Efficient and passive learning of networked dynamical systems driven by non-white exogenous inputs

Harish Doddi[1] **Deepjyoti Deka**[2] **Saurav Talukdar**[3] **Murti Salapaka**[4]
[1]Department of Mechanical Engineering, University of Minnesota Twin Cities
[2]Theoretical Division T-5, Los Alamos National Laboratory
[3]Google Inc.
[4]Department of Electrical & Computer Engineering, University of Minnesota Twin Cities

Abstract

We consider a networked linear dynamical system with p agents/nodes. We study the problem of learning the underlying graph of interactions/dependencies from observations of the nodal trajectories over a time-interval T. We present a regularized non-casual consistent estimator for this problem and analyze its sample complexity over two regimes: (a) where the interval T consists of n i.i.d. observation windows of length T/n (restart and record), and (b) where T is one continuous observation window (consecutive). Using the theory of M-estimators, we show that the estimator recovers the underlying interactions, in either regime, in a time-interval that is logarithmic in the system size p. To the best of our knowledge, this is the first work to analyze the sample complexity of learning linear dynamical systems *driven by unobserved not-white wide-sense stationary (WSS) inputs.*

1 INTRODUCTION

A networked linear dynamical system (LDS) is a network of agents/nodes, each of whose state evolves over time (in discrete or continuous steps) as a *linear* function of an external excitation and the states of its neighboring nodes in the network. The framework of LDS has been used to model dynamics in systems biology (Porreca et al., 2008; Koh et al., 2009), financial markets (Sandefur, 1990), energy (Inchauspe et al., 2015), transportation (Stathopoulos and Karlaftis, 2003) and other critical networks (Ascione et al., 2013; Kroutikova et al., 2007). Learning the dependencies, or topology learning, in a networked LDS is crucial for inference of influence pathways and subsequent control for the corresponding networks. As such, strategies to recover the underlying network structure from nodal time-series in LDS have been researched and can be classified into two categories: active and passive. Active learning involves efficient manipulation or interventions of nodal dependencies and injecting exogenous inputs into the LDS to infer the edges in the network by identifying the resulting changes (Dankers et al., 2015). Passive methods, on the other hand, use historical or streaming time-series of nodal states to infer the underlying topology. Our work falls within the domain of passive structure estimation. Very few works discuss learning such systems but are limited to the asymptotic regime (infinite sample limit). Examples include Materassi and Salapaka (2012); Talukdar et al. (2015, 2020).

Prior Work: Tractable passive topology learning in networked LDS and Vector Auto-Regressive processes (VAR) has been shown using the framework of l_1-regularized regression ((Basu et al., 2015; Loh et al., 2012) and references therein), where the focus is on extending the results from the static Lasso or Graphical Lasso (Tibshirani, 1996; Friedman et al., 2008; Meinshausen and Bühlmann, 2006) to one with correlated samples, by showing that properties such as Restricted strong convexity hold. A similar approach for continuous time stochastic differential equation has been studied in Bento et al. (2010). A graphical model for VAR processes, without performance guarantees, has been proposed in Songsiri et al. (2010). Least squared regression based identification of unstable dynamical systems using a single trajectory has been studied in Simchowitz et al. (2018); Faradonbeh et al. (2018). However, these algorithms rely on the assumption that unobserved exogenous inputs to the system are i.i.d. or white Gaussian noise, or that the exogenous inputs are observed (Fattahi and Sojoudi, 2018; Fattahi et al.,

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

2019).

Temporally correlated inputs: Learning networks excited by temporally correlated inputs is necessary to extend prior work restricted to learning under i.i.d inputs. Examples of systems excited by colored inputs include power grids, thermal networks of buildings Talukdar et al. (2020); Materassi and Innocenti (2010), as well as time-series of air quality, stock market, and magnetoencephalography datasets (Dahlhaus, 2000; Tank et al., 2015; Bach and Jordan, 2004).

