)(I_p-\Theta/\nu)^{l-1}(\boldsymbol{d}\Theta/\nu)\Big)\Big] \\ &+ \sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=1}^{k}\sum_{i=1}^{l-1}\Big[\mathrm{tr}\Big((I_p-\Theta/\nu)^{k-l}(\widehat{S}^{-1}/\nu)(I_p-\Theta/\nu)^{i-1}(\boldsymbol{d}\Theta/\nu)(I_p-\Theta/\nu)^{l-1-i}(\boldsymbol{d}\Theta/\nu)\Big)\Big]. \end{aligned}$$

By Hessian identification, we thus deduce

$$\begin{aligned} &\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta) \\ &= \sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=1}^{k}\sum_{j=1}^{k-l}\frac{1}{2}\Big[\Big\{(I_p-\Theta/\nu)^{j-1}\otimes(I_p-\Theta/\nu)^{k-l-j}(\widehat{S}^{-1}/\nu)(I_p-\Theta/\nu)^{l-1}\Big\} \\ &\quad +\Big\{(I_p-\Theta/\nu)^{j-1}\otimes(I_p-\Theta/\nu)^{l-1}(\widehat{S}^{-1}/\nu)(I_p-\Theta/\nu)^{k-l-j}\Big\}\Big]/\nu^2 \\ &+ \sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=1}^{k}\sum_{i=1}^{l-1}\frac{1}{2}\Big[\Big\{(I_p-\Theta/\nu)^{k-l}(\widehat{S}^{-1}/\nu)(I_p-\Theta/\nu)^{i-1}\otimes(I_p-\Theta/\nu)^{l-1-i}\Big\} \\ &\quad +\Big\{(I_p-\Theta/\nu)^{i-1}(\widehat{S}^{-1}/\nu)(I_p-\Theta/\nu)^{k-l}\otimes(I_p-\Theta/\nu)^{l-1-i}\Big\}\Big]/\nu^2. \end{aligned}$$

Consequently:

$$
\begin{aligned}
f_T(\Theta) &= \mathrm{vec}(\Gamma)^\top \nabla^2_{\theta\theta^\top} \mathbb{L}_{T,0}(\Theta)\mathrm{vec}(\Gamma) \\
&= \mathrm{vec}(\Gamma)^\top \Big(\sum_{k=1}^{\infty} \frac{1}{k} \sum_{l=1}^{k} \sum_{j=1}^{k-l} \frac{1}{2} \Big[\Big\{ (I_p - \Theta/\nu)^{j-1} \otimes (I_p - \Theta/\nu)^{k-l-j} (\widehat{S}^{-1}/\nu)(I_p - \Theta/\nu)^{l-1} \Big\} \\
&\quad + \Big\{ (I_p - \Theta/\nu)^{j-1} \otimes (I_p - \Theta/\nu)^{l-1} (\widehat{S}^{-1}/\nu)(I_p - \Theta/\nu)^{k-l-j} \Big\} \Big] / \nu^2 \\
&+ \sum_{k=1}^{\infty} \frac{1}{k} \sum_{l=1}^{k} \sum_{i=1}^{l-1} \frac{1}{2} \Big[\Big\{ (I_p - \Theta/\nu)^{k-l} (\widehat{S}^{-1}/\nu)(I_p - \Theta/\nu)^{i-1} \otimes (I_p - \Theta/\nu)^{l-1-i} \Big\} \\
&\quad + \Big\{ (I_p - \Theta/\nu)^{i-1} (\widehat{S}^{-1}/\nu)(I_p - \Theta/\nu)^{k-l} \otimes (I_p - \Theta/\nu)^{l-1-i} \Big\} \Big] / \nu^2 \Big) \mathrm{vec}(\Gamma).
\end{aligned}
$$

Using the properties of the minimum eigenvalue of the Kronecker product, of the product of positive definite matrices and convergence of geometric series, we obtain

$$
\begin{aligned}
& f_T(\Theta) \\
&\geq \frac{1}{\nu^2} \|\Gamma\|_F^2 \Big\{ \sum_{k=1}^{\infty} \frac{1}{k} \sum_{l=1}^{k} \sum_{j=1}^{k-l} \lambda_{\min}(I_p - \Theta/\nu)^{k-2} \lambda_{\min}(\widehat{S}^{-1}/\nu) + \sum_{k=1}^{\infty} \frac{1}{k} \sum_{l=1}^{k} \sum_{i=1}^{l-1} \lambda_{\min}(I_p - \Theta/\nu)^{k-2} \lambda_{\min}(\widehat{S}^{-1}/\nu) \Big\} \\
&\geq \frac{1}{\nu^2} \|\Gamma\|_F^2 \lambda_{\min}(\widehat{S}^{-1}/\nu) \Big\{ \sum_{k=1}^{\infty} (k-1) \lambda_{\min}(I_p - \Theta/\nu)^{k-2} \Big\} \\
&\geq \frac{1}{\nu^2} \|\Gamma\|_F^2 \lambda_{\min}(\widehat{S}^{-1}/\nu) \Big[1/(1 - \lambda_{\min}(I_p - \Theta/\nu))^2 \Big] \\
&\geq \|\Gamma\|_F^2 \lambda_{\min}(\widehat{S}^{-1}/\nu) \{\lambda_{\max}(\Theta_0) + 1\}^{-2},
\end{aligned}
$$

since $1 - \lambda_{\min}(I_p - \Theta/\nu) \leq \lambda_{\max}(\Theta/\nu)$. We then deduce

$$
(\theta - \theta_0)^\top \nabla^2_{\theta\theta^\top} \mathbb{L}_{T,\eta}(\bar{\Theta})(\theta - \theta_0) \geq \|\mathrm{vec}(\Theta) - \mathrm{vec}(\Theta_0)\|_2^2 \lambda_{\min}(\widehat{S}^{-1}/\nu) \{\lambda_{\max}(\Theta_0) + 1\}^{-2}.
$$

The RSC parameters are $\alpha_1 = \alpha_2 = \lambda_{\min}(\widehat{S}^{-1}/\nu)\{\lambda_{\max}(\Theta_0) + 1\}^{-2}$, and $\tau_1 = \tau_2 = 0$.

Let us evaluate the probability of satisfying (14) for $\eta = 0$, where the gradient is given by

$$
\nabla_\Theta \mathbb{L}_{T,0}(\Theta) = \Big(I_d - (\widehat{S}^{-1}/\nu)(\Theta/\nu)^{-1} \Big)/\nu = \Big(I_d - \widehat{S}^{-1}\Theta^{-1} \Big)/\nu.
$$

Thus, we have

$$
\|\nabla_\Theta \mathbb{L}_{T,0}(\Theta_0)\|_{\max} = \|(\Theta_0 - \widehat{S}^{-1})\Theta_0^{-1}/\nu\|_{\max} \leq \|\Theta_0^{-1}/\nu\|_s \|\Theta_0 - \widehat{S}^{-1}\|_s.
$$

Using the same steps as in the least squares case, we deduce $\|\Theta_0 - \widehat{S}^{-1}\|_{\max} \leq Ld\sqrt{\frac{\log(d^2)}{T}}$, with high probability and for $L > 0$ large enough. Then:

$$\|\nabla_\Theta \mathbb{L}_{T,0}(\Theta_0)\|_{\max} \leq Ld\sqrt{\frac{\log(d^2)}{T}}\|\Theta_0^{-1}/\nu\|_s,$$

with probability at least $1 - O(\exp\big(-\log(d^2)\big)) - o(\exp\big(-\log(d^2)\big))$.

□

Proof of Corollary 3.6. We first establish the RSC condition. The gradient is defined as

$$\nabla_\theta \mathbb{L}_T(\Theta) = \frac{1}{2}\text{vec}(\Theta\widehat{S} + \widehat{S}\Theta - 2I_d), \;\; \nabla_\Theta \mathbb{L}_T(\Theta) = \frac{1}{2}\Big(\Theta\widehat{S} + \widehat{S}\Theta - 2I_d\Big).$$

As for the Hessian, by identification, we have

$$\nabla^2_{\theta\theta^\top} \mathbb{L}_T(\Theta) = \frac{1}{2}\Big(\widehat{S} \otimes I_d + I_d \otimes \widehat{S}\Big).$$

For some $\Theta_1 \in \Omega$ and $u \in [0,1]$, let $\Theta = \Theta_0 + u\Gamma$ with $\Gamma = \Theta_1 - \Theta_0$. Then

$$f_T(\Theta) := \text{vec}(\Gamma)^\top \nabla^2_{\theta\theta^\top} \mathbb{L}_T(\Theta)\text{vec}(\Gamma) \geq \|\Gamma\|_F^2 \frac{1}{2}\Big[\lambda_{\min}(\widehat{S} \otimes I_d + I_d \otimes \widehat{S})\Big] \geq \|\Gamma\|_F^2 \lambda_{\min}(\widehat{S}).$$

For $\|\Gamma\|_F^2 = \|\Theta - \Theta_0\|_F^2$, we thus deduce

$$(\theta - \theta_0)^\top \nabla^2_{\theta\theta^\top} \mathbb{L}_T(\overline{\Theta})(\theta - \theta_0) \geq \lambda_{\min}(\widehat{S})\|\theta - \theta_0\|_2^2,$$

for any $\overline{\Theta}$ located between Θ and Θ_0. The RSC condition is then satisfied with coefficients $\alpha_1 = \lambda_{\min}(\widehat{S})$ and $\alpha_2 = \alpha_1$, $\tau_1 = \tau_2 = 0$.

