Testing for high-dimensional network parameters in auto-regressive models

Lili Zheng[1] and Garvesh Raskutti[1]

Abstract

High-dimensional auto-regressive models provide a natural way to model influence between M actors given multi-variate time series data for T time intervals. While there has been considerable work on network estimation, there is limited work in the context of inference and hypothesis testing. In particular, prior work on hypothesis testing in time series has been restricted to linear Gaussian auto-regressive models. From a practical perspective, it is important to determine suitable statistical tests for connections between actors that go beyond the Gaussian assumption. In the context of *high-dimensional* time series models, confidence intervals present additional estimators since most estimators such as the Lasso and Dantzig selectors are biased which has led to *de-biased* estimators. In this paper we address these challenges and provide convergence in distribution results and confidence intervals for the multi-variate AR(p) model with sub-Gaussian noise, a generalization of Gaussian noise that broadens applicability and presents numerous technical challenges. The main technical challenge lies in the fact that unlike Gaussian random vectors, for sub-Gaussian vectors zero correlation does not imply independence. The proof relies on using an intricate truncation argument to develop novel concentration bounds for quadratic forms of dependent sub-Gaussian random variables. Our convergence in distribution results hold provided $T = \Omega((s \vee \rho)^2 \log^2 M)$, where s and ρ refer to sparsity parameters which matches existed results for hypothesis testing with i.i.d. samples. We validate our theoretical results with simulation results for both block-structured and chain-structured networks.

1 Introduction

Vector autoregressive models arise in a number of applications including macroeconomics (see e.g.Ang and Piazzesi [2003],Hansen [2003],Shan [2005]), computational neuroscience (see e.g.Goebel et al. [2003],Seth et al. [2015],Harrison et al. [2003], Bressler et al. [2007]), and many others (see e.g.Michailidis and dAlché Buc [2013],Fujita et al. [2007]). Recent years has seen substantial

---

[1]Department of Statistics, University of Wisconsin-Madison

development in the theory and methodology of high-dimensional auto-regressive models with respect to parameter estimation (see e.g. Song and Bickel [2011],Basu et al. [2015],Davis et al. [2016],Medeiros and Mendes [2016], Mark B. and R. [2018]). In particular if there are M dependent time series (e.g. voxels in the brain, actors in a social network, measurements at different spatial locations), *time series network* models allow us to model temporal dependence between actors/nodes in a network.

More precisely, consider the following time series auto-regressive network model with lag p,

$$X_{t+1} = \sum_{j=1}^{p} A^*(j)X_{t+1-j} + \epsilon_t, \tag{1}$$

where $\{X_t\}_{t=0}^{T} \in \mathbb{R}^M$ is the time series data we have access to, $\{A^*(j) \in \mathbb{R}^{M \times M}, j = 1, \ldots, p\}$ are the network parameters of interest and $\epsilon_t \in \mathbb{R}^M$ is zero-mean noise. We are considering the high-dimensional setting where the number of nodes M in the network is much larger than the sample size T. Prior work in Basu et al. [2015] has addressed the question of how to estimate the network parameter A^* with Gaussian noise ϵ_t under sparsity assumptions and various structural constraints. In this paper, we focus on *inference and hypothesis testing* for the parameter A^* given the data $(X_t)_{t=0}^{T}$.

In high-dimensional statistics, there has recently been a growing body of work on confidence intervals and hypothesis testing under structural assumptions such as sparsity. Since the widely used Lasso estimator for sparse linear regression is asymptotically biased, one-step estimators based on bias-correction have been studied in works such as Zhang and Zhang [2014], Van de Geer et al. [2014] and Javanmard and Montanari [2014] which are referred to as LDPE, de-sparsifying and de-biasing estimator respectively. Low-dimensional components of these estimators have asymptotic normality and thus can be used for constructing hypothesis testing and confidence intervals.

In this paper, we adopt the framework of Ning and Liu (Ning et al. [2017]) who propose a high dimensional test statistic based on score function, called the decorrelated score function which we briefly describe here. Formally, consider a statistical model $\mathcal{P} = \{\mathbb{P}_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \Omega\}$ with high-dimensional parameter vector $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma}^\top)^\top \in \mathbb{R}^d$. Suppose we are interested in the scalar parameter θ and $\boldsymbol{\gamma} \in \mathbb{R}^{d-1}$ is the nuisance parameter. Suppose data $\{\boldsymbol{U}_i, i = 1, \ldots, n\}$ are i.i.d. data following distribution $\mathbb{P}_{\boldsymbol{\beta}}$, then the negative log-likelihood function is defined as

$$\ell(\theta, \boldsymbol{\gamma}) = -\frac{1}{n} \sum_{i=1}^{n} \log f(\boldsymbol{U}_i; \theta, \boldsymbol{\gamma}).$$

It is known that the score function $\sqrt{n}\nabla_\theta \ell(0, \boldsymbol{\gamma}^*)$ is asymptotically normal if the true parameter $\boldsymbol{\beta}^* = (0, \boldsymbol{\gamma}^*)$. If $\boldsymbol{\gamma}^*$ is substituted by some estimator $\hat{\boldsymbol{\gamma}}$, the estimation induced error can be approximated as the following:

$$\sqrt{n}\nabla_\theta \ell(0, \hat{\boldsymbol{\gamma}}) - \sqrt{n}\nabla_\theta \ell(0, \boldsymbol{\gamma}^*) \approx \sqrt{n}\nabla^2_{\theta\gamma} \ell(0, \boldsymbol{\gamma}^*)(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*),$$

when $\hat{\gamma} - \gamma^*$ is small enough. Although $\hat{\gamma} - \gamma^*$ converge to 0 with properly chosen $\hat{\gamma}$, e.g. Lasso estimator, $\sqrt{n}\nabla^2_{\theta\gamma}\ell(0, \gamma^*)(\hat{\gamma} - \gamma^*)$ would not vanish if $\mathbb{E}_{\boldsymbol{\beta}}\left(\nabla^2_{\theta\gamma}\ell(0, \gamma^*)\right) \neq 0$. This fact motivates the decorrelated score function:

$$S(\theta, \boldsymbol{\gamma}) = \nabla_\theta \ell(\theta, \boldsymbol{\gamma}) - \mathbf{I}_{\theta\gamma}\mathbf{I}^{-1}_{\gamma\gamma}\nabla_\gamma \ell(\theta, \boldsymbol{\gamma}),$$

with Fisher information matrix $\mathbf{I} = \mathbb{E}_{\boldsymbol{\beta}}\left(\nabla^2 \ell(\boldsymbol{\beta})\right)$. One can check that

$$\mathbb{E}\left(\nabla_\gamma S(\theta, \boldsymbol{\gamma})\right) = 0.$$

Both $\boldsymbol{\gamma}$ and $\mathbf{I}_{\theta\gamma}\mathbf{I}^{-1}_{\gamma\gamma}$ are substituted by some estimator, and it is shown in Ning et al. [2017] that the decorrelated score function is asymptotically normal.

In the linear regression case, the test statistic generated by the decorrelated score function in Ning et al. [2017] is equivalent to that constructed by de-biased estimator in Van de Geer et al. [2014]. However, Ning et al. [2017] allow a more general form, and thus is easier to adapt to the time series case. In fact Neykov et al.Neykov et al. [2018] consider amongst other examples, high-dimensional time series with Gaussian error innovations. While Gaussian error innovations are widely used, many time series models include data that has bounded range or discrete data, for which the Gaussian distribution is not a natural fit. In this paper, we address the more general and technically challenging setting in which the noise ϵ_t is sub-Gaussian.

One of the important technical challenges in going from the Gaussian to the sub-Gaussian case is that dependent Gaussian vectors can be rotated to be independent, while such a result does not hold for sub-Gaussian vectors. Prior work in Wong et al. [2016] addresses this challenges by imposing stationarity and β-mixing conditions. In order to avoid these conditions, we develop novel concentration bounds for sub-Gaussian random vectors.

In this paper, we investigate the hypothesis testing and confidence region with respect to a low-dimensional component of parameter matrices $\{A^*(j), j = 1, \ldots, p\}$ for sub-Gaussian data, using the testing framework in Ning et al. [2017]. Our major contributions are as follows:

- Extending theoretical results in Ning et al. [2017] for high-dimensional hypothesis testing from Gaussian to sub-Gaussian temporal dependent data (VAR model), both under null and alternative hypothesis. We also show that our techniques lead to similar results to Neykov et al.Neykov et al. [2018] in the Gaussian case but under less restrictive conditions;

- A novel concentration bound for quadratic forms of sub-Gaussian time series data. Note that unlike Gaussian vectors which can be rotated to be independent, sub-Gaussian vectors can not which present additional technical challenges. Our analysis also leads to estimators for covariance and regression parameters for time series data under sub-Gaussian assumptions which are of independent interest.

- We also construct semi-parametric efficient confidence region for multivariate parameters with fixed dimension;

- Finally we support our theoretical guarantees with a simulation study on bounded noise, which is sub-Gaussian but not Gaussian.

1.1 Related Work

In the literature on inference for high-dimensional VAR models, most work focuses on the estimation problem. Song and Bickel (Song and Bickel [2011]) investigate penalized least squares algorithms for different penalties, with some externally imposed assumptions on the temporal dependence. Theoretical guarantees on Dantzig type and Lasso type estimators are studied in Han et al. [2015] and Basu et al. [2015], but with Gaussian noise. Barigozzi and Brownlees (Barigozzi and Brownlees [2018]) consider the inference for stationary dependence structure built among variables, other than the parameters in the VAR model. In our work, we control the error bounds of Lasso and Dantzig type estimators for parameter matrices, with sub-Gaussian noise. Then we establish asymptotic distribution of test statistic based on this.

In the high-dimensional hypothesis testing literature, there is some work regarding to testing for high-dimensional mean vector (Srivastava [2009]), covariance matrices (Chen et al. [2010],Zhang et al. [2013]) and independence among variables (Schott [2005]). While for testing on regression parameters, most work assumes i.i.d samples. Lockhart et al. [2014], Taylor et al. [2014] and Lee et al. [2016] proposes methods to test whether a covariate should be selected conditioning on the selection of some other covariates. A penalized score test depending on the tuning parameter λ is considered in Voorman et al. [2014]. Our work follows the a line of work by Zhang and Zhang [2014], Van de Geer et al. [2014], Javanmard and Montanari [2014] and Ning et al. [2017], the de-sparsifying or decorrelated literature. We construct a VAR version of decorrelated score test proposed by Ning et al. [2017]. Chen and Wu (Chen and Wu [2018]) tackles the hypothesis testing problem for time series data as well, but they are testing the trend in a time series, instead of the autoregressive parameter which encodes the influence structure among variables.

As mentioned earlier, our work is most closely related to the prior work of Neykov et al.Neykov et al. [2018], which provides a hypothesis testing framework with high-dimensional Gaussian time series as a special case. In our work, we consider the more general and technically challenging case of sub-Gaussian vector auto-regressive models. Throughout this paper, we provide a comparison to results derived in this work for the Gaussian case.

1.2 Organization of the Paper

Section 2 explains the problem set up and proposes our test statistic. Theoretical guarantee is shown in section 3. Specifically, section 3.1 and 3.2 present the weak convergence rate of test statistic under the null and alternative hypothesis $\mathcal{H}_0$ and $\mathcal{H}_A$. Section 3.3 propose some feasible estimators, which satisfy the assumptions required and can be plugged into the test statistic. Section 3.4 considers the case when the variance of noise are unknown, and we construct a confidence region for multivariate parameter vectors in Section 3.5. We consider the special case of the AR(1) model with Gaussian noise, a detailed comparison with Neykov et al. [2018] is provided in section 3.6. Section 4 provides simulation results and section 5 includes the proofs for the two main theorems. Much of the proof is deferred to Appendices.

1.3 Notation

We define the following norms for vectors and matrices: For a vector $u = (u_1, \ldots, u_d)^\top \in \mathbb{R}^d$, we define the p-norm where $p \geq 1, \|u\|_p = \left(\sum_{i=1}^d u_i^p\right)^{\frac{1}{p}}$. For a matrix $U \in \mathbb{R}^{m \times n}$, the ℓ_p norm and Frobenius norm of U is defined as $\|U\|_p = \sup_v \frac{\|Uv\|_p}{\|v\|_p}$, $\|U\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n U_{ij}^2\right)^{\frac{1}{2}}$. We also use notation $\|U\|_{1,1}$ to denote the ℓ_1 penalty on U, which is $\sum_{i=1}^m \sum_{j=1}^n |U_{i,j}|$. Furthermore, if U is symmetric the trace norm of U is $\|U\|_{\mathrm{tr}} = \mathrm{tr}(\sqrt{U^2})$.

Throughout the paper, we assume that the entries of noise vectors $\{\epsilon_{ti}, 1 \leq i \leq M\}_{t=-\infty}^\infty$ are independent sub-Gaussian variables with constant scale factor. A univariate centered random variable X has a sub-Gaussian distribution with scale factor τ if

$$M_X(t) \triangleq \mathbb{E}\left[\exp(tX)\right] \leq \exp(\tau^2 t^2 / 2).$$

2 Problem Setup

We consider a general vector auto-regressive time series with lag p, where p is known and finite and independent of T or other dimensions:

$$X_{t+1} = \sum_{j=1}^p A(j) X_{t-j+1} + \epsilon_t, \tag{2}$$

where $X_t \in \mathbb{R}^M$, $\epsilon_t \in \mathbb{R}^M$ is zero-mean entry-wise independent sub-Gaussian noise with identity covariance matrix, and $A(j) \in \mathbb{R}^{M \times M}, j = 1, \cdots, p$ are parameters of interest. Define the matrix $A^* = (A(1), \cdots, A(p)) \in \mathbb{R}^{M \times pM}$ and $\mathcal{X}_t = (X_t^\top, \cdots, X_{t-p+1}^\top)^\top \in \mathbb{R}^{pM}$, then we can also write (2) as

$$X_{t+1} = A^* \mathcal{X}_t + \epsilon_t. \tag{3}$$

For notational convenience, we assume that time series data X_t has time range $1 - p \leq t \leq T$.

Based on data $(X_t)_{t=1-p}^T$, we test the hypothesis of whether a subset of entries in A^* are 0. Let A_i^* be the ith row vector of A^*. Without loss of generality, suppose the entries we test are in rows $1, \cdots, k$. Define $D_m \subset \{1, \cdots, pM\}$ as the columns we test in mth row with $d_m = |D_m|$, and $D = \{(i,j) : 1 \leq i \leq k, j \in D_i\}$, with $d = |D| = \sum_{m=1}^k d_m$. We test the null hypothesis:

$$\mathcal{H}_0 : \widetilde{A}_D = 0 \tag{4}$$

where $\widetilde{A}_D = ((A_1^*)_{D_1}^\top, \cdots, (A_k^*)_{D_k}^\top)^\top \in \mathbb{R}^d$. We also assume that d is finite and not increasing with T. In the work of of Neykov et al.Neykov et al. [2018], d is assumed to be 1.

2.1 Stationary distribution

Since we are developing a hypothesis testing framework based on the decorrelated score test, it is important to specify a stationary distribution for X_t Using standard notation from autoregressive time series models, define the polynomial $\mathcal{A}(z) = I_M - \sum_{j=1}^p A(j) z^j$, where I_M is an $M \times M$ identity matrix, and z is a complex number. To guarantee the existence of a stationary solution to (3), we assume

$$\det(\mathcal{A}(z)) \neq 0, \quad |z| \leq 1.$$

Then we can write

$$(\mathcal{A}(z))^{-1} = \sum_{j=0}^\infty \Psi_j z^j,$$

where $\Psi_j, j \geq 0$ are all real valued matrices which are polynomial functions of $A(i), 1 \leq i \leq p$. Note that in the special case where $p = 1$, $\Psi_j = (A^*)^j$.

It can be shown that the unique stationary solution to (2) is

$$X_t = \sum_{j=0}^\infty \Psi_j \epsilon_{t-j-1},$$

and the covariance matrix Σ of X_t satisfies

$$\Sigma = \mathrm{Cov}(X_t) = \sum_{j=0}^\infty \Psi_j \Psi_j^\top. \tag{5}$$

2.2 Decorrelated Score Function

Using the frameworks developed in Ning et al. [2017] for independent design, we consider the decorrelated score test. First we define the *score function* $S(A^*) \in \mathbb{R}^{M \times M}$, with each entry defined as follows:

$$[S(A^*)]_{jk} = -\frac{1}{T} \sum_{t=0}^{T-1} (X_{t+1,j} - a_j^{*\top} \mathcal{X}_t) \mathcal{X}_{tk} = -\frac{1}{T} \sum_{t=0}^{T-1} \epsilon_{t,j} \mathcal{X}_{tk}.$$

As pointed out in Ning et al. [2017], the standard score function is infeasible and we need to consider the *decorrelated score function*

$$S = (S_1^\top, S_2^\top, \cdots, S_k^\top)^\top \in \mathbb{R}^d,$$

with each $S_m \in \mathbb{R}^{d_m}$ corresponding to the tested row (m, D_m):

$$S_m = -\frac{1}{T} \sum_{t=0}^{T-1} \epsilon_{t,m} (\mathcal{X}_{t,D_m} - w_m^{*\top} \mathcal{X}_{t,D_m^c}),$$

where $\mathcal{X}_{t,D_m} \in \mathbb{R}^{d_m}$ is composed of the entries of $\mathcal{X}_t$ whose indices are within set D_m. $\mathcal{X}_{t,D_m^c} \in \mathbb{R}^{pM-d_m}$ is also defined similarly and $w_m^* \in \mathbb{R}^{(pM-d_m) \times d_m}$ is chosen to satisfy

$$\mathrm{Cov}(X_{t,D_m} - w_m^{*\top} \mathcal{X}_{t,D_m^c}, \mathcal{X}_{t,D_m^c}) = 0. \tag{6}$$

Specifically, w_m^* is defined as a function of $\Upsilon = \mathrm{Cov}(\mathcal{X}_t) \in \mathbb{R}^{pM \times pM}$:

$$w_m^* = (\Upsilon_{D_m^c, D_m^c})^{-1} \Upsilon_{D_m^c, D_m}. \tag{7}$$

2.3 Test Statistic

Based on the decorrelated score function S_m, we first define the statistic $V_{T,m} \in \mathbb{R}^{d_m}$:

$$V_{T,m} \triangleq \sqrt{T} (\Upsilon^{(m)})^{-\frac{1}{2}} S_m,$$

with $\Upsilon^{(m)} \in \mathbb{R}^{d_m \times d_m}$ being defined as:

$$\begin{aligned}
\Upsilon^{(m)} &\triangleq \mathrm{Cov}(\mathcal{X}_{t,D_m} - w_m^{*\top} \mathcal{X}_{t,D_m^c}) \\
&= \mathrm{Cov}(\mathcal{X}_{t,D_m} | \mathcal{X}_{t,D_m^c}) \\
&= \Upsilon_{D_m,D_m} - \Upsilon_{D_m,D_m^c} (\Upsilon_{D_m^c,D_m^c})^{-1} \Upsilon_{D_m^c,D_m}.
\end{aligned} \tag{8}$$

Let V_T be the d-dimensional vector concatenated by $V_{T,m}$'s:

$$V_T = (V_{T,1}^\top, \cdots, V_{T,k}^\top)^\top.$$

One of the main results of the paper is to show that V_T is asymptotically Gaussian. Define $U_T = \|V_T\|_2^2$, then U_T is asymptotically χ_d^2. Since we do not know ϵ_t, w_m^*, and $\Upsilon^{(m)}$, we later define estimators for these quantities. Formally, we define our test statistic $\widehat{U}_T$ as

$$\widehat{U}_T = T \sum_{m=1}^{k} \widehat{S}_m^\top \left(\widehat{\Upsilon^{(m)}}\right)^{-1} \widehat{S}_m, \tag{9}$$

where $\widehat{\Upsilon^{(m)}} \in \mathbb{R}^{d_m \times d_m}$ is an estimator for $\Upsilon^{(m)}$ and $\widehat{S}_m \in \mathbb{R}^{d_m}$ is defined as

$$\widehat{S}_m = -\frac{1}{T} \sum_{t=0}^{T-1} \left(X_{t+1,m} - (\widehat{A}_m)_{D_m^c}^\top \mathcal{X}_{t,D_m^c}\right) (\mathcal{X}_{t,D_m} - \hat{w}_m^\top \mathcal{X}_{t,D_m^c}),$$

with $\widehat{A}_m \in \mathbb{R}^{pM}$ and $\hat{w}_m \in \mathbb{R}^{(pM-d_m)\times d_m}$ estimating A_m^* and w_m^*. Here we are not worried about the invertible issue of $\widehat{\Upsilon^{(m)}}$, since $\Upsilon^{(m)}$ is a low dimensional covariance matrix. To guarantee a good estimation of the high-dimensional parameter A_m^* and w_m^*, we impose sparsity conditions upon them. Specifically, for each $1 \leq m \leq M$, $1 \leq i \leq k$ define

$$\rho_m \triangleq \|A_m^*\|_0, \quad s_i \triangleq \|w_i^*\|_0, \tag{10}$$

and note that they both depend on A^*.

The sparsity of w_m^* can be implied by the sparsity of Υ^{-1}, which is a common condition in high-dimensional hypothesis testing literature (e.g. see Van de Geer et al. [2014]). Specifically, the following Lemma shows that when lag $p = 1$ and A^* is symmetric, the sparsity of w_m^* is implied by the sparsity of A^*:

Lemma 2.1. *If $p = 1$, $A^* \in \mathbb{R}^{M \times M}$ is symmetric, then s_m defined in (10) satisfies*

$$s_m \leq d_m^2 \max_{1 \leq i \leq M} \rho_i, \quad \text{for } 1 \leq m \leq k.$$

The proof for Lemma 2.1 is included in Appendix E.

3   Theoretical guarantee

In this section, we present uniform convergence results for test statistic $\widehat{U}_T$ under $\mathcal{H}_0$ and $\mathcal{H}_A$, with A^* and estimators satisfying conditions. We also provide feasible estimators, and prove that they satisfy corresponding conditions in Section 3.3. Unknown variance and confidence region construction is discussed in Section 3.4 and 3.5. In Section 3.6 we provide consequences of our theory under AR(1) model with Gaussian noise and compare our results with Neykov et al.Neykov et al. [2018].

Recall that the null hypothesis is

$$\mathcal{H}_0 : \widetilde{A}_D = 0, \tag{11}$$

with $\widetilde{A}_D \in \mathbb{R}^d$ being concatenated by $(A_1^*)_{D_1}, \ldots, (A_k^*)_{D_k}$. While for the alternative hypothesis, like in Ning et al. [2017], we consider

$$\mathcal{H}_A : \widetilde{A}_D = T^{-\phi}\Delta, \tag{12}$$

with some constant $\phi > 0$ and constant vector $\Delta \in \mathbb{R}^d$. Write

$$\Delta = (\Delta_1^\top, \cdots . \Delta_k^\top)^\top,$$

where each $\Delta_m \in \mathbb{R}^{d_m}$. The reason why $T^{-\phi}\Delta$ instead of Δ is considered in (12) is that we expect the test to be more sensitive as sample size increases. We will see how the value of ϕ influences the convergence of $\widehat{U}_T$ in Theorem 3.2.

We still assume ϵ_{ti}'s are i.i.d. sub-Gaussian random variables, and also consider a special case, where $\epsilon_t \sim \mathcal{N}(0, I)$. We compare our result in the Gaussian case to results in Neykov et al.Neykov et al. [2018].

First we define the sets Ω_0 and Ω_1 of feasible parameter matrices A^* under $\mathcal{H}_0$ and $\mathcal{H}_A$ respectively. To control the stability of $\{X_t\}$ in model (3), we impose the condition:

$$\sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} \|\Psi_{i+j}\|_2^2 \right)^{\frac{1}{2}} \leq \beta, \tag{13}$$

for some constant $\beta > 0$. In the case $p = 1$, condition (13) reduces to

$$\sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} \left\|(A^*)^{i+j}\right\|_2^2 \right)^{\frac{1}{2}} \leq \beta, \tag{14}$$

which is implied by $\|A^*\|_2 \leq 1 - \epsilon$ for some $0 < \epsilon < 1$, a typical condition assumed (see e.g. Neykov et al. [2018]). Then define sets Ω_0 and Ω_1 for any $\beta, \rho, s, M, T, \phi > 0$, set D of size d and vector $\Delta = (\Delta_1^\top, \cdots, \Delta_k^\top)^\top \in \mathbb{R}^d$:

$$\Omega_0 = \{A^* \in \mathbb{R}^{M \times pM} : \widetilde{A}_D = 0, \sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} \|\Psi_{i+j}\|_2^2 \right)^{\frac{1}{2}} \leq \beta, \tag{15}$$

$$\max_m \rho_m(A^*) \leq \rho, \max_m s_m(A^*) \leq s\},$$

$$\Omega_1 = \{A^* \in \mathbb{R}^{M \times pM} : \widetilde{A}_D = T^{-\phi}\Delta, \sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} \|\Psi_{i+j}\|_2^2 \right)^{\frac{1}{2}} \leq \beta, \tag{16}$$

$$\max_m \rho_m(A^*) \leq \rho, \max_m s_m(A^*) \leq s\}.$$

Note here $\rho_m(A^*)$ and $s_m(A^*)$ are still functions of A^*, since Υ is determined by A^*. Clearly we need reliable estimators for $\widehat{A}_m$, $\hat{w}_m$ and $\widehat{\Sigma^{(m)}}$ with $1 \leq m \leq k$, to guarantee the weak convergence of $\widehat{U}_T$. We present the following assumptions for these estimators, which we will verify in section 3.3. Note that constants C may depend on p, d, β and τ, but do not depend on either M or T.

Assumption 3.1 (Estimation Error for A_m^*). *For each $A^* \in \Omega_0 \cup \Omega_1$,*

$$\left\|\widehat{A}_m - A_m^*\right\|_1 \leq C\rho_m \sqrt{\frac{\log M}{T}}, \quad \left\|\widehat{A}_m - A_m^*\right\|_2 \leq C\sqrt{\frac{\rho_m \log M}{T}},$$

$$(\widehat{A}_m - A_m^*)^\top \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_t \mathcal{X}_t^\top \right) (\widehat{A}_m - A_m^*) \leq C\frac{\rho_m \log M}{T}, \tag{17}$$

9

hold for $1 \leq m \leq k$, with probability at least $1 - c_1 \exp\{-c_2 \log M\}$.

These are standard error bounds for Lasso estimator and Dantzig Selector with independent design. In this paper we verify Assumption 3.1 in section 3.3 and the remaining two assumptions when we have dependent sub-Gaussian random variables, as we do for our vector auto-regressive model setting.

Assumption 3.2 (Estimation Error for w_m^*). *For each $A^* \in \Omega_0 \cup \Omega_1$:*

$$\|\hat{w}_m - w_m^*\|_1 \leq C s_m \sqrt{\frac{\log M}{T}},$$

$$tr\left[(\hat{w}_m - w_m^*)^\top \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_{t,D_m^c} \mathcal{X}_{t,D_m^c}^\top\right)(\hat{w}_m - w_m^*)\right] \leq C \frac{s_m \log M}{T}, \tag{18}$$

hold for $1 \leq m \leq k$, with probability at least $1 - c_1 \exp\{-c_2 \log M\}$.

Similar to Assumption 3.1, we will show that both Lasso estimator and Dantzig selector under model (3) satisfy Assumption 3.2.

Assumption 3.3 (Estimation Error for $\Upsilon^{(m)}$). *For each $A^* \in \Omega_0 \cup \Omega_1$,*

$$\left\|\Upsilon^{(m)\frac{1}{2}} \left(\widehat{\Upsilon^{(m)}}\right)^{-1} \Upsilon^{(m)\frac{1}{2}} - I\right\|_\infty \leq C \frac{(s \vee \rho) \log M}{\sqrt{T}}, \tag{19}$$

hold for $1 \leq m \leq k$, with probability at least $1 - c_1 \exp\{-c_2 \log M\}$.

Note that $\Upsilon^{(m)} \in \mathbb{R}^{d_m \times d_m}$ is a low-dimensional matrix, and thus it is computationally feasible to use the sample covariance matrix of $X_{t,D_m} - \hat{w}_m^\top X_{t,D_m^c}$ as an estimator for $\widehat{\Upsilon^{(m)}}$. We show in section 3.3 that, as long as $\hat{w}_m$ is a reliable estimator for w_m^*, $\widehat{\Upsilon^{(m)}}$ would satisfy a tighter bound than (19). This looser bound in Assumption 3.3 actually allows more choices for estimators for $(\Upsilon^{(m)})^{-1}$, as shown in section 3.5.

3.1 Uniform convergence under null hypothesis

Based on these assumptions, we have the following main theorem.

Theorem 3.1. *Consider the model (3) with i.i.d. sub-Gaussian noise ϵ_{ti} with sub-Gaussian parameter τ. If Assumptions 3.1-3.3 are satisfied, and $(\rho \vee s) \log M = o(\sqrt{T})$, then $\widehat{U}_T$ defined in (9) satisfies*

$$\sup_{x \in \mathbb{R}, A^* \in \Omega_0} \left|\mathbb{P}(\widehat{U}_T \leq x) - F_d(x)\right|$$

$$\leq \frac{C_1}{T^{\frac{1}{8}}} + C_2 \left(\frac{(s \vee \rho) \log M}{\sqrt{T}}\right)^{\frac{1}{2}} + \frac{C_3}{M^{C_4}}, \tag{20}$$

when $T > C$ for some constant C. Here the constants C_i's depend on p, d, β, τ.

Theorem 3.1 proves weak convergence of $\widehat{U}_T$ to χ_d^2. The uniform convergence rate can be understood as follows: the first term is due to the rate obtained by martingale CLT, where we require $T^{-\frac{1}{8}}$ rather than $T^{-\frac{1}{2}}$ due to the dependence; the remaining two terms arise from estimation error, with the second one being the error bounds, and third being the probability that the error bounds do not hold. If we assume Gaussianity, we can improve the first term in the rate of convergence from $T^{-\frac{1}{8}}$ to $T^{-\frac{1}{4}+\alpha}$ for any $\alpha > 0$. To the best of our knowledge, ours is the first work that formally attempts to characterize the rates of convergence.

Remark 3.1. *Compared to the theoretical result for independent design in Ning et al. [2017], the only additional condition we add is $\sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} \|\Psi_{i+j}\|_2^2 \right)^{\frac{1}{2}} \leq \beta$, which is used to control the strength of dependence uniformly. Also, we consider multivariate testing which is more general, and derive the explicit convergence rate.*

Remark 3.2. *The test statistic proposed in Van de Geer et al. [2014] and Javanmard and Montanari [2014] for the independent design share similar ideas with our test statistic. Instead of imposing a sparsity assumption upon w_m^*, Van de Geer et al. [2014] assumes Υ^{-1} to be row wise sparse. This is actually equivalent to the sparsity assumption on w_m^* in the univariate case. Javanmard and Montanari [2014] does not require the sparsity condition on Υ^{-1}, but it is hard to extend their theory to the time series setting, due to a difficulty in applying the martingale CLT.*

Remark 3.3. *The theoretical guarantee we obtained here, is more general and stronger than the result achieved in Neykov et al. [2018]. A more detailed comparison is presented in section 3.6.*

3.2 Uniform convergence under alternative hypothesis

Recall the definition of Ω_A in (16). The following theorem establishes the asymptotic behavior of $\widehat{U}_T$ for $A^* \in \Omega_A$, with different values of ϕ. First define

$$\widetilde{\Delta} = (\widetilde{\Delta}_1^\top, \cdots, \widetilde{\Delta}_k^\top)^\top, \quad \widetilde{\Delta}_m = (\Upsilon^{(m)})^{\frac{1}{2}} \Delta_m, \tag{21}$$

where $\Upsilon^{(m)}$ is defined in (8).