On learning networked LDS with temporally correlated but unobserved inputs, Dahlhaus (2000); Jung et al. (2015); Tank et al. (2015) relate the Conditional Independence Graph (CIG) to the support structure of the inverse Power Spectral Density (PSD) of the states. However this is insufficient for true topology recovery as the CIG includes additional edges, Materassi and Salapaka (2012). Talukdar et al. (2020) presents a consistent algorithm for exact recovery in this setting using non-causal regression (Wiener filter), that forms the starting point for the analysis in this article. Quinn et al. (2015) recovers the underlying topology in networked LDS using the framework of directed mutual information. However, these works do not provide for guarantees in the finite sample regime, aside from numerical examples.

The overarching goal of this work is thus to provide a structure learning algorithm for networked LDS driven by temporally-correlated inputs, with guarantees on its performance for finite lengths of state trajectories. We present a *regularized Wiener filter estimator* for this problem and determine the observation window T necessary to guarantee correct estimation over two regimes: (a) where T consists of n i.i.d. observation windows of length T/n (restart and record), and (b) where T is one continuous observation window (consecutive).

The rest of the article is organized as follows. In Section 2.2, we describe the mathematical model of networked linear dynamical system and our consistent learning algorithm. The main results are presented in Theorems 2.1, 2.3 and 2.2. Section 3 contains results on M-estimators used in the proof of our theorems, with sketches of proofs in Section 4. Section 6 contains simulation results, and Section 7 summarizes the article and includes potential extensions and generalizations.

2 MAIN RESULTS

Consider a graph $G = (V, E)$ of $p+1$ nodes in set $V = \{1, ..., p+1\}$ and undirected edge set $E \subset V \times V$. We denote the set of two-hop neighbors in G by set E_M, where $E_M = \{(i,j) | (ij) \in E \text{ or } \exists k, \text{ s.t } (ik), (jk) \in E\}$ (see Fig. 1). Note that $E_M \setminus E$ is the set of 'strict' two-hop neighbors in the graph G, that do not form edges in E. Each node $i \in V$ is associated with a real-valued scalar state variable $\{x_i(k), k \in \mathbb{Z}\}$ that evolves in discrete time [1] according to the following linear dynamical equation:

$$x_i(k+1) = h_{ii}x_i(k) + \sum_{(ij)\in E, j\neq i} h_{ij}x_j(k) + e_i(k), \quad (1)$$

where, $\{e_i(k), k \in \mathbb{Z}\}$, is an exogenous input. While samples of $x_i(k)$ are correlated in time due to the system dynamics, prior work on guaranteed learning of networked LDS include only temporally uncorrelated or white excitations/inputs $e_i(k)$. In this work, we consider $e(k)_{k\in\mathbb{Z}} = [e_1(k)...e_{p+1}(k)]^T$ to be a zero-mean **Wide-Sense Stationary (WSS)** Gaussian process, uncorrelated across nodes, i.e., $\forall k_1, k_2, \tau \in \mathbb{Z}$, $\mathbb{E}[e(k_1)] = \mathbb{E}[e(k_2)] = 0$, and $\mathbb{E}[e(k_1+\tau)e(k_1)^T] = \mathbb{E}[e(k_2+\tau)e(k_2)^T]$. The time-series vector $x(k)_{k\in\mathbb{Z}} = [x_1(k)...x_{p+1}(k)]^T \in \mathbb{R}^{p+1}$ is thus a zero mean jointly Gaussian WSS processes.

The frequency domain representation of Eq. 1 is obtained by taking the Z-transform ($\mathcal{Z}[.]$) on both sides of Eq. 1. Substituting $z = e^{\iota f}$ for a frequency $f \in [0, 2\pi)$, and rearranging for $X_i(f) := \mathcal{Z}[x_i]|_{z=e^{\iota f}}$, we obtain the following:

$$X_i(f) = \sum_{(ij)\in E, j\neq i} H_{ij}(f)X_j(f) + P_i(f), \text{ where,} \quad (2)$$
$$H_{ij}(f) := [\mathcal{Z}[h_{ij}](z - \mathcal{Z}[h_{ii}])^{-1}]|_{z=e^{\iota f}}, (ij) \in E,$$
$$P_i(f) = [\mathcal{Z}[e_i](z - \mathcal{Z}[h_{ii}])^{-1}]|_{z=e^{\iota f}}.$$

Here, $H_{ij}(f)$ is a linear time-invariant filter. Note that each edge $(ij) \in E$ corresponds to non-zero transfer functions H_{ij} and H_{ji}, that may be different.