We now evaluate the probability of satisfying (17). We have

$$\|\nabla_\Theta \mathbb{L}_T(\Theta_0)\|_{\max} = \|\frac{1}{2}\Theta_0\Big(\widehat{S} - \Theta_0^{-1}\Big) + \frac{1}{2}\Big(\widehat{S} - \Theta_0^{-1}\Big)\Theta_0\|_{\max} \leq \|\Theta_0\|_s\|\widehat{S} - \Sigma_x\|_s.$$

Hence, by Lemma 3.2, we have

$$\|\nabla_\Theta \mathbb{L}_T(\Theta_0)\|_{\max} \leq L\|\Theta_0\|_s d\sqrt{\frac{\log(d^2)}{T}},$$

for $L > 0$ large enough. Hence, for a sample size $T \geq M\|\Theta_0\|_s^2 d^2 \log(d^2)\max(R^2, k_0) \vee L\log(d^2)^{2/\gamma - 1}$, then (18) holds with probability at least $1 - O(\exp(-\log(d^2))) - o(\exp(-\log(d^2)))$ □

To establish the proofs for support recovery in subsection 3.3, the key point is to show that strict dual feasibility holds. To do so and for each sparse estimator case, following the PDW construction of Wainwright (2009) or Ravikumar et al. (2011), we construct a theoretical estimator optimized over the true subset (hence it is not possible to compute it empirically).

Proof of Corollary 3.7. The oracle estimator defined in (19) becomes for the Stein's loss:

$$\widehat{\Theta}^{\mathrm{g},\mathcal{O}} := \underset{\theta\in\mathbb{R}^{|\mathcal{A}|}:\Theta\in\Omega}{\arg\min}\Big\{\mathbb{L}_{T,1}(\Theta)\Big\} = \underset{\theta\in\mathbb{R}^{|\mathcal{A}|}:\Theta\in\Omega}{\arg\min}\Big\{\mathrm{tr}\Big(\widehat{S}\Theta\Big) - \log(|\Theta|)\Big\},$$

Proof of point (i). We highlight that our approach differs from the proof methods of Theorem 1 of Ravikumar et al. (2011) or Corollary 4 of Loh and Wainwright (2017) dedicated to Gaussian loss based sparse precision matrix, which are based on the Brouwer's fixed point Theorem, in that our proof strategy shares the same spirit as in the proof of Corollaries 2 and 3 respectively dedicated to Linear regression with corrupted covariates and GLM of Loh and Wainwright (2017) to construct the estimator $\widehat{\Theta}_{\mathcal{A}}$ such that $\mathrm{supp}(\widehat{\Theta}) \subseteq \mathcal{A}$ and $\widehat{\Theta}_{\mathcal{A}}$ is a zero subgradient point of

$$\widehat{\Theta}^{\mathrm{g}}_{\mathcal{A}} = \underset{\Theta:\Theta\in\Omega,\mathrm{supp}(\Theta)\subseteq\mathrm{supp}(\Theta_0)}{\arg\min}\Big\{\mathbb{L}_{T,1}(\Theta) + \boldsymbol{p}(\lambda_T,\theta)\Big\},$$

where $\mathbb{L}_{T,1}(.)$ and Ω are defined as in (5) for $\eta = 1$, and μ-amenable penalty functions (hence LASSO, SCAD and MCP). The latter amenability property is a key point since we require the incoherence condition in the μ-amenable case. We show that strict dual feasibility holds for such statistical problem. Now by the zero gradient condition (23) of the PDW step, we obtain

$$\nabla_\theta\mathbb{L}_{T,1}(\widehat{\Theta}) - \nabla_\theta\mathbb{L}_{T,1}(\Theta_0) + \nabla_\theta\mathbb{L}_{T,1}(\Theta_0) - \nabla\boldsymbol{q}(\lambda_T,\mathrm{vec}(\widehat{\Theta}^{\mathrm{g}})) + \lambda_T\widehat{z} = 0.$$

This implies

$$\widehat{\boldsymbol{K}}\mathrm{vec}(\widehat{\Theta}^{\mathrm{g}} - \Theta_0) + \nabla_\theta\mathbb{L}_{T,1}(\Theta_0) - \nabla\boldsymbol{q}(\lambda_T,\mathrm{vec}(\widehat{\Theta}^{\mathrm{g}})) + \lambda_T\widehat{z} = 0,$$

with $\widehat{\boldsymbol{K}} = \int_0^1 \nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0 + u(\widehat{\Theta}^{\mathrm{g}} - \Theta_0))du$. Equivalently, we have

$$\begin{pmatrix}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}} & \widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}^c}\\ \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}} & \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}^c}\end{pmatrix}\begin{pmatrix}\mathrm{vec}(\widehat{\Theta}^{\mathrm{g}} - \Theta_0)_{\mathcal{A}}\\ \mathbf{0}\end{pmatrix} + \begin{pmatrix}\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}} - \nabla\boldsymbol{q}(\lambda_T,\mathrm{vec}(\widehat{\Theta}^{\mathrm{g}})_{\mathcal{A}})\\ \nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}^c} - \nabla\boldsymbol{q}(\lambda_T,\mathrm{vec}(\widehat{\Theta}^{\mathrm{g}})_{\mathcal{A}^c})\end{pmatrix} + \lambda_T\begin{pmatrix}\widehat{z}_{\mathcal{A}}\\ \widehat{z}_{\mathcal{A}^c}\end{pmatrix} = \mathbf{0}.$$

Consequently, we obtain

$$\begin{aligned}\widehat{z}_{\mathcal{A}^c} &= \tfrac{1}{\lambda_T}\Big\{\nabla \boldsymbol{q}(\lambda_T, \mathrm{vec}(\widehat{\Theta}^{\mathrm{g}}))_{\mathcal{A}^c} - \nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\Big[\nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}} \\ &\quad -\nabla \boldsymbol{q}(\lambda_T, \mathrm{vec}(\widehat{\Theta}^{\mathrm{g}})_{\mathcal{A}}) + \lambda_T \widehat{z}_{\mathcal{A}}\Big]\Big\}.\end{aligned}$$

Using the regularity condition (v), $\nabla \boldsymbol{q}(\lambda_T, \mathrm{vec}(\widehat{\Theta}^{\mathrm{g}})_{\mathcal{A}^c}) = \nabla \boldsymbol{q}(\lambda_T, \mathbf{0}_{\mathcal{A}^c}) = \mathbf{0}_{\mathcal{A}^c}$. This implies

$$\widehat{z}_{\mathcal{A}^c} = \frac{1}{\lambda_T}\Big\{ - \nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\Big[\nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}} - \nabla \boldsymbol{q}(\lambda_T, \mathrm{vec}(\widehat{\Theta}^{\mathrm{g}})_{\mathcal{A}}) + \lambda_T \widehat{z}_{\mathcal{A}}\Big]\Big\}.$$

Taking the ℓ_∞-norm, we obtain

$$\begin{aligned}\|\widehat{\boldsymbol{z}}_{\mathcal{A}^c}\|_\infty &\leq \tfrac{1}{\lambda_T}\| - \nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty \\ &+ \tfrac{1}{\lambda_T}\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\big(\lambda_T \widehat{z}_{\mathcal{A}} - \nabla_\theta \boldsymbol{q}(\lambda_T, \widehat{\theta}_\Theta)_{\mathcal{A}}\big)\|_\infty \\ &\leq \tfrac{1}{\lambda_T}\| - \nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty + \|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty,\end{aligned}$$

where we used $\|\lambda_T \widehat{z}_{\mathcal{A}} - \nabla_\theta \boldsymbol{q}(\lambda_T, \mathrm{vec}(\widehat{\Theta}^{\mathrm{g}})_{\mathcal{A}})\|_\infty = \|\nabla_\theta \boldsymbol{q}(\lambda_T, \mathrm{vec}(\widehat{\Theta}^{\mathrm{g}})_{\mathcal{A}})\|_\infty \leq \lambda_T$ from Lemma 8 of Loh and Wainwright (2017). Furthermore, we have

$$\| - \nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty \leq \|\nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}^c}\|_\infty + \|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty.$$

First, let us consider $\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty$. To do so, we first control the following quantity:

$$\begin{aligned}&\|\widehat{\boldsymbol{K}} - \nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)\|_\infty \\ &= \|\int_0^1 \big(\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0 + s(\widehat{\Theta}^{\mathrm{g}} - \Theta_0)) - \nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)\big)ds\|_\infty \\ &\leq \int_0^1 \|\big(\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0 + s(\widehat{\Theta}^{\mathrm{g}} - \Theta_0)) - \nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)\big)\|_\infty ds \\ &= \int_0^1 \|\Big\{\Big(\Theta_0 + s(\widehat{\Theta}^{\mathrm{g}} - \Theta_0)\Big)^{-1} \otimes \Big(\Theta_0 + s(\widehat{\Theta}^{\mathrm{g}} - \Theta_0)\Big)^{-1}\Big\} - \Big\{\Theta_0^{-1} \otimes \Theta_0^{-1}\Big\}\|_\infty ds.\end{aligned}$$

Now for any $s \in [0,1]$, using the consistency result from Corollary 3.3, we have

$$\|\Big(\Theta_0 + s(\widehat{\Theta}^{\mathrm{g}} - \Theta_0)\Big) - \Theta_0\|_\infty = s\|\widehat{\Theta}^{\mathrm{g}} - \Theta_0\|_\infty \leq d\|\widehat{\Theta}^{\mathrm{g}} - \Theta_0\|_{\max} \leq d\|\widehat{\Theta}^{\mathrm{g}} - \Theta_0\|_F \leq Ld\sqrt{k_0}\sqrt{\frac{\log(d^2)}{T}},$$

for L sufficiently large. Hence, by the same reasoning as before on the norm on the inverse matrix difference:

$$\|\Big(\Theta_0 + s(\widehat{\Theta}^{\mathrm{g}} - \Theta_0)\Big)^{-1} - \Theta_0^{-1}\|_\infty \leq Ld\sqrt{k_0}\sqrt{\frac{\log(d^2)}{T}}.$$