Theorem 3.2. *Consider the model (3) with i.i.d. sub-Gaussian noise ϵ_{ti} and sub-Gaussian parameter τ. If Assumptions 3.1-3.3 are satisfied, and $(\rho \vee s) \log M = o(\sqrt{T})$, then when $T > C$ for some constant C,*

(1) $\phi = \frac{1}{2}$

$$\sup_{x \in \mathbb{R}, A^* \in \Omega_1} \left| \mathbb{P}(\widehat{U}_T \leq x) - F_{d, \|\widetilde{\Delta}\|_2^2}(x) \right|$$

$$\leq \frac{C_1}{T^{\frac{1}{8}}} + C_2 \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{\frac{1}{2}} + \frac{C_3}{M^{C_4}}. \tag{22}$$

(2) $0 < \phi < \frac{1}{2}$

$$\sup_{A^* \in \Omega_1} |\mathbb{P}(\widehat{U}_T \leq x)|$$
$$\leq \frac{C_1}{T^{\frac{1}{8}}} + \frac{C_2}{M^{C_3}} + C_4 \exp\{-C_5 T^{\frac{1}{2}-\phi} + C_6 \sqrt{x}\}. \tag{23}$$

(3) $\phi > \frac{1}{2}$

$$\sup_{x \in \mathbb{R}, A^* \in \Omega_1} \left| \mathbb{P}(\widehat{U}_T \leq x) - F_d(x) \right|$$
$$\leq \frac{C_1}{T^{\frac{1}{8}}} + C_2 \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{\frac{1}{2}} + \frac{C_3}{M^{C_4}} + C_3 T^{\frac{1-2\phi}{3}}. \tag{24}$$

Here C_i's are constants depending on $p, d, \beta, \Delta, \tau$.

Theorem 3.2 shows the threshold value of ϕ for $\mathcal{H}_A$ to be detectable. When $\phi > \frac{1}{2}$, we cannot distinguish $\mathcal{H}_0$ and $\mathcal{H}_A$ since under both cases $\widehat{U}_T$ converges to χ_d^2; When $\phi < \frac{1}{2}$, $\widehat{U}_T$ diverges to $+\infty$ in probability, thus it would be very easy to detect $\mathcal{H}_A$; When $\phi = \frac{1}{2}$, $\widehat{U}_T$ converges to a non-central χ_d^2 with noncentrality parameter determined by constant vector Δ and $\Upsilon = \text{Cov}(\mathcal{X}_t)$, which implies the power of the test. Note here, (23) holds also for the trivial case $\phi < 0$, since we do not use the fact $\phi > 0$ in the proof.

Remark 3.4. *Theorem 3.2 is also consistent with the threshold value of ϕ given by Ning et al. [2017] for linear regression with i.i.d samples. However, Ning et al. [2017] assumes additional conditions on the scaling of sample size, number of covariates and sparsity of w_m^* for proving asymptotic power. Our conditions are exactly the same as the ones for $\mathcal{H}_0$, due to a more specific model and careful analysis.*

3.3 Feasible Estimators

Both the estimation of w_m^* and A^* can be viewed as high-dimensional sparse regression problems, thus we can use the Lasso or Dantzig selector. Formally, define

$$\widehat{A}^{(L)} = \arg\min_{A \in \mathbb{R}^{M \times pM}} \frac{1}{T} \sum_{t=0}^{T-1} \|X_{t+1} - A\mathcal{X}_t\|_2^2 + \lambda_A \|A\|_{1,1}, \tag{25}$$

as the Lasso estimator for A^*, and

$$\widehat{A}^{(D)} = \arg\min_{A \in \mathbb{R}^{M \times pM}} \|A\|_{1,1}, \quad \text{s.t.} \quad \left\| \frac{1}{T} \sum_{t=0}^{T-1} (X_{t+1} - A\mathcal{X}_t)\mathcal{X}_t^\top \right\|_\infty \leq \lambda_A, \tag{26}$$

as the Dantzig selector estimator for A^*. Similarly, for $1 \leq m \leq k$, define

$$\hat{w}_m^{(L)} = \arg\min_{w \in \mathbb{R}^{(pM-d_m) \times d_m}} \frac{1}{T} \sum_{t=0}^{T-1} \|\mathcal{X}_{t,D_m} - w^\top \mathcal{X}_{t,D_m^c}\|_2^2 + \lambda_w \|w\|_{1,1}, \tag{27}$$

and

$$\hat{w}_m^{(D)} = \arg\min_{w \in \mathbb{R}^{(pM-d_m) \times d_m}} \|w\|_{1,1}, \quad \text{s.t.} \quad \left\| \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{X}_{t,D_m} - w^\top \mathcal{X}_{t,D_m^c}) \mathcal{X}_{t,D_m^c}^\top \right\|_\infty \le \lambda_w. \quad (28)$$

While for estimating $\Upsilon^{(m)}$, since this is a low dimensional covariance matrix for $\mathcal{X}_{t,D_m} - w_m^{*\top} \mathcal{X}_{t,D_m^c}$, we can directly use sample covariance of $\mathcal{X}_{t,D_m} - \hat{w}_m^\top \mathcal{X}_{t,D_m^c}$ as $\widehat{\Upsilon^{(m)}}$:

$$\widehat{\Upsilon^{(m)}} = \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{X}_{t,D_m} - \hat{w}_m^\top \mathcal{X}_{t,D_m^c})(\mathcal{X}_{t,D_m} - \hat{w}_m^\top \mathcal{X}_{t,D_m^c})^\top, \quad (29)$$

for $1 \le m \le k$. Here $\hat{w}_m$ in the definition of (29) is either $\hat{w}_m^{(L)}$ or $\hat{w}_m^{(D)}$.

As shown in the following, estimators (25) to (29) all satisfy Assumptions 3.1 to 3.3, under the model setting stated in (3):

Lemma 3.1. *If $\widehat{A} = \widehat{A}^{(L)}$, or $\widehat{A} = \widehat{A}^{(D)}$, which are defined as in (25) and (26) with $\lambda_A \asymp \sqrt{\frac{\log M}{T}}$, then $\widehat{A}$ satisfies Assumption 3.1 when $T > C\rho \log M$.*

Lemma 3.2. *If $\hat{w}_m = \hat{w}_m^{(L)}$ or $\hat{w}_m = \hat{w}_m^{(D)}$, which are defined as in (27) and (28) with $\lambda_w \asymp \sqrt{\frac{\log M}{T}}$, then $\hat{w}_m$'s satisfy Assumption 3.2 when $T > Cs \log M$.*

Lemma 3.3. *If $\widehat{\Upsilon^{(m)}}$'s are defined as in (29), where $\hat{w}_m$ satisfies (18) with probability at least $1 - c_1 \exp\{-c_2 \log M\}$, then*

$$\left\| \Upsilon^{(m)\frac{1}{2}} \left(\widehat{\Upsilon^{(m)}} \right)^{-1} \Upsilon^{(m)\frac{1}{2}} - I \right\|_\infty \le C \sqrt{\frac{\log M}{T}},$$

with probability at least $1 - c_1 \exp\{-c_2 \log M\}$, when $T > Cs^2 \log M$.

Note here Lemma 3.3 is stronger than Assumption 3.3. The proof of these Lemmas are deferred to Appendix A. By these lemmas and Theorem 3.1, 3.2, we arrive at following Corollary.

Corollary 3.1. *Under model (3) with i.i.d sub-Gaussian noise ϵ_{ti} with parameter τ, if $\widehat{A} = \widehat{A}^{(L)}$ or $\widehat{A}^{(D)}$, $\hat{w}_m = \hat{w}_m^{(L)}$ or $\hat{w}_m^{(D)}$, and $\widehat{\Upsilon^{(m)}}$'s are defined as in (29) for $1 \le m \le k$ with $\lambda_A \asymp \lambda_w \asymp \sqrt{\frac{\log M}{T}}$, then if $(\rho \vee s) \log M = o(\sqrt{T})$ and $T > C$ for some constant $C > 0$, bounds (20) to (24) from Theorems 3.1 and 3.2 hold.*

3.4 Variance Estimation

In this section, we consider the case where $\sigma^{*2} = \text{Var}(\epsilon_{ti})$ is unknown under model (3). Actually, if $\sigma^* \ne 1$ is known, it is straightforward to extend Theorem 3.1 to Theorem 3.2 for $\widehat{U}_T$ defined as follows:

$$\widehat{U}_T = T \sum_{m=1}^k \widehat{S}_m^\top (\widehat{\Upsilon^{(m)}})^{-1} \widehat{S}_m / \sigma^{*2}. \quad (30)$$

This follows since if we consider $Y_t = X_t/\sigma^*$, time series data Y_t would satisfy the same model but with unit variance noise.

When σ^{*2} is unknown, we apply the estimator

$$\hat{\sigma}^2 = \frac{1}{MT} \sum_{t=0}^{T-1} \|X_{t+1} - \widehat{A}\mathcal{X}_t\|_2^2, \tag{31}$$

and define the test statistic

$$\widetilde{U}_T = T \sum_{m=1}^{k} \widehat{S}_m^\top (\widehat{\Upsilon^{(m)}})^{-1} \widehat{S}_m / \hat{\sigma}^2. \tag{32}$$

We show that $\widetilde{U}_T$ has the same convergence results we derive for the unit variance noise case.

Theorem 3.3. *Consider the model* (3) *with i.i.d. sub-Gaussian noise ϵ_{ti} of variance $\sigma^{*2} = Var(\epsilon_{ti}) \geq \sigma_0^2 > 0$ and scale factor $\tau\sigma^*$. Then Theorem 3.1 and 3.2 hold for $\widetilde{U}_T$ under each corresponding condition, and constants C_i's also depend on σ_0.*

Theorem 3.3 shows that when we have to estimate the unknown σ^{*2}, test statistic $\widetilde{U}_T$ maintains the same asymptotic behavior as $\widehat{U}_T$ under the known variance case, given that all the assumptions for estimation errors are satisfied and σ^* is lower bounded by some constant.

Remark 3.5. *With sub-Gaussian noise ϵ_{ti}, if we still assume the scale factor $\tau\sigma^*$ of ϵ_{ti} to be bounded by constant, then Lemma 3.1 to 3.3 would still hold. Thus the assumptions imposed on estimation errors of $\widehat{A}$, $\hat{w}_m$ and $\widehat{\Upsilon^{(m)}}$ are all satisfied. However, if we don't assume σ^* to be bounded, then the tuning parameters λ_A and λ_w have to scale with σ^*.*

Remark 3.6. *Neykov et al. [2018] proposes another estimator for the variance of ϵ_{ti}, based on the fact that $\Sigma = A\Sigma A^\top + Cov(\epsilon_t)$. Both these estimators are consistent and lead to convergence in distribution results.*

3.5 Semi-parametric Optimal Confidence Region

In this section, we construct a confidence region for $\widetilde{A}_D$, under model (3) with unknown noise variance σ^{*2}. Similar to Ning et al. [2017], we consider the one-step estimator $\hat{a}(m)$ for each $(A_m^*)_{D_m}$, based on the decorrelated score function:

$$\hat{a}(m) = (\widehat{A}_m)_{D_m} - \left(\widehat{\Upsilon^{(m)}}\right)^{-1} \widetilde{S}_m, \tag{33}$$

where $\widehat{A}_m$ is any estimator satisfying the Assumptions 3.1 on error bounds for $\widehat{A}_m - A_m^*$, and both the Lasso or Dantzig Estimator for A_m^* are suitable. $\widehat{\Upsilon^{(m)}}$ takes the form:

$$\widehat{\Upsilon^{(m)}} = \frac{1}{T} \sum_{t=0}^{T-1} \left(\mathcal{X}_{t,D_m} - \hat{w}_m^\top \mathcal{X}_{t,D_m^c} \right) \mathcal{X}_{t,D_m}^\top, \tag{34}$$

which is another estimator for $\Upsilon^{(m)}$, and

$$\widetilde{S}_m = -\frac{1}{T}\sum_{t=0}^{T-1}\left(X_{t+1,m} - \widehat{A}_m^\top \mathcal{X}_t\right)\left(\mathcal{X}_{t,D_m} - \hat{w}_m^\top \mathcal{X}_{t,D_m^c}\right).$$

We will show that $\hat{a}(m) - (A_m^*)_{D_m}$ is asymptotically Gaussian with covariance matrix $(\Upsilon^{(m)})^{-1}$. Thus we construct the following confidence region for $\widetilde{A}_D$, with asymptotic confidence coefficient $1 - \alpha$:

$$CR(\alpha) = \left\{\theta = (\theta_1^\top, \ldots, \theta_k^\top)^\top : \theta_m \in \mathbb{R}^{d_m},\right.$$
$$\left.\frac{T}{\hat{\sigma}^2}\sum_{m=1}^{k}(\hat{a}(m) - \theta_m)^\top \widehat{\Upsilon^{(m)}}(\hat{a}(m) - \theta_m) \leq \chi_d^2(1 - \alpha)\right\}. \tag{35}$$

This is a d dimensional elliptical ball with center vector $(\hat{a}(1)^\top, \ldots \hat{a}(k)^\top)^\top$. The following theorem shows the weak convergence result of

$$\widehat{R}_T \triangleq \frac{T}{\hat{\sigma}^2}\sum_{m=1}^{k}(\hat{a}(m) - (A_m^*)_{D_m})^\top \widehat{\Upsilon^{(m)}}(\hat{a}(m) - (A_m^*)_{D_m}). \tag{36}$$

Theorem 3.4. *Under model (3) with i.i.d. sub-Gaussian noise ϵ_{ti} with variance $\sigma^{*2} = Var(\epsilon_{ti}) \geq \sigma_0^2 > 0$ and sub-Gaussian parameter $\tau\sigma^*$, then Theorem 3.1 and 3.2 hold for $\widehat{R}_T$ under each corresponding condition, and the constants C_i's also depend on σ_0.*

Remark 3.7. *In the definition of one-step estimator $\hat{a}(m)$, we use $\widetilde{\Upsilon^{(m)}}$ instead of $\widehat{\Upsilon^{(m)}}$ for theoretical convenience. Theorem 3.4 would still hold true if $\hat{a}(m)$ is defined as $(\widehat{A}_m)_{D_m} - \left(\widehat{\Upsilon^{(m)}}\right)^{-1}\widetilde{S}_m$.*

Remark 3.8. *We have exactly the same theoretical result for $\widetilde{U}_T$ and $\widehat{R}_T$, and this is due to the close relationship between these two quantities. In particular,*

$$\widehat{R}_T = T\sum_{m=1}^{k}\widehat{S}_m^\top\left(\widetilde{\Upsilon^{(m)}}^\top\right)^{-1}\widehat{\Upsilon^{(m)}}\left(\widetilde{\Upsilon^{(m)}}\right)^{-1}\widehat{S}_m/\hat{\sigma}^2,$$

compared to $\widetilde{U}_T = T\sum_{m=1}^{k}\widehat{S}_m^\top(\widehat{\Upsilon^{(m)}})^{-1}\widehat{S}_m/\hat{\sigma}^2$. We show in the proof of Theorem 3.4 that $\left(\widetilde{\Upsilon^{(m)}}^\top\right)^{-1}\widehat{\Upsilon^{(m)}}\left(\widetilde{\Upsilon^{(m)}}\right)^{-1}$ also satisfies Assumption 3.3 as an estimator for $(\Upsilon^{(m)})^{-1}$.

Remark 3.9. *The one-step estimator $\hat{a}(m)$ is asymptotically unbiased, and shares a similar form to the de-biased estimator proposed by Zhang and Zhang [2014], Van de Geer et al. [2014]. The de-biased estimator in Van de Geer et al. [2014] would take the following form under our setting:*

$$\widehat{b}_m = (\widehat{A}_m)_{D_m} + \widehat{\Theta}_{D_m,\cdot}\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{X}_t\left(X_{t+1,m} - \mathcal{X}_t^\top\widehat{A}_m\right),$$

where $\widehat{\Theta}$ is computed by node-wise regression, as an estimator for Υ^{-1}. When $d_m = |D_m| = 1$, this is essentially the same as our estimator $\hat{a}(m)$, but would be slightly different in the multivariate case. Note that the asymptotic covariance matrix for $\hat{a}(m)$ equals to the partial information matrix $I^*(A_{m,D_m}|A_{m,D_m^c})$, and thus is semi-parametric efficient, while $\hat{b}_m$ is only efficient when it is a scalar.

Remark 3.10. *$\widehat{R}_T$ is also very similar to the test statistic proposed by Neykov et al. [2018] for VAR model with lag 1. The only difference lies in the estimation of $Var(\epsilon_{ti})$, and they only consider Dantzig selector for estimating A^* and w_m^*. We will provide a detailed comparison between their theoretical result with ours in section 3.6.*

3.6 Special case: AR(1) with Gaussian noise

Our theoretical guarantee covers VAR models with lag p and sub-Gaussian noise, of which AR(1) model and Gaussian noise are special cases. Here we explain the consequences of our result under this special case and provide comparison with Neykov et al. [2018].

When we consider lag $p = 1$, the constraint for A^* becomes

$$\sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} \left\| (A^*)^{i+j} \right\|_2^2 \right)^{\frac{1}{2}} \leq \beta, \max_m \rho_m(A^*) \leq \rho, \max_m s_m(A^*) \leq s,$$

with $(\rho \vee s) \log M = o(\sqrt{T})$. The two sparsity conditions and sample size requirement are included in the conditions Neykov et al. [2018] proposes. In addition, they assume the following:

$$\|A^*\|_1 \leq C, \|A^*\|_2 \leq 1 - \varepsilon, \left\| \Sigma^{-1} \right\|_1 \leq C.$$

for some $0 < \varepsilon < 1$. Note that we don't require these conditions, among which the first and third are quite strong, and the second one $\|A^*\|_2 \leq 1 - \varepsilon$ is sufficient for our condition $\sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} \left\| (A^*)^{i+j} \right\|_2^2 \right)^{\frac{1}{2}} \leq \beta$. This follows since if $\|A^*\|_2 \leq 1 - \varepsilon$,

$$\sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} \left\| (A^*)^{i+j} \right\|_2^2 \right)^{\frac{1}{2}} \leq \sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} \|A^*\|_2^{2(i+j)} \right)^{\frac{1}{2}} \leq \frac{\sum_{i=0}^{\infty} (1-\varepsilon)^i}{\sqrt{1 - (1-\varepsilon)^2}} \leq \left(2\varepsilon - \varepsilon^2 \right)^{-\frac{1}{2}}.$$

Until now the discussion focuses on the case where ϵ_{ti} are i.i.d. sub-Gaussian noise of scale factor $C\sigma^*$, with $(\sigma^*)^2$ being the variance of ϵ_{ti} and lower bounded by some constant. Thus our setting covers the case where $\epsilon_t \sim \mathcal{N}(0, (\sigma^*)^2 I)$ with $\sigma^* \geq c$. If $\epsilon_t \sim \mathcal{N}(0, \Psi)$ with $\Psi_{ii} \geq c$ as assumed in Neykov et al. [2018], we can still prove the same theoretical guarantee, under even weaker condition based on spectral density, due to established concentration bounds in Basu et al. [2015].

4 Numerical Experiments

In this section, we provide a simulation study to validate our theoretical results. For simplicity, our simulation is based on the AR(1) model:

$$X_{t+1} = A^* X_t + \epsilon_t, \quad t = 0, \ldots, T, \tag{37}$$

where $A^* \in \mathbb{R}^{M \times M}$ is set to be row-wise sparse. Symmetricity is not required in our theory, but in order to ensure the sparsity of w_m^*, we focus on symmetric matrices under $\mathcal{H}_0$, and slightly asymmetric ones under $\mathcal{H}_A$. The eigenvalues of A^* all fall in the unit circle of the complex plane, which ensures the existence of stationary solution to this model. White noise ϵ_{ti} is simulated as independent Uniform$(-1, 1)$ in order to satisfy the sub-Gaussianity condition. Other distributions were also used but not reported since the results were very similar.

To consider multi-variate test sets, throughout the simulation we test the index set D with $d = |D| = 6$, which involves three different rows and two columns in each row:

$$D = \{(1, 3), (1, 5), (3, 3), (3, 4), (5, 4), (5, 8)\}.$$

The null hypothesis takes the form $\mathcal{H}_0 : \widetilde{A}_D = \mu$ with some d-dimensional vector μ. Correspondingly, we consider alternative hypothesis $\mathcal{H}_A : \widetilde{A}_D = \mu + T^{-\phi} \Delta$, with Δ randomly selected from d-dimensional Gaussian distribution, and ϕ ranges from 0.25 to 1.2.

Under $\mathcal{H}_0$, we generate A^* with different row-wise sparsity levels and structures, and for each A^*, vector μ may differ depending on the corresponding $\widetilde{A}_D$. Under $\mathcal{H}_A$, A^* are still the same matrices as under $\mathcal{H}_0$, but only adding the tested indices $\widetilde{A}_D$ by $T^{-\phi} \Delta$. The experiments are repeated under different settings of A^*, Δ, M, T and ϕ.

We use Lasso estimators defined in (25), (27) for the estimation of A^* and w_m^*, $1 \leq m \leq k$, and tuning parameters λ_A, λ_w are selected using cross validation. In cross validation, the training sets are composed of consecutive time series data, with the remaining 10% of the original data set being testing sets. Under $\mathcal{H}_0$, 1000 simulations are carried out under each parameter setting, while under $\mathcal{H}_A$, we have 100 simulations. In the following sections, we look into false positive rates (FPR) and true positive rates (TPR) of test statistics $\widetilde{U}_T$ and $\widehat{R}_T$ as defined in (32) and (36), when we set the level of test as $\alpha = 0.05$.

4.1 Under the Null Hypothesis

(1) Varying sparsity

Here we summarize the experiments with randomly generated A^*, that are symmetric and row-wise sparse, with different sparsity levels ρ defined in (10). Figure 1 shows how FPR of $\widetilde{U}_T$ and $\widehat{R}_T$ averaged over 1000 experiments vary with $\sqrt{T}$. We can see that when T
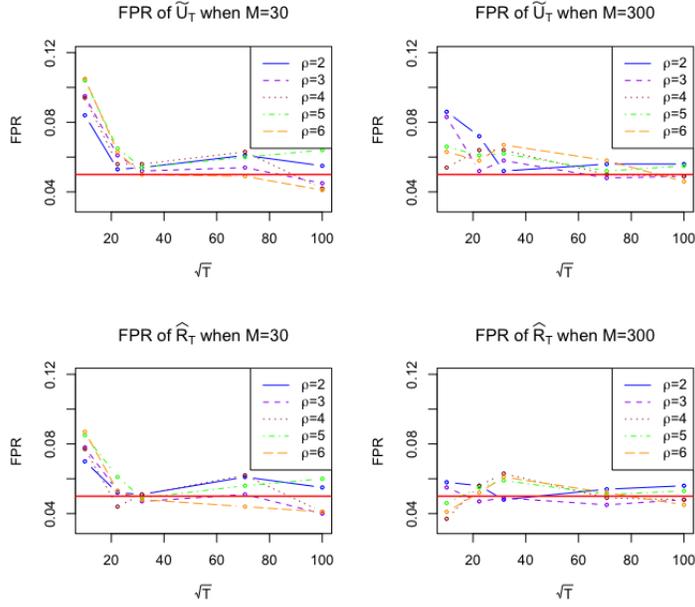
Figure 1: False positive rate (FPR) of $\widetilde{U}_T$ and $\widehat{R}_T$ v.s. $\sqrt{T}$, with various dimension M and sparsity level ρ. The red line is the significance level $\alpha = 0.05$.

increases to about 500, the FPR becomes stable and close to $\alpha = 0.05$ regardless of ρ, M, choice between $\widetilde{U}_T$ and $\widehat{R}_T$.

When the sample size T is small, the test tends to be conservative, which is the consequence of estimating variance σ^{*2} and covariances $\Upsilon^{(m)}$'s. In the simulation we use naive estimators for these two quantities, as defined in (31) and (29) which tend to be smaller than the true parameters. This is because we usually fit noise in the regression, as noticed by Fan et al. [2012]. As shown in these two figures, $\widehat{R}_T$ is less conservative than $\widetilde{U}_T$ when T is small, since the magnitude of $\widetilde{\Upsilon}^{(m)}$ is larger than $\widehat{\Upsilon}^{(m)}$, which makes $\left(\widetilde{\Upsilon}^{(m)\top}\right)^{-1} \widehat{\Upsilon}^{(m)} \left(\widetilde{\Upsilon}^{(m)}\right)^{-1}$ probably a better estimator for $\Upsilon^{(m)}$. We also summarize the FPR when the variance σ^{*2} of ϵ_{ti} is known in Figure 2. We can see from these figures that $\widehat{U}_T$ is still a little conservative when T is small, while $\widehat{R}_T$ with $\hat{\sigma}^2$ substituted by σ^{*2} is not conservative.

(2) Different Graph Structures

If we consider the M actors in the time series as nodes in a network, and a nonzero A_{ij}^* represents an directed edge from j to i, then each matrix A^* corresponds to a M-dimensional directed graph. We experiment with different structures of A^*, which also correspond to different graph structure, including block graph or chain graph. Specifically, we consider
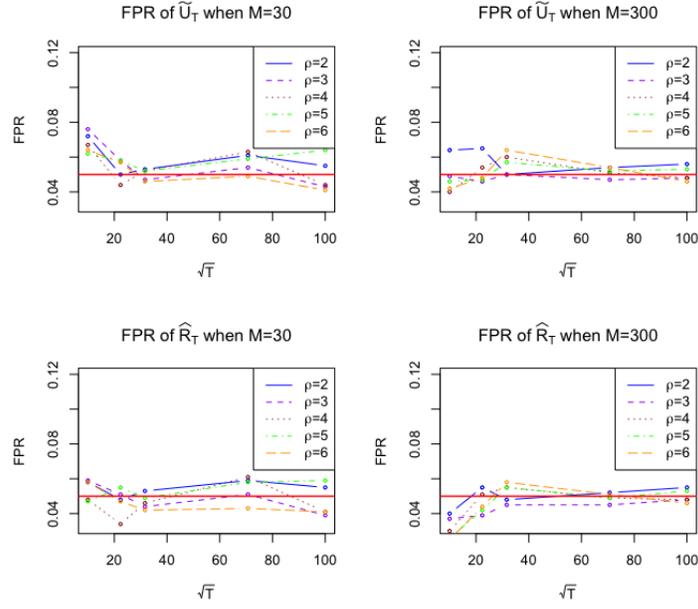
18

Figure 2: FPR of $\widetilde{U}_T$ and $\widehat{R}_T$ when residual variance is known.

matrices with ℓ_2 norm equal to 0.75:

$$A^{(1)} = \begin{pmatrix} 1/4 & 1/2 & 0 & 0 & \cdots & 0 & 0 \\ 1/2 & 1/4 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1/4 & 1/2 & \cdots & \vdots & \vdots \\ 0 & 0 & 1/2 & 1/4 & \cdots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & \cdots & 1/4 & 1/2 \\ 0 & 0 & \cdots & \cdots & \cdots & 1/2 & 1/4 \end{pmatrix},$$

which is a block graph;

$$A^{(2)} = \begin{pmatrix} c & c & 0 & \cdots & \cdots & 0 \\ c & 0 & c & \cdots & \cdots & 0 \\ 0 & c & 0 & c & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & c & 0 & c \\ 0 & \cdots & \cdots & \cdots & c & 0 \end{pmatrix},$$

with constant c chosen to ensure $\left\| A^{(2)} \right\|_2 = 0.75$, which is a chain graph; and $A^{(3)}$ being randomly generated symmetric matrix of sparsity level $\rho = 2$, and largest eigenvalue equal to 0.75. Figure 3 shows the difference among these three different structures. We can see
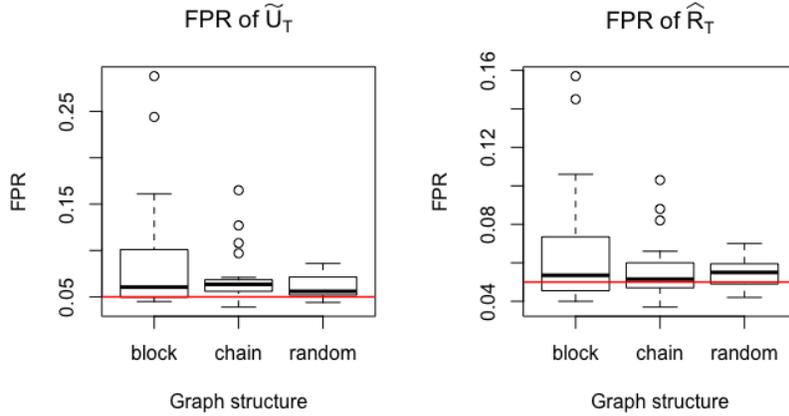
19

Figure 3: FPR under different graph structure. Block refers to $A^{(1)}$, chain refers to $A^{(2)}$ and random refers to $A^{(3)}$.

that block graph is less accurate than the other two, which is due to a larger variance for each $X_{t,D_m} - w_m^{*\top} X_{t,D_m^c}$. Investigating the question of how graph structure theoretically influences testing performance remains an open and interesting direction.

4.2 Alternative Hypothesis

First we look into how the true positive rate (TPR) varies with $\|T^{-\phi}\Delta\|_2$, since we set $\mathcal{H}_A$ as $\widetilde{A}_D = \mu + T^{-\phi}\Delta$ and $\|T^{-\phi}\Delta\|_2$ may be viewed as a measure of distance from the null hypothesis. Fig. 4 only presents the simulation results when $A^* = A^{(1)}$ and $M = 300$, while the other choices of A^* and M generate very similar results. We can see from these two figures that as $\|T^{-\phi}\Delta\|_2$ increases, TPR approaches 1. The slope increases when sample size T gets larger, or when the test statistic changes from $\widehat{R}_T$ to $\widetilde{U}_T$. This aligns with intuition, since when T increases, we are supposed to distinguish between $\mathcal{H}_0$ and $\mathcal{H}_A$ better, and $\widetilde{U}_T$ is more conservative than $\widehat{R}_T$ as we show in subsection 4.1.