Given time-series of $x(k)$, we define the lagged correlation matrix $R_x(\tau)$ for $\tau \in \mathbb{Z}$, and its Discrete Time Fourier Transform (DTFT), namely, power spectral density Φ_x, at frequency f as

$$R_x(\tau) = \mathbb{E}(x(\tau)x^T(0)),$$
$$\Phi_x = \mathcal{F}\{R_x(\tau)\} = \lim_{m\to\infty} \sum_{\tau=-m}^{m} R_x(\tau)e^{-\iota f\tau}. \quad (3)$$

Topology Learning: Consider n state trajectories of all the nodes in V for the graph $G = (V, E)$ excited by *unobserved* WSS (temporally correlated) inputs, such that the r^{th} state trajectory (x^r) has N samples. Let $T = n \times N$ be equal to the total observation window. For the r^{th} state trajectory, define the Discrete Fourier Transform (DFT)[2] is

$$X_i^r = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} x_i^r(k)e^{-\iota fk}, X_{\bar{i}}^r = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} x_{\bar{i}}^r(k)e^{-\iota fk}, \quad (4)$$

[1] we discuss extension to continuous time and higher order models in Section 7

[2] computed at frequency $f = \frac{2\pi l}{N}$, $l \in \{0, \cdots, N-1\}$ unless explicitly mentioned

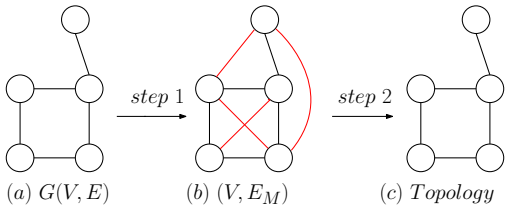

Figure 1: Topology Learning: In step 1, the two-hop neighborhood set E_M is estimated using Lemma 2.1(a). In step 2, strict two-hop neighbors (red colored edges) are eliminated from E_M using Lemma 2.1(b).

where, $r \in \{1, \cdots, n\}$, $x^r_{\bar{i}} = [x^r_1, \cdots, x^r_{i-1}, x^r_{i+1}, \cdots, x^r_{p+1}]^T$. Construct $\mathcal{Y} = [X^1_i, \cdots, X^n_i]^T \in \mathbb{C}^n$ and $\mathcal{X} = [X^1_{\bar{i}}, \cdots, X^n_{\bar{i}}]^T \in \mathbb{C}^{n\times p}$ respectively. We assume that $\mathcal{X}$ and $\mathcal{Y}$ are column-normalized, that is,

$$\frac{\|\mathcal{Y}\|_2}{\sqrt{n}} \leq 1, \quad \frac{\|\mathcal{X}(*, l)\|_2}{\sqrt{n}} \leq 1, \ \forall l \in \{1, \cdots, p\}. \tag{5}$$

Column-normalization is a common data pre-processing step encountered in practice to scale the nodal state trajectories prior to regression. Eq. 5 is not necessary for Eq. 1 to be stable. For any quantity $\beta \in \mathbb{C}$, we use $\Re(\beta)$ and $\Im(\beta)$ to denote its real and imaginary components.

We list the following result from Talukdar et al. (2020) that enables consistent estimation of all edges in E (as described in Figure 1), using nodal state trajectories.

Lemma 2.1 (Talukdar et al. (2020)). *For $i \in V$ of a well-posed networked LDS, the Wiener filter W_i in Eq. 6 satisfies (a) $W_i[j] \neq 0$ if and only if $(ij) \in E_M$ (b) for $(ij) \in E_M$, $\Im(W_i[j]) \neq 0$ if and only if (ij) is a true edge in G.*

$$W_i = \lim_{n,N\to\infty} \arg\min_{\beta\in\mathbb{C}^p} \frac{1}{2n}\|\mathcal{Y} - \mathcal{X}\beta\|_2^2. \tag{6}$$