Thus, using Lemma 13 of Loh and Wainwright (2017) on the upper bound of $\|A \otimes A - B \otimes B\|_\infty$, we obtain

$$\|\widehat{\boldsymbol{K}} - \nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)\|_\infty \leq Ld\sqrt{k_0}\sqrt{\frac{\log(d^2)}{T}},$$

which provides the bound that hold with high probability with $L > 0$ when restricting to $\mathcal{A}$:

$$\upsilon_1 := \|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}} - \nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}^c\mathcal{A}}\|_\infty \leq L\sqrt{k_0^2\frac{\log(d^2)}{T}}, \;\; \upsilon_2 := \|\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}} - (\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0))^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty \leq L\sqrt{k_0^2\frac{\log(d^2)}{T}},$$

by union bound and since $\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}} - \nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}\mathcal{A}}\|_\infty \leq L\sqrt{k_0^2\frac{\log(d^2)}{T}}$. Now we bound $\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}}\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty$ to prove the success of the PDW method. To do so, we consider the expansion

$$\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}}\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty \leq M_1 + M_2, \;\; \text{where:}$$

$$\begin{aligned} M_1 &:= \|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty, \\ M_2 &:= \|\Big\{\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\Big\}\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty. \end{aligned}$$

First, let us highlight that $\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)] = \nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)$ since it involves no empirical estimator such as the sample variance covariance matrix. Then, we have

$$M_1 \leq \|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty\|\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty \leq \omega K\sqrt{\frac{\log(d^2)}{T}},$$

with high probability and using the incoherence condition $\|\boldsymbol{K}_{0,\mathcal{A}^c\mathcal{A}}\boldsymbol{K}^{-1}_{0,\mathcal{A}\mathcal{A}}\| \leq \omega$. As for M_2, we have

$$\begin{aligned} M_2 \leq & \Big\{\|\Big(\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\Big)\Big(\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\Big)\|_\infty \\ & + \|\Big(\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\Big)\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty \\ & + \|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\Big(\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\Big)\|_\infty\Big\}\|\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty \\ \leq & \Big\{\upsilon_1\upsilon_2 + \upsilon_1\|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty + \upsilon_2\|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\|_\infty\Big\}\|\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty \\ \leq & \Big\{\upsilon_1\upsilon_2 + \upsilon_1\|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty + \upsilon_2\|\Sigma_x\|^2_\infty\Big\}\|\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty \\ \leq & K\sqrt{k_0^2\|\Sigma_x\|^4_\infty\frac{\log(d^2)}{T}}\|\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty, \end{aligned}$$

with high probability and K sufficiently large, where we used $\|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\|_\infty \leq \|\Sigma_x\|^2_\infty$. Then:

$$\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}}\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty \leq L\sqrt{k_0^2\|\Sigma_x\|^4_\infty\frac{\log(d^2)}{T}}\sqrt{\frac{\log(d^2)}{T}},$$

with high probability under the scaling behaviour $T > Mk_0^2\|\Sigma_x\|^4_\infty\log(d^2)$ with M sufficiently large. More-

over, using the incoherence condition and

$$
\begin{aligned}
&\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \\
&\leq \|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty + \|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty \\
&\leq L\sqrt{k_0^2\|\Sigma_x\|_\infty^4\frac{\log(d^2)}{T}} + \omega.
\end{aligned}
$$

Thus we have for $L > 0$ sufficiently large

$$
\|\boldsymbol{z}_{\mathcal{A}^c}\|_\infty \leq \frac{1}{\lambda_T}\Big(L_1\sqrt{k_0^3\|\Sigma_x\|_\infty^4\frac{\log(d^2)}{T}} + L_2\sqrt{\frac{\log(d^2)}{T}}\Big) + L_3\sqrt{k_0^3\|\Sigma_x\|_\infty^4\frac{\log(d^2)}{T}} + \omega,
$$

for $L_1, L_2, L_3 > 0$. Then strict dual feasibility of Theorem 5.1 is satisfied when $\frac{1}{1-\omega}L\sqrt{\frac{\log(d^2)}{T}} \leq \lambda_T$, under the scaling $T > Ck_0^2\|\Sigma_x\|_\infty^4\log(d^2)$.

We now focus on the ℓ_∞ bound to apply point (i) of Theorem 5.2. We have

$$
\begin{aligned}
&\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty \\
&\leq \|\{\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\}\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty + \|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty \\
&\leq \|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty\|\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty + \|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\nabla_\theta\mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty \\
&\leq C_1\sqrt{k_0^3\frac{\log(d^2)}{T}}\sqrt{\frac{\log(d^2)}{T}} + C_2\sqrt{\|\Sigma_x\|_\infty^4\frac{\log(d^2)}{T}},
\end{aligned}
$$

with $C_1, C_2 > 0$ and using our previous upper bounds. We proved

$$
\begin{aligned}
\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - (\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty &\leq \sqrt{k_0}\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - (\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]_{\mathcal{A}\mathcal{A}})^{-1}\|_s \\
&\leq L\sqrt{k_0^2\tfrac{\log(d^2)}{T}} \leq \beta_\infty.
\end{aligned}
$$

Hence

$$
\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \leq \|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - (\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty + \|(\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,1}(\Theta_0)]_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty \leq 2\beta_\infty.
$$

Consequently, by part (i) of Theorem 5.2, we obtain for $\tilde{L} > 0$

$$
\|\widehat{\Theta}^{\mathrm{g}} - \Theta_0\|_{\max} \leq \tilde{L}\sqrt{\frac{\log(d^2)}{T}} + \lambda_T\beta_\infty.
$$

Proof of point (ii). The same approach as in the proof of (i) can be applied. Since the regularizer is assumed to be (μ, ζ)-amenable, we have by Lemma 5 of Loh and Wainwright (2017) that $\lambda_T\widehat{\boldsymbol{z}}_{\mathcal{A}} -$

$\nabla_\theta \boldsymbol{q}(\lambda_T, \text{vec}(\widehat{\Theta})_{\mathcal{A}}) = 0$. Hence we have

$$\|\widehat{\boldsymbol{z}}_{\mathcal{A}^c}\|_\infty \leq \frac{1}{\lambda_T}\| - \nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty.$$

Following the same steps as in the proof of part (i), by upper bounding $\|\nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}^c}\|_\infty$ and $\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty$ we establish strict dual feasibility by Proposition 5.3. Then the remainder follows from part (ii) of Theorem 5.2. Point (ii) highlights the gain of non-convex penalties: the incoherence condition can be relaxed. □

Proof of Corollary 3.8. We remind that the oracle estimator is defined as

$$\widehat{\Theta}^{\text{ls},\mathcal{O}} = \underset{\theta\in\mathbb{R}^{|\mathcal{A}|}:\Theta\in\Omega}{\arg\min}\Big\{\mathbb{L}_T(\Theta)\Big\} = \underset{\theta\in\mathbb{R}^{|\mathcal{A}|}:\Theta\in\Omega}{\arg\min}\Big\{\|\Theta - \widehat{S}^{-1}\|_F^2\Big\}.$$

We first establish strict dual feasibility for

$$\widehat{\Theta}^{\text{ls}}_{\mathcal{A}} = \underset{\Theta:\Theta\in\Omega,\text{supp}(\Sigma)\subseteq\text{supp}(\Theta_0)}{\arg\min}\Big\{\mathbb{L}_T(\Theta) + \boldsymbol{p}(\lambda_T,\theta)\Big\},$$

where $\mathbb{L}_T(.)$ and Ω are defined as in (8). We only focus on (μ,ζ)-amenable penalty functions: indeed, should we consider μ-amenable penalty functions, the incoherence condition is not satisfied. We have

$$\widehat{z}_{\mathcal{A}^c} = \frac{1}{\lambda_T}\Big\{ - \nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\Big[\nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}} - \nabla \boldsymbol{q}(\lambda_T, \text{vec}(\widehat{\Theta}^{\text{ls}})_{\mathcal{A}}) + \lambda_T \widehat{z}_{\mathcal{A}}\Big]\Big\}.$$

Taking the ℓ_∞-norm, we obtain

$$\begin{aligned}\|\widehat{\boldsymbol{z}}_{\mathcal{A}^c}\|_\infty &\leq \tfrac{1}{\lambda_T}\| - \nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty \\ &+ \tfrac{1}{\lambda_T}\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\big(\lambda_T\widehat{z}_{\mathcal{A}} - \nabla_\theta \boldsymbol{q}(\lambda_T, \text{vec}(\widehat{\Theta}^{\text{ls}})_{\mathcal{A}})\big)\|_\infty \\ &\leq \tfrac{1}{\lambda_T}\| - \nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty,\end{aligned}$$

where we used $\lambda_T\widehat{z}_{\mathcal{A}} - \nabla_\theta \boldsymbol{q}(\lambda_T, \text{vech}(\widehat{\Theta})_{\mathcal{A}}) = 0$ from Lemma 8 of Loh and Wainwright (2017). Then:

$$\|\widehat{z}_{\mathcal{A}^c}\|_\infty \leq \frac{1}{\lambda_T}\|2\text{vec}(\widehat{S}^{-1} - \Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}2\text{vec}(\Theta_0 - \widehat{S}^{-1})_{\mathcal{A}}\|_\infty.$$

Since $\widehat{\boldsymbol{K}} = I_d$, we have with probability at least $1 - O(\exp\big(-\log(d^2)\big)) - o(\exp\big(-\log(d^2)\big))$ that

$$\|\widehat{z}_{\mathcal{A}^c}\|_\infty \leq \frac{2}{\lambda_T}\Big(\|\text{vec}(\widehat{S}^{-1} - \Theta_0)_{\mathcal{A}^c}\|_\infty + \|\text{vec}(\Theta_0 - \widehat{S}^{-1})_{\mathcal{A}}\|_\infty\Big) \leq \frac{L}{\lambda_T}\Big(\sqrt{(d^2-k_0)\frac{\log(d^2)}{T}} + \sqrt{k_0\frac{\log(d^2)}{T}}\Big),$$

using the arguments in the proof of Corollary 3.4 on the upper bound of the gradient. Provided $\lambda_T > L\sqrt{\{(d^2-k_0)\vee k_0\}\frac{\log(d^2)}{T}}$, strict dual feasibility holds and support recovery is satisfied by Theorem 5.1 of Loh and Wainwright.