We also check the influence of ϕ. Figure 5 reveals how TPR changes when T increases, if we set $\left\|\widetilde{\Delta}\right\|_2$ and ϕ fixed. If $\phi < 0.5$, TPR converges to 1 very quickly, while if $\phi > 0.5$, TPR converges to 0.05, but the convergence is slower when ϕ or $\left\|\widetilde{\Delta}\right\|_2$ increases. When $\phi = 0.5$, Theorem 3.3 and 3.4 states that $\widetilde{U}_T$ and $\widehat{R}_T$ would converge to $\chi_{d,\|\widetilde{\Delta}\|_2^2}$, thus the TPR should converge to some value between 0.05 and 1, depending on d and $\left\|\widetilde{\Delta}\right\|_2^2$. The black lines in figure 5 indicate this convergence value, but since the test tends to be conservative when T is not large enough, TPR when $\phi = 0.5$ is usually above the black line. The conservative issue is more severe under $\mathcal{H}_A$ since the deviation $\widetilde{\Delta}$ is also multiplied by the estimated variances, which exaggerates
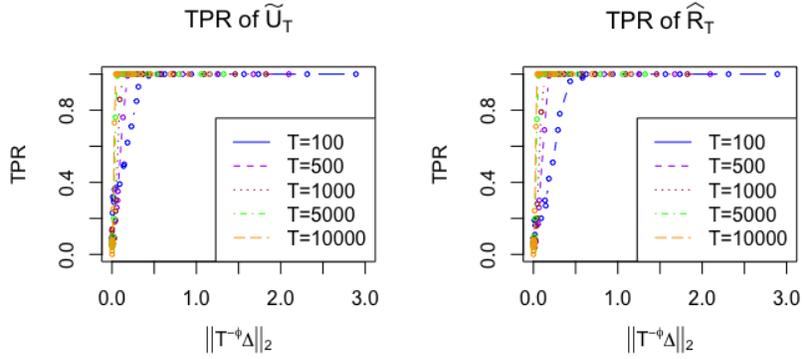
20

Figure 4: True positive rate of $\widetilde{U}_T$ and $\widehat{R}_T$, when $A^* = A^{(1)}$ and $M = 300$

the conservative tendency. However, this may not be a big concern under $\mathcal{H}_A$, since we always want the TPR to be large.

5 Proof Overview

One of the main contributions of this work is the proof technique, which addresses a number of technical challenges and develops novel concentration bounds for dependent sub-Gaussian random vectors. In this section, we present and discuss key lemmas for the proof and provide the main steps for proving Theorems 3.1 and 3.2, deferring the more technically intensive steps to the supplement.

5.1 Key Lemmas

The major technical challenge lies in proving the following two concentration bounds for dependent sub-Gaussian random vectors.

Lemma 5.1 (Deviation Bound for A^*). *Under model* (3), *when* ϵ_{ti} *are sub-Gaussian noise with scale factor* τ, *and* $A^* \in \Omega_0 \cup \Omega_1$,

$$\mathbb{P}\left(\left\| \frac{1}{T} \sum_{t=0}^{T-1} \epsilon_t \mathcal{X}_t^\top \right\|_\infty > C\sqrt{\frac{\log M}{T}} \right) \leq c_1 \exp\{-c_2 \log M\},$$

When $T \geq C \log M$.

Lemma 5.1 is a standard deviation bound for proving estimation error bound of Lasso type or Dantzig selector type estimators. We apply this lemma both in the proof of Theorem 3.1, 3.2 and Lemma 3.1.
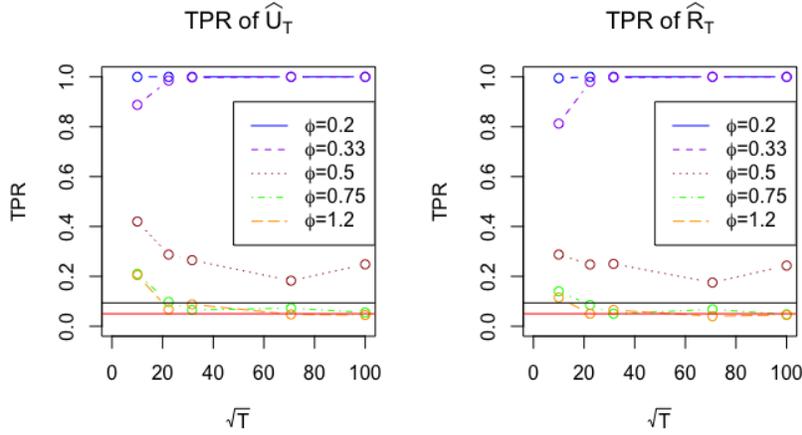
Figure 5: TPR of $\widetilde{U}_T$ and $\widehat{R}_T$ when $\left\|\tilde{\Delta}\right\|_2 = 1$, $A^* = A^{(1)}$. Results for different graph size M from 30 to 300 are combined together and average TPR is taken. Red line is significance level α, the value that TPR should converge to when $\phi < 0.5$; while the black line is the convergence point specified in Theorem 3.2 when $\phi = 0.5$.

Lemma 5.2. *Under model* (3), *when ϵ_{ti} are sub-Gaussian noise with constant scale factor τ, and $A^* \in \Omega_0 \cup \Omega_1$, if $B \in \mathbb{R}^{pM \times pM}$ is a symmetric matrix, we have*

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{X}_t^\top B \mathcal{X}_t - tr(B\Upsilon)\right| > \delta\right) \leq c_1 \exp\left\{-c_2 T \min\left\{\frac{\delta}{\|B\|_2}, \frac{\delta^2}{\|B\|_{tr}\|B\|_2}\right\}\right\}.$$

Lemma 5.2 provides concentration bound for the sample average of general quadratic form $\mathcal{X}_t^\top B \mathcal{X}_t$, and is very helpful in proving martingale CLT under our setting, REC, Lemma 3.3, etc.

In the Gaussian case, both these lemmas follow from prior work in Basu et al. [2015] which relies on the fact that dependent Gaussian vectors can be rotated to be independent. Since dependent sub-Gaussian random variables cannot be rotated to be independent (only uncorrelated), we exploit the independence of ϵ_t by representing each $\mathcal{X}_t$ by linear function of the infinite series $\{\epsilon_i\}_{i=-\infty}^{i=t}$ and then use a careful truncation argument. We analyze sufficiently many terms in the summation, and control the infinite residues.

5.2 Proof of Theorem 3.1

Proof. Suppose $A^* \in \Omega_0$. We will use C_i, c_i to refer to constants that only depend on p, d, β, τ (not M or T), and different constants might share the same notation.

The proof can be divided into two major parts: showing the convergence of U_T to χ_d^2, and

bounding the estimation error $\left|\widehat{U}_T - U_T\right|$. Formally, for any $\varepsilon > 0$,

$$\mathbb{P}(\widehat{U}_T \leq x) - F_d(x)$$
$$\leq \mathbb{P}(U_T \leq x + \varepsilon) + \mathbb{P}(\left|\widehat{U}_T - U_T\right| > \varepsilon) - F_d(x)$$
$$\leq |\mathbb{P}(U_T \leq x + \varepsilon) - F_d(x + \varepsilon)| + F_d(x + \varepsilon) - F_d(x) + \mathbb{P}\left(\left|\widehat{U}_T - U_T\right| > \varepsilon\right),$$

and

$$F_d(x) - \mathbb{P}(\widehat{U}_T \leq x)$$
$$= \mathbb{P}(\widehat{U}_T > x) - (1 - F_d(x))$$
$$\leq \mathbb{P}(U_T > x - \varepsilon) + \mathbb{P}(\left|\widehat{U}_T - U_T\right| > \varepsilon) - 1 + F_d(x)$$
$$\leq |F_d(x - \varepsilon) - \mathbb{P}(U_T \leq x - \varepsilon)| + F_d(x) - F_d(x - \varepsilon) + \mathbb{P}\left(\left|\widehat{U}_T - U_T\right| > \varepsilon\right),$$

which implies

$$\left|\mathbb{P}(\widehat{U}_T \leq x) - F_d(x)\right|$$
$$\leq \sup_{y \in \mathbb{R}} |\mathbb{P}(U_T \leq y) - F_d(y)| + F_d(x + \varepsilon) - F_d(x - \varepsilon) + \mathbb{P}\left(\left|\widehat{U}_T - U_T\right| > \varepsilon\right). \tag{38}$$

In the following, we provide bounds on each of the three terms. The following lemma shows the uniform weak convergence rate of $\|V_T + \mu\|_2^2$ to $\chi^2_{d,\|\mu\|_2^2}$, of which the convergence of $U_T = \|V_T\|_2^2$ to χ^2_d is a special case.

Lemma 5.3 (Convergence Rate of $\|V_T + \mu\|_2^2$). *Under model (3) with ϵ_{ti} being sub-Gaussian noise of scale factor τ, then for any $A^* \in \Omega_0$, $\forall \mu \in \mathbb{R}^d$,*

$$\sup_{x \in \mathbb{R}} \left|\mathbb{P}(\|V_T + \mu\|_2^2 \leq x) - F_{d,\|\mu\|_2^2}(x)\right| \leq C(\|\mu\|_2) T^{-\frac{1}{8}}, \tag{39}$$

when $T > C$ for some absolute constant C, where $C(\|\mu\|_2)$ is a constant depending on and is non-decreasing with respect to $\|\mu\|_2$.

This Lemma is proved in section C, by applying a uniform martingale central limit theorem result. Thus, by Lemma 5.3, if $T > C$ for some constant C,

$$\sup_{y \in \mathbb{R}} |\mathbb{P}(U_T \leq y) - F_d(y)| \leq CT^{-\frac{1}{8}}.$$

Meanwhile,

$$F_d(x + \varepsilon) - F_d(x - \varepsilon) \leq C_2 \varepsilon$$

since χ^2_d has bounded density.

Now we only need to choose a proper ε and bound $\mathbb{P}\left(\left|\widehat{U}_T - U_T\right| > \varepsilon\right)$.

$$
\begin{aligned}
\left|\widehat{U}_T - U_T\right| &\leq \sum_{m=1}^{k} \left| T\widehat{S}_m^{\top}(\widehat{\Upsilon^{(m)}})^{-1}\widehat{S}_m - \|V_{T,m}\|_2^2 \right| \\
&\leq \sum_{m=1}^{k} \left| T\widehat{S}_m^{\top}\left((\widehat{\Upsilon^{(m)}})^{-1} - (\Upsilon^{(m)})^{-1}\right)\widehat{S}_m + \left\|\sqrt{T}(\Upsilon^{(m)})^{-\frac{1}{2}}\widehat{S}_m\right\|_2^2 - \|V_{T,m}\|_2^2 \right| \\
&\leq \sum_{m=1}^{k} \left\| \Upsilon^{(m)\frac{1}{2}}\left(\widehat{\Upsilon^{(m)}}\right)^{-1}\Upsilon^{(m)\frac{1}{2}} - I \right\|_{\infty} \left\|\sqrt{T}(\Upsilon^{(m)})^{-\frac{1}{2}}\widehat{S}_m\right\|_1^2 \\
&\quad + \left\|\sqrt{T}(\Upsilon^{(m)})^{-\frac{1}{2}}(\widehat{S}_m - S_m)\right\|_2^2 + 2\|V_{T,m}\|_2\left\|\sqrt{T}(\Upsilon^{(m)})^{-\frac{1}{2}}(\widehat{S}_m - S_m)\right\|_2.
\end{aligned}
\tag{40}
$$

Define $E_m = \sqrt{T}(\Upsilon^{(m)})^{-\frac{1}{2}}\left(\widehat{S}_m - S_m\right)$, then (40) turns into

$$
\begin{aligned}
\left|\widehat{U}_T - U_T\right| &\leq \sum_{m=1}^{k} \|E_m\|_2^2 + 2\|V_{T,m}\|_2\|E_m\|_2 \\
&\quad + \left\| \Upsilon^{(m)\frac{1}{2}}\left(\widehat{\Upsilon^{(m)}}\right)^{-1}\Upsilon^{(m)\frac{1}{2}} - I \right\|_{\infty}\left(\|V_{T,m}\|_2 + \|E_m\|_2\right)^2.
\end{aligned}
\tag{41}
$$

We can bound $\|V_{T,m}\|_2$ using Lemma 5.3 and $\left\|\Upsilon^{(m)\frac{1}{2}}\left(\widehat{\Upsilon^{(m)}}\right)^{-1}\Upsilon^{(m)\frac{1}{2}} - I\right\|_{\infty}$ using Lemma 19, while for bounding the estimation induced error $\|E_m\|_2$, we first apply the following lemma to bound the eigenvalues of $\Upsilon^{(m)}$.

Lemma 5.4. *Consider the model (2) with independent noise ϵ_{ti} of unit variance, A^* satisfies (13), then the eigenvalues of Υ can be bounded as follows:*

$$
0 < C_1(\beta) \leq \Lambda_{\min}(\Upsilon) \leq \Lambda_{\max}(\Upsilon) \leq C_2(\beta).
$$

Lemma 5.4 is proved based on established results in Basu et al. [2015]. Note that we assumed unit variance in Theorem 3.1 and 3.2, so we can apply Lemma 5.4 here. Since $\left(\Upsilon^{(m)}\right)^{-1} = \left(\Upsilon^{-1}\right)_{D_m,D_m}$, applying Lemma 5.4 would lead us to the following:

$$
\begin{aligned}
\Lambda_{\min}\left((\Upsilon^{(m)})^{-1}\right) &\geq \Lambda_{\min}(\Upsilon^{-1}) = \Lambda_{\max}(\Upsilon)^{-1} \geq C, \\
\Lambda_{\max}\left((\Upsilon^{(m)})^{-1}\right) &\leq \Lambda_{\max}(\Upsilon^{-1}) = \Lambda_{\min}(\Upsilon)^{-1} \leq C.
\end{aligned}
\tag{42}
$$

Thus we have

$$
\|E_m\|_2 \leq C\sqrt{T}\left\|\widehat{S}_m - S_m\right\|_2,
$$

with

$$\widehat{S}_m - S_m = (\hat{w}_m - w_m^*)^\top \frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_{t,D_m^c} \epsilon_{t,m}$$

$$+ \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{X}_{t,D_m} - w_m^{*\top} \mathcal{X}_{t,D_m^c}) \mathcal{X}_{t,D_m^c}^\top \left((\widehat{A}_m)_{D_m^c} - (A_m^*)_{D_m^c} \right) \tag{43}$$

$$- (\hat{w}_m - w_m^*)^\top \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_{t,D_m^c} \mathcal{X}_{t,D_m^c}^\top \right) \left((\widehat{A}_m)_{D_m^c} - (A_m^*)_{D_m^c} \right).$$

The following two lemmas provide bounds for $\left\| \frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_{t,D_m^c} \epsilon_{t,m} \right\|_\infty$, and

$$\left\| \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{X}_{t,D_m} - w_m^{*\top} \mathcal{X}_{t,D_m^c}) \mathcal{X}_{t,D_m^c}^\top \right\|_\infty.$$

Lemma 5.5. *When $T \geq C \log M$,*

$$\mathbb{P}\left(\left\| \frac{1}{T} \sum_{t=0}^{T-1} \epsilon_t \mathcal{X}_t^\top \right\|_\infty > C\sqrt{\frac{\log M}{T}} \right) \leq c_1 \exp\{-c_2 \log M\}.$$

Lemma 5.1 is a common condition in high-dimensional regression problems, and is usually referred to as deviation bound. We will prove it in Section C.

Lemma 5.6 (Deviation Bound for w_m^*). *With probability at least $1 - c_1 \exp\{-c_2 \log M\}$, for all $1 \leq m \leq k$,*

$$\left\| \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{X}_{t,D_m} - w_m^{*\top} \mathcal{X}_{t,D_m^c}) \mathcal{X}_{t,D_m^c}^\top \right\|_\infty \leq C\sqrt{\frac{\log M}{T}}.$$

Lemma 5.6 can also be viewed as a deviation bound, if we consider a regression problem with $\mathcal{X}_{t,D_m}$ as response and $\mathcal{X}_{t,D_m^c}$ as covariates. This is also proved in Section C. Applying Assumptions 3.1 and 3.2, with probability at least $1 - c_1 \exp\{-c_2 \log M\}$,

$$\|E_m\|_2 \leq C \frac{(s_m \vee \rho_m) \log M}{\sqrt{T}} + \sqrt{T} Q_1^{\frac{1}{2}} Q_2^{\frac{1}{2}} \leq C \frac{(s_m \vee \rho_m) \log M}{\sqrt{T}},$$

where

$$Q_1 = \left(\left(\widehat{A}_m \right)_{D_m^c} - (A_m^*)_{D_m^c} \right)^\top \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_{t,D_m^c} \mathcal{X}_{t,D_m^c}^\top \right) \left(\left(\widehat{A}_m \right)_{D_m^c} - (A_m^*)_{D_m^c} \right)$$

$$Q_2 = \mathrm{tr}\left[(\hat{w}_m - w_m^*)^\top \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_{t,D_m^c} \mathcal{X}_{t,D_m^c}^\top \right) (\hat{w}_m - w_m^*) \right],$$

25

and Assumption 3.1 and 3.2 implies $Q_1 \leq C\frac{\rho_m \log M}{T}$ and $Q_2 \leq C\frac{s_m \log M}{T}$. The former is not straightforward: to see why it holds true, let $\hat{h}_m = \widehat{A}_m - A_m^*$ and $H = \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{X}_t \mathcal{X}_t^\top$, then we have

$$
\begin{aligned}
Q_1 =& \frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{X}_{t,D_m^c}^\top\left(\hat{h}_m\right)_{D_m^c}\right]^2 \\
=& \frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{X}_t^\top \hat{h}_m - \mathcal{X}_{t,D_m}^\top\left(\hat{h}_m\right)_{D_m}\right]^2 \\
\leq& \frac{2}{T}\sum_{t=0}^{T-1}\left[\left(\mathcal{X}_t^\top \hat{h}_m\right)^2 + \left(\mathcal{X}_{t,D_m}^\top\left(\hat{h}_m\right)_{D_m}\right)^2\right] \\
=& 2\hat{h}_m^\top H \hat{h}_m + 2\left(\hat{h}_m\right)_{D_m}^\top H_{D_m,D_m}\left(\hat{h}_m\right)_{D_m} \\
\leq& C\frac{\rho_m \log M}{T}.
\end{aligned}
\tag{44}
$$

Here we apply Assumption 3.1, and the fact that

$$
\begin{aligned}
&\left(\hat{h}_m\right)_{D_m}^\top H_{D_m,D_m}\left(\hat{h}_m\right)_{D_m} \\
\leq& d_m \|H\|_\infty \|\hat{h}_m\|_2^2 \\
\leq& d_m\left(\|H - \Upsilon\|_\infty + \Lambda_{\max}(\Upsilon)\right)\frac{\rho_m \log M}{T} \\
\leq& C\frac{\rho_m \log M}{T}.
\end{aligned}
$$

The last inequality is due to Lemma 5.4 and the following lemma:

Lemma 5.7. *With probability at least $1 - c_1\exp\{-c_2\log M\}$,*

$$
\left\|\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{X}_t \mathcal{X}_t^\top - \Upsilon\right\|_\infty \leq C\sqrt{\frac{\log M}{T}}.
$$

Therefore, by taking a union bound, we show that

$$
\|E_m\|_2 \leq C\frac{(s_m \vee \rho_m)\log M}{\sqrt{T}},
$$

for any $1 \leq m \leq k$, with probability at least $1 - c_1\exp\{-c_2\log M\}$.

Meanwhile, by applying Lemma 5.3, one can show that for $y > \sqrt{5d}$,

$$
\begin{aligned}
\mathbb{P}\left(\|V_{T,m}\|_2 > y\right) \leq& CT^{-\frac{1}{8}} + 1 - F_d(y^2) \\
\leq& CT^{-\frac{1}{8}} + \exp\{-(y^2 - d)/4\} \\
\leq& CT^{-\frac{1}{8}} + Cy^{-2},
\end{aligned}
\tag{45}
$$

26

where the second inequality is due to a χ_d^2 tail bound established in Laurent and Massart [2000] (see Lemma 1 in Laurent and Massart [2000]), and the third inequality comes from the fact that, $\forall$ constant $C_1 > 0$, $\exists$ constant C_2 such that

$$\sup_{y \geq 0} y^2 e^{-C_1 y^2} \leq C_2.$$

Let $y = \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{-\frac{1}{4}}$ and plug it into (41), then with Assumption 3.3, we can show that with probability at least

$$1 - c_1 \exp\{-c_2 \log M\} - c_3 T^{-\frac{1}{8}} - c_4 \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{\frac{1}{2}},$$

the following holds:

$$\left| \widehat{U}_T - U_T \right| \leq C_1 \frac{(s \vee \rho) \log M}{\sqrt{T}} \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{-\frac{1}{2}} + C_2 \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{\frac{3}{4}}$$

$$\leq C \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{\frac{1}{2}},$$

if $(s \vee \rho) \log M = o(\sqrt{T})$ and $T > C$ for some constant C. Therefore, applying (38) with $\varepsilon = C \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{\frac{1}{2}}$,

$$\left| \mathbb{P}(\widehat{U}_T \leq x) - F_d(x) \right| \leq C_1 T^{-\frac{1}{8}} + C_2 \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{\frac{1}{2}} + C_3 \exp\{-c \log M\}.$$

Since constants C_i only depend on d, β and τ, this bound also holds for supremum over $A^* \in \Omega_0$ and $x \in \mathbb{R}$. Note that for a clear presentation, we are not showing the sharpest bound, which can be obtained by choosing a different y. $\qquad\square$

5.3 Proof of Theorem 3.2

proof of Theorem 3.2. We prove this case by case. We will use C_i, c_i to refer to constants that only depend on d, β, Δ, ϕ, and different constants might share the same notation.

Similar from the proof of Theorem 3.1, the major part of the proof is devoted to bounding $\left| \widehat{U}_T - \|V_T + \mu\|_2^2 \right|$ with high probability for some vector $\mu \in \mathbb{R}^d$.

(1) $\phi = \frac{1}{2}$

Suppose $A^* \in \Omega_1$. Using similar deduction as in the proof of Theorem 3.1, for any $\varepsilon > 0$,

$$\left| \mathbb{P}(\widehat{U}_T \leq x) - F_{d, \|\widetilde{\Delta}\|_2^2}(x) \right|$$

$$\leq \sup_{y \in \mathbb{R}} \left| \mathbb{P}\left(\|V_T - \widetilde{\Delta}\|_2^2 \leq y \right) - F_{d, \|\widetilde{\Delta}\|_2^2}(y) \right| \qquad (46)$$

$$+ F_{d, \|\widetilde{\Delta}\|_2^2}(x + \varepsilon) - F_{d, \|\widetilde{\Delta}\|_2^2}(x - \varepsilon) + \mathbb{P}\left(\left| \widehat{U}_T - \left\| V_T - \widetilde{\Delta} \right\|_2^2 \right| > \varepsilon \right).$$

(a) Bounding the first two terms

The first term is the convergence rate of $\|V_T - \widetilde{\Delta}\|_2^2$ to $\chi^2_{d,\|\widetilde{\Delta}\|_2^2}$. By Lemma 5.3,

$$\sup_{y \in \mathbb{R}} \left| \mathbb{P}\left(\|V_T - \widetilde{\Delta}\|_2^2 \le y \right) - F_{d,\|\widetilde{\Delta}\|_2^2}(y) \right| \le C(\|\widetilde{\Delta}\|_2) T^{-\frac{1}{8}} \le C\|\Delta\|_2 T^{-\frac{1}{8}}.$$

The last inequality is due to

$$\|\widetilde{\Delta}\|_2^2 = \sum_{m=1}^{k} \|\widetilde{\Delta}_m\|_2^2 \le \sum_{m=1}^{k} \Lambda_{\max}\left(\Upsilon^{(m)} \right) \|\Delta\|_2^2,$$

and an upper bound for $\Lambda_{\max}\left(\Upsilon^{(m)} \right)$ in (42).

Bounding the second term in (46) is not straightforward as bounding $F_d(x+\varepsilon) - F_d(x-\varepsilon)$ in the proof of Theorem 3.1, since $\widetilde{\Delta}$ is not a constant vector when A^* takes different values in Ω_1^*. We only have a uniform bound of $\left\| \widetilde{\Delta} \right\|_2$ as shown above. One can show that

$$F_{d,\|\widetilde{\Delta}\|_2^2}(x+\varepsilon) - F_{d,\|\widetilde{\Delta}\|_2^2}(x-\varepsilon) = \mathbb{P}\left(\left\| Z + \widetilde{\Delta} \right\|_2^2 \in (x-\varepsilon, x+\varepsilon] \right)$$

$$\le \begin{cases} C(d)\left((x+\varepsilon)^{\frac{d}{2}} - (x-\varepsilon)^{\frac{d}{2}} \right) e^{-(\sqrt{x-\varepsilon} - \|\widetilde{\Delta}\|_2)^2/2}, & \sqrt{x-\varepsilon} \ge 2\|\widetilde{\Delta}\|_2 \\ C(d)\left((x+\varepsilon)^{\frac{d}{2}} - (x-\varepsilon)^{\frac{d}{2}} \right), & \sqrt{x-\varepsilon} < 2\|\widetilde{\Delta}\|_2 \end{cases},$$

where Z is a d-dimensional standard Gaussian random vector with density $\phi(z) = C(d)\exp\{-\|z\|_2^2/2\}$. The last inequality holds because that, for any set $\mathcal{C} \subset \mathbb{R}^d$,

$$\mathbb{P}\left(Z \in \mathcal{C} \right) \le \sup_{z \in \mathcal{C}} \phi(z) \int_{z \in \mathcal{C}} \mathrm{d}z.$$

Suppose $0 < \varepsilon \le 1$, then if $\sqrt{x-\varepsilon} \ge 2\|\widetilde{\Delta}\|_2$,

$$\left((x+\varepsilon)^{\frac{d}{2}} - (x-\varepsilon)^{\frac{d}{2}} \right) \exp\left\{ -(\sqrt{x-\varepsilon} - \|\widetilde{\Delta}\|_2)^2/2 \right\}$$

$$\le d\varepsilon (x+\varepsilon)^{\frac{d}{2}-1} \exp\{-(x-\varepsilon)/8\}$$

$$\le d\varepsilon e^{\frac{\varepsilon}{4}} \sup_{y \ge 0} y^{\frac{d}{2}-1} \exp\{-y/8\} \le C(d)\varepsilon,$$

otherwise,

$$\left((x+\varepsilon)^{\frac{d}{2}} - (x-\varepsilon)^{\frac{d}{2}} \right) \le d\varepsilon (x+\varepsilon)^{\frac{d}{2}-1} \le C(d)\varepsilon.$$

Thus,

$$F_{d,\|\widetilde{\Delta}\|_2^2}(x+\varepsilon) - F_{d,\|\widetilde{\Delta}\|_2^2}(x-\varepsilon) \le C(d)\varepsilon.$$

(b) Bounding $\left| \widehat{U}_T - \left\| V_T - \widetilde{\Delta} \right\|_2^2 \right|$

Similar from (41) in the proof of Theorem 3.1, it is straightforward to show that

$$\left| \widehat{U}_T - \left\| V_T - \widetilde{\Delta} \right\|_2^2 \right|$$

$$\leq \sum_{m=1}^{k} \|E_m\|_2^2 + 2 \left\| V_{T,m} - \widetilde{\Delta}_m \right\|_2 \|E_m\|_2 \tag{47}$$

$$+ \left\| \Upsilon^{(m)\frac{1}{2}} \left(\widehat{\Upsilon^{(m)}} \right)^{-1} \Upsilon^{(m)\frac{1}{2}} - I \right\|_\infty \left(\|V_{T,m} - \widetilde{\Delta}_m\|_2 + \|E_m\|_2 \right)^2,$$

where $E_m = \sqrt{T}(\Upsilon^{(m)})^{-\frac{1}{2}} \widehat{S}_m - V_{T,m} + \widetilde{\Delta}_m$. To bound $\|E_m\|_2$, note that

$$V_{T,m} - \widetilde{\Delta}_m = \sqrt{T}(\Upsilon^{(m)})^{-\frac{1}{2}} S_m - \widetilde{\Delta}_m = \sqrt{T}(\Upsilon^{(m)})^{-\frac{1}{2}} (S_m - \Upsilon^{(m)}(A_m^*)_{D_m}),$$

and

$$S_m - \Upsilon^{(m)}(A_m^*)_{D_m} = \left[\frac{1}{T} \sum_t (\mathcal{X}_{t,D_m} - w_m^{*\top} \mathcal{X}_{t,D_m^c}) \mathcal{X}_{t,D_m}^\top - \Upsilon^{(m)} \right] (A_m^*)_{D_m}$$

$$- \frac{1}{T} \sum_{t=0}^{T-1} \left(\mathcal{X}_{t+1,m} - (A_m^*)_{D_m^c}^\top \mathcal{X}_{t,D_m^c} \right) \left(\mathcal{X}_{t,D_m} - w_m^{*\top} \mathcal{X}_{t,D_m^c} \right)$$

$$= \widetilde{S}_m + W_m^* \left(\frac{1}{T} \sum_t \mathcal{X}_t \mathcal{X}_{t,D_m}^\top - \Upsilon_{\cdot,D_m} \right) (A_m^*)_{D_m},$$

with $\widetilde{S}_m \in \mathbb{R}^{d_m}$ and $W_m^* \in \mathbb{R}^{d_m \times M}$ defined as follows:

$$\widetilde{S}_m = -\frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{X}_{t+1,m} - (A_m^*)_{D_m^c}^\top \mathcal{X}_{t,D_m^c})(\mathcal{X}_{t,D_m} - w_m^{*\top} \mathcal{X}_{t,D_m^c}),$$

$$(W_m^*)_{\cdot,D_m} = I_{d_m \times d_m}, \quad (W_m^*)_{\cdot,D_m^c} = w_m^{*\top}. \tag{48}$$

Therefore,

$$\|E_m\|_2 \leq \left\| \sqrt{T}(\Upsilon^{(m)})^{-\frac{1}{2}}(\widehat{S}_m - \widetilde{S}_m) \right\|_2$$

$$+ \left\| (\Upsilon^{(m)})^{-\frac{1}{2}} W_m^* \left(\frac{1}{T} \sum_t \mathcal{X}_t \mathcal{X}_{t,D_m}^\top - \Upsilon_{\cdot,D_m} \right) \Delta_m \right\|_2$$

$$\leq C\sqrt{T} \left\| \widehat{S}_m - \widetilde{S}_m \right\|_2 + C\sqrt{d_m} \max_i \|(W_m^*)_{i\cdot}\|_1 \left\| \frac{1}{T} \sum_t \mathcal{X}_t \mathcal{X}_t^\top - \Upsilon \right\|_\infty.$$

The last inequality applies (42). Meanwhile,

$$\max_i \|(W_m^*)_{i\cdot}\|_1 = 1 + \max_i \|(w_m^*)_{\cdot i}\|_1$$

$$\leq 1 + \max_i \sqrt{s_m} \|(w_m^*)_{\cdot i}\|_2$$

$$\leq 1 + \sqrt{s_m} \Lambda_{\min}(\Upsilon_{D_m^c, D_m^c})^{-1} \max_i \|\Upsilon_{\cdot i}\|_2 \tag{49}$$

$$\leq 1 + C\sqrt{s_m} \max_i \sqrt{(\Upsilon^2)_{ii}}$$

$$\leq 1 + C\sqrt{s_m} \Lambda_{\max}(\Upsilon) \leq C\sqrt{s_m}.$$

The first equality and second inequality come from the definition of W_m^* and w_m^*; the third inequality is because that $\|\Upsilon_{\cdot i}\|_2^2 = (\Upsilon^2)_{ii}$; the fourth inequality is due to that $(\Upsilon^2)_{ii} = e_i^\top \Upsilon^2 e_i \leq \Lambda_{\max}(\Upsilon)^2$; and the last inequality is obtained from Lemma 5.4. Applying Lemma 5.7 leads us to

$$\|E_m\|_2 \leq C\sqrt{\frac{s_m \log M}{T}} + C\sqrt{T}\|\widehat{S}_m - \widetilde{S}_m\|_2.$$

We can write $\widehat{S}_m - \widetilde{S}_m$ as

$$\widehat{S}_m - \widetilde{S}_m = (\hat{w}_m - w_m^*)^\top \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{X}_{t,D_m^c}\left(\epsilon_{t,m} + T^{-\frac{1}{2}}\Delta_m^\top \mathcal{X}_{t,D_m}\right)$$

$$+ \left(\left(\widehat{A}_m\right)_{D_m^c} - (A_m^*)_{D_m^c}\right)^\top \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{X}_{t,D_m^c}\left(\mathcal{X}_{t,D_m} - w_m^{*\top}\mathcal{X}_{t,D_m^c}\right)^\top$$

$$- \left(\left(\widehat{A}_m\right)_{D_m^c} - (A_m^*)_{D_m^c}\right)^\top \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{X}_{t,D_m^c}\mathcal{X}_{t,D_m^c}^\top(\hat{w}_m - w_m^{*\top}).$$

Note that

$$\left\|\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{X}_{t,D_m^c}\mathcal{X}_{t,D_m}^\top\right\|_\infty \leq \left\|\frac{1}{T}\sum_t \mathcal{X}_t\mathcal{X}_t^\top - \Upsilon\right\|_\infty + \|\Upsilon\|_\infty$$

$$\leq C\sqrt{\frac{\log M}{T}} + \|\Upsilon\|_2 \leq C, \tag{50}$$

due to Lemma 5.4 and 5.7, which further implies

$$\left\|(\hat{w}_m - w_m^*)^\top \frac{1}{T}\sum_{t=0}^{T-1}\mathcal{X}_{t,D_m^c}\mathcal{X}_{t,D_m}^\top \Delta_m\right\|_2 \leq C\|\hat{w}_m - w_m^*\|_1.$$

Applying Assumption 3.1 to 3.3, Lemma 5.1, 5.6, one can show that with probability at least $1 - c_1\exp\{-c_2\log M\}$,

$$\|E_m\|_2 \leq C\frac{(s_m \vee \rho_m)\log M}{\sqrt{T}}, \tag{51}$$

with the same arguments as bounding $\|\widehat{S}_m - S_m\|_2$ under $\mathcal{H}_0$.