The proof of Lemma 2.1 (see Talukdar et al. (2020) for details) follows by showing that $W_i[j] = -[\Phi_x^{-1}(i,i)]^{-1}\Phi_x^{-1}(i,j)$. The result then follows from algebraic properties of Φ_x^{-1} (inverse power spectral density) derived from Eq. 2. It is worth noting that, in the time-domain, Eq. 6 is equivalent to a *non-causal* regression of the time-series, termed as "Wiener filter" (Materassi and Salapaka, 2012). This is effectively a *non-causal extension* of the connection between the inverse covariance matrix and the neighborhood regression used in learning static Gaussian graphical models (Friedman et al., 2008; Meinshausen and Bühlmann, 2006; Ravikumar et al., 2008).

For the finite sample regime, we study the problem of estimating edges $\hat{E}$ such that $\mathbb{P}[\hat{E} = E] \geq 1 - \epsilon$ for any user-defined threshold $\epsilon \in (0, .5)$. Estimating Φ_x and then inverting it requires significant amount of data in the high dimensional setting. Instead, we use a regularized version of Eq. 6 as our graph estimator.

2.1 Regularized Wiener Filter Estimator

We propose a Regularized Wiener Filter Estimator $\hat{W}_i$ for a node $i \in V$ as follows:

$$\hat{W}_i(\lambda) = \arg\min_{\beta\in\mathbb{C}^p} \frac{1}{2n}\|\mathcal{Y} - \mathcal{X}\beta\|_2^2 + \lambda\|\beta\|_1, \tag{7}$$

where, $\lambda > 0$ is the regularization parameter. As $\beta \in \mathbb{C}^p$, $\|\beta\|_1$ is equal to the $1,2$-group norm over $[\Re(\beta)\ \Im(\beta)]$. For thresholds τ_1, τ_2, we construct sets

$$\hat{E}_M := \{(ij) | |\hat{W}_i[j]| + |\hat{W}_j[i]| \geq \tau_1\},$$
$$\hat{E} := \{(ij) | (ij) \in \hat{E}_M, |\Im(\hat{W}_i[j])| + |\Im(\hat{W}_j[i])| \geq \tau_2\}. \tag{8}$$

In the remaining of the article, we find sufficient conditions on n, N and λ and fix thresholds such that $\mathbb{P}[\hat{E} = E] \geq 1 - \epsilon$, for given $\epsilon \in (0, 0.5)$.

We consider two settings for the state trajectories:

(i) **Restart & Record (i.i.d):** The n trajectories of length N each are independent. Here, we start recording and then stop recording after collecting N measurements. For the next trajectory, we restart the recording again with a random state initialization and collect the measurements. Hence, it is a process of restart and record, and $\{x^r\}_{r=1}^N$ are i.i.d. trajectories.

(ii) **Consecutive (non i.i.d):** In the second and more realistic setting, we consider the n state trajectories to be consecutive, i.e., $\{x^r\}_{r=1}^N$ correspond to N-length intervals from a single larger trajectory of length $n \times N$. See Figure. 2 for the two settings considered in this article.

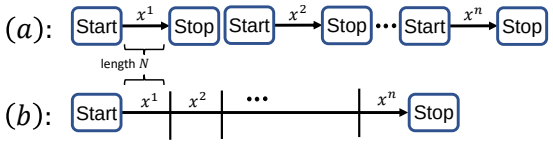

Figure 2: (a) i.i.d trajectories are generated using restart & record (b) a single trajectory is generated for the non i.i.d, consecutive setting.

2.2 Main Results

The error in topology learning (see Eq. 8) arises due to the finite N in computing $\mathcal{X}$ in Eq. 4, as well as the finite n in estimating W_i in Eq. 7. For our analysis, we consider the following non-zero parameters of the LDS

over graph $G = (V, E)$.

$$L = \lambda_{min}(\Phi_x^{-1});\ U = \lambda_{max}(\Phi_x^{-1});\ d = \max_{i \in V} \deg_{E_M}(i);$$
$$C > 0, \delta > 1,\ \text{s.t.} \|R_x(\tau)\|_2 \leq C\delta^{-|\tau|},\ \tau \in \mathbb{Z};$$
$$m_i = \min_{j|(ij)\in E} |\Im(W_i[j])|,\ m = \arg\min_{i \in V} m_i. \quad (9)$$