As for the ℓ_∞-bound, using $\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty = 1$, we have

$$\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_T(\Theta)_{\mathcal{A}}\|_\infty = \|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} 2\mathrm{vec}(\Theta_0 - \widehat{S}^{-1})_{\mathcal{A}}\|_\infty \leq L\sqrt{k_0 \frac{\log(d^2)}{T}},$$

for $L > 0$ large enough, with probability $1 - O(\exp\big(-\log(d^2)\big)) - o(\exp\big(-\log(d^2)\big))$. □

Proof of Corollary 3.9. The oracle estimator for the von Neumann divergence is defined as

$$\widehat{\Theta}^{\mathrm{vn},\mathcal{O}} = \underset{\theta\in\mathbb{R}^{|\mathcal{A}|}:\Theta\in\Omega^{\mathrm{vn}}}{\arg\min} \Big\{\mathbb{L}_{T,0}(\Theta)\Big\} = \underset{\theta\in\mathbb{R}^{|\mathcal{A}|}:\Theta\in\Omega^{\mathrm{vn}}}{\arg\min} \Big\{\mathrm{tr}\Big(\Theta - \log(\Theta/\nu)\widehat{S}^{-1}\Big)/\nu\Big\}, \text{ with } \Omega^{\mathrm{vn}} \text{ as in (11).}$$

Proof of point (i). The PDW construction is based on the estimator

$$\widehat{\Theta}^{\mathrm{vn}}_{\mathcal{A}} = \underset{\Theta:\Theta\in\Omega^{\mathrm{vn}},\mathrm{supp}(\Sigma)\subseteq\mathrm{supp}(\Theta_0)}{\arg\min} \Big\{\mathbb{L}_{T,0}(\Theta) + \boldsymbol{p}(\lambda_T,\theta)\Big\},$$

where $\mathbb{L}_{T,0}(.)$ and Ω^{vn} are defined as in (11). The proof relies on the same steps as in the proof of Corollary 3.7: we need to show that strict dual feasibility holds for the previous criterion satisfied in the presence of μ-amenable penalty functions. Now based on the same steps as in point (i) of the proof of Corollary 3.7:

$$\begin{aligned}
\|\widehat{\boldsymbol{z}}_{\mathcal{A}^c}\|_\infty &\leq \tfrac{1}{\lambda_T}\| - \nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty \\
&\quad + \tfrac{1}{\lambda_T}\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\big(\lambda_T\widehat{z}_{\mathcal{A}} - \nabla_\theta \boldsymbol{q}(\lambda_T, \mathrm{vec}(\widehat{\Theta}^{\mathrm{vn}})_{\mathcal{A}})\big)\|_\infty \\
&\leq \tfrac{1}{\lambda_T}\| - \nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty + \|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty.
\end{aligned}$$

Moreover, we have

$$\| - \nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty \leq \|\nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}^c}\|_\infty + \|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty.$$

First, let us consider $\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty$. To do so, we first control the following quantity:

$$\begin{aligned}
&\widehat{\boldsymbol{K}} - \nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0) \\
&= \int_0^1 \big(\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0 + s(\widehat{\Theta}^{\mathrm{vn}} - \Theta_0)) - \nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)\big)ds \\
&= \int_0^1 s\nabla_\theta\big(\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_s)\big)\mathrm{vec}\big(\widehat{\Theta}^{\mathrm{vn}} - \Theta_0\big)ds,
\end{aligned}$$

where we used the mean value theorem with Θ_s lying between Θ_0 and $s(\widehat{\Theta} - \Theta_0)$. Such expansion is necessary due to the high non-linearity with respect to the parameters of the Hessian function. Let $w = d^2$,

for any $(\boldsymbol{u},\boldsymbol{v})\in\mathbb{R}^w\times\mathbb{R}^w$, we have

$$\begin{aligned}
&|\boldsymbol{u}^\top\{\widehat{\boldsymbol{K}}-\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)\}\boldsymbol{v}|\\
&= |\int_0^1\{s\sum_{a,b,c=1}^{w}\big(\nabla^3_{abc}\mathbb{L}_{T,0}(\Theta_s)\mathrm{vec}(\widehat{\Theta}^{\mathrm{vn}}-\Theta_0)_c\boldsymbol{u}_a\boldsymbol{v}_b\big)\}ds|\\
&\leq \int_0^1 s|\{\sum_{a,b,c=1}^{w}\big(\nabla^3_{abc}\mathbb{L}_{T,0}(\Theta_s)\mathrm{vec}(\widehat{\Theta}^{\mathrm{vn}}-\Theta_0)_c\boldsymbol{u}_a\boldsymbol{v}_b\big)\}|ds,
\end{aligned}$$

where ∇^3_{abc} refers to the third order derivative with respect to θ. Let us now derive the third order derivative of $\mathbb{L}_{T,0}(.)$. Element by element, for any $a,b,c=1,\cdots,w$, we have

$$\begin{aligned}
&\partial^3_{abc}\mathbb{L}_{T,0}(\Theta)\\
&= -\sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=1}^{k}\sum_{j=1}^{k-l}\mathrm{tr}\Big(\sum_{x=1}^{j-1}(I_d-\Theta/\nu)^{x-1}(\partial_c\Theta/\nu)(I_d-\Theta/\nu)^{j-1-x}(\partial_b\Theta/\nu)(I_d-\Theta/\nu)^{k-l-j}(\widehat{S}^{-1}/\nu)(I_d-\Theta/\nu)^{l-1}(\partial_a\Theta/\nu)\\
&\quad+(I_d-\Theta/\nu)^{j-1}(\partial_b\Theta/\nu)\sum_{y=1}^{k-l-j}(I_d-\Theta/\nu)^{y-1}(\partial_c\Theta/\nu)(I_d-\Theta/\nu)^{k-l-j-y}(\widehat{S}^{-1}/\nu)(I_d-\Theta/\nu)^{l-1}(\partial_a\Theta/\nu)\\
&\quad+(I_d-\Theta/\nu)^{j-1}(\partial_b\Theta/\nu)(I_d-\Theta/\nu)^{k-l-j}(\widehat{S}^{-1}/\nu)\sum_{z=1}^{l-1}(I_d-\Theta/\nu)^{z-1}(\partial_c\Theta/\nu)(I_d-\Theta/\nu)^{l-1-z}(\partial_a\Theta/\nu)\Big)\\
&\quad-\sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=1}^{k}\sum_{i=1}^{l-1}\mathrm{tr}\Big(\sum_{x=1}^{k-l}(I_d-\Theta/\nu)^{x-1}(\partial_c\Theta/\nu)(I_d-\Theta/\nu)^{k-l-x}(\widehat{S}^{-1}/\nu)(I_d-\Theta/\nu)^{i-1}(\partial_b\Theta/\nu)(I_d-\Theta/\nu)^{l-1-i}(\partial_a\Theta/\nu)\\
&\quad+(I_d-\Theta/\nu)^{k-l}(\widehat{S}^{-1}/\nu)\sum_{y=1}^{i-1}(I_d-\Theta/\nu)^{y-1}(\partial_c\Theta/\nu)(I_d-\Theta/\nu)^{i-1-y}(\partial_b\Theta/\nu)(I_d-\Theta/\nu)^{l-1-i}(\partial_a\Theta/\nu)\\
&\quad+(I_d-\Theta/\nu)^{k-l}(\widehat{S}^{-1}/\nu)(I_d-\Theta/\nu)^{i-1}(\partial_b\Theta/\nu)\sum_{z=1}^{l-1-i}(I_d-\Theta/\nu)^{z-1}(\partial_c\Theta/\nu)(I_d-\Theta/\nu)^{l-1-i-1}(\partial_a\Theta/\nu)\Big)\\
&:= \sum_{r=1}^{6}\Delta_{r,abc}.
\end{aligned}$$

Now by the Cauchy-Schwartz inequality, we have

$$|\nabla_\theta\Big\{\boldsymbol{u}^\top\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_s)\boldsymbol{v}\Big\}\mathrm{vec}(\widehat{\Theta}-\Theta_0)|^2\leq\Big\{\sum_{a,b,c=1}^{w}\partial^3_{abc}\mathbb{L}_{T,0}(\Theta_s)^2\boldsymbol{u}_a^2\boldsymbol{v}_b^2\Big\}\|\widehat{\Theta}^{\mathrm{vn}}-\Theta_0\|_F^2.$$

Regarding the third order derivative, taking the supremum on unit vectors $\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^{\mathrm{card}(\mathcal{A})}$, using the Frobenius norm, we have the upper bound:

$$\sum_{a,b,c=1}^{w}\partial^3_{abc}\mathbb{L}_{T,0}(\Theta_s)^2\boldsymbol{u}_a^2\boldsymbol{v}_b^2\leq k_0\Big\{\sqrt{k_0}\frac{1}{\nu^3}\|\widehat{S}^{-1}/\nu\|_F\sum_{k=1}^{\infty}\|I_d-\Theta_s/\nu\|_s^{k-3}(k^2-3k+2)\Big\}^2,$$

where we used $\|(I_d-\Theta_s/\nu)^l\|_F\leq\sqrt{k_0}\|(I_d-\Theta_s/\nu)^l\|_s\leq\sqrt{k_0}\|I_d-\Theta_s/\nu\|_s^l$ since we restrict to the $\mathcal{A}$ block.