While for $\left\|V_{T,m} - (\Upsilon^{(m)})^{\frac{1}{2}}\Delta_m\right\|_2$, applying Lemma 5.3 leads us to

$$\mathbb{P}\left(\left\|V_{T,m} - (\Upsilon^{(m)})^{\frac{1}{2}}\Delta_m\right\|_2 > y\right)$$

$$\leq C_1 T^{-\frac{1}{8}} + 1 - F_{d,\|\widetilde{\Delta}\|_2^2}(y^2)$$

$$= C_1 T^{-\frac{1}{8}} + \mathbb{P}\left(\|Z + \widetilde{\Delta}\|_2^2 > y^2\right)$$

$$\leq C_1 T^{-\frac{1}{8}} + \mathbb{P}\left(\|Z\|_2^2 > (y - C\|\Delta\|_2)^2\right),$$

for any $y \geq 0$, where $Z \sim \mathcal{N}(0, I_d)$. We apply the tail bound for χ_d^2 (Lemma 1 in Laurent and Massart [2000]) as in (45), and obtain

$$\mathbb{P}\left(\|Z\|_2^2 > (y - C\|\Delta\|_2)^2\right) \leq C\left(y - C\|\Delta\|_2\right)^{-2} \leq Cy^{-2},$$

when $y > C$ for some constant C. Let $y = \left(\frac{(s \vee \rho)\log M}{\sqrt{T}}\right)^{-\frac{1}{4}}$, and plug $\left\|V_{T,m} - (\Upsilon^{(m)})^{\frac{1}{2}}\Delta_m\right\|_2 \leq y$, (51) and (19) into (47), one can show that

$$\left|\widehat{U}_T - \left\|V_T - \widetilde{\Delta}\right\|_2^2\right|$$

$$\leq C_1 \left(\frac{(s \vee \rho)\log M}{\sqrt{T}}\right)^{\frac{3}{4}} + C_2 \frac{(s \vee \rho)\log M}{\sqrt{T}} \left(\frac{(s \vee \rho)\log M}{\sqrt{T}}\right)^{-\frac{1}{2}}$$

$$\leq C \left(\frac{(s \vee \rho)\log M}{\sqrt{T}}\right)^{\frac{1}{2}},$$

with probability at least

$$1 - c_1 \exp\{-c_2 \log M\} - c_3 T^{-\frac{1}{8}} - c_4 \left(\frac{(s \vee \rho)\log M}{\sqrt{T}}\right)^{\frac{1}{2}},$$

if $(s \vee \rho)\log M = o(T)$ and $T > C$.

Therefore, applying (46) with $\varepsilon = C\left(\frac{(s \vee \rho)\log M}{\sqrt{T}}\right)^{\frac{1}{2}}$ leads to

$$\left|\mathbb{P}(\widehat{U}_T \leq x) - F_d(x)\right|$$

$$\leq C_1 T^{-\frac{1}{8}} + C_2 \left(\frac{(s \vee \rho)\log M}{\sqrt{T}}\right)^{\frac{1}{2}} + C_3 \exp\{-C_4 \log M\}.$$

Since constants C_i only depend on d, β, Δ, τ, this bound also holds for supremum over $A^* \in \Omega_1$ and $x \in \mathbb{R}$.

(2) $0 < \phi < \frac{1}{2}$

First we provide a lower bound for $\widehat{U}_T$ with high probability. Since bounds in Assumption 3.1 to 3.3, Lemma 5.1 to 5.7 hold with probability at least $1 - c_1 \exp\{-c_2 \log M\}$, we apply these bounds directly in following deduction. Meanwhile, we always assume $(\rho \vee s)\log M =$

$o(\sqrt{T})$ and $T > C$ for desired constant C. With these conditions, one can show that

$$
\begin{aligned}
\widehat{U}_T &= \sum_{m=1}^{k} T\widehat{S}_m^{\top}(\widehat{\Upsilon^{(m)}})^{-1}\widehat{S}_m \\
&\geq \sum_{m=1}^{k} T\|\Upsilon^{(m)-\frac{1}{2}}\widehat{S}_m\|_2^2 \left(1 - d_m \left\|\Upsilon^{(m)\frac{1}{2}}(\widehat{\Upsilon^{(m)}})^{-1}\Upsilon^{(m)\frac{1}{2}} - I\right\|_{\infty}\right) \\
&\geq CT \sum_{m=1}^{k} \left\|(\Upsilon^{(m)})^{-\frac{1}{2}}\widehat{S}_m\right\|_2^2 \\
&\geq C \left(\left(T \sum_{m=1}^{k} \left\|(\Upsilon^{(m)})^{-\frac{1}{2}}(\widehat{S}_m - S_m)\right\|_2^2\right)^{\frac{1}{2}} - \|V_T\|_2\right)^2 .
\end{aligned}
\tag{52}
$$

The third line is due to Assumption 3.3, which implies $\left\|\Upsilon^{(m)\frac{1}{2}}(\widehat{\Upsilon^{(m)}})^{-1}\Upsilon^{(m)\frac{1}{2}} - I\right\|_{\infty}$ converges to 0 under our scaling $(\rho \vee s)\log M = o(\sqrt{T})$.

We provide a lower bound for $\left\|(\Upsilon^{(m)})^{-\frac{1}{2}}(\widehat{S}_m - S_m)\right\|_2^2$ in the following. First write $\widehat{S}_m - S_m$ as

$$
\begin{aligned}
\widehat{S}_m - S_m =& (\hat{w}_m - w_m^*)^{\top} \left(\frac{1}{T}\sum_{t=0}^{T-1} \epsilon_{t,m}\mathcal{X}_{t,D_m^c}\right) \\
&- \frac{1}{T}\sum_{t=0}^{T-1}(\mathcal{X}_{t,D_m} - \hat{w}_m^{\top}\mathcal{X}_{t,D_m^c})\mathcal{X}_{t,D_m}^{\top}(A_m^*)_{D_m} \\
&+ \frac{1}{T}\sum_{t=0}^{T-1}(\mathcal{X}_{t,D_m} - \hat{w}_m^{\top}\mathcal{X}_{t,D_m^c})\mathcal{X}_{t,D_m^c}^{\top}((\widehat{A}_m)_{D_m^c} - (A_m^*)_{D_m^c}) \\
\triangleq& E_m^{(1)} + E_m^{(2)} + E_m^{(3)},
\end{aligned}
$$

we find the upper bounds for $\left\|E_m^{(1)}\right\|_2, \left\|E_m^{(3)}\right\|_2$ and lower bound for $\left\|E_m^{(2)}\right\|_2$ in the following. Applying Assumption 3.2 and Lemma 5.1 provides an upper bound for $\left\|E_m^{(1)}\right\|_2$:

$$
\|E_m^{(1)}\|_2 \leq \|\hat{w}_m - w_m^*\|_1 \left\|\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{X}_t\epsilon_t^{\top}\right\|_{\infty} \leq \frac{s_m \log M}{T}.
$$

Since

$$
\begin{aligned}
\left\|E_m^{(3)}\right\|_2 \leq& \left\|(\hat{w}_m - w_m^*)^{\top}\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{X}_{t,D_m^c}\mathcal{X}_{t,D_m^c}^{\top}((\widehat{A}_m)_{D_m^c} - (A_m^*)_{D_m^c})\right\|_2 \\
&+ \sqrt{d_m}\left\|\frac{1}{T}\sum_{t=0}^{T-1}(\mathcal{X}_{t,D_m} - w_m^{*\top}\mathcal{X}_{t,D_m^c})\mathcal{X}_{t,D_m^c}^{\top}\right\|_{\infty}\left\|(\widehat{A}_m)_{D_m^c} - (A_m^*)_{D_m^c}\right\|_1 ,
\end{aligned}
$$

then using the same argument as bounding $\|\widehat{S}_m - S_m\|_2$ when proving Theorem 3.1, we have

$$
\left\|E_m^{(3)}\right\|_2 \leq C\frac{(s_m \vee \rho_m)\log M}{T}.
$$

To lower bound $\|E_m^{(2)}\|_2$, first note that

$$\left\|\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathcal{X}_{t,D_m}-\hat{w}_m^\top\mathcal{X}_{t,D_m}\right)\mathcal{X}_{t,D_m^c}^\top-\Upsilon^{(m)}\right\|_\infty$$

$$\leq\max_i\|(W_m^*)_{i\cdot}\|_1\left\|\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{X}_t\mathcal{X}_t^\top-\Upsilon\right\|_\infty+\|\hat{w}_m-w_m^*\|_1\left\|\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{X}_t\mathcal{X}_t^\top\right\|_\infty \qquad (53)$$

$$\leq Cs_m\sqrt{\frac{\log M}{T}},$$

where we apply (49), Lemma 5.7, Assumption 3.2, and bound $\left\|\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{X}_t\mathcal{X}_t^\top\right\|_\infty$ using the same argument as in (50). Thus,

$$\|E_m^{(2)}\|_2\geq T^{-\phi}\left\|\Upsilon^{(m)}\Delta_m\right\|_2-Cs_m\sqrt{\frac{\log M}{T}}T^{-\phi}\geq CT^{-\phi},$$

since Δ_m is a constant vector, and $\Lambda_{\min}(\Upsilon^{(m)}$ is lower bounded by constant as in (42).

Applying these bounds for $\|E_m^{(i)}\|_2, 1\leq i\leq 3$, one can show that,

$$T\sum_{m=1}^k\left\|(\Upsilon^{(m)})^{-\frac{1}{2}}(\widehat{S}_m-S_m)\right\|_2^2\geq\sum_{m=1}^k\left(C_1T^{\frac{1}{2}-\phi}-C_2\frac{(s\vee\rho)\log M}{\sqrt{T}}\right)^2\geq CT^{1-2\phi}.$$

Plug this into (52) and apply Lemma 5.3, we have

$$\mathbb{P}(\widehat{U}_T\leq x)\leq C\exp\{-c\log M\}+\mathbb{P}\left(\|V_T\|_2\geq C_1T^{\frac{1}{2}-\phi}-C_2\sqrt{x}\right)$$

$$\leq C_1\exp\{-c\log M\}+C_2T^{-\frac{1}{8}}+1-F_d((C_3T^{\frac{1}{2}-\phi}-C_4\sqrt{x})^2)$$

$$\leq C_1\exp\{-c\log M\}+C_2T^{-\frac{1}{8}}+C_3\exp\{-(C_3T^{\frac{1}{2}-\phi}-C_4\sqrt{x})^2\},$$

where in the last line we apply the χ_d^2 tail bound as in (45). Since the constants here only depend on d,β,Δ,τ, this bound holds when taking supremum over $A^*\in\Omega_1$ and $x\in\mathbb{R}$.

(3) $\phi>\frac{1}{2}$

The proof of this case is similar to that of Theorem 3.1. The only thing different lies in the choice of ε and bounding $\mathbb{P}\left(\left|\widehat{U}_T-U_T\right|>\varepsilon\right)$. The bound (41) for $\left|\widehat{U}_T-U_T\right|$ still holds here, with $E_m=\sqrt{T}(\Upsilon^{(m)})^{-\frac{1}{2}}(\widehat{S}_m-S_m)$. We directly apply the bounds in Assumptions

3.1 to 3.3, and Lemma 5.1 to Lemma 5.7 in the following. First we write

$$
\begin{aligned}
\widehat{S}_m - S_m =& (\hat{w}_m - w_m^*)^\top \frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_{t,D_m^c} \epsilon_{t,m} \\
&+ \frac{1}{T} \sum_{t=0}^{T-1} \left(\mathcal{X}_{t,D_m} - w_m^{*\top} \mathcal{X}_{t,D_m^c} \right) \mathcal{X}_{t,D_m^c}^\top \left((\widehat{A}_m)_{D_m^c} - (A_m^*)_{D_m^c} \right) \\
&- (\hat{w}_m - w_m^*)^\top \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_{t,D_m^c} \mathcal{X}_{t,D_m^c}^\top \right) \left((\widehat{A}_m)_{D_m^c} - (A_m^*)_{D_m^c} \right) \\
&- T^{-(1+\phi)} \sum_{t=0}^{T-1} (\mathcal{X}_{t,D_m} - \hat{w}_m^\top \mathcal{X}_{t,D_m^c}) \mathcal{X}_{t,D_m}^\top \Delta_m.
\end{aligned}
$$

Note here that the first three terms are exactly the same as in (43), and thus can be bounded as in the proof of Theorem 3.1. We only have to tackle the last term. By (53), one can show that,

$$
\left\| \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{X}_{t,D_m} - \hat{w}_m^\top \mathcal{X}_{t,D_m^c}) \mathcal{X}_{t,D_m}^\top \Delta_m \right\|_2 \leq \left\| \Upsilon^{(m)} \Delta_m \right\|_2 + C s_m \sqrt{\frac{\log M}{T}} \leq C,
$$

Thus, going through the same arguments as bounding $\left\| \widehat{S}_m - S_m \right\|_2$ under $\mathcal{H}_0$, we have

$$
\| E_m \|_2 \leq C_1 \frac{(s \vee \rho) \log M}{\sqrt{T}} + C_2 T^{\frac{1}{2} - \phi},
$$

with probability at least $1 - C \exp\{-c \log M\}$. Recall that in (45), when $y > C$ for some constant C,

$$
\mathbb{P}(\|V_{T,m}\|_2 \geq y) \leq C_1 T^{-\frac{1}{8}} + C_2 y^{-2}.
$$

Let $y = \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{-\frac{1}{4}} \wedge T^{\frac{2\phi - 1}{6}}$, then by (41) one can show that

$$
\begin{aligned}
&\left| \widehat{U}_T - U_T \right| \\
\leq& C_1 \frac{(s \vee \rho) \log M}{\sqrt{T}} \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{-\frac{1}{2}} + C_2 \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{\frac{3}{4}} + C_3 T^{\frac{1-2\phi}{3}} \\
\leq& C_1 \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{\frac{1}{2}} + C_2 T^{\frac{1-2\phi}{3}},
\end{aligned}
$$

with probability at least

$$
1 - c_1 \exp\{-c_2 \log M\} - c_3 T^{-\frac{1}{8}} - c_4 \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{\frac{1}{2}} - c_5 T^{\frac{1-2\phi}{3}},
$$

if $(s \vee \rho) \log M = o(\sqrt{T})$ and $T > C$ for some constant C. Therefore, applying (38) with $\varepsilon = C_1 \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{\frac{1}{2}} + C_2 T^{\frac{1-2\phi}{3}}$,

$$
\left| \mathbb{P}(\widehat{U}_T \leq x) - F_d(x) \right|
$$

34

$$\leq C_1 T^{-\frac{1}{8}} + C_2 \left(\frac{(s \vee \rho) \log M}{\sqrt{T}} \right)^{\frac{1}{2}} + C_3 T^{\frac{1-2\phi}{3}} + C_4 \exp\{-C_5 \log M\}.$$

Since constants C_i only depend on d, β, τ, Δ, this bound also holds for supremum over $A^* \in \Omega_1$ and $x \in \mathbb{R}$.

$\square$

6 Conclusion

In this paper, we have provided theoretical guarantees for hypothesis tests for sparse high-dimensional auto-regressive models with sub-Gaussian innovations. Specific upper bounds for the convergence rates of test statistics are given. Importantly, our results go beyond the Gaussian assumption and do not rely on mixing assumptions. As a consequence of our theory, we also develop novel concentration bounds for quadratic forms of dependent sub-Gaussian random variables using a careful truncation argument.

It would be of interest to consider other variance estimation method, e.g., scaled Lasso Sun and Zhang [2012], or cross-validation based method Fan et al. [2012], and establish corresponding theoretical guarantee. There also remain a number of open questions/challenges including extensions to generalized linear models, heavy-tailed innovations and incorporating hidden variables under time series setting.

Acknowledgements

References

A. Ang and M. Piazzesi. A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. *Journal of Monetary economics*, 50(4):745–787, 2003.

M. Barigozzi and C. T. Brownlees. Nets: Network estimation for time series. 2018.

S. Basu, G. Michailidis, et al. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

S. L. Bressler, C. G. Richter, Y. Chen, and M. Ding. Cortical functional network organization from autoregressive modeling of local field potential oscillations. *Statistics in medicine*, 26 (21):3875–3885, 2007.

L. Chen and W. B. Wu. Testing for trends in high-dimensional time series. *Journal of the American Statistical Association*, (just-accepted):1–37, 2018.

S. X. Chen, L.-X. Zhang, and P.-S. Zhong. Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association*, 105(490):810–819, 2010.

R. A. Davis, P. Zang, and T. Zheng. Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096, 2016.

J. Fan, S. Guo, and N. Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):37–65, 2012.

A. Fujita, J. R. Sato, H. M. Garay-Malpartida, R. Yamaguchi, S. Miyano, M. C. Sogayar, and C. E. Ferreira. Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC systems biology*, 1(1):39, 2007.

R. Goebel, A. Roebroeck, D.-S. Kim, and E. Formisano. Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping. *Magnetic resonance imaging*, 21(10):1251–1261, 2003.

I. Grama and E. Haeusler. An asymptotic expansion for probabilities of moderate deviations for multivariate martingales. *Journal of Theoretical Probability*, 19(1):1–44, 2006.

R. M. Gray et al. Toeplitz and circulant matrices: A review. *Foundations and Trends® in Communications and Information Theory*, 2(3):155–239, 2006.

F. Han, H. Lu, and H. Liu. A direct estimation of high dimensional stationary vector autoregressions. *The Journal of Machine Learning Research*, 16(1):3115–3150, 2015.

P. R. Hansen. Structural changes in the cointegrated vector autoregressive model. *Journal of Econometrics*, 114(2):261–295, 2003.

M. Hardy. Combinatorics of partial derivatives. *the electronic journal of combinatorics*, 13(1): 1, 2006.

L. Harrison, W. D. Penny, and K. Friston. Multivariate autoregressive modeling of fmri time series. *Neuroimage*, 19(4):1477–1491, 2003.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

J. D. Lee, D. L. Sun, Y. Sun, J. E. Taylor, et al. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

R. G. Mark B. and W. R. Network estimation from point process data. *IEEE Trans. of Info. Theory*, 2018. To appear.

M. C. Medeiros and E. F. Mendes. 1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics*, 191(1):255–271, 2016.

G. Michailidis and F. dAlché Buc. Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Mathematical biosciences*, 246(2):326–334, 2013.

M. Neykov, Y. Ning, J. S. Liu, H. Liu, et al. A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statistical Science*, 33(3):427–443, 2018.

Y. Ning, H. Liu, et al. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

M. Rudelson, R. Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

J. R. Schott. Testing for complete independence in high dimensions. *Biometrika*, 92(4):951–956, 2005.

A. K. Seth, A. B. Barrett, and L. Barnett. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297, 2015.

J. Shan. Does financial development lead'economic growth? a vector auto-regression appraisal. *Applied Economics*, 37(12):1353–1367, 2005.

S. Song and P. J. Bickel. Large vector auto regressions. *arXiv preprint arXiv:1106.3915*, 2011.

M. S. Srivastava. A test for the mean vector with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis*, 100(3):518–532, 2009.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

J. Taylor, R. Lockhart, R. J. Tibshirani, and R. Tibshirani. Post-selection adaptive inference for least angle regression and the lasso. *arXiv preprint*, 2014.

S. Van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

A. Voorman, A. Shojaie, and D. Witten. Inference in high dimensions with the penalized score test. *arXiv preprint arXiv:1401.2678*, 2014.

K. C. Wong, Z. Li, and A. Tewari. Lasso guarantees for time series estimation under subgaussian tails and β-mixing. *arXiv preprint arXiv:1602.04265*, 2016.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

R. Zhang, L. Peng, R. Wang, et al. Tests for covariance matrix with fixed or divergent dimension. *The Annals of Statistics*, 41(4):2075–2096, 2013.

A   Proof of Lemmas in Section 3.3

Proof of Lemma 3.1. We prove the error bounds for each $\widehat{A}_m$ and then take a union bound. Without loss of generality, we consider the estimation of $A_1^* \in \mathbb{R}^M$. With a little abuse of notation, let $S = \mathrm{supp}(A_1^*)$, $\hat{h} = \widehat{A}_1 - A_1^*$, $S = \mathrm{supp}(A_1^*)$, and $H = \frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_t \mathcal{X}_t^\top$ (S is not the decorrelated score function we defined in section 9). We would like to bound $\|\hat{h}\|_1$, $\|\hat{h}\|_2$ and $\hat{h}^\top H \hat{h}$ under two cases separately:

(1) $\widehat{A} = \widehat{A}^{(L)}$.

   Here we adopt the standard proof framework for Lasso. By (25) we know that $\widehat{A}_1 \in \mathbb{R}^M$ satisfies

$$\widehat{A}_1 = \arg\min_{\beta \in \mathbb{R}^M} \frac{1}{T} \sum_{t=0}^{T-1} (X_{t+1,1} - \mathcal{X}_t^\top \beta)^2 + \lambda_A \|\beta\|_1,$$

   which implies

$$\frac{1}{T} \sum_{t=0}^{T-1} (X_{t+1,1} - \mathcal{X}_t^\top \widehat{A}_1)^2 + \lambda_A \|\widehat{A}_1\|_1 \le \frac{1}{T} \sum_{t=0}^{T-1} (X_{t+1,1} - \mathcal{X}_t^\top A_1^*)^2 + \lambda_A \|A_1^*\|_1.$$

Rearranging the terms, we have

$$\hat{h}^\top H \hat{h} \le 2\hat{h}^\top \left(\frac{1}{T}\sum_{t=0}^{T-1}\epsilon_{t,1}\mathcal{X}_t\right) + \lambda_A\|A_1^*\|_1 - \lambda_A\|\widehat{A}_1\|_1$$

$$\le 2\left\|\frac{1}{T}\sum_{t=0}^{T-1}\epsilon_t\mathcal{X}_t^\top\right\|_\infty \|\hat{h}\|_1 + \lambda_A\|\hat{h}_S\|_1 - \lambda_A\|\hat{h}_{S^c}\|_1.$$

The last line is due to that

$$\|A_1^*\|_1 - \|\widehat{A}_1\|_1 = \|(A_1^*)_S\|_1 - \|(\widehat{A}_1)_S\|_1 - \|(\widehat{A}_1)_{S^c}\|_1$$
$$= \|(A_1^*)_S\|_1 - \|(\widehat{A}_1)_S\|_1 - \|\hat{h}_{S^c}\|_1$$
$$\le \|\hat{h}_S\|_1 - \|\hat{h}_{S^c}\|_1.$$

By Lemma 5.1, with probability at least $1 - c_1\exp\{-c_2\log M\}$,

$$\left\|\frac{1}{T}\sum_{t=0}^{T-1}\epsilon_t\mathcal{X}_t^\top\right\|_\infty \le \frac{1}{4}\lambda_A = C\sqrt{\frac{\log M}{T}}.$$

Meanwhile, since H is positive semi-definite,

$$0 \le \hat{h}^\top H \hat{h} \le \frac{3\lambda_A}{2}\|\hat{h}_S\|_1 - \frac{\lambda_A}{2}\|\hat{h}_{S^c}\|_1,$$
$$\|\hat{h}_{S^c}\|_1 \le 3\|\hat{h}_S\|_1.$$

We have the following restricted eigenvalue condition for H.

Lemma A.1. *Under the model specified in (3) with independent sub-Gaussian noise ϵ_{ti} of constant scale factor, and $A^* \in \Omega_0 \cup \Omega_1$, for any set $J \subset \{1, 2, \cdots, pM\}$, positive integer $\kappa > 0$, H satisfies the following REC:*

$$\inf\{v^\top H v : v \in \mathcal{C}(J, \kappa), \|v\|_2 \le 1\} \ge C_1 > 0,$$

with probability at least $1 - 2\exp\{-cT\}$, when $|J|\log pM \le C_2 T$. Here $\mathcal{C}(J, \kappa) = \{v : \|v_{J^c}\|_1 \le \kappa\|v_J\|_1\}$, constant C_1 depends on β, c and C_2 depend on κ and β.

Here $\hat{h} \in \mathcal{C}(S, 3)$, $|S| = \rho_1$, by Lemma A.1, when $T \ge C\rho\log M$,

$$\hat{h}^\top H \hat{h} \ge C\|\hat{h}\|_2^2,$$

with probability at least $1 - 2\exp\{-cT\}$, when $T > C\rho\log M$. Thus

$$\|\hat{h}\|_2^2 \le C\hat{h}^\top H \hat{h} \le C\lambda_A\|\hat{h}_S\|_1 \le C\sqrt{\frac{\rho_1\log M}{T}}\|\hat{h}\|_2, \tag{54}$$

which implies

$$\|\hat{h}\|_2 \le C\sqrt{\frac{\rho_1\log M}{T}}, \quad \hat{h}^\top H \hat{h} \le C\frac{\rho_1\log M}{T},$$

$$\|\hat{h}\|_1 \le 4\|\hat{h}_S\|_1 \le 4\sqrt{\rho_1}\|\hat{h}\|_2 \le C\rho_1\sqrt{\frac{\log M}{T}},$$

with probability at least $1 - c_1\exp\{-c_2\log M\}$.

(2) $\widehat{A} = \widehat{A}^{(D)}$.

Here we adopt the standard proof framework for Dantzig selector. By (26),

$$\widehat{A}_1 = \underset{\beta \in \mathbb{R}^M}{\arg\min} \|\beta\|_1, \quad \text{s.t.} \quad \left\| \frac{1}{T} \sum_{t=0}^{T-1} (X_{t+1,1} - \mathcal{X}_t^\top \beta) \mathcal{X}_t \right\|_\infty \leq \lambda_A. \tag{55}$$

By Lemma 5.1, when $T \geq C \log M$, with probability at least $1 - c_1 \exp\{-c_2 \log M\}$,

$$\left\| \frac{1}{T} \sum_{t=0}^{T-1} (X_{t+1,1} - \mathcal{X}_t^\top A_1^*) \mathcal{X}_t \right\|_\infty = \left\| \frac{1}{T} \sum_{t=0}^{T-1} \epsilon_{t,1} \mathcal{X}_t \right\|_\infty \leq \lambda_A,$$

which implies

$$\|H\hat{h}\|_\infty \leq C \sqrt{\frac{\log M}{T}}.$$

Meanwhile, by (55),

$$\|\widehat{A}_1\|_1 \leq \|A_1^*\|_1, \quad \|\hat{h}_{S^c}\|_1 \leq \|\hat{h}_S\|_1.$$

Here $\hat{h} \in \mathcal{C}(S, 1)$, $|S| = \rho_1$, by Lemma A.1, when $T \geq C\rho \log M$,

$$\hat{h}^\top H \hat{h} \geq C \|\hat{h}\|_2^2,$$

with probability at least $1 - 2\exp\{-cT\}$, when $T > C\rho \log M$. Thus

$$\|\hat{h}\|_2^2 \leq C \hat{h}^\top H \hat{h} \leq \|H\hat{h}\|_\infty \|\hat{h}\|_1 \leq C \sqrt{\frac{\log M}{T}} \|\hat{h}\|_1 \leq C \sqrt{\frac{\rho_1 \log M}{T}} \|\hat{h}\|_2, \tag{56}$$

which implies

$$\|\hat{h}\|_2 \leq C \sqrt{\frac{\rho_1 \log M}{T}}, \quad \hat{h}^\top H \hat{h} \leq C \frac{\rho_1 \log M}{T},$$

$$\|\hat{h}\|_1 \leq 4\|\hat{h}_S\|_1 \leq 4\sqrt{\rho_1} \|\hat{h}\|_2 \leq C \rho_1 \sqrt{\frac{\log M}{T}},$$

with probability at least $1 - c_1 \exp\{-c_2 \log M\}$.

Therefore, after taking a union bound over $m = 1, \cdots, k$, proof complete. $\qquad \square$

Proof of Lemma 3.2. Without loss of generality, we consider the estimation of $(w_1^*)_{\cdot,1}$ and then take a union bound. Let $v^* = (w_1^*)_{\cdot,1}$, $\hat{v} = (\hat{w}_1)_{\cdot,1}$, $\hat{h} = \hat{v} - v^* \in \mathbb{R}^{M-d_1}$ and $S = \text{supp}(v^*)$. Then we prove upper bounds for $\|\hat{h}\|_1$ and $\hat{h}^\top H_{D_1^c, D_1^c} \hat{h}$ with high probability under two cases.