Note that under persistently exciting inputs, Φ_e is a positive definite matrix almost surely at all frequencies (Materassi and Salapaka, 2012). Further, G is a connected network. Hence, under standard well-posedness assumptions, $(\mathbb{I} - H)$ and Φ_P^{-1} in Eq. 2 are full-ranked and $L \geq 0$. Using norm bounds for matrix products, L and U can be bounded in terms of maximum and minimum eigen-values of $(\mathbb{I}-H^*)(\mathbb{I}-H)$ and Φ_P^{-1}. C, δ relate to the rate of decay of temporal correlation in the system states. Higher values of C and δ^{-1} imply greater temporal correlation. d, the maximum degree due to edges in E_M, is upper-bounded by the square of the maximum nodal degree in $G = (V, E)$.

The following two theorems bound the errors in estimating W_i by regression (Eq. 7), for restart & record (i.i.d), and consecutive (non-i.i.d) trajectories respectively.

Theorem 2.1 (*restart & record-* squared error)**.** *Let* $\epsilon_1 > 0$, $i \in V$, $4\sqrt{\frac{3\log(8p/\epsilon_1)}{nL}} \leq \lambda \leq \frac{m_i}{1536U\sqrt{d}}$, $N \geq \frac{4CU\delta^{-1}}{(1-\delta^{-1})^2}$, *and* $n \geq max\{\frac{1}{c}\log\frac{4c'}{\epsilon_1}, (3456)^2(\frac{U}{L} + 0.5)\log(2p)d, 3(6144)^2\frac{U^2}{L}d\log(\frac{8p}{\epsilon_1})(\frac{1}{m_i})^2\}$ *where the n trajectories are i.i.d. Then* $\|\hat{W}_i(\lambda) - W_i\|_2 \leq \frac{m_i}{2}$ *holds with a probability of at least* $1 - \epsilon_1$. c, c' *are universal positive constants and* U, L, C, δ, m_i, d *are defined in Eq. 9.*

Theorem 2.2 (*consecutive-* squared error)**.** *Let* $\epsilon_1 > 0$ *such that* $\epsilon_1 \geq \frac{8}{p^2}$ *and* $i \in V$. *For* $4\sqrt{\frac{(3+24\sqrt{3}UC(\delta-1)^{-1})\log(8p/\epsilon_1)}{nL}} \leq \lambda \leq \frac{m_i}{1536U\sqrt{d}}$, $N \geq \frac{4CU\delta^{-1}}{(1-\delta^{-1})^2}$, *and* $n \geq \max\{33^2\log p[\frac{U}{L} + 0.5 + 4\sqrt{8}\frac{CU}{\delta-1}]^2, 2\log(\frac{8p^2}{p^2\epsilon_1-8}), (3 + 24\sqrt{3}UC(\delta-1)^{-1})(6144)^2\frac{U^2}{L}d\log(\frac{8p}{\epsilon_1})(\frac{1}{m_i})^2\}$, *where the n trajectories are non-i.i.d. Then* $\|\hat{W}_i(\lambda) - W_i\|_2 \leq \frac{m_i}{2}$ *holds with a probability of at least* $1-\epsilon_1$, *where c,* c' *are universal positive constants.* U, L, C, δ, m_i, d *are defined in Eq. 9.*

Using Theorems 2.1 and 2.2, we give the correctness of the thresholding procedure listed in Eq. 8.