Now rewriting this series as a combination of geometric series, we deduce

$$
\begin{aligned}
&\sum_{a,b,c=1}^{w} \partial^3_{abc}\mathbb{L}_{T,0}(\Theta_s)^2\boldsymbol{u}_a^2\boldsymbol{v}_b^2 \\
&\leq \quad k_0\Big\{\frac{1}{\nu^3}\sqrt{k_0}\|\widehat{S}^{-1}/\nu\|_F\Big[\|I_d-\Theta_s/\nu\|_s^{-1}\sum_{k=2}^{\infty}k(k-1)\|I_d-\Theta_s/\nu\|_s^{k-2}-2\|I_d-\Theta_s/\nu\|_s^{-2}\sum_{k=1}^{\infty}k\|I_d-\Theta_s/\nu\|_s^{k-1} \\
&\qquad +2\|I_d-\Theta_s/\nu\|_s^{-2}\sum_{k=0}^{\infty}\|I_d-\Theta_s/\nu\|_s^{k}\Big]\Big\}^2 \\
&\leq \quad k_0\Big\{\frac{1}{\nu^3}\sqrt{k_0}\|\widehat{S}^{-1}/\nu\|_F\Big[\|I_d-\Theta_s/\nu\|_s^{-1}\frac{2}{(1-\|I_d-\Theta_s/\nu\|_s)^3}-2\|I_d-\Theta_s/\nu\|_s^{-2}\frac{1}{(1-\|I_d-\Theta_s/\nu\|_s)^2} \\
&\qquad +2\frac{1}{1-\|I_d-\Theta_s/\nu\|_s}\Big]\Big\}^2 \\
&= \quad 2k_0\Big\{\frac{1}{\nu^3}\sqrt{k_0}\frac{\|\widehat{S}^{-1}/\nu\|_F}{1-\|I_d-\Theta_s/\nu\|_s}\Big[\|I_d-\Theta_s/\nu\|_s^{-1}\frac{1}{(1-\|I_d-\Theta_s/\nu\|_s)^2} \\
&\qquad -\|I_d-\Theta_s/\nu\|_s^{-2}\frac{2}{(1-\|I_d-\Theta_s/\nu\|_s)}+1\Big]\Big\}^2 \\
&:= \quad k_0\Big\{\frac{1}{\nu^3}\sqrt{k_0}\|\widehat{S}^{-1}/\nu\|_F\phi(\Theta_s,\nu)\Big\}^2.
\end{aligned}
$$

Now let us bound $1-\|I_d-\Theta_s/\nu\|_s$. For ν large enough ensuring that the spectral norm of Θ/ν is smaller than 1, we have $\|I_d-\Theta_s/\nu\|_s = 1-\lambda_{\min}(\Theta_s/\nu)$. The inequality $\frac{1}{1-\|I_d-\Theta_s/\nu\|_s} \leq \lambda_{\min}(\Theta_s/\nu)^{-1}$ then holds. Moreover,

$$\lambda_{\min}(\Theta_s/\nu) \geq \lambda_{\min}(\Theta_0/\nu) + s\lambda_{\min}((\widehat{\Theta}^{\mathrm{vn}}-\Theta_0)/\nu) \geq \lambda_{\min}(\Theta_0/\nu) - \frac{s}{\nu}\|\widehat{\Theta}^{\mathrm{vn}}-\Theta_0\|_F.$$

Thus using the upper bound with respect to $\|\widehat{\Theta}^{\mathrm{vn}}-\Theta_0\|_F$ from Corollary 3.5, we obtain

$$\lambda_{\min}(\Theta_s/\nu) \geq \lambda_{\min}(\Theta_0/\nu) - \frac{s}{\nu}L\sqrt{k_0\frac{\log(d^2)}{T}},$$

for $L>0$ large enough. We deduce $\lambda_{\min}(\Theta_s/\nu) \geq \lambda_{\min}(\Theta_0/\nu) - \epsilon_{T,d} > 0$, with $\epsilon_{T,d} \to 0$ as $T,d \to \infty$ for a suitable scaling behaviour (T,d,k_0). Consequently, we deduce $(1-\|I_d-\Theta_s/\nu\|_s)^{-1} \leq (\lambda_{\min}(\Theta_0/\nu)-\epsilon_{T,d})^{-1}$. Now let us treat $\|I_d-\Theta_s/\nu\|_s^{-1}$. We have

$$\|I_d-\Theta_s/\nu\|_s = 1-\lambda_{\min}(\Theta_s/\nu) \geq 1-\lambda_{\max}(\Theta_s/\nu) \geq 1-\lambda_{\max}(\Theta_0/\nu)-\frac{1}{\nu}\|\widehat{\Theta}^{\mathrm{vn}}-\Theta_0\|_F \geq 1-\lambda_{\max}(\Theta_0/\nu)-\epsilon_{T,d},$$

with $\epsilon_{T,d} \to 0$ as $T,d \to \infty$ for a suitable scaling behaviour (T,d,k_0). As a consequence, we obtain

$$\|I_d-\Theta_s/\nu\|_s^{-1} \leq \Big(1-\lambda_{\max}(\Theta_0/\nu)-\epsilon_{T,d}\Big)^{-1}.$$

As for $\|\widehat{S}^{-1}\|_F$, we have $\|\widehat{S}^{-1}\|_F \leq \|\widehat{S}^{-1} - \Sigma_x^{-1}\|_F + \|\Sigma_x^{-1}\|_F$. Using the same arguments as in the proof of Corollary 3.4 for bounding $\|\widehat{S}^{-1} - \Sigma_x^{-1}\|_F$, we have $\|\widehat{S}^{-1} - \Sigma_x^{-1}\|_F \leq \sqrt{k_0}\|\widehat{S}^{-1} - \Sigma_x^{-1}\|_s \leq K\sqrt{k_0^2\frac{\log(d^2)}{T}}$, with high probability and for $K > 0$ large enough. Putting all the pieces together, we deduce

$$
\begin{aligned}
&\Big\{\sum_{a,b,c=1}^{w} \partial^3_{abc}\mathbb{L}_{T,0}(\Theta_s)^2\boldsymbol{u}_a^2\boldsymbol{v}_b^2\Big\}\|\widehat{\Theta}^{\mathrm{vn}} - \Theta_0\|_F^2 \\
&\leq\ Lk_0\Big[\frac{1}{\nu^4}\sqrt{k_0}\Big(K\sqrt{k_0^2\frac{\log(d^2)}{T}} + \|\Theta_0\|_F\Big)\Big(\lambda_{\min}(\Theta_0/\nu) - \epsilon_{T,d}\Big)^{-1} \\
&\quad\times\Big\{\Big(1 - \lambda_{\max}(\Theta_0/\nu) - \epsilon_{T,d}\Big)^{-1}\Big(\lambda_{\min}(\Theta_0/\nu) - \epsilon_{T,d}\Big)^{-2} - \Big(1 - \lambda_{\max}(\Theta_0/\nu) - \epsilon_{T,d}\Big)^{-2} \\
&\quad\Big(\lambda_{\min}(\Theta_0/\nu) - \epsilon_{T,d}\Big)^{-1} + 1\Big\}\Big]^2 k_0\frac{\log(d^2)}{T} \\
&\leq\ Lk_0\Big[\frac{1}{\nu^4}\sqrt{k_0}\Big(K\sqrt{k_0^2\frac{\log(d^2)}{T}} + \|\Theta_0\|_F\Big)\Big(\lambda_{\min}(\Theta_0/\nu) - \epsilon_{T,d}\Big)^{-2}\Big(1 - \lambda_{\max}(\Theta_0/\nu) - \epsilon_{T,d}\Big)^{-1} \\
&\quad\times\Big\{\Big(\lambda_{\min}(\Theta_0/\nu) - \epsilon_{T,d}\Big)^{-1} - \Big(1 - \lambda_{\max}(\Theta_0/\nu) - \epsilon_{T,d}\Big)^{-1} \\
&\quad+\Big(\lambda_{\min}(\Theta_0/\nu) - \epsilon_{T,d}\Big)\Big(1 - \lambda_{\max}(\Theta_0/\nu) - \epsilon_{T,d}\Big)\Big\}\Big]^2 k_0\frac{\log(d^2)}{T}.
\end{aligned}
$$

with high probability for $L > 0$ large enough. Now, let us treat $\epsilon_{T,d} \to 0$ under a suitable scaling (T, d). Using $\lambda_{\min}(\Theta_0/\nu)^{-1} = \nu\|\Sigma_x\|_s$, for $\tilde{L} > 0$ large enough, with high probability, the upper bound becomes

$$
\begin{aligned}
&\Big\{\sum_{a,b,c=1}^{w} \partial^3_{abc}\mathbb{L}_{T,0}(\Theta_s)^2\boldsymbol{u}_a^2\boldsymbol{v}_b^2\Big\}\|\widehat{\Theta}^{\mathrm{vn}} - \Theta_0\|_F^2 \\
&\quad\leq\ \tilde{L}k_0\Big[\frac{1}{\nu^4}\sqrt{k_0}\Big(K\sqrt{k_0^2\frac{\log(d^2)}{T}} + \|\Theta_0\|_F\Big)\nu^3\|\Sigma_x\|_s^2\Big(1 - \|\Theta_0/\nu\|_s\Big)^{-1}\Big]^2 k_0\frac{\log(d^2)}{T}.
\end{aligned}
$$