(1) $\hat{w}_m = \hat{w}_m^{(L)}$.

Looking into the definition (27) of $\hat{w}_1$, it is clear that the optimization can be viewed as d_1 separate optimization problems, in terms of each column of $\hat{w}_1$. Thus

$$\hat{v} = \arg \min_{v \in \mathbb{R}^{M-d_1}} \frac{1}{T} \sum_{t=0}^{T-1} \left((\mathcal{X}_{t,D_1})_1 - \mathcal{X}_{t,D_1^c}^\top v \right)^2 + \lambda_w \|v\|_1.$$

The following proof is almost identical to the proof in Lemma 3.1 under $\widehat{A} = \widehat{A}^{(L)}$, except some difference in notation and application of Lemmas. One can show that,

$$\frac{1}{T}\sum_{t=0}^{T-1}\left((\mathcal{X}_{t,D_1})_1 - \mathcal{X}_{t,D_1^c}^\top \hat{v}\right)^2 + \lambda_w\|\hat{v}\|_1$$

$$\leq \frac{1}{T}\sum_{t=0}^{T-1}\left((\mathcal{X}_{t,D_1})_1 - \mathcal{X}_{t,D_1^c}^\top v^*\right)^2 + \lambda_w\|v^*\|_1,$$

Rearranging the inequality gives us

$$\hat{h}^\top H_{D_1^c,D_1^c}\hat{h} \leq 2\hat{h}^\top \left(\frac{1}{T}\sum_{t=0}^{T-1}\left((\mathcal{X}_{t,D_1})_1 - \mathcal{X}_{t,D_1^c}v^*\right)\mathcal{X}_{D_1^c}\right) + \lambda_w\|v^*\|_1 - \lambda_w\|\hat{v}\|_1$$

$$\leq 2\left\|\frac{1}{T}\sum_{t=0}^{T-1}(\mathcal{X}_{t,D_1} - w_1^{*\top}\mathcal{X}_{t,D_1^c})\mathcal{X}_{D_1^c}^\top\right\|_\infty \|\hat{h}\|_1 + \lambda_w\|\hat{h}_S\|_1 - \lambda_A\|\hat{h}_{S^c}\|_1.$$

By Lemma 5.6, with probability at least $1 - c_1\exp\{-c_2\log M\}$,

$$\left\|\frac{1}{T}\sum_{t=0}^{T-1}(\mathcal{X}_{t,D_1} - w_1^{*\top}\mathcal{X}_{t,D_1^c})\mathcal{X}_{D_1^c}^\top\right\|_\infty \leq \frac{1}{4}\lambda_w = C\sqrt{\frac{\log M}{T}},$$

which implies,

$$0 \leq \hat{h}^\top H_{D_1^c,D_1^c}\hat{h} \leq \frac{3\lambda_w}{2}\|\hat{h}_S\|_1 - \frac{\lambda_w}{2}\|\hat{h}_{S^c}\|_1,$$

$$\|\hat{h}_{S^c}\|_1 \leq 3\|\hat{h}_S\|_1.$$

Let $\tilde{h} \in \mathbb{R}^M$ be defined as the following:

$$\tilde{h}_{D_1} = 0, \quad \tilde{h}_{D_1^c} = \hat{h}, \tag{57}$$

By Lemma A.1, when $T \geq Cs\log M$, with probability at least $1 - 2\exp\{-cT\}$,

$$\|\hat{h}\|_2^2 = \|\tilde{h}\|_2^2 \leq C\tilde{h}^\top H\tilde{h} = 2\hat{h}^\top H_{D_1^c,D_1^c}\hat{h} \leq C\lambda_w\|\hat{h}_S\|_1 \leq C\sqrt{\frac{s_1\log M}{T}}\|\hat{h}\|_2,$$

which implies

$$\hat{h}^\top H\hat{h} \leq C\frac{s_1\log M}{T},$$

and

$$\|\hat{h}\|_1 \leq 4\|\hat{h}_S\|_1 \leq 4\sqrt{s_1}\|\hat{h}\|_2 \leq Cs_1\sqrt{\frac{\log M}{T}},$$

with probability at least $1 - c_1\exp\{-c_2\log M\}$.

41

(2) $\hat{w}_m = \hat{w}_m^{(D)}$.

By (28),

$$\hat{v} = \underset{v \in \mathbb{R}^{M-d_1}}{\arg\min} \|v\|_1, \quad \text{s.t.} \quad \left\| \frac{1}{T} \sum_{t=0}^{T-1} \left((\mathcal{X}_{t,D_1})_1 - v^\top \mathcal{X}_{t,D_1^c} \right) \mathcal{X}_{t,D_1^c} \right\|_\infty \le \lambda_w. \tag{58}$$

This proof is also pretty similar to the proof of Lemma 3.1 under the case where $\widehat{A} = \widehat{A}^{(D)}$.

By Lemma 5.6,

$$\left\| \frac{1}{T} \sum_{t=0}^{T-1} \left((\mathcal{X}_{t,D_1})_1 - v^{*\top} \mathcal{X}_{t,D_1^c} \right)_1 \mathcal{X}_{t,D_1^c} \right\|_\infty \le \lambda_w = C\sqrt{\frac{\log M}{T}},$$

with probability at least $1 - c_1 \exp\{-c_2 \log M\}$. Thus,

$$\left\| H_{D_1^c,D_1^c}^\top \hat{h} \right\|_\infty \le C\sqrt{\frac{\log M}{T}}.$$

Meanwhile, by (58),

$$\|\hat{v}\|_1 = \|\hat{v}_S\|_1 + \|\hat{v}_{S^c}\|_1 \le \|v^*\|_1 = \|v_S^*\|_1,$$

which further implies

$$\left\| \hat{h}_{S^c} \right\|_1 \le \left\| \hat{h}_S \right\|_1. \tag{59}$$

Recall the definition of $\tilde{h}$ in (57), then by Lemma A.1, (59) and (57), when $T \ge Cs \log M$,

$$\begin{aligned}
\|\hat{h}\|_2^2 = \|\tilde{h}\|_2^2 &\le C\tilde{h}^\top H \tilde{h} \\
&= C\hat{h}^\top H_{D_1^c,D_1^c} \hat{h} \\
&\le C \left\| \hat{h} \right\|_1 \left\| H_{D_1^c,D_1^c}^\top \hat{h} \right\|_\infty \\
&\le C\sqrt{\frac{\log M}{T}} \|\hat{h}_S\|_1 \\
&\le C\sqrt{\frac{s_1 \log M}{T}} \|\hat{h}\|_2,
\end{aligned}$$

which implies

$$\hat{h}^\top H_{D_1^c,D_1^c} \hat{h} \le C\frac{s_1 \log M}{T},$$

and

$$\|\hat{h}\|_1 \le C\sqrt{s_1}\|\hat{h}\|_2 \le Cs_1\sqrt{\frac{\log M}{T}},$$

with probability at least $1 - c_1 \exp\{-c_2 \log M\}$.

Since

$$\|(\hat{w}_1) - (w_1^*)\|_1 = \sum_{j=1}^{d_1} \|(\hat{w}_1)_{\cdot,j} - (w_1^*)_{\cdot,j}\|_1,$$

and

$$\operatorname{tr}\left\{ (\hat{w}_1 - w_1^*)^\top \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_{t,D_m^c} \mathcal{X}_{t,D_m^c}^\top \right) (\hat{w}_1 - w_1^*) \right\}$$

$$= \sum_{j=1}^{d_1} ((\hat{w}_1)_{\cdot,j} - (w_1^*)_{\cdot,j})^\top \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_{t,D_1^c} \mathcal{X}_{t,D_1^c}^\top \right) ((\hat{w}_1)_{\cdot,j} - (w_1^*)_{\cdot,j}),$$

taking a union bound over $\{\hat{w}_m : m = 1, \cdots, k\}$ and all columns of $\hat{w}_m$, proof is complete. $\square$

Proof of Lemma 3.3. The following established result can be applied here:

Lemma A.2. *For any invertible matrix B, if $B + \Delta$ is also invertible, then*

$$\|(B + \Delta)^{-1} - B^{-1}\|_2 \le \frac{\|B^{-1}\|_2^2 \|\Delta\|_2}{1 - \|B^{-1}\|_2 \|\Delta\|_2}. \tag{60}$$

Since $\|I\|_2 = 1$, one can show that for $1 \le m \le k$,

$$\left\| \Upsilon^{(m)\frac{1}{2}} \widehat{\Upsilon^{(m)}}^{-1} \Upsilon^{(m)\frac{1}{2}} - I \right\|_\infty \le \left\| \Upsilon^{(m)\frac{1}{2}} \widehat{\Upsilon^{(m)}})^{-1} \Upsilon^{(m)\frac{1}{2}} - I \right\|_2 \le \frac{\|\Delta\|_2}{1 - \|\Delta\|_2},$$

where $\Delta = \Upsilon^{(m)-\frac{1}{2}} \widehat{\Upsilon^{(m)}} \Upsilon^{(m)-\frac{1}{2}} - I$. Due to (42),

$$\|\Delta\|_2 \le \left(\Lambda_{\min} \Upsilon^{(m)} \right)^{-1} \left\| \widehat{\Upsilon^{(m)}} - \Upsilon^{(m)} \right\|_2$$

$$\le C \left\| \widehat{\Upsilon^{(m)}} - \Upsilon^{(m)} \right\|_F \le d_m \left\| \widehat{\Upsilon^{(m)}} - \Upsilon^{(m)} \right\|_\infty.$$

In the following we bound $\left\| \widehat{\Upsilon^{(m)}} - \Upsilon^{(m)} \right\|_\infty$. Write $\widehat{\Upsilon^{(m)}} - \Upsilon^{(m)}$ as

$$\widehat{\Upsilon^{(m)}} - \Upsilon^{(m)} = W_m^* \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_t \mathcal{X}_t^\top - \Upsilon \right) W_m^{*\top}$$

$$- (\hat{w}_m - w_m^*)^\top \frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_{t,D_m^c} (\mathcal{X}_{t,D_m} - w_m^{*\top} \mathcal{X}_{t,D_m^c})^\top$$

$$- \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{X}_{t,D_m} - w_m^{*\top} \mathcal{X}_{t,D_m^c}) \mathcal{X}_{t,D_m^c}^\top (\hat{w}_m - w_m^*)$$

$$+ (\hat{w}_m - w_m^*)^\top \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_{t,D_m^c} \mathcal{X}_{t,D_m^c}^\top \right) (\hat{w}_m - w_m^*)$$

$$\triangleq E_1^{(m)} - E_2^{(m)} - \left(E_2^{(m)} \right)^\top + E_3^{(m)},$$

43

where W_m^* is defined as in (48). Actually,

$$\|E_1^{(m)}\|_\infty = \max_{i,j} \left| W_{m,i\cdot}^* \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_t \mathcal{X}_t^\top - \Upsilon \right) W_{m,j\cdot}^{*\top} \right|$$

$$= \max_{i,j} \left| \frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_t^\top W_{m,i\cdot}^{*\top} W_{m,j\cdot}^* \mathcal{X}_t - \mathrm{tr}(W_{m,i\cdot}^{*\top} W_{m,j\cdot}^* \Upsilon) \right|,$$

which is the maximum over deviations of some quadratic forms from their expectation. The following lemma provides a bound for quadratic form $\frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_t^\top B \mathcal{X}_t$, with $B \in \mathbb{R}^{M \times M}$ being any symmetric matrix.

By Lemma 5.2, we only need to bound the trace norm and operator norm of

$$\frac{1}{2} \left((W_m^*)_{i\cdot}^\top (W_m^*)_{j\cdot} + (W_m^*)_{j\cdot}^\top (W_m^*)_{i\cdot} \right).$$

The following lemma establishes the relationship between $\|\cdot\|_{\mathrm{tr}}$ and $\|\cdot\|_2$ for symmetric matrices.

Lemma A.3. *For any symmetric matrix U of rank r, $\|U\|_{\mathrm{tr}} \leq r\|U\|_2$.*

Since $\frac{1}{2} \left((W_m^*)_{i\cdot}^\top (W_m^*)_{j\cdot} + (W_m^*)_{j\cdot}^\top (W_m^*)_{i\cdot} \right)$ is of rank 2,

$$\left\| \frac{1}{2} \left((W_m^*)_{i\cdot}^\top (W_m^*)_{j\cdot} + (W_m^*)_{j\cdot}^\top (W_m^*)_{i\cdot} \right) \right\|_{\mathrm{tr}}$$
$$\leq 2 \left\| \frac{1}{2} \left(((W_m^*)_{i\cdot})^\top (W_m^*)_{j\cdot} + (W_m^*)_{j\cdot}^\top (W_m^*)_{i\cdot} \right) \right\|_2 \qquad (61)$$
$$\leq 2 \left\| (W_m^*)_{i\cdot}^\top (W_m^*)_{j\cdot} \right\|_2 = 2\|(W_m^*)_{i\cdot}\|_2 \|(W_m^*)_{j\cdot}\|_2.$$

Meanwhile, similar from (49), we bound $\max_i \|(W_m^*)_{i\cdot}\|_2^2$ by

$$\|(W_m^*)_{i\cdot}\|_2^2 = 1 + \|(w_m^*)_{\cdot,i}\|_2^2$$
$$\leq 1 + \Lambda_{\max}(\Upsilon_{D_m^c, D_m^c}^{-1})^2 \|\Upsilon_{\cdot,i}\|_2^2 \qquad (62)$$
$$\leq 1 + \Lambda_{\min}(\Upsilon)^{-2} \Lambda_{\max}(\Upsilon)^2 \leq C,$$

where the second inequality is due to that $\|\Upsilon_{\cdot,i}\|_2^2 = (\Upsilon^2)_{ii} \leq \Lambda_{\max}(\Upsilon^2) \leq \Lambda_{\max}(\Upsilon)^2$. Thus, both the trace norm and ℓ_2 norm of $\frac{1}{2} \left(W_{m,i\cdot}^{*\top} W_{m,j\cdot}^* + W_{m,j\cdot}^{*\top} W_{m,i\cdot}^* \right)$ can be bounded by constant, and applying Lemma 5.2 gives us

$$\mathbb{P} \left(\|E_1^{(m)}\|_\infty > C \sqrt{\frac{\log M}{T}} \right) \leq c_1 \exp\{-c_2 \log M\}.$$

Meanwhile, by Lemma 5.6 and Assumption 3.2, with probability at least $1 - c_1 \exp\{-c_2 \log M\}$,

$$\|E_2^{(m)}\|_\infty \leq \left\| \frac{1}{T} \sum_{t=0}^{T-1} (\mathcal{X}_{t,D_m} - w_m^{*\top} \mathcal{X}_{t,D_m^c}) \mathcal{X}_{t,D_m^c} \right\|_\infty \|\hat{w}_m - w_m^*\|_1$$

$$\leq C\frac{s_m \log M}{T},$$

and

$$\begin{aligned}
\left\|E_3^{(m)}\right\|_\infty &= \max_{i,j}\left|(E_3^{(m)})_{ij}\right| = \max_{i,j}(\hat{w}_m - w_m^*)_{\cdot,i}^\top H_{D_m^c,D_m^c}(\hat{w}_m - w_m^*)_{\cdot,j} \\
&\leq \max_l (\hat{w}_m - w_m^*)_{\cdot,l}^\top H_{D_m^c,D_m^c}(\hat{w}_m - w_m^*)_{\cdot,l} \\
&\leq \mathrm{tr}\left\{(\hat{w}_m - w_m^*)^\top H_{D_m^c,D_m^c}(\hat{w}_m - w_m^*)\right\} \\
&\leq C\frac{s_m \log M}{T}.
\end{aligned}$$

Here the second line is because that $H_{D_m^c,D_m^c} = \frac{1}{T}\sum_{t=0}^{T-1}\mathcal{X}_{t,D_m^c}\mathcal{X}_{t,D_m^c}^\top$ is symmetric and positive semi-definite, thus we can apply Cauchey-Schwartz inequality. When $T \geq Cs^2 \log M$.

$$\frac{s_m \log M}{T} \leq \sqrt{\frac{\log M}{T}},$$

which implies

$$\left\|\widehat{\Upsilon^{(m)}} - \Upsilon^{(m)}\right\|_2 \leq \|\widehat{\Upsilon^{(m)}} - \Upsilon^{(m)}\|_F \leq d_m\|\widehat{\Upsilon^{(m)}} - \Upsilon^{(m)}\|_\infty \leq C\sqrt{\frac{\log M}{T}}.$$

Therefore, take a union bound over $1 \leq m \leq k$, with probability at least $1 - c_1 \exp\{-c_2 \log M\}$,

$$\left\|\Upsilon^{(m)\frac{1}{2}}\widehat{\Upsilon^{(m)}}^{-1}\Upsilon^{(m)\frac{1}{2}} - I\right\|_\infty \leq C\sqrt{\frac{\log M}{T}}.$$

when $T \geq Cs^2 \log M$. $\qquad\square$

B   Proof of Theorem 3.3 and Theorem 3.4

Proof of Theorem 3.3. Now we consider model (3), with unknown $\sigma^{*2} = \mathrm{Var}(\epsilon_{ti}) \geq \sigma_0^2$. Under this model, we use the notation $\widehat{U}_T$ for the quantity defined in the following:

$$\widehat{U}_T = T\sum_{m=1}^k \widehat{S}_m^\top(\widehat{\Upsilon^{(m)}})^{-1}\widehat{S}_m/\sigma^{*2}.$$

As explained in Section 3.4, $\widehat{U}_T$ satisfies Theorem 3.1 and 3.2 under each corresponding condition. We show in the following that we only need to control the estimation error of $\hat{\sigma}^2$. Note that for any $0 < \delta < 1$,

$$\mathbb{P}\left(\widetilde{U}_T \leq x\right) \leq \mathbb{P}\left(\widehat{U}_T \leq \frac{x}{1-\delta}\right) + \mathbb{P}\left(\frac{\sigma^{*2}}{\hat{\sigma}^2} < 1 - \delta\right),$$

and

$$\mathbb{P}\left(\widetilde{U}_T > x\right) \leq \mathbb{P}\left(\widehat{U}_T > \frac{x}{1+\delta}\right) + \mathbb{P}\left(\frac{\sigma^{*2}}{\hat{\sigma}^2} > 1 + \delta\right).$$

For any distribution function $F(x)$,

$$\left| \mathbb{P}\left(\widetilde{U}_T \leq x \right) - F(x) \right| \leq \sup_y \left| \mathbb{P}\left(\widehat{U}_T \leq y \right) - F(y) \right| + \sup_y \left| F(y) - F(y(1-\delta)) \right|$$
$$+ \mathbb{P}\left(\hat{\sigma}^2 < \frac{\sigma^{*2}}{1+\delta} \right) + \mathbb{P}\left(\hat{\sigma}^2 > \frac{\sigma^{*2}}{1-\delta} \right).$$

Recall that Theorem 3.1 and 3.2 establish bounds for $\mathbb{P}\left(\widehat{U}_T \leq x \right) - F_d(x)$ under $\mathcal{H}_0$, or under $\mathcal{H}_A$ with $\phi > \frac{1}{2}$, for $\mathbb{P}\left(\widehat{U}_T \leq x \right) - F_{d,\|\widetilde{\Delta}\|_2^2}(x)$ when $\phi = \frac{1}{2}$, and for $\mathbb{P}\left(\widehat{U}_T \leq x \right)$ when $0 < \phi < \frac{1}{2}$. Thus we only need to bound $\mathbb{P}\left(\hat{\sigma}^2 < \frac{\sigma^{*2}}{1+\delta} \right)$, $\mathbb{P}\left(\hat{\sigma}^2 > \frac{\sigma^{*2}}{1-\delta} \right)$ and $\sup_y |F(y) - F(y(1-\delta))|$ with $F(x) = F_d(x)$ or $F(x) = F_{d,\|\widetilde{\Delta}\|_2^2}(x)$. Since $0 < \delta < 1$,

$$\mathbb{P}\left(\hat{\sigma}^2 < \frac{\sigma^{*2}}{1+\delta} \right) + \mathbb{P}\left(\hat{\sigma}^2 > \frac{\sigma^{*2}}{1-\delta} \right) \leq \mathbb{P}\left(|\hat{\sigma}^2 - \sigma^{*2}| > \frac{\delta \sigma^{*2}}{2} \right) \leq \mathbb{P}\left(|\hat{\sigma}^2 - \sigma^{*2}| > \frac{\delta \sigma_0^2}{2} \right).$$

Meanwhile,

$$\hat{\sigma}^2 - \sigma^{*2} = \frac{1}{MT} \sum_{t=0}^{T-1} \left\| X_{t+1} - \widehat{A}\mathcal{X}_t \right\|_2^2 - \sigma^{*2}$$

$$= \frac{1}{MT} \sum_{t=0}^{T-1} \|\epsilon_t\|_2^2 - \sigma^{*2} + \frac{1}{MT} \sum_{t=0}^{T-1} \left\| (\widehat{A} - A^*)\mathcal{X}_t \right\|_2^2$$

$$+ \frac{2}{MT} \sum_{t=0}^{T-1} \left| \epsilon_t^\top (\widehat{A} - A^*)\mathcal{X}_t \right|$$

$$= \frac{1}{MT} \sum_{t=0}^{T-1} \|\epsilon_t\|_2^2 - \sigma^{*2} + \frac{1}{M} \sum_{i=1}^{M} (\widehat{A}_i - A_i^*)^\top H(\widehat{A}_i - A_i^*)$$

$$+ \frac{2}{M} \sum_{i=1}^{M} (\widehat{A}_i - A_i^*)^\top \left(\frac{1}{T} \sum_{t=0}^{T-1} \epsilon_{ti}\mathcal{X}_t \right).$$

By Assumption 3.1 and Lemma 5.1, with probability at least $1 - c_1 \exp\{-c_2 \log M\}$,

$$\frac{1}{M} \sum_{i=1}^{M} (\widehat{A}_i - A_i^*)^\top H(\widehat{A}_i - A_i^*) \leq C \frac{\rho \log M}{T} \leq C \sqrt{\frac{\rho \log M}{T}},$$

and

$$\frac{2}{M} \sum_{i=1}^{M} (\widehat{A}_i - A_i^*)^\top H \left(\frac{1}{T} \sum_{t=0}^{T-1} \epsilon_{ti}\mathcal{X}_t \right) \leq 2 \max_i \left\| \widehat{A}_i - A_i^* \right\|_1 \left(\frac{1}{T} \sum_{t=0}^{T-1} \epsilon_t \mathcal{X}_t^\top \right)$$

$$\leq C \frac{\rho \log M}{T} \leq C \sqrt{\frac{\rho \log M}{T}}.$$

Also, since ϵ_{ti} are independent sub-Gaussian random variables with scale factor $C\sigma^*$, the first term can be bounded by Bernstein type inequality of sub-exponential random variables(see

proposition 5.16 in Vershynin [2010]):

$$\mathbb{P}\left(\left|\frac{1}{MT}\sum_{t=0}^{T-1}\|\epsilon_t\|_2^2 - \sigma^{*2}\right| > \frac{\delta\sigma^{*2}}{2}\right) \leq 2\exp\left\{-cMT\min\{\delta^2,\delta\}\right\}.$$

Let $\delta = C\sqrt{\frac{\rho\log M}{T}}$, then

$$\mathbb{P}\left(\hat{\sigma}^2 < \frac{\sigma^{*2}}{1+\delta}\right) + \mathbb{P}\left(\hat{\sigma}^2 > \frac{\sigma^{*2}}{1-\delta}\right)$$

$$\leq 2\exp\left\{-c_1\rho M\log M\right\} + c_2\exp\{-c_3\log M\}.$$

While for $\sup_x F_{d,\|\mu\|_2^2}(x) - F_{d,\|\mu\|_2^2}(x(1-\delta))$ with any $\mu \in \mathbb{R}^d$ satisfying $\|\mu\|_2 \leq C$, if $\delta < \frac{1}{2}$,

$$F_{d,\|\mu\|_2^2}(x) - F_{d,\|\mu\|_2^2}(x(1-\delta))$$

$$=\mathbb{P}\left(\|Z+\mu\|_2^2 \in (x(1-\delta), x]\right)$$

$$\leq C(d)\left(x^{\frac{d}{2}} - (x(1-\delta))^{\frac{d}{2}}\right)\sup_{\|z+\mu\|_2^2\in(x(1-\delta),x]} e^{-\|z\|_2^2/2}$$

$$\leq C(d)\delta x^{\frac{d}{2}}\exp\left\{-\frac{1}{2}\left(\sqrt{x(1-\delta)}-\|\mu\|_2\right)^2 \mathbb{1}(\sqrt{x(1-\delta)}\geq\|\mu\|_2)\right\}.$$

Here $Z \in \mathbb{R}^d$ is a standard Gaussian random vector, the third line is due to that the density of Z is $(2\pi)^{-\frac{d}{2}}e^{-\|z\|_2^2/2}$, and the fourth line applies the fact that when $0 < \delta < \frac{1}{2}$,

$$\left[1-(1-\delta)^{\frac{d}{2}}\right] \leq \frac{d}{2}\sup_{\xi\in(1-\delta,1)}\xi^{\frac{d}{2}-1}\delta = \frac{d}{2}(1-\delta)^{(\frac{d}{2}-1)\mathbb{1}(d\leq 2)}\delta \leq C(d)\delta.$$

Meanwhile, when $\sqrt{x(1-\delta)} < \|\mu\|_2$,

$$x^{\frac{d}{2}} \leq \frac{\|\mu\|_2^d}{(1-\delta)^{\frac{d}{2}}} \leq C(d),$$

and when $\sqrt{x(1-\delta)} \geq \|\mu\|_2$,

$$x^{\frac{d}{2}}\exp\left\{-\frac{1}{2}\left(\sqrt{x(1-\delta)}-\|\mu\|_2\right)^2 \mathbb{1}(\sqrt{x(1-\delta)}\geq\|\mu\|_2)\right\}$$

$$\leq \sup_{y\geq 0}(y+C)^d e^{-y^2/2} \leq C(d),$$

which implies

$$F_{d,\|\mu\|_2^2}(x) - F_{d,\|\mu\|_2^2}(x(1-\delta)) \leq C(d)\delta.$$

To see why all the bounds for $\widehat{U}_T$ still hold for $\widetilde{U}_T$, note that we only need to add $C\sqrt{\frac{\rho\log M}{T}} + 2\exp\left\{-c_1\rho M\log M\right\} + c_2\exp\{-c_3\log M\}$ to the bounds under $\mathcal{H}_0$, and under $\mathcal{H}_A$ when $\phi \geq \frac{1}{2}$, which only changes the constant factors of the previous bounds. For the bound under $\mathcal{H}_A$ when $0 < \phi < \frac{1}{2}$, we substitute x by $\frac{x}{1-\delta}$ with $\delta = C\sqrt{\frac{\log M}{T}}$, and add $2\exp\left\{-c_1\rho M\log M\right\} + c_2\exp\{-c_3\log M\}$, which only changes the constant factors as well. Therefore, all the conclusions for $\widehat{U}_T$ in Theorem 3.1 and 3.2 still hold for $\widetilde{U}_T$ under each corresponding condition. $\qquad\square$

Proof of Theorem 3.4. First we show the connection between R_T and $\widetilde{U}_T$. Note that

$$\widetilde{S}_m = -\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathcal{X}_{t,D_m} - \hat{w}_m^\top\mathcal{X}_{t,D_m^c}\right)\left(X_{t+1,m} - \hat{A}_m^\top\mathcal{X}_t\right)$$

$$= \widehat{S}_m + \left[\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathcal{X}_{t,D_m} - \hat{w}_m^\top\mathcal{X}_{t,D_m^c}\right)\mathcal{X}_{t,D_m}^\top\right]\left(\left(\hat{A}_m\right)_{D_m} - (A_m^*)_{D_m}\right)$$

$$= \widehat{S}_m + \widetilde{\Upsilon^{(m)}}\left(\left(\hat{A}_m\right)_{D_m} - (A_m^*)_{D_m}\right),$$

which implies

$$\hat{a}(m) - (A_m^*)_{D_m} = \widehat{(A_m)}_{D_m} - (A_m^*)_{D_m} - \left(\widetilde{\Upsilon^{(m)}}\right)^{-1}\widetilde{S}_m = -\left(\widetilde{\Upsilon^{(m)}}\right)^{-1}\widehat{S}_m.$$

Thus

$$R_T = \frac{T}{\hat{\sigma}^2}\sum_{m=1}^{k}(\hat{a}(m) - (A_m^*)_{D_m})^\top\widehat{\Upsilon^{(m)}}(\hat{a}(m) - (A_m^*)_{D_m})$$

$$= \frac{T}{\hat{\sigma}^2}\sum_{m=1}^{k}\widehat{S}_m^\top\left(\widetilde{\Upsilon^{(m)}}^\top\right)^{-1}\widehat{\Upsilon^{(m)}}\left(\widetilde{\Upsilon^{(m)}}\right)^{-1}\widehat{S}_m,$$

and the only difference between R_T and $\widetilde{U}_T$ is that we substitute $\left(\widetilde{\Upsilon^{(m)}}\right)^{-1}$ by $\left(\widetilde{\Upsilon^{(m)}}^\top\right)^{-1}\widehat{\Upsilon^{(m)}}\left(\widetilde{\Upsilon^{(m)}}\right)^{-1}$.

We only need to prove that $\left(\widetilde{\Upsilon^{(m)}}^\top\right)^{-1}\widehat{\Upsilon^{(m)}}\left(\widetilde{\Upsilon^{(m)}}\right)^{-1}$ satisfies Assumption 3.3. The argument is very similar to the proof of Lemma 3.3, but we need to bound $\left\|\widehat{\Upsilon^{(m)}}\left(\widetilde{\Upsilon^{(m)}}\right)^{-1}\widetilde{\Upsilon^{(m)}}^\top - \Upsilon^{(m)}\right\|_\infty$ instead of $\left\|\widehat{\Upsilon^{(m)}} - \Upsilon^{(m)}\right\|_\infty$ here.