Theorem 2.3 (structure learning)**.** *Let* $\epsilon > 0$, *and* $N \geq \frac{4CU\delta^{-1}}{(1-\delta^{-1})^2}$, *with constants* U, L, C, δ, m, d *defined in Eq. 9 and universal positive constants* c, c'. *Construct an undirected edge set* $\hat{E}_M$ *and* $\hat{E}$ *as per Eq. 8 with thresholds* $\tau_1 = \tau_2 = m$. *Then* $E = \hat{E}$ *holds with a probability of at least* $1 - \epsilon$, *if*
(a) 'restart & record' (i.i.d.): $\frac{m}{1536U\sqrt{d}} \geq \lambda \geq 4\sqrt{\frac{3\log(8p^2/\epsilon)}{nL}}$, *and* $n \geq max\{\frac{1}{c}\log\frac{4c'p}{\epsilon}, (3456)^2(\frac{U}{L} + 0.5)\log(2p)d, 3(6144)^2\frac{U^2}{L}d\log\frac{8p^2}{\epsilon}\frac{1}{m^2}\}$.
(b) 'consecutive' (non-i.i.d.): $\epsilon \geq \frac{8}{p}$, $\frac{m}{1536U\sqrt{d}} \geq \lambda \geq 4\sqrt{\frac{(3+24\sqrt{3}UC(\delta-1)^{-1})\log(8p^2/\epsilon)}{nL}}$, *and* $n \geq max\{33^2\log p[\frac{U}{L} + 0.5 + 4\sqrt{8}\frac{CU}{\delta-1}]^2, 2\log(\frac{8p^2}{p\epsilon-8}), (3 + 24\sqrt{3}UC(\delta-1)^{-1})(6144)^2\frac{U^2}{L}d\log(\frac{8p^2}{\epsilon})\frac{1}{m^2}\}$.

The proofs of Theorems 2.1, 2.2 and 2.3 are provided in Section 4. These proofs are based on the theory of M-estimators Negahban et al. (2012), for the complex-valued regression problem. It is worth mentioning that when the n trajectories are consecutive, i.e., they correspond to a single time-series, the DFT coefficients computed in Eq. 4 are correlated, as against being i.i.d. in the 'restart & record' setting. The derivation of sample complexity in the 'consecutive' setting requires concentration results for correlated Gaussian variables, which are more involved and less sharp than comparable results in the i.i.d. setting, as discussed later. In the next section, we present the theory of M-estimators (Negahban et al., 2012) in the complex domain, necessary to prove our results for correct structure recovery.

3 M-ESTIMATOR BASED ANALYSIS OF REGULARIZED WIENER FILTER

The regularized Wiener filter estimator Eq. 7 belongs to a class of regularized M-estimators. Note that the regularizer $(\|.\|_1)$ in Eq. 7 satisfies decomposability property with respect to the following complex-valued subspaces: $\mathcal{M} = \{v \in \mathbb{C}^p | v[j] = 0 \text{ if } W_i[j] = 0\ \}, \mathcal{M}^\perp = \{v \in \mathbb{C}^p | v[j] = 0 \text{ if } W_i[j] \neq 0\ \}$ for a node $i \in V$. That is, $\|v\|_1 = \|v_{\mathcal{M}}\|_1 + \|v_{\mathcal{M}^\perp}\|_1$, where $v_{\mathcal{M}}$, $v_{\mathcal{M}^\perp}$ are the projections of v on $\mathcal{M}$ and $\mathcal{M}^\perp$. We follow the approach in Negahban et al. (2012) to bound the error

$$\hat{\Delta} := \hat{W}_i(\lambda) - W_i. \quad (10)$$

Negahban et al. (2012) states that two conditions are sufficient to control the error $\|\hat{\Delta}\|_2$.

First condition (λ choice): $\lambda \geq \frac{2}{n}\|\mathcal{X}^H(\mathcal{Y} - \mathcal{X}W_i)\|_\infty$. (11)

Eq. 11 ensures that $\hat{\Delta}$, defined in Eq. 10, belongs to the set

$$\mathcal{D}(W_i) = \{\Delta \in \mathbb{C}^p\ |\|\Delta_{\mathcal{M}^\perp}\|_1 \leq 3\|\Delta_{\mathcal{M}}\|_1\}. \quad (12)$$

Second condition (restricted eigenvalue property):

$$\frac{1}{n}\|\mathcal{X}\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2, \forall \Delta \in \mathcal{D}(W_i). \quad (13)$$

The following proposition, similar to Theorem 1 in Negahban et al. (2012), bounds the error $\|\hat{\Delta}\|_2$.