Thus, taking the supremum with respect to unit vector $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{\mathcal{A}}$, since $\|\Theta_0\|_F$ is of order k_0, we obtain

$$
\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}} - \nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}\mathcal{A}}\|_s \leq C\sqrt{\frac{1}{\nu^2}k_0^7\frac{\log(d^2)}{T}\|\Sigma_x\|_s^4\Big(1 - \|\Theta_0/\nu\|_s\Big)^{-2}}. \tag{30}
$$

for $C > 0$ sufficiently large. Furthermore, we have

$$
\begin{aligned}
&\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}\mathcal{A}}\|_s \\
&\quad\leq\ \|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}} - \nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}\mathcal{A}}\|_s + \|\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}\mathcal{A}}\|_s.
\end{aligned}
$$

Let us control for $\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}\mathcal{A}}$. First, using the formula we obtained for

$\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta)$ in the proof of Corollary 3.5, the population level Hessian $\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]$ is

$$\begin{aligned}
&\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)] \\
&= \sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=1}^{k}\sum_{j=1}^{k-l}\frac{1}{2}\Big[\Big\{(I_p-\Theta_0/\nu)^{j-1}\otimes(I_p-\Theta_0/\nu)^{k-l-j}(\Big\{\mathbb{E}[\widehat{S}^{-1}]\Big\}/\nu)(I_p-\Theta_0/\nu)^{l-1}\Big\} \\
&\quad +\Big\{(I_p-\Theta_0/\nu)^{j-1}\otimes(I_p-\Theta_0/\nu)^{l-1}(\Big\{\mathbb{E}[\widehat{S}^{-1}]\Big\}/\nu)(I_p-\Theta_0/\nu)^{k-l-j}\Big\}\Big]/\nu^2 \\
&+ \sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=1}^{k}\sum_{i=1}^{l-1}\frac{1}{2}\Big[\Big\{(I_p-\Theta_0/\nu)^{k-l}(\Big\{\widehat{S}^{-1}-\mathbb{E}[\widehat{S}^{-1}]\Big\}/\nu)(I_p-\Theta_0/\nu)^{i-1}\otimes(I_p-\Theta_0/\nu)^{l-1-i}\Big\} \\
&\quad +\Big\{(I_p-\Theta_0/\nu)^{i-1}(\Big\{\mathbb{E}[\widehat{S}^{-1}]\Big\}/\nu)(I_p-\Theta_0/\nu)^{k-l}\otimes(I_p-\Theta_0/\nu)^{l-1-i}\Big\}\Big]/\nu^2. \qquad (31)
\end{aligned}$$

Now by the properties on the Kronecker product, we obtain

$$\begin{aligned}
&\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}\mathcal{A}} \\
&= \sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=1}^{k}\sum_{j=1}^{k-l}\frac{1}{2}\Big[\Big\{(I_p-\Theta_0/\nu)^{j-1}\otimes(I_p-\Theta_0/\nu)^{k-l-j}(\Big\{\widehat{S}^{-1}-\mathbb{E}[\widehat{S}^{-1}]\Big\}/\nu)(I_p-\Theta_0/\nu)^{l-1}\Big\} \\
&\quad +\Big\{(I_p-\Theta_0/\nu)^{j-1}\otimes(I_p-\Theta_0/\nu)^{l-1}(\Big\{\widehat{S}^{-1}-\mathbb{E}[\widehat{S}^{-1}]\Big\}/\nu)(I_p-\Theta_0/\nu)^{k-l-j}\Big\}\Big]/\nu^2 \\
&+ \sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=1}^{k}\sum_{i=1}^{l-1}\frac{1}{2}\Big[\Big\{(I_p-\Theta_0/\nu)^{k-l}(\Big\{\widehat{S}^{-1}-\mathbb{E}[\widehat{S}^{-1}]\Big\}/\nu)(I_p-\Theta_0/\nu)^{i-1}\otimes(I_p-\Theta_0/\nu)^{l-1-i}\Big\} \\
&\quad +\Big\{(I_p-\Theta_0/\nu)^{i-1}(\Big\{\widehat{S}^{-1}-\mathbb{E}[\widehat{S}^{-1}]\Big\}/\nu)(I_p-\Theta_0/\nu)^{k-l}\otimes(I_p-\Theta_0/\nu)^{l-1-i}\Big\}\Big]/\nu^2.
\end{aligned}$$

First we consider $\mathbb{E}[\widehat{S}^{-1}]$. By a Taylor expansion, $\widehat{S}^{-1}=\Theta_0+\Sigma_x^{-1}\Big(\Sigma_x-\widehat{S}\Big)\Sigma_x^{-1}+u_T$, where $u_T=o_p(1)$. Thus, assuming $\mathbb{E}[\widehat{S}^{-1}]<\infty$, then $\mathbb{E}[\widehat{S}^{-1}]-\Theta_0=o(1)$ for a sufficiently large T. Using the properties of the multiplicative norm of the Kronecker product, we obtain for T large enough

$$\begin{aligned}
&\|\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}\mathcal{A}}\|_s \\
&\leq \frac{1}{\nu^3}\sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=1}^{k}\sum_{j=1}^{k-l}\|I_p-\Theta_0/\nu\|_s^{k-2}\|\widehat{S}^{-1}-\Sigma_x^{-1}\|_s+\frac{1}{\nu^3}\sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=1}^{k}\sum_{i=1}^{l-1}\|I_p-\Theta_0/\nu\|_s^{k-2}\|\widehat{S}^{-1}-\Sigma_x^{-1}\|_s \\
&\leq \frac{1}{\nu^3}\|\widehat{S}^{-1}-\Sigma_x^{-1}\|_s\sum_{k=1}^{\infty}(k-1)\|I_d-\Theta_0/\nu\|_s^{k-2}.
\end{aligned}$$

With high probability, we established $\|\widehat{S}^{-1}-\Sigma_x^{-1}\|_s\leq K\sqrt{k_0\frac{\log(d^2)}{T}}$ when restricted to the $\mathcal{A}$ block. As a consequence,

$$\begin{aligned}
\|\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}\mathcal{A}}\|_s &\leq \tfrac{1}{\nu^3}K\sqrt{k_0\tfrac{\log(d^2)}{T}}\Big(1-\|I_d-\Theta_0/\nu\|_s\Big)^{-2} \\
&\leq \tfrac{1}{\nu^3}K\sqrt{k_0\tfrac{\log(d^2)}{T}}\lambda_{\min}(\Theta_0/\nu)^{-2} \\
&\leq K\sqrt{\tfrac{1}{\nu^2}\|\Sigma_x\|_s^4 k_0\tfrac{\log(d^2)}{T}}.
\end{aligned}$$

for $L > 0$ sufficiently large since $\|I_d - \Theta_0/\nu\|_s < 1$. Thus, putting the pieces together, we deduce

$$\begin{aligned}
&\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}\mathcal{A}}\|_s \\
&\leq \quad C\sqrt{\frac{1}{\nu^2}k_0^7\frac{\log(d^2)}{T}\|\Sigma_x\|_s^4\Big(1-\|\Theta_0/\nu\|_s\Big)^{-2}} + K\sqrt{\frac{1}{\nu^2}\|\Sigma_x\|_s^4 k_0\frac{\log(d^2)}{T}}.
\end{aligned}$$

Thus we obtain

$$\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}\mathcal{A}}\|_s \leq L\sqrt{\frac{1}{\nu^2}k_0^7\frac{\log(d^2)}{T}\|\Sigma_x\|_s^4\Big(1-\|\Theta_0/\nu\|_s\Big)^{-2}},$$

for L a sufficiently large constant. Now, using the same trick as in the proof of Corollary 3.4 for controlling for $\|\Theta_0^{-1} - \widehat{S}\|_s$, we obtain

$$\|\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_s \leq L\sqrt{\frac{1}{\nu^2}k_0^7\frac{\log(d^2)}{T}\|\Sigma_x\|_s^4\Big(1-\|\Theta_0/\nu\|_s\Big)^{-2}}. \tag{32}$$

Based on the same arguments for deriving (30), by union bound, we have

$$\max_{i\in\mathcal{A}^c}\|e_i^\top\big(\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}} - \nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}^c\mathcal{A}}\big)\|_2 \leq C\sqrt{\frac{1}{\nu^2}k_0^7\frac{\log(d^2)}{T}\|\Sigma_x\|_s^4\Big(1-\|\Theta_0/\nu\|_s\Big)^{-2}},$$

which implies

$$\max_{i\in\mathcal{A}^c}\|e_i^\top\big(\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\big)\|_2 \leq C\sqrt{\frac{1}{\nu^2}k_0^7\frac{\log(d^2)}{T}\|\Sigma_x\|_s^4\Big(1-\|\Theta_0/\nu\|_s\Big)^{-2}}. \tag{33}$$