Let $E = \widetilde{\Upsilon^{(m)}} - \widehat{\Upsilon^{(m)}}$, then

$$\widetilde{\Upsilon^{(m)}}\left(\widehat{\Upsilon^{(m)}}\right)^{-1}\widetilde{\Upsilon^{(m)}}^\top$$

$$= \left(\widehat{\Upsilon^{(m)}} + E\right)\left(\widehat{\Upsilon^{(m)}}\right)^{-1}\left(\widehat{\Upsilon^{(m)}} + E^\top\right)$$

$$= \widehat{\Upsilon^{(m)}} + E + E^\top + E\left(\widehat{\Upsilon^{(m)}}\right)^{-1}E^\top.$$

Recall that when proving Lemma 3.3, we already upper bound $\left\|\widehat{\Upsilon^{(m)}} - \Upsilon^{(m)}\right\|_\infty$ by $C\sqrt{\frac{\log M}{T}}$ with probability at least $1 - c_1\exp\{-c_2\log M\}$. Thus for any vector $u \in \mathbb{R}^{d_m}$ s.t $\|u\|_2 = 1$,

$$u^\top\widehat{\Upsilon^{(m)}}u = u^\top\Upsilon^{(m)}u + u^\top\left(\widehat{\Upsilon^{(m)}} - \Upsilon^{(m)}\right)u$$

$$\geq \Lambda_{\min}\left(\Upsilon^{(m)}\right) - d_m\left\|\widehat{\Upsilon^{(m)}} - \Upsilon^{(m)}\right\|_\infty \geq C,$$

48

which implies $\Lambda_{\max}\left(\left(\widehat{\Upsilon^{(m)}}\right)^{-1}\right) \leq C$, and $\left\|E\left(\widehat{\Upsilon^{(m)}}\right)^{-1}E^\top\right\|_\infty \leq Cd_m\|E\|_\infty$. We bound $\|E\|_\infty$ in the following. One can show that

$$
\begin{aligned}
\|E\|_\infty &= \left\|\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathcal{X}_{t,D_m} - \hat{w}_m^\top \mathcal{X}_{t,D_m^c}\right)\mathcal{X}_{t,D_m^c}^\top \hat{w}_m\right\|_\infty \\
&\leq \left\|\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathcal{X}_{t,D_m} - w_m^{*\top}\mathcal{X}_{t,D_m^c}\right)\mathcal{X}_{t,D_m^c}^\top\right\|_\infty \left(\|w_m^*\|_1 + \|\hat{w}_m - w_m^*\|_1\right) \\
&\quad + \max_{i,j}\left|\left((\hat{w}_m - w_m^*)\right)_{\cdot i}^\top \frac{1}{T}\sum_{t=0}^{T-1}\left(\mathcal{X}_{t,D_m^c}\mathcal{X}_{t,D_m^c}^\top\right)\left((\hat{w}_m - w_m^*)\right)_{\cdot j}\right| \\
&\quad + \left\|\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{X}_{t,D_m^c}\mathcal{X}_{t,D_m^c}^\top w_m^*\right\|_\infty \|\hat{w}_m - w_m^*\|_1.
\end{aligned}
$$

Applying (42), (62), Lemma 5.7, we have

$$
\begin{aligned}
&\left\|\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{X}_{t,D_m^c}\mathcal{X}_{t,D_m^c}^\top w_m^*\right\|_\infty \\
&\leq \left\|\Upsilon_{D_m^c,D_m^c}w_m^*\right\|_\infty + \left\|\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{X}_{t,D_m^c}\mathcal{X}_{t,D_m^c}^\top - \Upsilon_{D_m^c,D_m^c}\right\|_\infty \|w_m^*\|_1 \\
&\leq \Lambda_{\max}(\Upsilon)\max_i\|(w_m^*)_{\cdot,i}\|_2 + C\frac{s_m\log M}{T} \leq C.
\end{aligned}
\tag{63}
$$

Thus, with Lemma 5.6, Assumption 3.2, and (63), we show that with probability at least $1 - c_1\exp\{-c_2\log M\}$,

$$
\|E\|_\infty \leq C\sqrt{\frac{\log M}{T}} + C\frac{s_m\log M}{T} + Cs_m\sqrt{\frac{\log M}{T}} \leq Cs_m\sqrt{\frac{\log M}{T}}.
$$

Therefore, using the same arguments as in the proof of Lemma 3.3,

$$
\begin{aligned}
&\left\|\Upsilon^{(m)\frac{1}{2}}\widetilde{\Upsilon^{(m)}}\left(\widehat{\Upsilon^{(m)}}\right)^{-1}\widetilde{\Upsilon^{(m)}}^\top \Upsilon^{(m)\frac{1}{2}} - I\right\|_2 \\
&\leq C\left\|\widetilde{\Upsilon^{(m)}}\left(\widehat{\Upsilon^{(m)}}\right)^{-1}\widetilde{\Upsilon^{(m)}}^\top - \Upsilon^{(m)}\right\|_2 \\
&\leq Cd_m\left\|\widetilde{\Upsilon^{(m)}}\left(\widehat{\Upsilon^{(m)}}\right)^{-1}\widetilde{\Upsilon^{(m)}}^\top - \Upsilon^{(m)}\right\|_\infty \\
&\leq C\left\|\widehat{\Upsilon^{(m)}} - \Upsilon^{(m)}\right\|_\infty + C\|E\|_\infty \\
&\leq Cs_m\sqrt{\frac{\log M}{T}}.
\end{aligned}
$$

By Lemma A.2,

$$
\left\|\Upsilon^{(m)-\frac{1}{2}}\left(\widetilde{\Upsilon^{(m)}}^\top\right)^{-1}\widehat{\Upsilon^{(m)}}\left(\widetilde{\Upsilon^{(m)}}\right)^{-1}\Upsilon^{(m)-\frac{1}{2}} - I\right\|_\infty \leq Cs_m\sqrt{\frac{\log M}{T}}.
$$

$\square$

C Proof of Lemmas in Section 5

Proof of Lemma 5.3. Let

$$
\xi_{T,t} = -\frac{1}{\sqrt{T}}\begin{pmatrix} \epsilon_{t,1}(\Upsilon^{(1)})^{-\frac{1}{2}}W_1^*\mathcal{X}_t \\ \vdots \\ \epsilon_{t,k}(\Upsilon^{(k)})^{-\frac{1}{2}}W_k^*\mathcal{X}_t \end{pmatrix}.
$$

Define filtration $\mathcal{F}_{T,t} = \sigma(X_{-p+1}, X_{-p+2}, \cdots, X_{t+1})$, then $(\xi_{Tt}, \mathcal{F}_{Tt})_{0 \leq t \leq T-1}$ is a martingale difference sequence, and $V_T = \sum_{t=0}^{T-1}\xi_{T,t}$. To bound the convergence rate, we are going to use a modified version of Lemma 4 in *Grama and Haeusler (2006)*.

Lemma C.1. *Let $(\xi_{ni}, \mathcal{F}_{ni})_{0 \leq i \leq n}$ be a martingale difference sequence taking values in $\mathbb{R}^d$. Let $X_k^n = \sum_{i=1}^k \xi_{ni}$, and $\langle X^n \rangle_k = \sum_{i=1}^k a_{ni} \triangleq \sum_{i=1}^k \mathbb{E}(\xi_{ni}\xi_{ni}^\top | \mathcal{F}_{n,i-1})$. Define $R_\delta^{n,d} = L_\delta^{n,d} + N_\delta^{n,d}$,*

$$
L_\delta^{n,d} = \sum_{i=1}^n \mathbb{E}\|\xi_{ni}\|_2^{2+2\delta}, N_\delta^{n,d} = \mathbb{E}\| \langle X^n \rangle_n - I\|_{\mathrm{tr}}^{1+\delta}.
$$

Then $\forall \mu \in \mathbb{R}^d, r \geq 0, 0 < \delta \leq \frac{1}{2}$, when $R_\delta^{n,d} \leq 1$,

$$
\mathbb{P}(\|X_n^n + \mu\|_2 \geq r) - \mathbb{P}(\|Z + \mu\|_2 \geq r) \leq C(\|\mu\|_2, d, \delta)\left(R_\delta^{n,d}\right)^{\frac{1}{3+2\delta}},
$$

where $Z_{d \times 1} \sim \mathcal{N}(0, I)$, $C(\|\mu\|_2, d, \delta)$ is non-decreasing as $\|\mu\|_2$ increases.

By Lemma C.1, to bound $\sup_{x>0,}\left|\mathbb{P}(\|V_T + \mu\|_2^2 \leq x) - F_{d,\|\mu\|_2^2}(x)\right|$, we only need to bound $R_\delta^{T,d} = L_\delta^{T,d} + N_\delta^{T,d}$.

$$
\begin{aligned}
L_\delta^{T,d} &= \sum_{t=0}^{T-1} \mathbb{E}\left(\|\xi_{T,t}\|_2^{2+2\delta}\right) \\
&\leq CT^{-(1+\delta)}\sum_{t=1}^T \mathbb{E}\left(\sum_{m=1}^k \|W_m^*\mathcal{X}_t\|_2^2 \epsilon_{t,m}^2\right)^{1+\delta} \\
&\leq CT^{-(1+\delta)}\sum_{t=0}^{T-1} k^\delta \sum_{m=1}^k \mathbb{E}\left(|\epsilon_{t,m}|^{2+2\delta}\|W_m^*\mathcal{X}_t\|_2^{2+2\delta}\right) \\
&= T^{-\delta}k^\delta C(\delta)\sum_{m=1}^k \mathbb{E}\left(\|W_m^*\mathcal{X}_0\|_2^{2+2\delta}\right)
\end{aligned}
$$

Here the second line is due to $\Lambda_{\min}(\Upsilon^{(m)}) \geq 1$, and the third line is due to $f(x) = x^{1+\delta}$ is a convex function. More specifically,

$$
\left(\sum_{m=1}^k \|W_m^*\mathcal{X}_t\|_2^2 \epsilon_{t,K}^2\right)^{1+\delta} \leq \frac{1}{k}\sum_{m=1}^k \left(k\|W_m^*\mathcal{X}_t\|_2^2 \epsilon_{t,K}^2\right)^{1+\delta} = k^\delta \sum_{m=1}^k \left(\|W_m^*\mathcal{X}_t\|_2^{2+2\delta}\epsilon_{t,K}^{2+2\delta}\right).
$$

While for the last line, since $\epsilon_{t,m}$ is sub-Gaussian with parameter τ, $\mathbb{E}|\epsilon_{t,m}|^{2+2\delta} \leq C(\delta)$. Note that d, β, τ are all viewed as constants here. Due to the sub-Gaussianity of $\epsilon_{t,i}$'s, we have the following lemma.

Lemma C.2.
$$\mathbb{E}\left(\|W_m^* \mathcal{X}_t\|_2^q\right)^{\frac{1}{q}} \leq Cq \quad \text{for all } q \geq 1.$$

Therefore,
$$\mathbb{E}\left(\|W_m^* \mathcal{X}_0\|_2^{2+2\delta}\right) \leq C(\delta),$$

which implies
$$L_\delta^{T,d} \leq C(\delta)T^{-\delta}.$$

While for $N_\delta^{T,d}$, since

$$\sum_{t=0}^{T-1} \mathbb{E}\left(\xi_{T,t}\xi_{T,t}^\top | \mathcal{F}_{T,t-1}\right) - I$$

$$= \begin{pmatrix} (\Upsilon^{(1)})^{-\frac{1}{2}} B_1 (\Upsilon^{(1)})^{-\frac{1}{2}} & \cdots & \cdots & 0 \\ 0 & (\Upsilon^{(2)})^{-\frac{1}{2}} B_2 (\Upsilon^{(2)})^{-\frac{1}{2}} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & (\Upsilon^{(k)})^{-\frac{1}{2}} B_k (\Upsilon^{(k)})^{-\frac{1}{2}} \end{pmatrix},$$

where $B_m = W_m^* \left(\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{X}_t \mathcal{X}_t^\top - \Upsilon\right) W_m^{*\top}$,

$$N_\delta^{T,d} = \mathbb{E}\left(\left(\sum_{m=1}^k \left\|(\Upsilon^{(m)})^{-\frac{1}{2}} B_m (\Upsilon^{(m)})^{-\frac{1}{2}}\right\|_{\text{tr}}\right)^{1+\delta}\right)$$

$$\leq \mathbb{E}\left(\left(\sum_{m=1}^k d_m \left\|(\Upsilon^{(m)})^{-\frac{1}{2}} B_m (\Upsilon^{(m)})^{-\frac{1}{2}}\right\|_2\right)^{1+\delta}\right)$$

$$\leq \mathbb{E}\left(\left(\sum_{m=1}^k d_m^2 \|B_m\|_\infty\right)^{1+\delta}\right),$$

where the second line is because that $(\Upsilon^{(m)})^{-\frac{1}{2}} B_m (\Upsilon^{(m)})^{-\frac{1}{2}}$ is of rank at most d_m, and we can apply Lemma A.3; the last line is due to

$$\|B_m\|_2 = \sup_{\|u\|_2=1} \|B_m u\|_2 \leq \sup_{\|u\|_2=1} \sqrt{d_m}\|B_m u\|_\infty \leq \sup_{\|u\|_2=1} \sqrt{d_m}\|B_m\|_\infty \|u\|_1 = d_m \|B_m\|_\infty.$$

Since
$$(B_m)_{ij} = \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{X}_t^\top (W_m^*)_{i\cdot}^\top (W_m^*)_{j\cdot} \mathcal{X}_t - \text{tr}\left((W_m^*)_{i\cdot}^\top (W_m^*)_{j\cdot}\Upsilon\right),$$

by Lemma 5.2, we only need to bound the operator norm and trace norm of

$$\frac{1}{2}\left((W_m^*)_{i\cdot}^\top (W_m^*)_{j\cdot} + (W_m^*)_{j\cdot}^\top (W_m^*)_{i\cdot}\right).$$

By (61) and (62), we have the following:

$$\left\|\frac{1}{2}\left((W_m^*)_{i\cdot}^\top (W_m^*)_{j\cdot} + (W_m^*)_{j\cdot}^\top (W_m^*)_{i\cdot}\right)\right\|_{\mathrm{tr}}$$
$$\leq 2\left\|\frac{1}{2}\left((W_m^*)_{i\cdot}^\top (W_m^*)_{j\cdot} + (W_m^*)_{j\cdot}^\top (W_m^*)_{i\cdot}\right)\right\|_2 \leq C.$$

Therefore, applying Lemma 5.2 leads us to

$$\mathbb{P}\left(\left(\sum_{m=1}^k d_m^2 \|B_m\|_\infty\right)^{1+\delta} > x\right)$$
$$\leq \sum_{m=1}^k \mathbb{P}\left(\|B_m\|_\infty > \frac{x^{\frac{1}{1+\delta}}}{d^2}\right)$$
$$\leq c_1 \exp\left\{-c_2 T \min\left\{x^{\frac{1}{1+\delta}}, x^{\frac{2}{1+\delta}}\right\}\right\},$$

which implies

$$N_\delta^{T,d} \leq \int_0^\infty \mathbb{P}\left(\left(\sum_{m=1}^k d_m^2 \|B_m\|_\infty\right)^{1+\delta} > x\right) dx$$
$$\leq \int_0^\infty c_1 \exp\left\{-c_2 T \min\left\{x^{\frac{2}{1+\delta}}, x^{\frac{1}{1+\delta}}\right\}\right\} dx$$
$$\leq C(\delta)\left(\int_0^1 u^\delta \exp\{-cTu^2\}du + \int_1^\infty u^\delta \exp\{-cTu\}du\right)$$
$$\leq C(\delta)\left(T^{-\frac{1+\delta}{2}}\Gamma\left(\frac{1+\delta}{2}\right) + T^{-1-\delta}\Gamma(1+\delta)\right)$$
$$\leq C(\delta)T^{-\frac{1+\delta}{2}}.$$

Thus,

$$R_\delta^{T,d} = N_\delta^{T,d} + L_\delta^{T,d} \leq C(\delta)\left(T^{-\delta} + T^{-\frac{1+\delta}{2}}\right).$$

By Lemma C.1, for any $x \geq 0$, $\mu \in \mathbb{R}^d$, and $0 \leq \delta \leq \frac{1}{2}$, when $T > C(\delta)$,

$$\left|\mathbb{P}\left(\|V_T + \mu\|_2^2 \leq x\right) - F_{d,\|\mu\|_2^2}(x)\right| \leq C(\|\mu\|_2, \delta)\left(R_\delta^{T,d}\right)^{\frac{1}{3+2\delta}}.$$

The best rate is achieved when $\delta = \frac{1}{2}$, and thus when $T > C$,

$$\sup_{x \geq 0}\left|\mathbb{P}\left(\|V_T + \mu\|_2^2 \leq x\right) - F_{d,\|\mu\|_2^2}(x)\right| \leq C(\|\mu\|_2)T^{-\frac{1}{8}},$$

$\square$

52

Proof of Lemma 42. We prove the lower and upper bounds for eigenvalues of Υ, by establishing a connection between our stability condition (13) and another spectral density based condition proposed in Basu et al. [2015]. First we introduce the following lemma, which is a direct result of proposition 2.3 and (2.6) in Basu et al. [2015] under our setting.

Lemma C.3. *Under the model specified in* (3) *with independent noise ϵ_{ti} of unit variance, the eigenvalues of Υ can be bounded as follows:*

$$(\mu_{\max}(\mathcal{A}))^{-1} \leq \Lambda_{\min}(\Upsilon) \leq \Lambda_{\max}(\Upsilon) \leq (\mu_{\min}(\mathcal{A}))^{-1},$$

where $\mu_{\min}(\mathcal{A}) = \min_{|z|=1} \Lambda_{\min}(\mathcal{A}^*(z)\mathcal{A}(z))$, and $\mu_{\max}(\mathcal{A}) = \max_{|z|=1} \Lambda_{\max}(\mathcal{A}^*(z)\mathcal{A}(z))$.

By Lemma C.3, we only need to prove that condition (13) implies a lower bound for $\mu_{\min}(\mathcal{A})$ and upper bound for $\mu_{\max}(\mathcal{A})$. First note that

$$
\begin{aligned}
\mu_{\min}(\mathcal{A}) &= \min_{|z|=1} \Lambda_{\min}(\mathcal{A}(z)\mathcal{A}^*(z)) \\
&= \min_{|z|=1} \inf_{u} \frac{\|\mathcal{A}^*(z)u\|_2^2}{\|u\|_2^2} \\
&= \min_{|z|=1} \inf_{v} \frac{\|v\|_2^2}{\left\|(\mathcal{A}^*(z))^{-1}v\right\|_2^2} \\
&= \min_{|z|=1} \left(\left\|(\mathcal{A}^*(z))^{-1}\right\|_2\right)^{-2},
\end{aligned}
$$

where the last equality is due to that $\left\|(\mathcal{A}^*(z))^{-1}\right\|_2 = \sup_v \frac{\|(\mathcal{A}^*(z))^{-1}v\|_2}{\|v\|_2}$. Meanwhile, for any $|z| = 1$,

$$\left\|(\mathcal{A}^*(z))^{-1}\right\|_2 = \left\|\mathcal{A}^{-1}(z)\right\|_2 = \left\|\sum_{j=0}^{\infty} \Psi_j z^j\right\|_2 \leq \sum_{j=0}^{\infty} \|\Psi_j\|_2 \leq \beta,$$

where we apply condition (13) in the last inequality. Thus $\mu_{\min}(\mathcal{A}) \geq \beta^{-2}$.

While for bounding $\mu_{\max}(\mathcal{A})$, we start by bounding $\|A_n\|_2$ for $0 \leq n \leq p$. Here we define $A_0 = I_{M \times M}$, and $A_n = 0$ for all $n > p$. Since

$$I = \mathcal{A}^{-1}(z)\mathcal{A}(z) = \left(\sum_{j=0}^{\infty} \Psi_j z^j\right)\left(\sum_{i=0}^{p} A_i z^i\right) = \sum_{n=0}^{\infty}\left(\sum_{i=0}^{\infty} \Psi_i A_{n-i}\right) z^n,$$

one can show that $\Psi_0 = I$, and $\sum_{i=0}^{n} \Psi_i A_{n-i} = 0$ for $n \geq 1$. Thus

$$A_n = -\sum_{i=1}^{n} \Psi_i A_{n-i} \text{ for } n \geq 1,$$

and $\|A_n\|_2 \leq \sum_{i=1}^{n} \|\Psi_i\|_2 \|A_{n-i}\|_2$. We have the following claim:

$$\text{For } 0 \leq n \leq p, \quad \|A_n\|_2 \leq \beta^n \vee 1. \tag{64}$$

This can be proved by induction. It is clear that $\|A_0\|_2 = \|I\|_2 = \beta^0$, and if (64) holds for $0 \le n = k \le p$,

$$\|A_{k+1}\|_2 \le \sum_{i=1}^{n} \|\Psi_i\|_2(\beta^{n-i} \vee 1) \le \beta \max_i (\beta^{n-i} \vee 1) \le \beta^n \vee 1.$$

Therefore, $\mu_{\max}(\mathcal{A})$ can be bounded in the following:

$$
\begin{aligned}
\mu_{\max}(\mathcal{A}) &= \max_{|z|=1} \Lambda_{\max}\left(\mathcal{A}(z)\mathcal{A}^*(z)\right) \\
&= \max_{|z|=1} \|\mathcal{A}^*(z)\|_2^2 \\
&\le \left(\sum_{i=0}^{p} \|A_i\|_2\right)^2 \\
&\le \left(\frac{\beta^{p+1}-1}{\beta-1}\right)^2 \mathbf{1}(\beta > 1) + (p+1)^2 \mathbf{1}(0 \le \beta \le 1).
\end{aligned}
$$

With Lemma C.3, we conclude that

$$C_1(\beta) \le \Lambda_{\min}(\Upsilon) \le \Lambda_{\max}(\Upsilon) \le C_2(\beta),$$

where $C_1(\beta) = \left(\frac{1-\beta}{1-\beta^{p+1}}\right)^2 \mathbf{1}(\beta > 1) + (p+1)^{-2}\mathbf{1}(0 \le \beta \le 1)$, and $C_2(\beta) = \beta^2$. $\qquad\square$

Proof of Lemma 5.1. Recall that $X_t = \sum_{j=0}^{\infty} \Psi_j \epsilon_{t-j-1}$. Define $\Psi_j^{(p)} \in \mathbb{R}^{pM \times M}$ as the following:

$$
\Psi_j^{(p)} = \begin{pmatrix} \Psi_j \mathbf{1}(j \ge 0) \\ \vdots \\ \Psi_{j-p+1}\mathbf{1}(j-p+1 \ge 0) \end{pmatrix}, \tag{65}
$$

then we can also write $\mathcal{X}_t$ as an infinite sum $\mathcal{X}_t = \sum_{j=0}^{\infty} \Psi_j^{(p)} \epsilon_{t-j-1}$. Without loss of generality, we consider the first entry of $\frac{1}{T}\sum_{t=0}^{T-1} \epsilon_t \mathcal{X}_t^\top$:

$$\frac{1}{T}\sum_{t=0}^{T-1} \epsilon_{t,1} \sum_{j=0}^{\infty} (\Psi_j)_{1\cdot} \epsilon_{t-j-1}. \tag{66}$$

In the following, we tackle the infinite sum in (66), by focusing our analysis on the finite sum and let the residue converges to 0. Rigorously, for any positive integer m, let

$$\tilde{\epsilon} = (\epsilon_{-m-1}^\top, \ldots, \epsilon_{T-1}^\top)^\top, \quad \eta^{(t)} = ((\Psi_{t+m})_{1\cdot}, \ldots, (\Psi_0)_{1,\cdot}, 0, \ldots, 0)^\top \in \mathbb{R}^{(T+m+1)M},$$

and $e^{(t)} \in \mathbb{R}^{(T+m+1)M}$ satisfying $e_i^{(t)} = \mathbf{1}(i = (t+m)M+1)$, then we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \epsilon_{t,1} \sum_{j=0}^{\infty} (\Psi_j)_{1\cdot} \epsilon_{t-j-1} \\
=&\tilde{\epsilon}^\top \left(\frac{1}{T}\sum_{t=0}^{T-1} e^{(t)} \eta^{(t)\top}\right) \tilde{\epsilon} + \frac{1}{T}\sum_{t=0}^{T-1} \epsilon_{t,1} \sum_{j=t+m+1}^{\infty} (\Psi_j)_{1\cdot} \epsilon_{t-j-1} \\
\triangleq& E_1 + E_2.
\end{aligned}
$$

We will let m be sufficiently large in later argument. The following arguments are devided into two parts: bounding E_1 and E_2.

(1) Bounding E_1

Since all entries of $\tilde{\epsilon}$ are independent sub-Gaussian with constant parameter, we can apply the following Hanson-Wright inequality:

Lemma C.4. *Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent components X_i which satisfy $\mathbb{E}(X_i) = 0$ and $\|X_i\|_{\psi_2} \leq K$. Let A be an $n \times n$ matrix. Then, for every $t \geq 0$,*

$$\mathbb{P}\left(|X^\top A X - \mathbb{E}X^\top A X| > t\right) \leq 2\exp\left\{-c\min\left(\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|_2}\right)\right\}$$

This lemma is a result in Rudelson et al. [2013].By Lemma C.4, we only need to bound the norms of $\frac{1}{T}\sum_{t=0}^{T-1} e^{(t)}\eta^{(t)\top}$.

First note that

$$\left\|\frac{1}{T}\sum_{t=0}^{T-1} e^{(t)}\eta^{(t)\top}\right\|_2 = \sup_{\|u\|_2=\|v\|_2=1} \frac{1}{T}\sum_{t=0}^{T-1} u^\top e^{(t)}\eta^{(t)\top}v.$$

For any $u, v \in \mathbb{R}^{(T+m+1)M}$ with unit ℓ_2 norm, one can show that

$$\frac{1}{T}\sum_{t=0}^{T-1} u^\top e^{(t)}\eta^{(t)\top}v$$

$$= \frac{1}{T}\sum_{t=0}^{T-1} u_{(t+m)M+1}\sum_{i=0}^{t+m}(\Psi_{t+m-i})_{1.}v^{(i+1)}$$

$$\leq \frac{1}{T}\sum_{t=0}^{T-1} u_{(t+m)M+1}\sum_{i=0}^{t+m}\alpha_{t+m-i}\|v^{(i+1)}\|_2$$

$$\leq \frac{1}{T}(u_{mM+1}, \cdots, u_{(T+m-1)M+1})\Gamma\begin{pmatrix}\|v^{(1)}\|_2 \\ \vdots \\ \|v^{(T+m)}\|_2\end{pmatrix}$$

$$\leq \frac{\|\Gamma\|_2}{T},$$

where $v^{(i)} = (v_{(i-1)M+1}, \ldots, v_{iM})^\top$, $\alpha_i = \|\Psi_i\|_2 \geq \|(\Psi_i)_{1.}\|_2$, and $\Gamma \in \mathbb{R}^{T\times(T+m)}$ is a matrix with each entry $\Gamma_{ij} = \alpha_{m+i-j}1(m+i-j \geq 0)$. Since Γ is a Toeplitz matrix, we will use the following lemma to bound its ℓ_2 norm.

Lemma C.5. *Let $f(\lambda)$ be a Fourier series defined as $f(\lambda) = \sum_{t=-\infty}^{\infty} t_k \exp\{ik\lambda\}$, with $\sum_{k=-\infty}^{\infty}|t_k| < \infty$. We define a sequence of Toeplitz matrices T_n with $(T_n)_{i,j} = t_{i-j}$, then the operator norm of T_n is bounded by*

$$\|T_n\|_2 \leq 2\,\mathrm{ess}\sup f.$$

where $\mathrm{ess}\sup f$ the essential supremum.

This is actually Lemma 4.1 in Gray et al. [2006], and we directly apply it here. By Lemma C.5,

$$\|\Gamma\|_2 \le 2 \sup_\lambda \left| \sum_{k=-m}^{\infty} \alpha_{m+k} e^{ik\lambda} \right| \le 2 \sum_{k=0}^{\infty} \alpha_k \le \sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} \alpha_{i+j}^2 \right)^{\frac{1}{2}} \le \beta.$$

Thus $\left\| \frac{1}{T} \sum_{t=0}^{T-1} e^{(t)} \eta^{(t)\top} \right\|_2 \le \frac{\beta}{T}$. While for the Frobenius norm, we have

$$
\begin{aligned}
\left\| \frac{1}{T} \sum_{t=0}^{T-1} e^{(t)} \eta^{(t)\top} \right\|_F^2 &= \mathrm{tr} \left(\left(\frac{1}{T} \sum_{t=0}^{T-1} \eta^{(t)} e^{(t)\top} \right) \left(\frac{1}{T} \sum_{l=0}^{T-1} e^{(t)} \eta^{(t)\top} \right) \right) \\
&= \frac{1}{T^2} \sum_{t=0}^{T-1} \|\eta^{(t)}\|_2^2 \\
&\le \frac{1}{T^2} \sum_{t=0}^{T-1} \sum_{i=0}^{t+m} \alpha_i^2 \le \frac{\beta^2}{T}.
\end{aligned}
$$

Therefore, by Lemma C.4, for any $\delta > 0$,

$$\mathbb{P}\left(|E_1| > \delta\right) \le 2 \exp\left\{ -cT \min\{\delta, \delta^2\} \right\}.$$

(2) Bounding E_2

First note that

$$
\begin{aligned}
|E_2| &= \left| \frac{1}{T} \sum_{t=0}^{T-1} \epsilon_{t,1} \sum_{j=t+m+1}^{\infty} (\Psi_j)_{1\cdot} \epsilon_{t-j-1} \right| \\
&\le \frac{1}{2T} \sum_{t=0}^{T-1} \epsilon_{t,1}^2 + \frac{1}{2T} \sum_{t=0}^{T-1} \left(\sum_{j=t+m+1}^{\infty} (\Psi_j)_{1\cdot} \epsilon_{t-j-1} \right)^2.
\end{aligned}
$$

Recall the definition of $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$ in the proof of Lemma C.2. Since $\|\epsilon_{t,1}^2\|_{\psi_1} \le 2\|\epsilon_{t,1}\|_{\psi_2}^2 \le 2\tau^2$,

$$\mathbb{P}\left(\left| \frac{1}{2T} \sum_{t=0}^{T-1} \epsilon_{t,1}^2 \right| > \delta \right) \le 2 \exp\{ -cT \min\{\delta, \delta^2\} \},$$

by Bernstein type inequality of sub-exponential random variables(see proposition 5.16 in Vershynin [2010]).

Now we bound the second term $\frac{1}{2T} \sum_{t=0}^{T-1} \left(\sum_{j=t+m+1}^{\infty} (\Psi_j)_{1\cdot} \epsilon_{t-j-1} \right)^2$. Since

$$\left| \sum_{j=t+m+1}^{\infty} (\Psi_j)_{1\cdot} \epsilon_{t-j-1} \right| \le \sum_{j=t+m+1}^{\infty} \alpha_j \|\epsilon_{t-j-1}\|_2,$$

one can show that

$$
\left\| \frac{1}{2T} \sum_{t=0}^{T-1} \left(\sum_{j=t+m+1}^{\infty} (\Psi_j)_{1.}\epsilon_{t-j-1} \right)^2 \right\|_{\psi_1}
$$

$$
\leq \left\| \sum_{j=t+m+1}^{\infty} (\Psi_j)_{1.}\epsilon_{t-j-1} \right\|_{\psi_2}^2
$$

$$
\leq CM\tau^2 \left(\sum_{j=t+m+1}^{\infty} \alpha_j \right)^2,
$$

where we apply the fact that $\|\|\epsilon_t\|_2\|_{\psi_2} \leq C\sqrt{M}\tau$, which is shown in the proof of Lemma C.2. Thus we have

$$
\mathbb{P} \left(\frac{1}{2T} \sum_{t=0}^{T-1} \left(\sum_{j=t+m+1}^{\infty} (\Psi_j)_{1.}\epsilon_{t-j-1} \right)^2 > \delta \right)
$$

$$
\leq C \exp \left\{ -\frac{c\delta}{M\tau^2 \left(\sum_{j=t+m+1}^{\infty} \alpha_j \right)^2} \right\}.
$$

due to the tail bound of sub-exponential r.v. (also see Vershynin [2010]). Since

$$
\sum_{i=0}^{\infty} \alpha_i \leq \sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} \alpha_{i+j}^2 \right)^{\frac{1}{2}} \leq \beta,
$$

$$
\lim_{m \to \infty} \left(\sum_{j=t+m+1}^{\infty} \alpha_j \right)^2 = 0.
$$

Let m be sufficiently large such that $\left(\sum_{j=t+m+1}^{\infty} \alpha_j \right)^2 \leq \frac{1}{MT}$, then we arrive at the following

$$
\mathbb{P} \left(\frac{1}{T} \sum_{t=0}^{T-1} \epsilon_{t,1}(\mathcal{X}_t)_1 \right) \leq C \exp\{-cT \min\{\delta, \delta^2\}\}.
$$

Let $\delta = C\sqrt{\log MT}$ and take a union bound over the pM^2 entries of $\frac{1}{T} \sum_{t=0}^{T-1} \epsilon_t \mathcal{X}_t^\top$, the conclusion follows.

$\square$

Proof of Lemma 5.6. Without loss of generality, consider

$$
\frac{1}{T} \sum_{t=0}^{T-1} \left(\mathcal{X}_{t,D_m} - w_m^{*\top} \mathcal{X}_{t,D_m^c} \right)_i \mathcal{X}_{t,j}
$$

57

for any $1 \leq i \leq d_m$, and $j \in D_m^c$. Similar from the proof of Lemma 5.6, We can write it as a quadratic form

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_t^\top \frac{1}{2} \left((W_m^*)_{i \cdot}^\top e_j^\top + e_j (W_m^*)_{i \cdot} \right) \mathcal{X}_t,$$

where W_m^* is defined as in (48). Since $\frac{1}{2} \left((W_m^*)_{i \cdot}^\top e_j^\top + e_j (W_m^*)_{i \cdot} \right)$ is of rank 2, and we have bounded $\|(W_m^*)_{i \cdot}\|_2$ in (62), applying Lemma A.3 leads to

$$\left\| \frac{1}{2} \left((W_m^*)_{i \cdot}^\top e_j^\top + e_j (W_m^*)_{i \cdot} \right) \right\|_{\mathrm{tr}} \leq 2 \left\| \frac{1}{2} \left((W_m^*)_{i \cdot}^\top e_j^\top + e_j (W_m^*)_{i \cdot} \right) \right\|_2$$

$$\leq \|(W_m^*)_{i \cdot}\|_2 \leq C.$$

Applying Lemma 5.2, and taking a union bound over all entries of

$$\frac{1}{T} \sum_{t=0}^{T-1} \left(\mathcal{X}_{t, D_m} - w_m^{* \top} \mathcal{X}_{t, D_m^c} \right) \mathcal{X}_t,$$

the conclusion follows. $\qquad \square$

Proof of Lemma 5.7. Similar from the proof of Lemma 5.1, we consider $\left| \frac{1}{T} \sum_{t=0}^{T-1} X_{ti} X_{tj} - \Upsilon_{ij} \right|$. Since

$$\frac{1}{T} \sum_{t=0}^{T-1} X_{ti} X_{tj} = \frac{1}{T} \sum_{t=0}^{T-1} X_t^\top \left(\frac{1}{2} (e_i e_j^\top + e_j e_i^\top) \right) X_t,$$

by Lemma 5.2, we need to bound norms of $\frac{1}{2}(e_i e_j^\top + e_j e_i^\top)$, which is of rank at most 2. One can show that

$$\left\| \frac{1}{2} (e_i e_j^\top + e_j e_i^\top) \right\|_{\mathrm{tr}} \leq 2 \left\| \frac{1}{2} (e_i e_j^\top + e_j e_i^\top) \right\|_2 \leq 2 \|e_i\|_2 \|e_j\|_2,$$

with Lemma A.3. Therefore, by taking a union bound, it is clear that

$$\left\| \frac{1}{T} \sum_{t=0}^{T-1} X_t X_t^\top - \Upsilon \right\|_\infty \leq C \sqrt{\frac{\log M}{T}},$$

with probability at least $1 - c_1 \exp\{-c_2 \log M\}$. $\qquad \square$

D   Proof of Lemmas in Section A and Appendix C

Proof of Lemma C.1. Here we adopt the proof framework for Lemma 4 in Grama and Haeusler [2006], but with some small adjustments. First we construct a new martingale difference sequence $(m_{nk}, \mathcal{G}_{nk})_{1 \leq k \leq n+1}$, sum of whose covariances equal to $I_{d \times d}$. Random projections are used for construction. The following lemma on random projections is stated as Lemma 3 in Grama and Haeusler [2006].

Lemma D.1. *Let V and $a_1, \cdots, a_n$ be positive semi-definite $d \times d$ matrices. Set $A_k = a_1 + \cdots + a_k$, for $k = 1, \cdots, n$. Then there exist a sequence of integers $1 \leq \tau_1 \leq \cdots \leq \tau_d \leq n$ and a corresponding sequence $\mathcal{S}_1 \supseteq \cdots \supseteq \mathcal{S}_d$ of subspaces of $\mathbb{R}^d$ such that, with P_k defined as the projection matrix of subspace $\mathcal{S}_i$, for $\tau_i \leq k < \tau_{i+1}$ (where $\tau_0 = 1, \tau_{d+1} = n + 1, \mathcal{S}_0 = \mathbb{R}^d$), the following statements hold true for $k = 1, \cdots, n$:*
(a)$V - \widehat{A}_k$ is non-negative definite, where $\widehat{A}_k = P_1 a_1 P_1 + \cdots + P_k a_k P_k$;
(b)$x^\top (\widehat{A}_k - A_k) x = 0$, for all $x \in \Pi_k \triangleq \{P_k x : x \in \mathbb{R}^d\}$;
(c)$x^\top (\widehat{A}_k - V + \alpha_k I) x \geq 0$ for all $x \in \Pi_k^\top$, where $\alpha_k = \max\{\|a_{\tau_j}\|_2 : \tau_j \leq k\}$.
Meanwhile, P_k is determined by $a_1, \cdots, a_k$ and V.

Given this claim, m_{nk} can be constructed as follows:

Recall the martingale sequence we consider is $(\xi_{nk}, \mathcal{F}_{nk})_{1 \leq k \leq n+1}$, and $a_{nk} = \mathbb{E}\left(\xi_{nk} \xi_{nk}^\top\right)$. Apply the fact with $V = I$, $a_k = a_{nk}$, and let $\{P_{nk}\}_{k=1}^n$ be the corresponding projection matrices. Let $D_n = I - \sum_{k=1}^n P_{nk} a_{nk} P_{nk}$, which is non-negative definite. Define

$$M_k^n = \sum_{k=1}^n m_{nk}, \quad 1 \leq k \leq n+1,$$

where

$$m_{nk} = P_{nk} \xi_{nk}, \text{ for } 1 \leq k \leq n, \quad m_{n,n+1} = D_n^{\frac{1}{2}} \eta_{n,n+1}.$$

Since $P_{nk} \in \mathcal{F}_{n,k-1}$, $m_{nk} \in \mathcal{F}_{nk}$ for $1 \leq k \leq n$. Thus $(m_{nk}, \mathcal{G}_{nk})$ is also a martingale difference sequence with $\mathcal{G}_{nk} = \mathcal{F}_{nk}$, when $1 \leq k \leq n$, and $\mathcal{G}_{n,n+1} = \sigma(F_{nn}, \eta_{n,n+1})$. Meanwhile,

$$\langle M^n \rangle_{n+1} = \sum_{k=1}^{n+1} \mathbb{E}(m_{nk} m_{nk}^\top | \mathcal{F}_{n,k-1}) = I_{d \times d}.$$

This construction is from Grama and Haeusler [2006]. They also prove that, for any $\varepsilon, \delta > 0$,

$$\mathbb{P}\left(\|X_n^n - M_{n+1}^n\|_2 \geq \varepsilon\right) \leq C(d, \delta) \varepsilon^{-2-2\delta} \left(L_\delta^{n,d} + N_\delta^{n,d}\right), \tag{67}$$

Since

$$
\begin{aligned}
& -\mathbb{P}(\|X_n^n - M_{n+1}^n\|_2 > \varepsilon) - \mathbb{P}(\|Z + \mu\|_2 \geq r + 2\varepsilon) \\
& + \mathbb{P}(\|M_{n+1}^n + \mu\|_2 \geq r + \varepsilon) - \mathbb{P}(Z \in [r, r + 2\varepsilon)) \\
\leq & \mathbb{P}(\|X_n^n + \mu\|_2 \geq r) - \mathbb{P}(\|Z + \mu\|_2 \geq r) \\
\leq & \mathbb{P}(\|M_{n+1}^n + \mu\|_2 \geq r - \varepsilon) - \mathbb{P}(\|Z + \mu\|_2 \geq r - 2\varepsilon) \\
& + \mathbb{P}(\|X_n^n - M_{n+1}^n\|_2 > \varepsilon) + \mathbb{P}(Z \in [r - 2\varepsilon, r)),
\end{aligned}
\tag{68}
$$

for any $\mu \in \mathbb{R}^d, r \geq 0, \varepsilon > 0$, we need to bound

$$\mathbb{E}(1(\|Z + \mu\|_2 \geq r + 2\varepsilon)) - \mathbb{E}(1(\|M_{n+1}^n + \mu\|_2 \geq r + \varepsilon))$$

and
$$\mathbb{E}(1(\|M_{n+1}^n + \mu\|_2 \geq r - \varepsilon)) - \mathbb{E}(1(\|Z + \mu\|_2 \geq r - 2\varepsilon)).$$

The following functions are defined as a smooth relaxation for indicator function. Let

$$f_*(z) = \int_{-\infty}^{z-\frac{1}{2}} \phi(t)\mathrm{d}t, \text{ with } \phi(t) = \frac{1}{C}\exp\{-\frac{4}{1-4t^2}\}1_{(-\frac{1}{2},\frac{1}{2})}(t), \tag{69}$$

where C is a normalizing constant s.t. $\int \phi(t)dt = 1$. Then we have $f_*(z) = 0$ if $z \leq 0$, $0 \leq f_*(z) \leq 1$ if $0 \leq z \leq 1$, and $f_*(z) = 1$ if $z \geq 1$. $f_*(z)$ is infinitely many times differentiable on $\mathbb{R}$, and since $f_*(z)$ is constant when $z \leq 0$ or $z \geq 1$, for any fixed order, the derivative of $f_*(z)$ is bounded. For any $z \in \mathbb{R}^d$, let

$$f_{l,\mu,r,\varepsilon}(z) = f_*(g_{l,\mu,r,\varepsilon}(z)), \tag{70}$$

where

$$g_{1,\mu,r,\varepsilon}(z) = \frac{\|z + \mu\|_2 - r - \varepsilon}{\varepsilon}, \quad g_{2,\mu,r,\varepsilon}(z) = \frac{\|z + \mu\|_2 - r + 2\varepsilon}{\varepsilon}. \tag{71}$$

In the following proof, we will denote $f_{l,\mu,r,\varepsilon}(z)$ and $g_{l,\mu,r,\varepsilon}(z)$ as $f_l(z)$ and $g_l(z)$, $l = 1, 2$ for brevity. Therefore,

$$\mathbb{E}(1(\|Z + \mu\|_2 \geq r + 2\varepsilon)) - \mathbb{E}(1(\|M_{n+1}^n + \mu\|_2 \geq r + \varepsilon)) \leq \mathbb{E}(f_1(Z) - f_1(M_{n+1}^n)),$$

$$\mathbb{E}(1(\|M_{n+1}^n + \mu\|_2 \geq r - \varepsilon)) - \mathbb{E}(1(\|Z + \mu\|_2 \geq r - 2\varepsilon)) \leq \mathbb{E}(f_2(M_{n+1}^n) - f_1(Z)).$$

Thus,

$$|\mathbb{P}(\|X_n^n + \mu\|_2 \geq r) - \mathbb{P}(\|Z + \mu\|_2 \geq r)|$$
$$\leq \max_{l=1,2}|\mathbb{E}(f_l(M_{n+1}^n) - f_l(Z))| + \mathbb{P}(\|X_n^n - M_{n+1}^n\|_2 > \varepsilon)$$
$$+ \mathbb{P}(\|Z + \mu\|_2 \in [r - 2\varepsilon, r + 2\varepsilon]).$$

Actually, when $r \leq 3\varepsilon$, the right hand side of (68) can be substituted by

$$\mathbb{P}(\|Z + \mu\|_2 < 3\varepsilon),$$

and

$$|\mathbb{P}(\|X_n^n + \mu\|_2 \geq r) - \mathbb{P}(\|Z + \mu\|_2 \geq r)|$$
$$\leq \max\{|\mathbb{E}(f_1(M_{n+1}^n) - f_1(Z))| + \mathbb{P}(\|X_n^n - M_{n+1}^n\|_2 > \varepsilon) \tag{72}$$
$$+ \mathbb{P}(\|Z + \mu\|_2 \in [r, r + 2\varepsilon]), \mathbb{P}(\|Z + \mu\|_2 \in [0, 3\varepsilon))\}.$$

To bound $\mathbb{E}(f_l(M_{n+1}^n) - f_l(Z))$, we will use the following lemma.

Lemma D.2. *For $f_l(\cdot)$ defined as in (70),*

$$\left|\sum_{1 \leq i_1, \cdots, i_k \leq d} y_{i_1} \cdots y_{i_k} \frac{\partial^k}{\partial z_{i_1} \cdots \partial z_{i_k}} f_l(z)\right| \leq C(k)\varepsilon^{-k}\|y\|_2^k, \tag{73}$$

for any $k \in \mathbb{Z}^*$, $y, z \in \mathbb{R}^d$, when $l = 1$, or when $l = 2$ and $r > 3\varepsilon$.

60

The proof of this lemma is deferred to Appendix E. In the following proof, we will always assume the condition $l = 1$ or $l = 2$ and $r > 3\varepsilon$ hold. Therefore, for any $m \in \mathbb{Z}^*$,

$$\left| f_l(z + y) - f_l(y) - \sum_{k=1}^{m} \sum_{1 \le i_1, \cdots, i_k \le d} y_{i_1} \cdots y_{i_k} \frac{\partial^k}{\partial z_{i_1} \cdots \partial z_{i_k}} f_l(z) \right|$$

$$= \left| \sum_{1 \le i_1, \cdots, i_{m+1} \le d} y_{i_1} \cdots y_{i_{m+1}} \frac{\partial^{m+1}}{\partial u_{i_1} \cdots \partial u_{i_{m+1}}} f_l(u) \right|$$

$$\le C(m+1) \varepsilon^{-m-1} \|y\|_2^{m+1},$$

where $u = z + t_1 y$ for some $0 \le t_1 \le 1$. Meanwhile,

$$\left| f_l(z + y) - f_l(y) - \sum_{k=1}^{m} \sum_{1 \le i_1, \cdots, i_k \le d} y_{i_1} \cdots y_{i_k} \frac{\partial^k}{\partial z_{i_1} \cdots \partial z_{i_k}} f_l(z) \right|$$

$$= \left| \sum_{1 \le i_1, \cdots, i_m \le d} y_{i_1} \cdots y_{i_{(m)}} \frac{\partial^m}{\partial v_{i_1} \cdots \partial v_{i_m}} f_l(v) \right.$$

$$\left. - \sum_{1 \le i_1, \cdots, i_m \le d} y_{i_1} \cdots y_{i_m} \frac{\partial^m}{\partial z_{i_1} \cdots \partial z_{i_m}} f_l(z) \right|$$

$$\le 2 C(m) \varepsilon^{-m} \|y\|_2^m,$$

where $v = z + t_2 y$ for some $0 \le t_2 \le 1$. Thus, for any $\delta > 0$,

$$\left| f_l(z + y) - f_l(y) - \sum_{k=1}^{\lceil 2+2\delta \rceil - 1} \sum_{1 \le i_1, \cdots, i_k \le d} y_{i_1} \cdots y_{i_k} \frac{\partial^k}{\partial z_{i_1} \cdots \partial z_{i_k}} f_l(z) \right|$$

$$\le C(\delta) \max\left\{ \varepsilon^{-\lceil 2+2\delta \rceil + 1} \|y\|_2^{\lceil 2+2\delta \rceil - 1}, \varepsilon^{-\lceil 2+2\delta \rceil} \|y\|_2^{\lceil 2+2\delta \rceil} \right\}$$

$$\le C(\delta) \varepsilon^{-2-2\delta} \|y\|_2^{2+2\delta}.$$

Let $\tilde{w}_{nk}$, $1 \le k \le n$ be i.i.d. standard Gaussian random vectors that are independent of $\mathcal{G}_{n,n+1}$, $w_{nk} = (b_{nk})^{\frac{1}{2}} \tilde{w}_{nk}$, for $k = 1, \cdots, n+1$, where $b_{nk} = \mathbb{E}(m_{nk} m_{nk}^{\top} | \mathcal{G}_{n,k-1})$. Define

$$W_{n+2}^n = 0, \quad W_k^n = \sum_{i=k}^{n+1} w_{ni}, \quad 1 \le k \le n+1.$$

Then W_1^n follows standard Gaussian distribution. Let $U_k^n = M_{k-1}^n + W_{k+1}^n$, then

$$\left| \mathbb{E}(f_l(M_{n+1}^n) - f_l(Z)) \right|$$

$$= \left| \mathbb{E}(f_l(M_{n+1}^n) - f_l(W_1^n)) \right|$$

$$= \left| \sum_{k=1}^{n+1} \mathbb{E}(f_l(U_k^n + m_{nk}) - f_l(U_k^n + w_{nk})) \right|$$

$$\leq \sum_{k=1}^{n+1} \left| \mathbb{E}(f_l(U_k^n + m_{nk}) - f_l(U_k^n) \right.$$

$$- \sum_{j=1}^{\lceil 2+2\delta \rceil - 1} \sum_{1 \leq i_1, \cdots, i_j \leq d} (m_{nk})_{i_1} \cdots (m_{nk})_{i_j} \frac{\partial^j}{\partial z_{i_1} \cdots \partial z_{i_j}} f_l(U_k^n)) \Bigg|$$

$$+ \sum_{k=1}^{n+1} \left| \mathbb{E}(f_l(U_k^n + w_{nk}) - f_l(U_k^n) \right.$$

$$- \sum_{j=1}^{\lceil 2+2\delta \rceil - 1} \sum_{1 \leq i_1, \cdots, i_j \leq d} (w_{nk})_{i_1} \cdots (w_{nk})_{i_j} \frac{\partial^j}{\partial z_{i_1} \cdots \partial z_{i_j})} f_l(U_k^n)) \Bigg|$$

$$\leq \sum_{k=1}^{n+1} C(\delta) \varepsilon^{-2-2\delta} \mathbb{E}(\|m_{nk}\|_2^{2+2\delta}).$$

Generally this inequality holds for $\delta \in (0, \frac{1}{2}]$, since w_{nk} and m_{nk} have the same second order moments, which justifies the fourth line. By the proof of Lemma 4 in Grama and Haeusler [2006],

$$\sum_{k=1}^{n+1} \mathbb{E}(\|m_{nk}\|_2^{2+2\delta}) \leq C(d, \delta)(L_\delta^{n,d} + N_\delta^{n,d}),$$

thus

$$\left| \mathbb{E} \left(f_l(M_{n+1}^n) - f_l(z) \right) \right| \leq C(d, \delta) \varepsilon^{-2-2\delta} R_\delta^{n,d}. \tag{74}$$

Now we only need to bound $\mathbb{P} \left(\|Z + \mu\|_2 \in [r - 2\varepsilon, r + 2\varepsilon] \right)$ and $\mathbb{P} \left(\|Z + \mu\|_2 \in [0, 3\varepsilon) \right)$. Assume $\varepsilon \leq 1$, then

$$\mathbb{P} \left(\|Z + \mu\|_2 \in [0, 3\varepsilon) \right) = \mathbb{P} \left(Z \in \mathbb{B}_{3\varepsilon}(-\mu) \right) \leq C(d) \varepsilon^d \leq C(d) \varepsilon.$$

Meanwhile,

$$\mathbb{P}(\|Z + \mu\|_2 \in [r - 2\varepsilon, r + 2\varepsilon])$$

$$= \mathbb{P}(Z \in \mathbb{B}_{r+2\varepsilon}(-\mu) \backslash \mathbb{B}_{r-2\varepsilon}(-\mu))$$

$$\leq \begin{cases} C(d) \left((r + 2\varepsilon)^d - (r - 2\varepsilon)^d \right), & r \leq 2\varepsilon + \|\mu\| \\ C(d) \exp\{-(r - 2\varepsilon - \|\mu\|_2)^2/2\} \left((r + 2\varepsilon)^d - (r - 2\varepsilon)^d \right), & r > 2\varepsilon + \|\mu\|_2 \end{cases}$$

$$\leq C(d, \|\mu\|_2) \varepsilon.$$

The last line is due to that

$$(r + 2\varepsilon)^d - (r - 2\varepsilon)^d \leq 4\varepsilon d(r + 2\varepsilon)^{d-1} \leq 4d\varepsilon(4 + \|\mu\|_2)^{d-1},$$

when $r \leq 2\varepsilon + \|\mu\|_2$, and

$$\exp\{-(r - 2\varepsilon - \|\mu\|_2)^2/2\} \left((r + 2\varepsilon)^d - (r - 2\varepsilon)^d\right)$$

$$\leq 4\varepsilon d \sup_{x>0}(x + 4\varepsilon + \|\mu\|_2)^{d-1} \exp\{-x^2/2\}$$

$$\leq 4d \sup_{x>0}(x + 4 + \|\mu\|_2)^{d-1} \exp\{-x^2/2\}\varepsilon.$$

Here clearly $C(d, \|\mu\|_2)$ is non-decreasing with respect to $\|\mu\|_2$. Therefore, by (72), (67) and (74), when $R_\delta^{n,d} \leq 1$, for any $\mu \in \mathbb{R}^d$, $r \geq 0$, $0 < \delta \leq \frac{1}{2}$, with $\varepsilon = (R_\delta^{n,d})^{\frac{1}{3+2\delta}}$,

$$\mathbb{P}(\|X_n^n + \mu\|_2 \geq r) - \mathbb{P}(\|Z + \mu\|_2 \geq r) \leq C(d, \delta, \|\mu\|_2) \left(R_\delta^{n,d}\right)^{\frac{1}{3+2\delta}},$$

where $C(d, \delta, \|\mu\|_2)$ is non-decreasing with respect to $\|\mu\|_2$.

$\square$

Proof of Lemma C.2. First we introduce the following two norms:
For any random variable X,

$$\|X\|_{\psi_1} = \sup_{p \geq 1} p^{-1} \mathbb{E}\left(|X|^p\right)^{\frac{1}{p}},$$

$$\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-\frac{1}{2}} \mathbb{E}\left(|X|^p\right)^{\frac{1}{p}}.$$

These two norms are related to sub-exponential and sub-Gaussian random variables, and the following lemma shows the connections between the two norms and the scale factor for sub-Gaussian r.v.

Lemma D.3. *For any sub-Gaussian r.v. X with scale factor τ, the following hold:*

$$c\tau \leq \|X\|_{\psi_2} \leq C\tau,$$

with some absolute constants c, C, and

$$\|X\|_{\psi_2}^2 \leq \|X\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2.$$

This is an established result in Vershynin [2010]. By Lemma D.3, bounding $\left\|\|W_m^* \mathcal{X}_t\|_2^2\right\|_{\psi_1}$ would be sufficient, and we start from bounding $\mathbb{E}\left(\exp\left\{\lambda\left(W_m^*\right)_{i\cdot} \mathcal{X}_t\right\}\right)$ for any $\lambda \in \mathbb{R}$. Recall that $\mathcal{X}_t = \Psi_j^{(p)} \varepsilon_{t-j-1}$, with $\Psi_j^{(p)}$ defined as in (65), we can write

$$\left(W_m^*\right)_{i\cdot} \mathcal{X}_t = \left(W_m^*\right)_{i\cdot} \sum_{k=0}^{\infty} \Psi_k^{(p)} \epsilon_{t-k-1} = \lim_{N \to \infty} \sum_{k=0}^{N} \left(W_m^*\right)_{i\cdot} \Psi_k^{(p)} \epsilon_{t-k-1},$$

$$\exp\left\{\lambda\left(W_m^*\right)_{i\cdot} \mathcal{X}_t\right\} = \lim_{N \to \infty} \exp\left\{\lambda \sum_{k=0}^{N} \left(W_m^*\right)_{i\cdot} \Psi_k^{(p)} \epsilon_{t-k-1}\right\},$$

63

and

$$\exp\left\{\lambda \sum_{k=0}^{N} (W_m^*)_{i\cdot} \Psi_k^{(p)} \epsilon_{t-k-1}\right\} \leq \exp\left\{|\lambda| \sum_{k=0}^{\infty} \|(W_m^*)_{i\cdot}\|_2 \, \tilde{\alpha}_k \, \|\epsilon_{t-k-1}\|_2\right\},$$

where $\tilde{\alpha}_k$ is defined as $\left\|\Psi_k^{(p)}\right\|_2$. The relationship between $\tilde{\alpha}_k$ and $\alpha_k = \|\Psi_k\|_2$ can be established as follows:

$$\tilde{\alpha}_k = \sup_{\|u\|_2=1} \left\|\Psi_k^{(p)} u\right\|_2 = \sup_{\|u\|_2=1} \left(\sum_{n=0}^{(p-1)\wedge j} \|\Psi_{k-n} u\|_2^2\right)^{\frac{1}{2}} \leq \left(\sum_{n=0}^{p-1} \alpha_{k-n}^2\right)^{\frac{1}{2}}, \tag{75}$$

if we define $\alpha_i = 0$ when $i < 0$. We now prove that $\exp\left\{|\lambda| \sum_{k=0}^{\infty} \|(W_m^*)_{i\cdot}\|_2 \, \tilde{\alpha}_k \, \|\epsilon_{t-k}\|_2\right\}$ is integrable so that we can use Dominated Convergence Theorem. Since ϵ_{ti}'s are all independent sub-Gaussian random variables with parameter τ,

$$\left\|\|\epsilon_t\|_2\right\|_{\psi_2} \leq \left\|\|\epsilon_t\|_2^2\right\|_{\psi_1}^{\frac{1}{2}} \leq \left(M\|\epsilon_{ti}^2\|_{\psi_1}\right)^{\frac{1}{2}} \leq C\sqrt{M}\tau, \tag{76}$$

where the second inequality is due to Minkowski's inequality. Thus,

$$\mathbb{E}\left(\exp\left\{|\lambda| \sum_{k=0}^{\infty} \|(W_m^*)_{i\cdot}\|_2 \, \tilde{\alpha}_k \, \|\epsilon_{t-k}\|_2\right\}\right)$$

$$= \lim_{N\to\infty} \mathbb{E}\left(\exp\left\{|\lambda| \sum_{k=0}^{N} \|(W_m^*)_{i\cdot}\|_2 \, \tilde{\alpha}_k \, \|\epsilon_{t-k}\|_2\right\}\right)$$

$$\leq \lim_{N\to\infty} \exp\left\{CM\lambda^2 \|(W_m^*)_{i\cdot}\|_2^2 \sum_{k=0}^{N} \tilde{\alpha}_k^2\right\} \leq \exp\left\{CM\lambda^2\right\},$$

where the first equality is due to Monotone Convergence Theorem, and the last line is due to (62) and the fact that

$$\sum_{k=0}^{N} \tilde{\alpha}_k^2 \leq \sum_{k=0}^{N}\sum_{n=0}^{p-1} \alpha_{k-n}^2 \leq p\sum_{k=0}^{N} \alpha_k^2 \leq \beta^2.$$

Therefore, by Dominated Convergence Theorem,

$$\mathbb{E}\left(\exp\left\{\lambda \, (W_m^*)_{i\cdot} \, \mathcal{X}_t\right\}\right)$$

$$= \lim_{N\to\infty} \mathbb{E}\left(\exp\left\{\lambda \sum_{k=0}^{N} (W_m^*)_{i\cdot} \Psi_k^{(p)} \epsilon_{t-k}\right\}\right)$$

$$\leq \exp\left\{C\lambda^2 \|(W_m^*)_{i\cdot}\|_2^2 \sum_{k=0}^{\infty} \tilde{\alpha}_k^2\right\}$$

$$= \exp\left\{C\lambda^2\right\}.$$

By Lemma D.3, $\|(W_m^*)_{i\cdot} \mathcal{X}_t\|_{\psi_2} \leq C$, and

$$\left\|\|W_m^* \mathcal{X}_t\|_2^2\right\|_{\psi_1} \leq \sum_{i=1}^{d_m} \left\|((W_m^*)_{i\cdot} \mathcal{X}_t)^2\right\|_{\psi_1} \leq 2\sum_{i=1}^{d_m} \|(W_m^*)_{i\cdot} \mathcal{X}_t\|_{\psi_2}^2 \leq C.$$

Thus

$$\mathbb{E}\left(\|W_m^* \mathcal{X}_t\|_2^p\right)^{\frac{1}{p}} \leq C\sqrt{p}.$$

$\square$

Proof of Lemma 5.2. Recall that $\mathcal{X}_t = \sum_{j=0}^{\infty} \Psi_j^{(p)} \epsilon_{t-j-1}$, where $\Psi_j^{(p)}$ is defined in (65). Similar from the proof of Lemma 5.1, for any positive integer m, we can write down $\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{X}_t^\top B \mathcal{X}_t$ as the following:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{X}_t^\top B \mathcal{X}_t &= \frac{1}{T}\sum_{t=0}^{T-1}\left(\sum_{j=0}^{\infty}\Psi_j^{(p)}\epsilon_{t-j-1}\right)^\top B \left(\sum_{j=0}^{\infty}\Psi_j^{(p)}\epsilon_{t-j-1}\right)\\
&= \frac{1}{T}\sum_{t=0}^{T-1}\left(\sum_{j=0}^{t+m-1}\Psi_j^{(p)}\epsilon_{t-j-1}\right)^\top B \left(\sum_{j=0}^{t+m-1}\Psi_j^{(p)}\epsilon_{t-j-1}\right)\\
&\quad + \frac{1}{T}\sum_{t=0}^{T-1}\left(\sum_{j=t+m}^{\infty}\Psi_j^{(p)}\epsilon_{t-j-1}\right)^\top B \left(\sum_{j=t+m}^{\infty}\Psi_j^{(p)}\epsilon_{t-j-1}\right)\\
&\quad + \frac{2}{T}\sum_{t=0}^{T-1}\left(\sum_{j=0}^{t+m-1}\Psi_j^{(p)}\epsilon_{t-j-1}\right)^\top B \left(\sum_{j=t+m}^{\infty}\Psi_j^{(p)}\epsilon_{t-j-1}\right)\\
&\triangleq E_1 + E_2 + E_3.
\end{aligned}$$

Then we can bound each E_i from its expectation separately, and m will be chosen to be sufficiently large later.