Now we are in a position to control the quantity:

$$\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}}\nabla_\theta\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty \leq M_1 + M_2, \text{ with}$$

$$\begin{aligned}
M_1 &:= \|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]^{-1}_{\mathcal{A}^c\mathcal{A}}\nabla_\theta\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty, \\
M_2 &:= \|\Big\{\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\Big\}\nabla_\theta\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty.
\end{aligned}$$

By assumption, the population level Hessian is bounded. As for the gradient, using Corollary 3.5, we obtain with high probability that $M_1 \leq C\sqrt{k_0\frac{\log(d^2)}{T}}\|\Theta_0^{-1}/\nu\|_s$. Regarding M_2, we have

$$M_2 \leq \max_{i\in\mathcal{A}^c}\|e_i^\top\{\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\}\|_2\|\nabla_\theta\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_2. \tag{34}$$

Restricting to the true support, the following holds:

$$\begin{aligned}
&\|\nabla_\theta\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_2 = \|\nabla_\Theta\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_F \\
&\leq \quad \sqrt{k_0}\|\nabla_\Theta\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_s \leq \tilde{L}\sqrt{k_0}\sqrt{k_0\frac{\log(d^2)}{T}}\|\Theta_0^{-1}/\nu\|_s = \tilde{L}\sqrt{k_0}\sqrt{\frac{1}{\nu^2}\|\Sigma_x\|_s^2 k_0\frac{\log(d^2)}{T}},
\end{aligned} \tag{35}$$

for $\tilde{L}$ a sufficiently large constant with high probability. Moreover

$$\begin{aligned}
&\|e_i^\top \{\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\}\|_2 \\
&\quad \leq \quad \|e_i^\top \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\Upsilon_1\|_2 + \|e_i^\top \Upsilon_2 \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_2 + \|e_i^\top \Upsilon_2\Upsilon_1\|_2, \qquad (36)
\end{aligned}$$

with $\Upsilon_1 := \widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}$, $\Upsilon_2 := \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}$. By inequalities (32) and (33), we obtain

$$\|\Upsilon_1\|_s \leq C\sqrt{\frac{1}{\nu^2}k_0^7\frac{\log(d^2)}{T}\|\Sigma_x\|_s^4\Big(1-\|\Theta_0/\nu\|_s\Big)^{-2}}, \ \max_{i\in\mathcal{A}^c}\|e_i^\top\Upsilon_2\|_2 \leq C\sqrt{\frac{1}{\nu^2}k_0^7\frac{\log(d^2)}{T}\|\Sigma_x\|_s^4\Big(1-\|\Theta_0/\nu\|_s\Big)^{-2}}.$$

Hence, using inequalities (34), (35) and (36), we obtain for a sufficiently large constant $\tilde{C}$ that

$$\begin{aligned}
&\|\Big\{\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}^c\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\Big\}\nabla_\theta\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty \\
&\quad \leq \quad \tilde{C}\sqrt{k_0}\sqrt{\frac{1}{\nu^2}k_0\frac{\log(d^2)}{T}\|\Sigma_x\|_s^2}\sqrt{\frac{1}{\nu^2}k_0^7\frac{\log(d^2)}{T}\|\Sigma_x\|_s^4\Big(1-\|\Theta_0/\nu\|_s\Big)^{-2}}.
\end{aligned}$$

Using the incoherence condition $\|\boldsymbol{K}_{0,\mathcal{A}^c\mathcal{A}}\boldsymbol{K}^{-1}_{0,\mathcal{A}\mathcal{A}}\|_\infty < \omega < 1$, we then obtain with high probability

$$\begin{aligned}
&\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}}\nabla_\theta\mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty \\
&\quad \leq \quad \omega K_1\sqrt{\frac{1}{\nu^2}\|\Sigma_x\|_s^2k_0^2\frac{\log(d^2)}{T}} + K_2\sqrt{\frac{1}{\nu^4}k_0^9\frac{\log(d^2)}{T}\|\Sigma_x\|_s^6\Big(1-\|\Theta_0/\nu\|_s\Big)^{-2}},
\end{aligned}$$

wit $K_1, K_2 > 0$ large enough. Moreover, using the incoherence condition and

$$\begin{aligned}
&\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty \\
&\quad \leq \quad \sqrt{k_0}\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty + \|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty \\
&\quad \leq \quad \tilde{C}\sqrt{\frac{1}{\nu^4}k_0^9\frac{\log(d^2)}{T}\|\Sigma_x\|_s^6\Big(1-\|\Theta_0/\nu\|_s\Big)^{-2}} + \omega.
\end{aligned}$$

Thus, putting the pieces together, we have for $L_1, L_2, L_3 > 0$ sufficiently large

$$\begin{aligned}
&\|\boldsymbol{z}_{\mathcal{A}^c}\|_\infty \\
&\quad \leq \quad \frac{1}{\lambda_T}\Big(L_1\sqrt{\frac{1}{\nu^4}k_0^9\frac{\log(d^2)}{T}\|\Sigma_x\|_s^6\Big(1-\|\Theta_0/\nu\|_s\Big)^{-2}} + L_2\sqrt{\frac{1}{\nu^2}(d^2-k_0)\|\Sigma_x\|_s^2\frac{\log(d^2)}{T}}\Big) \\
&\quad\quad +L_3\sqrt{\frac{1}{\nu^2}k_0^7\frac{\log(d^2)}{T}\|\Sigma_x\|_s^4\Big(1-\|\Theta_0/\nu\|_s\Big)^{-2}} + \omega.
\end{aligned}$$

Hence, strict dual feasibility of Theorem 5.1 is satisfied when

$$\frac{1}{1-\omega}L\sqrt{\frac{\log(d^2)}{T}} \leq \lambda_T,$$

under the scaling $T > C\frac{1}{\nu^2}\Big[(d^2-k_0) \vee \frac{k_0^9}{\nu^2}\|\Sigma_x\|_s^4\Big(1-\|\Theta_0/\nu\|_s\Big)^{-2}\Big]\|\Sigma_x\|_s^2\log(d^2)$.

We now focus on the ℓ_∞ bound. We have

$$\begin{aligned}
&\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty \\
&\leq \|\{\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}\mathcal{A}}^{-1}\}\nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty + \|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty \\
&\leq \sqrt{k_0}\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}\mathcal{A}}^{-1}\|_s\|\nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty + \|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}(\Theta_0)]_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty \\
&\leq C_1\sqrt{k_0}\sqrt{\frac{1}{\nu^2}k_0^7\frac{\log(d^2)}{T}\|\Sigma_x\|_s^4\Big(1-\|\Theta_0/\nu\|_s\Big)^{-2}}\sqrt{\frac{1}{\nu^2}\|\Sigma_x\|_s^2 k_0\frac{\log(d^2)}{T}} \\
&\quad + C_2\sqrt{\frac{1}{\nu^2}\|\Sigma_x\|_s^2 k_0\frac{\log(d^2)}{T}},
\end{aligned}$$

with $C_1, C_2 > 0$ using inequality (32). We proved

$$\begin{aligned}
&\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - (\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty \\
&\leq \sqrt{k_0}\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - (\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}\mathcal{A}})^{-1}\|_s \leq \sqrt{\frac{1}{\nu^2}k_0^7\frac{\log(d^2)}{T}\|\Sigma_x\|_s^4\Big(1-\|\Theta_0/\nu\|_s\Big)^{-2}} \leq \beta_\infty.
\end{aligned}$$

Hence

$$\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \leq \|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - (\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty + \|(\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_{T,0}(\Theta_0)]_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty \leq 2\beta_\infty.$$

Consequently, by part (i) of Theorem 5.2, we obtain for $\tilde{L} > 0$

$$\|\widehat{\Theta}^{\mathrm{vn}} - \Theta_0\|_{\max} \leq \tilde{L}\sqrt{\frac{\log(d^2)}{T}} + \lambda_T\beta_\infty.$$

Proof of point (ii). We follow the same proof as in point (ii) of Corollary 3.7. Hence we have

$$\|\widehat{\boldsymbol{z}}_{\mathcal{A}^c}\|_\infty \leq \frac{1}{\lambda_T}\| - \nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty.$$

Following the same steps as in the proof of part (i), by upper bounding $\|\nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}^c}\|_\infty$ and $\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta \mathbb{L}_{T,0}(\Theta_0)_{\mathcal{A}}\|_\infty$ we establish strict dual feasibility by Proposition 5.3. Then the remainder follows from part (ii) of Theorem 5.2.

□

Proof of Corollary 3.10. The oracle estimator is defined as

$$\widehat{\Theta}^{\mathrm{dt},\mathcal{O}} = \underset{\theta\in\mathbb{R}^{|\mathcal{A}|}:\Theta\in\Omega}{\arg\min}\Big\{\mathbb{L}_T(\Theta)\Big\} = \underset{\theta\in\mathbb{R}^{|\mathcal{A}|}:\Theta\in\Omega}{\arg\min}\Big\{\frac{1}{2}\mathrm{tr}\Big(\Theta^2\widehat{S}-\Theta\Big)\Big\}.$$

Proof of point (i). Strict dual feasibility is checked for

$$\widehat{\Theta}^{\mathrm{dt}} = \underset{\Theta:\Theta\in\Omega,\ \mathrm{supp}(\Theta)\subseteq\mathrm{supp}(\Theta_0)}{\arg\min}\Big\{\mathbb{L}_T(\Theta)+\boldsymbol{p}(\lambda_T,\theta)\Big\},$$

where $\mathbb{L}_T(.)$ and Ω are defined in (16). Following the same steps as in the Stein's case in point (i), we aim at bounding

$$\|\widehat{z}_{\mathcal{A}^c}\|_\infty \leq \frac{1}{\lambda_T}\| - \nabla_\theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty + \|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty,$$

with $\widehat{\boldsymbol{K}} = \int_0^1 \nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0+u(\widehat{\Theta}^{\mathrm{dt}}-\Theta_0))du$. Now we have

$$\| - \nabla_\theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty \leq \|\nabla_\theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}^c}\|_\infty + \|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty.$$

Let us highlight that the Hessian matrix does not depend on the parameter Θ. Proceeding as in the proof of point (i) of Corollary 3.7, we obtain

$$\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\nabla_\theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty \leq M_1+M_2, \text{ where}$$

$$\begin{aligned} M_1 &= \|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\nabla_\theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty, \\ M_2 &= \|\Big\{\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - \mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\Big\}\nabla_\theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty. \end{aligned}$$

Note that $\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)] = \frac{1}{2}\Big(\Sigma_x\otimes I_d + I_d\otimes\Sigma_x\Big)$. Then, we have

$$M_1 \leq \|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty\|\nabla_\theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty \leq \omega K\sqrt{\|\Theta_0\|_s^2 k_0\frac{\log(d^2)}{T}},$$

with high probability and using the incoherence condition, a condition identical to Zhang and Zou (2014), Theorem 2. As for M_2, we first consider:

$$\|\widehat{\boldsymbol{K}}-\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]\|_\infty = \frac{1}{2}\|\Big(\widehat{S}-\Sigma_x\otimes I_d\Big)+\Big(I_d\otimes\widehat{S}-\Sigma_x\Big)\|_\infty \leq \|\widehat{S}-\Sigma_x\|_\infty \leq d\|\widehat{S}-\Sigma_x\|_{\max} \leq L\sqrt{d^2\frac{\log(d^2)}{T}},$$

with high probability for $L>0$ large enough. Then let us consider the probability:

$$\begin{aligned}
&\forall \epsilon>0,\ \mathbb{P}(\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\|_\infty>\epsilon)\\
&\quad\leq\ \mathbb{P}(\max_{j\in\mathcal{A}^c}\sum_{l=1}^{k_0}|\widehat{\boldsymbol{K}}_{jl}-\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]_{jl}|>\epsilon)=\mathbb{P}(\max_{j\in\mathcal{A}^c}\sum_{l=1}^{k_0}|\widehat{S}_{jl}-\Sigma_{x,jl}|>\epsilon)\leq(d^2-k_0)\sum_{l=1}^{k_0}\mathbb{P}(|\widehat{S}_{jl}-\Sigma_{x,jl}|>\epsilon/k_0).
\end{aligned}$$

Then we deduce for $L>0$ large enough and with a suitable sample size that

$$\upsilon_1=\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\|_\infty,\ \upsilon_2=\|\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty,$$

are bounded by $L\sqrt{k_0\frac{\log(d^2)}{T}}$ with high probability. Then

$$\begin{aligned}
M_2\leq&\Big\{\|\Big(\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\Big)\Big(\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\Big)\|_\infty\\
&+\|\Big(\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\Big)\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty\\
&+\|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\Big(\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\Big)\|_\infty\Big\}\|\nabla_\Theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_{\max}\\
\leq&\ \Big\{\upsilon_1\upsilon_2+\upsilon_1\|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty+\upsilon_2\|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\|_\infty\Big\}\|\nabla_\Theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_{\max}\\
\leq&\ \Big\{\upsilon_1\upsilon_2+\upsilon_1\|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty+\upsilon_2\|\Sigma_x\|_\infty\Big\}\|\nabla_\Theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_{\max}\\
\leq&\ K\sqrt{k_0\|\Sigma_x\|^2_\infty\frac{\log(d^2)}{T}}\|\nabla_\Theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_{\max},
\end{aligned}$$

with high probability and for K sufficiently large, where we used $\|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\|_\infty\leq\|\Sigma_x\|_\infty$. Since $\|\nabla_\Theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_{\max}\leq K\sqrt{k_0\|\Theta_0\|^2_s\frac{\log(d^2)}{T}}$ with high probability, we conclude that

$$\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}}\nabla_\theta\mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty\leq L\sqrt{k_0\|\Sigma_x\|^2_\infty\frac{\log(d^2)}{T}}\sqrt{\|\Theta_0\|^2_s\frac{k_0\log(d^2)}{T}}.$$

with high probability under the scaling $T>Mk_0^2\Big\{\|\Sigma_x\|^2_\infty\vee\|\Theta_0\|^2_s\Big\}\log(d^2)$. Furthermore, by the incoherence condition

$$\begin{aligned}
&\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty\\
&\quad\leq\ \|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}}\widehat{\boldsymbol{K}}^{-1}_{\mathcal{A}\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty+\|\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta\theta^\top}\mathbb{L}_T(\Theta_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty\\
&\quad\leq\ L\sqrt{k_0\|\Sigma_x\|^2_\infty\frac{\log(d^2)}{T}}+\omega.
\end{aligned}$$

Thus we have for $L>0$ sufficiently large

$$\|\boldsymbol{z}_{\mathcal{A}^c}\|_\infty\leq\frac{1}{\lambda_T}\Big(L_1\sqrt{k_0^2\Big\{\|\Sigma_x\|^2_\infty\vee\|\Theta_0\|^2_s\Big\}\frac{\log(d^2)}{T}}+L_2\sqrt{(d^2-k_0)\|\Theta_0\|^2_s\frac{\log(d^2)}{T}}\Big)+L_3\sqrt{k_0\|\Sigma_x\|^2_\infty\frac{\log(d^2)}{T}}+\omega,$$

for $L_1, L_2, L_3 > 0$. Hence, then strict dual feasibility of Theorem 5.1 is satisfied when

$$\frac{1}{1-\omega} L \sqrt{\frac{\log(d^2)}{T}} \leq \lambda_T,$$

under the scaling $T > C \max\left\{k_0^2 \left[\|\Sigma_x\|_\infty^2 \vee \|\Theta_0\|_s^2\right], (d^2 - k_0)\|\Theta_0\|_s^2\right\} \log(d^2)$.

Let us consider the ℓ_∞ bound to apply point (i) of Theorem 5.2. We have

$$\begin{aligned}
&\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} \nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty \\
&\leq \|\{\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - \mathbb{E}[\nabla^2_{\theta\theta^\top} \mathbb{L}_T(\Theta_0)]_{\mathcal{A}\mathcal{A}}^{-1}\} \nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty + \|\mathbb{E}[\nabla^2_{\theta\theta^\top} \mathbb{L}_T(\Theta_0)]_{\mathcal{A}\mathcal{A}}^{-1} \nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty \\
&\leq \|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - \mathbb{E}[\nabla^2_{\theta\theta^\top} \mathbb{L}_T(\Theta_0)]_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \|\nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty + \|\mathbb{E}[\nabla^2_{\theta\theta^\top} \mathbb{L}_T(\Theta_0)]_{\mathcal{A}\mathcal{A}}^{-1} \nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty \\
&\leq C_1 \sqrt{k_0 \frac{\log(d^2)}{T}} \sqrt{k_0 \|\Theta_0\|_s^2 \frac{\log(d^2)}{T}} + C_2 \sqrt{k_0 \|\Theta_0\|_\infty^2 \frac{\log(d^2)}{T}},
\end{aligned}$$

with $C_1, C_2 > 0$ and using our previous upper bounds. We proved

$$\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - (\mathbb{E}[\nabla^2_{\theta\theta^\top} \mathbb{L}_T(\Theta_0)]_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty \leq L \sqrt{k_0 \frac{\log(d^2)}{T}} \leq \beta_\infty.$$

Hence

$$\|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \leq \|\widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} - (\mathbb{E}[\nabla^2_{\theta\theta^\top} \mathbb{L}_T(\Theta_0)]_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty + \|(\mathbb{E}[\nabla^2_{\theta\theta^\top} \mathbb{L}_T(\Theta_0)]_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty \leq 2\beta_\infty.$$

Consequently, by part (i) of Theorem 5.2, we obtain

$$\|\widehat{\Theta}^{\mathrm{g}} - \Theta_0\|_{\max} \leq \tilde{L} \sqrt{\frac{\log(d^2)}{T}} + \lambda_T \beta_\infty,$$

for $\tilde{L} > 0$.

Proof of point (ii). The same approach as in the proof of (i) can be applied. Since the regularizer is assumed to be (μ, ζ)-amenable, we have by Lemma 5 of Loh and Wainwright (2017) that $\lambda_T \widehat{\boldsymbol{z}}_{\mathcal{A}} - \nabla_\theta \boldsymbol{q}(\lambda_T, \mathrm{vec}(\widehat{\Theta})_{\mathcal{A}}) = 0$. Hence we have

$$\|\widehat{\boldsymbol{z}}_{\mathcal{A}^c}\|_\infty \leq \frac{1}{\lambda_T} \| - \nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}^c} + \widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}} \widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} \nabla_\theta \mathbb{L}_{T,1}(\Theta_0)_{\mathcal{A}}\|_\infty.$$

Following the same steps as in the proof of part (i), by upper bounding $\|\nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}^c}\|_\infty$ and $\|\widehat{\boldsymbol{K}}_{\mathcal{A}^c\mathcal{A}} \widehat{\boldsymbol{K}}_{\mathcal{A}\mathcal{A}}^{-1} \nabla_\theta \mathbb{L}_T(\Theta_0)_{\mathcal{A}}\|_\infty$, we establish strict dual feasibility by Proposition 5.3. Then the remainder follows from part (ii) of Theorem 5.2. □