(1) Bounding $E_1 - \mathbb{E}(E_1)$

Let $\Theta^{(t)} \in \mathbb{R}^{pM \times (T+m)M}$ and $\tilde{\epsilon} \in \mathbb{R}^{(T+m)M}$ be defined as

$$\Theta^{(t)} = \begin{pmatrix} \Psi_{t+m-1}^{(p)} & \cdots & \Psi_0^{(p)} & 0 & \cdots & 0 \end{pmatrix},$$

$$\tilde{\epsilon} = \begin{pmatrix} \epsilon_{-m}^\top & \cdots & \epsilon_{T-1}^\top \end{pmatrix}^\top.$$

Then $E_1 = \tilde{\epsilon}^\top \left(\frac{1}{T}\sum_{t=0}^{T-1}\Theta^{(t)\top}B\Theta^{(t)}\right)\tilde{\epsilon}$, and by Lemma C.4 we only need to bound the operator norm and Frobenius norm of $\frac{1}{T}\sum_{t=0}^{T-1}\Theta^{(t)\top}B\Theta^{(t)}$.

i. Bounding $\left\|\frac{1}{T}\sum_{t=0}^{T-1}\Theta^{(t)\top}B\Theta^{(t)}\right\|_2$

For any unit vector $u, v \in \mathbb{R}^{(t+m)M}$,

$$
\begin{aligned}
u^\top \frac{1}{T} \sum_{t=0}^{T-1} \Theta^{(t)\top} B \Theta^{(t)} v =& \frac{1}{T} \sum_{t=0}^{T-1} \sum_{i,j=1}^{t+m} u^{(i)\top} \Psi_{t+m-1}^{(p)\top} B \Psi_{t+m-j}^{(p)} v^{(j)} \\
=& \frac{1}{T} \sum_{i,j=1}^{T+m-1} u^{(i)\top} \left[\sum_{t=(i \vee j-m) \vee 0}^{T-1} \Psi_{t+m-1}^{(p)\top} B \Psi_{t+m-j}^{(p)} \right] v^{(j)} \\
\leq& \frac{1}{T} \sum_{i,j=1}^{T+m-1} \|u^{(i)}\|_2 \|v^{(j)}\|_2 \|B\|_2 \sum_{l=0}^{\infty} \left\| \Psi_{|i-j|+l}^{(p)} \right\|_2 \left\| \Psi_l^{(p)} \right\|_2,
\end{aligned}
$$

where $u^{(i)} = (u_{(i-1)M+1}, \ldots, u_{iM})$. Let $\tilde{\alpha}_i = \left\| \Psi_i^{(p)} \right\|_2$, and $\Gamma \in \mathbb{R}^{(t+m) \times (t+m)}$ be defined as $\Gamma_{ij} = \sum_{k=0}^{\infty} \tilde{\alpha}_{|i-j|+k} \tilde{\alpha}_k$, then

$$
\begin{aligned}
u^\top \frac{1}{T} \sum_{t=0}^{T-1} \Theta^{(t)\top} B \Theta^{(t)} v \leq& \frac{\|B\|_2}{T} (\|u^{(1)}\|_2, \ldots, \|u^{(t+m)}\|_2) \Gamma \begin{pmatrix} \|v^{(1)}\|_2 \\ \vdots \\ \|v^{(t+m)}\|_2 \end{pmatrix} \\
\leq& \frac{\|B\|_2 \Lambda_{\max}(\Gamma)}{T}.
\end{aligned}
$$

Thus we only need to bound $\Lambda_{\max}(\Gamma)$. Applying Lemma C.5, the largest eigenvalue of Toeplitz matrix Γ can be bounded by

$$
\begin{aligned}
\Lambda_{\max}(\Gamma) \leq& \operatorname{ess\,sup}_{\lambda} \left| \sum_{l=-\infty}^{\infty} \sum_{j=0}^{\infty} \tilde{\alpha}_{|l|+j} \tilde{\alpha}_j e^{il\lambda} \right| \\
\leq& \left| \sum_{l=-\infty}^{\infty} \sum_{j=0}^{\infty} \tilde{\alpha}_{|l|+j} \tilde{\alpha}_j \right| \\
\leq& 2 \sum_{l=0}^{\infty} \left(\sum_{j=0}^{\infty} \tilde{\alpha}_{l+j}^2 \right)^{\frac{1}{2}} \left(\sum_{j=0}^{\infty} \tilde{\alpha}_j^2 \right)^{\frac{1}{2}}.
\end{aligned}
$$

where the third inequality is due to Cauchey-Schwartz inequality. Due to (75), we can further obtain

$$
\begin{aligned}
\Lambda_{\max}(\Gamma) \leq& 2 \sum_{l=0}^{\infty} \left(\sum_{j=0}^{\infty} \sum_{n=0}^{p-1} \alpha_{l+j-n}^2 \right)^{\frac{1}{2}} \left(\sum_{j=0}^{\infty} \sum_{n=0}^{p-1} \alpha_{j-n}^2 \right)^{\frac{1}{2}} \\
\leq& 2p \left(\sum_{i=0}^{\infty} \alpha_{1-p+i}^2 \right)^{\frac{1}{2}} \sum_{l=0}^{\infty} \left(\sum_{i=0}^{\infty} \alpha_{l+1-p+i}^2 \right)^{\frac{1}{2}} \leq C(\beta).
\end{aligned}
$$

and we define $\alpha_i = 0$ when $i < 0$ for convenience. Therefore,

$$
\left\| \frac{1}{T} \sum_{t=0}^{T-1} \Theta^{(t)\top} B \Theta^{(t)} \right\|_2 \leq \frac{C\|B\|_2}{T}.
$$

66

ii. Bounding $\left\|\frac{1}{T}\sum_{t=0}^{T-1}\Theta^{(t)\top}B\Theta^{(t)}\right\|_F^2$

First note that

$$\left\|\frac{1}{T}\sum_{t=0}^{T-1}\Theta^{(t)\top}B\Theta^{(t)}\right\|_F^2 \le \frac{1}{T^2}\sum_{s,t=0}^{T-1}\left|\text{tr}\left(\Theta^{(s)\top}B\Theta^{(s)}\Theta^{(t)\top}B\Theta^{(t)}\right)\right|,$$

and if we write $B = P^\top\Lambda P$ with orthogonal P and diagonal Λ (since B is symmetric),

$$\left|\text{tr}\left(\Theta^{(s)\top}B\Theta^{(s)}\Theta^{(t)\top}B\Theta^{(t)}\right)\right|$$
$$= \left|\text{tr}\left(P\Theta^{(s)}\Theta^{(t)\top}B\Theta^{(t)}\Theta^{(s)\top}P^\top\Lambda\right)\right|$$
$$\le\|B\|_{\text{tr}}\left\|\Theta^{(s)}\Theta^{(t)\top}B\Theta^{(t)}\Theta^{(s)\top}\right\|_2$$
$$\le\|B\|_{\text{tr}}\|B\|_2\left\|\Theta^{(s)}\Theta^{(t)\top}\right\|_2^2.$$

Meanwhile, due to that $\tilde{\alpha}_i = \left\|\Psi_i^{(p)}\right\|_2$ and (75),

$$\sum_{s,t=0}^{T-1}\left\|\Theta^{(s)}\Theta^{(t)\top}\right\|_2^2 = \sum_{s,t=0}^{T-1}\left\|\sum_{i=1}^{t\wedge s+m}\Psi_{t+m-i}^{(p)}\Psi_{s+m-i}^{(p)}\right\|_2^2$$
$$\le \sum_{s,t=0}^{T-1}\left(\sum_{i=1}^{t\wedge s+m}\tilde{\alpha}_{t+m-i}\tilde{\alpha}_{s+m-i}\right)^2$$
$$= \sum_{s,t=0}^{T-1}\left(\sum_{i=0}^{(t\wedge s)+m-1}\tilde{\alpha}_i\tilde{\alpha}_{|t-s|+i}\right)^2$$
$$\le \sum_{s,t=0}^{T-1}\left(p\sum_{i=0}^{\infty}\alpha_i^2\right)\left(p\sum_{i=1-p}^{\infty}\alpha_{|t-s|+i}^2\right).$$

Note that $\sum_{i=0}^{\infty}\left(\sum_{j=0}^{\infty}\alpha_{i+j}^2\right)^{\frac{1}{2}} \le \beta$,

$$\sum_{s,t=0}^{T-1}\left\|\Theta^{(s)}\Theta^{(t)\top}\right\|_2^2 \le Cp^2\sum_{s,t=0}^{T-1}\left(\sum_{i=1-p}^{\infty}\alpha_{|t-s|+i}^2\right)$$
$$\le Cp^2\sum_{l=0}^{T-1}2(T-l)\left(\sum_{i=1-p}^{\infty}\alpha_{l+i}^2\right)$$
$$\le CT\sum_{l=0}^{\infty}\left(\sum_{i=0}^{\infty}\alpha_{l+i}^2\right)$$
$$\le CT\left(\sum_{l=0}^{\infty}\left(\sum_{i=0}^{\infty}\alpha_{l+i}^2\right)^{\frac{1}{2}}\right)^2 \le CT,$$

67

where the fourth line is due to Cauchey-Schwartz inequality. Therefore,

$$\left\|\frac{1}{T}\sum_{t=0}^{T-1}\Theta^{(t)\top}B\Theta^{(t)}\right\|_F^2 \leq \frac{C\|B\|_2\|B\|_{\mathrm{tr}}}{T}.$$

Now we apply Lemma C.4, and arrive at

$$\mathbb{P}\left(|E_1 - \mathbb{E}(E_1)| > \delta\right) \leq 2\exp\left\{-cT\min\left\{\frac{\delta}{\|B\|_2}, \frac{\delta^2}{\|B\|_2\|B\|_{\mathrm{tr}}}\right\}\right\}.$$

(2) Bounding $E_2 - \mathbb{E}(E_2)$

We will show that $|E_2 - \mathbb{E}(E_2)|$ vanishes when m is large enough. First we bound $\|E_2\|_{\psi_1}$. Since

$$|E_2| \leq \frac{1}{T}\sum_{t=0}^{T-1}\|B\|_2\left(\sum_{j=t+m}^{\infty}\tilde{\alpha}_j\|\epsilon_{t-j-1}\|_2\right)^2,$$

by (75) and (76),

$$\|E_2\|_{\psi_1} \leq \frac{2}{T}\sum_{t=0}^{T-1}\|B\|_2\left(\sum_{j=t+m}^{\infty}\tilde{\alpha}_j\left\|\|\epsilon_{t-j-1}\|_2\right\|_{\psi_2}\right)^2$$

$$\leq \frac{CM\|B\|_2}{T}\sum_{t=0}^{T-1}\left(\sum_{j=t+m}^{\infty}\tilde{\alpha}_j\right)^2$$

$$\leq CM\|B\|_2\left(\sum_{j=m}^{\infty}\tilde{\alpha}_j\right)^2$$

$$\leq CM\|B\|_2 p^2\left(\sum_{j=m-p}^{\infty}\alpha_j\right)^2.$$

Meanwhile,

$$|\mathbb{E}(E_2)| = \left|\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{tr}\left(B\sum_{j=t+m}^{\infty}\Psi_j^{(p)}\Psi_j^{(p)\top}\right)\right|$$

$$\leq \left|\frac{1}{T}\sum_{t=0}^{T-1}\|B\|_{\mathrm{tr}}\sum_{j=t+m}^{\infty}\tilde{\alpha}_j^2\right|$$

$$\leq p\|B\|_{\mathrm{tr}}\sum_{j=m-p}^{\infty}\alpha_j^2.$$

For any $\delta > 0$, let m be sufficiently large such that $\sum_{j=m-p}^{\infty}\alpha_j^2 < \frac{\delta}{2p\|B\|_{\mathrm{tr}}}$, $\|E_2\|_{\psi_1} \leq \frac{C\|B\|_2}{T}$, then by tail bound of sub-exponential random variable (see Vershynin [2010]),

$$\mathbb{P}\left(|E_2 - \mathbb{E}(E_2)| > \delta\right) \leq C\exp\left\{-\frac{c\delta T}{\|B\|_2}\right\}.$$

68

(3) Bounding $E_3 - \mathbb{E}(E_3)$

One can show that

$$|E_3| \leq \frac{2\|B\|_2}{T} \sum_{t=0}^{T-1} \sum_{j=t+m}^{\infty} \tilde{\alpha}_j \|\epsilon_{t-j-1}\|_2 \sum_{j=0}^{\infty} \tilde{\alpha}_j \|\epsilon_{t-j-1}\|_2,$$

and

$$\left\| \sum_{j=n}^{\infty} \tilde{\alpha}_j \|\epsilon_{t-j-1}\|_2 \right\|_{\psi_2} \leq C\sqrt{M}\tau \sum_{j=n}^{\infty} \tilde{\alpha}_j \leq Cp\sqrt{M}\tau \sum_{j=n-p}^{\infty} \alpha_j.$$

Thus

$$\begin{aligned}
\|E_3\|_{\psi_1} &\leq \frac{4\|B\|_2}{T} \sum_{t=0}^{T-1} \left\| \sum_{j=t+m}^{\infty} \tilde{\alpha}_j \|\epsilon_{t-j-1}\|_2 \right\|_{\psi_2} \left\| \sum_{j=0}^{\infty} \tilde{\alpha}_j \|\epsilon_{t-j-1}\|_2 \right\|_{\psi_2} \\
&\leq C\|B\|_2 \sqrt{M} p\tau \left(\sum_{j=m-p}^{\infty} \alpha_j \right) \left(\sum_{j=0}^{\infty} \alpha_j \right) \\
&\leq C\|B\|_2 \sqrt{M} \sum_{j=m-p}^{\infty} \alpha_j.
\end{aligned}$$

The first line is due to the following fact: For any two sub-Gaussian random variables X and Y, $\|XY\|_{\psi_1} \leq 2\|X\|_{\psi_2}\|Y\|_{\psi_2}$. We can prove this in the following:

$$\begin{aligned}
\sup_{q \geq 1} q^{-1} \left(\mathbb{E}|XY|^q \right)^{\frac{1}{q}} &\leq \sup_{q \geq 1} q^{-1} \left(\mathbb{E}|X|^{2q} \right)^{\frac{1}{2q}} \left(\mathbb{E}|Y|^{2q} \right)^{\frac{1}{2q}} \\
&\leq 2 \sup_{q \geq 1} q^{-\frac{1}{2}} \left(\mathbb{E}|X|^q \right)^{\frac{1}{q}} \sup_{q \geq 1} q^{-\frac{1}{2}} \left(\mathbb{E}|Y|^q \right)^{\frac{1}{q}} \\
&= 2\|X\|_{\psi_2}\|Y\|_{\psi_2},
\end{aligned}$$

where the first line applies Cauchey-Schwartz inequality. Thus, with large enough m, $\|E_3\|_{\psi_1} \leq \frac{\|B\|_2}{T}$. Also, $\mathbb{E}(E_3) = 0$, therefore implies the same bound for $E_3 - \mathbb{E}(E_3)$ as the one for $E_2 - \mathbb{E}(E_2)$:

$$\mathbb{P}\left(|E_3 - \mathbb{E}(E_3)| > \delta \right) \leq C \exp\left\{ -\frac{c\delta T}{\|B\|_2} \right\}.$$

In conclusion, for any $\delta > 0$, if we choose some m accordingly,

$$\begin{aligned}
\mathbb{P}\left(\left| \frac{1}{T} \sum_{t=0}^{T-1} \mathcal{X}_t^\top B \mathcal{X}_t - \text{tr}(B\Upsilon) \right| > \delta \right) \\
\leq \sum_{i=1}^{3} \mathbb{P}\left(|E_i - \mathbb{E}(E_i)| > \frac{\delta}{3} \right) \\
\leq C \exp\left\{ -cT \min\left\{ \frac{\delta}{\|B\|_2}, \frac{\delta^2}{\|B\|_2 \|B\|_{\text{tr}}} \right\} \right\}.
\end{aligned}$$

$\square$

Proof of Lemma A.1. Here we apply some results in Basu et al. [2015] with a little change in notation. These results simplifies the original problem to finding a upper bound for $\left|v^\top(H-\Upsilon)v\right|$ with any fixed unit vector v. Specifically, the following lemmas are useful:

Lemma D.4. *For any $J \subset \{1, \cdots, pM\}$, and $\kappa > 0$,*

$$\mathcal{C}(J,\kappa) \cap \{v \in \mathbb{R}^{pM} : \|v\|_2 \le 1\} \subset (\kappa + 2)cl\{conv\{\mathcal{K}(|J|)\}\},$$

where $\mathcal{K}(l) = \{v \in \mathbb{R}^{pM} : \|v\|_0 \le l, \|v\|_2 \le 1\}$ for any positive integer l.

Lemma D.5.

$$\sup_{v \in cl\{conv(\mathcal{K}(l))\}} \left|v^\top Dv\right| \le 3 \sup_{v \in \mathcal{K}(2l)} \left|v^\top Dv\right|.$$

Lemma D.6. *Consider a symmetric matrix $D \in \mathbb{R}^{pM \times pM}$. If for any vector $v \in \mathbb{R}^{pM}$ with $\|v\|_2 \le 1$, and any $\eta \ge 0$,*

$$\mathbb{P}\left(\left|v^\top Dv\right| > \eta\right) \le c_1 \exp\left\{-c_2 T \min\left\{\eta, \eta^2\right\}\right\},$$

then for any integer $l \ge 1$,

$$\mathbb{P}\left(\sup_{v \in \mathcal{K}(l)} \left|v^\top Dv\right| > \eta\right) \le c_1 \exp\left\{-c_2 T \min\left\{\eta, \eta^2\right\} + l \min\left\{\log(pM), \log(21epM/l)\right\}\right\}.$$

By Lemma D.4 and Lemma D.5,

$$\sup\left\{\left|v^\top(H-\Upsilon)v\right| : v \in \mathcal{C}(J,\kappa), \|v\|_2 \le 1\right\}$$
$$\le \sup\left\{\left|v^\top(H-\Upsilon)v\right| : v \in (\kappa+2)cl\{conv\{\mathcal{K}(|J|)\}\}\right\}$$
$$\le 3(\kappa+2)^2 \sup\left\{\left|v^\top(H-\Upsilon)v\right| : v \in \mathcal{K}(2|J|)\right\}.$$

For any unit vector $v \in \mathbb{R}^{pM}$,

$$v^\top(H-\Upsilon)v = \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{X}_t^\top vv^\top \mathcal{X}_t - \mathrm{tr}\left(vv^\top\Upsilon\right),$$

Thus $\left|v^\top(H-\Upsilon)v\right|$ can be bounded by Lemma 5.2.

$$\left\|vv^\top\right\|_{\mathrm{tr}} = \left\|vv^\top\right\|_2 = \|v\|_2^2 = 1,$$

which implies

$$\mathbb{P}\left(\left|v^\top(H-\Upsilon)v\right| > \eta\right) \le c_1 \exp\{-c_2 T \min\{\eta, \eta^2\}\}.$$

By Lemma D.6, when $|J|\log pM \le C(\eta)T$,

$$\sup\left\{\left|v^\top(H-\Upsilon)v\right| : v \in \mathcal{K}(2|J|)\right\} \le \eta,$$

with probability at least $1 - c_1 \exp\{-c_2 T \min\{\eta, \eta^2\}\}$. Let $\eta = [6(\kappa+2)^2]^{-1} \Lambda_{\min}(\Upsilon) \geq C(\kappa, \beta)$, then

$$
\begin{aligned}
&\inf\left\{ v^\top H v : v \in \mathcal{C}(J, \kappa), \|v\|_2 \leq 1 \right\} \\
&\geq \Lambda_{\min}(\Upsilon) - \sup\left\{ \left| v^\top (H - \Upsilon) v \right| : v \in \mathcal{C}(J, \kappa), \|v\|_2 \leq 1 \right\} \\
&\geq \frac{1}{2} \Lambda_{\min}(\Upsilon) \geq C(\beta),
\end{aligned}
$$

with probability at least $1 - c_1 \exp\{-c_2 T\}$, when $|J| \log pM \leq C(\kappa, \beta) T$, and c_2 depends on κ and β. Here we apply Lemma 5.4 to lower bound the eigenvalues of Υ. $\qquad \square$

E   Proof of Lemma D.2, 2.1, A.2, and A.3

Proof of Lemma D.2. Recall that $f_l(z) = f_*(g_l(z))$, with $f_*(z) = \int_{-\infty}^{z-\frac{1}{2}} \phi(z) dz$, $g_1(z) = (\|Z + \mu\|_2 - r - \varepsilon)/\varepsilon$, and $g_2(z) = (\|Z + \mu\|_2 - r + 2\varepsilon)/\varepsilon$. In order to bound the partial derivatives of composite function, we apply the following lemma which is a direct result of Proposition 1 and 2 in Hardy [2006].

Lemma E.1. *Suppose univariate function f and g: $\mathbb{R}^n \to \mathbb{R}$ have derivatives and partial derivatives of orders up to k, then $\forall \{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$,*

$$
\frac{\partial^k}{\partial x_{i_1} \cdots \partial x_{i_k}} f(g(x)) = \sum_{\pi \in \Pi(k)} f^{(|\pi|)}(g(x)) \prod_{B \in \pi} \frac{\partial^{|B|} g(x)}{\prod_{j \in B} \partial x_{i_j}},
$$

where $\Pi(k)$ is the set of partitions for $\{1, \cdots, k\}$, and $B \in \pi$ is a block in π. Formally,

$$
\Pi(k) = \{\{B_1, B_2, \cdots, B_n\} : B_i \cap B_j = \emptyset, \cup_i B_i = \{1, 2, \cdots, k\}\}.
$$

By Lemma E.1, we can write out the kth order partial derivatives of f_l:

$$
\frac{\partial^k}{\partial z_{i_1} \cdots \partial z_{i_k}} f_l(z) = \sum_{\pi \in \Pi(k)} f_*^{(|\pi|)}(g_l(z)) \prod_{B \in \pi} \frac{\partial^{|B|} g_l(z)}{\prod_{j \in B} \partial z_{i_j}}.
$$

Moreover, we can also write $g_l(z)$ as a composite function $\varphi_l(\psi(z))$, with $\varphi_1(x) = \frac{\sqrt{x} - r - \varepsilon}{\varepsilon}$, $\varphi_2(x) = \frac{\sqrt{x} - r + 2\varepsilon}{\varepsilon}$, and $\psi(z) = \|z + \mu\|_2^2$. Then applying Lemma E.1 on $g_l(z)$ gives us

$$
\frac{\partial^n}{\partial z_{i_1} \cdots \partial z_{i_n}} g_l(z) = \sum_{\pi \in \Pi(n)} \varphi_l^{(|\pi|)}(\psi(z)) \prod_{B \in \pi} \frac{\partial^{|B|} \psi(z)}{\prod_{j \in B} \partial z_{i_j}}. \tag{77}
$$

Note that

$$
\frac{\partial^{|B|} \psi(z)}{\prod_{j \in B} \partial z_{i_j}} = \begin{cases} z_{i_j} + \mu_{i_j} & \text{if } B = \{j\} \text{ for any } j \\ 1(i_j = i_l) & \text{if } B = \{j, l\} \text{ for any } j, l \\ 0 & \text{if } |B| > 2, \end{cases}
$$

71

which means that we only need to consider the partitions with all blocks of size 1 or 2, when calculating the partial derivative of $g_l(z)$ using (77). Also note that we need partitions for blocks within an original partition π, we define the following partition set $\mathcal{C}(\pi)$ for any partition $\pi = \{B_1, \ldots, B_n\}$ of size n:

$$\mathcal{C}(\pi) = \{\cup_{i=1}^n \tilde{\pi}_i : \tilde{\pi}_i \in \Pi(B_i) s.t. \forall C \in \tilde{\pi}_i, |C| \leq 2\}.$$

This set $\mathcal{C}(\pi)$ include the unions of partitions for each block B_i within π, and each block within the partition of B_i has size bounded by 2. Let $S(\tilde{\pi}) = \{i : \{i\} \in \tilde{\pi}\}$, and $P(\tilde{\pi}) = \{\{i,j\} : \{i,j\} \in \tilde{\pi}\}$, then the partial derivative of $f_l(z)$ can be expanded as

$$\frac{\partial^k}{\partial z_{i_1} \cdots \partial z_{i_k}} f_l(z) = \sum_{\substack{\pi \in \Pi(n) \\ \tilde{\pi} \in \mathcal{C}(\pi)}} f_*^{|\pi|}(g_l(z)) C(\pi, \tilde{\pi}) \frac{\Pi_{j \in S(\tilde{\pi})}(z_{i_j} + \mu_{i_j}) \Pi_{\{j,l\} \in P(\tilde{\pi})} 1(i_j = i_l)}{\varepsilon^{|\pi|} \|z + \mu\|_2^{2|\tilde{\pi}| - |\pi|}}, \tag{78}$$

where we apply the fact that $\varphi_l^{(k)}(x) = \frac{C(k)}{\varepsilon x^{k - \frac{1}{2}}}$. For each fixed $\pi \in \Pi(k)$ and $\tilde{\pi} \in \mathcal{C}(\pi)$,

$$\left| \sum_{1 \leq i_1, \cdots, i_k \leq d} y_{i_1} \cdots y_{i_k} \Pi_{j \in S(\tilde{\pi})}(z_{i_j} + \mu_{i_j}) \Pi_{\{j,l\} \in P(\tilde{\pi})} 1(i_j = i_l) \right|$$

$$= \left| \left(y^\top (z + \mu) \right)^{|S(\tilde{\pi})|} \|y\|_2^{2|P(\tilde{\pi})|} \right| \leq \|y\|_2^k \|z + \mu\|_2^{|S(\tilde{\pi})|},$$

then combine this with (78), we have

$$\left| \sum_{1 \leq i_1, \cdots, i_k \leq d} y_{i_1} \cdots y_{i_k} \frac{\partial^k}{\partial z_{i_1} \cdots \partial z_{i_k}} f_l(z) \right| \leq \sum_{\substack{\pi \in \Pi(n) \\ \tilde{\pi} \in \mathcal{C}(\pi)}} \frac{f_*^{(|\pi|)}(g_l(z)) C(\pi, \tilde{\pi}) \|y\|_2^k}{\varepsilon^{|\pi|} \|z + \mu\|_2^{k - |\pi|}}.$$

In addition, note that $f_*^{(k)}(x) = \phi^{(k-1)}(x - \frac{1}{2}) = 0$ when $x \leq 0$ or $x \geq 1$, and is bounded on $(0,1)$. Thus we only have to consider $\|z + \mu\|_2 > r + \varepsilon$ when $l = 1$ and $\|z + \mu\|_2 > r - 2\varepsilon$ when $l = 2$. If $r > 3\varepsilon$ and $l = 2$, $\|z + \mu\|_2 > r - 2\varepsilon > \varepsilon$. Therefore,

$$\left| \sum_{1 \leq i_1, \cdots, i_k \leq d} y^{(i_1)} \cdots y^{(i_k)} \frac{\partial^k}{\partial z^{(i_1)} \cdots \partial z^{(i_k)}} f_l(z) \right|$$

$$\leq \sum_{\pi \in \Pi(k)} \sum_{(S_i, P_i)_{i=1}^{|\pi|} \in \mathcal{C}(\pi)} \frac{C(|\pi|) \|y\|^k}{\varepsilon^k} \leq C(k) \varepsilon^{-k} \|y\|^k.$$

$\square$

Proof of Lemma 2.1. Note that

$$w_m^* = \Upsilon_{D_m^c, D_m^c}^{-1} \Upsilon_{D_m^c, D_m} = -(\Upsilon^{-1})_{D_m^c, D_m} \left[(\Upsilon^{-1})_{D_m, D_m} \right]^{-1}.$$

When A^* is symmetric, $\Upsilon^{-1} = I - (A^*)^2$, thus

$$w_m^* = \left((A^*)^2\right)_{D_m^c, D_m} \left[I - \left((A^*)^2\right)_{D_m, D_m}\right]^{-1} \in \mathbb{R}^{(M-d_m) \times d_m}.$$

It is clear that

$$s_m = \|w_m^*\|_0 \le d_m \left|\{i : (w_m^*)_{i\cdot} \ne 0\}\right| \le d_m \left|\{i : [(A^*)^2]_{i, D_m} \ne 0\}\right|.$$

Let $R_m = \left|\{i : [(A^*)^2]_{i, D_m} \ne 0\}\right|$ and $C_m = \{j : A_{j, D_m}^* \ne 0\}$, then

$$|C_m| \le d_m \max_{1 \le i \le M} \|a_i^*\|_0$$

and

$$R_m \subset \{i : \operatorname{supp}(A_{i\cdot}^*) \cap C_m \ne \emptyset\}.$$

Therefore,

$$s_m \le d_m |R_m| \le d_m \sum_{j \in C_m} |\operatorname{supp}(A_{\cdot j}^*)| \le d_m^2 (\max_{1 \le i \le M} \|a_i^*\|_0)^2.$$

$\square$

Proof of Lemma A.2. Let $Y = (B + \Delta)^{-1}$, then immediately we have $YB - I = -Y\Delta$, which is equivalent to $Y - B^{-1} = -Y\Delta B^{-1}$. Thus the ℓ_2 norm of $Y - B^{-1}$ can be bounded by $\|Y\|_2 \|\Delta\|_2 \|B^{-1}\|_2$. Moreover, note that $\|Y\|_2 \le \|Y - B^{-1}\| + \|B^{-1}\|$, we have

$$\|Y\|_2 \le \|B^{-1}\|_2 + \|Y\|_2 \|\Delta\|_2 \|B^{-1}\|_2,$$

and rearranging terms gives us

$$\|Y\|_2 \le \frac{\|B^{-1}\|_2}{1 - \|B^{-1}\|_2 \|\Delta\|_2}.$$

Therefore,

$$\|Y - B^{-1}\|_2 \le \frac{\|B^{-1}\|_2^2 \|\Delta\|_2}{1 - \|B^{-1}\|_2 \|\Delta\|_2}.$$

$\square$

Proof of Lemma A.3. First note that for any symmetric matrix U, we can write it as $U = P^\top \Lambda P$, with orthogonal matrix P and diagonal matrix Λ. By the definition of trace norm,

$$\|U\|_{\operatorname{tr}} = \operatorname{tr}\left(\sqrt{U^2}\right) = \operatorname{tr}\left(\sqrt{P^\top \Lambda^2 P}\right) = \operatorname{tr}\left(P^\top \sqrt{\Lambda^2} P\right) = \operatorname{tr}\left(\sqrt{\Lambda^2}\right).$$

If we denote the non-zero eigenvalues of U as $\lambda_1, \ldots, \lambda_r$, then

$$\|U\|_{\operatorname{tr}} = \operatorname{tr}\left(\sqrt{\Lambda^2}\right) \le r \max_i |\lambda_i| \le r \|U\|_2.$$

$\square$