Proposition 1. *For the regularized Wiener filter estimator defined in Eq. 7, $\|\hat{W}_i - W_i\|_2 \leq (\frac{3}{\kappa}\lambda\sqrt{d})$, whenever Eq. 11 and Eq. 13 hold.*

For completion, we outline a proof of Eq. 12 and Proposition 1 for complex-valued variables in Section 2 of the Supplementary material, following the real-valued analysis in Negahban et al. (2012).

We now show that Eq. 11 and Eq. 13 hold, for both *restart & record* (i.i.d.) and *consecutive* (non-i.i.d.) trajectories. These results are then used to prove Theorems 2.1 and 2.2.

Restart & record (i.i.d.) trajectories:

Lemma 3.1. *Suppose $\epsilon_3 > 0$. Let rows in $\{X_i^r\}_{r=1}^n$ and $\{X_{\tilde{i}}^r\}_{r=1}^n$ defined in Eq. 4 be i.i.d. If $\lambda \geq 4\sqrt{\frac{3\log(4p/\epsilon_3)}{nL}}$, then $\lambda \geq \frac{2}{n}\|\mathcal{X}^H(\mathcal{Y} - \mathcal{X}W_i)\|_\infty$ holds with a probability of at least $1-\epsilon_3$.*

Lemma 3.2. *Suppose $\epsilon_2 > 0$ be given. Let rows in $\{X_i^r\}_{r=1}^n$ and $\{X_{\tilde{i}}^r\}_{r=1}^n$ defined in Eq. 4 be i.i.d. If $n \geq max\{\frac{1}{c}\log\frac{2c'}{\epsilon_2}, (3456)^2(\frac{U}{L}+0.5)\log(2p)d\}$, $N \geq \frac{4CU\delta^{-1}}{(1-\delta^{-1})^2}$, then Eq. 13 holds with $\kappa = \frac{1}{256U}$, with a probability of at least $1-\epsilon_2$.*

The proofs for Lemmas 3.1 and 3.2 are provided in Section 5 and uses concentration bounds for Gaussian random variables.

Consecutive (non i.i.d.) trajectories:

Lemma 3.3. *Suppose $\epsilon_3 > 0$. Assume that both $\{X_i^r\}_{r=1}^n$ and $\{X_{\tilde{i}}^r\}_{r=1}^n$ defined in Eq. 4 are non i.i.d. If $\lambda \geq 4\sqrt{\frac{(3+24\sqrt{3}UC(\delta-1)^{-1})\log(4p/\epsilon_3)}{nL}}$, then $\lambda \geq \frac{2}{n}\|\mathcal{X}^H(\mathcal{Y} - \mathcal{X}W_i)\|_\infty$ holds with a probability of at least $1-\epsilon_3$.*

Lemma 3.4. *Suppose $\epsilon_2 > 0$ such that $\epsilon_2 \geq \frac{4}{p^2}$. Assume that both $\{X_i^r\}_{r=1}^n$ and $\{X_{\tilde{i}}^r\}_{r=1}^n$ defined in Eq. 4 are non i.i.d. Then, if $n \geq \max\{33^2\log p[\frac{U}{L} + 0.5 + 4\sqrt{8}\frac{CU}{\delta-1}]^2, 2\log(\frac{4p^2}{p^2\epsilon_2-4})\}$, $N \geq \frac{4CU\delta^{-1}}{(1-\delta^{-1})^2}$, then $\frac{1}{n}\|\mathcal{X}\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2$, holds for all $\Delta \in \mathcal{D}(W_i)$ with $\kappa = \frac{1}{256U}$, with a probability of at least $1-\epsilon_2$.*

The proofs for Lemmas 3.3 and 3.4 are provided in Section 3 of the Supplementary material.

4 PROOF OF MAIN THEOREMS

To prove the main theorems for structure learning, we use the M-estimator lemmas from the previous section for the regularized Wiener filter at each node, under both i.i.d. and non-i.i.d. trajectories, and then apply the Union bound for all nodes.

Proof of Theorem 2.1. For $n \geq max\{\frac{1}{c}\log\frac{4c'}{\epsilon_1}, (3456)^2(\frac{U}{L}+0.5)\log(2p)d\}$ and $N \geq \frac{4CU\delta^{-1}}{(1-\delta^{-1})^2}$, we apply Lemma 3.2 with $\epsilon_2 = \frac{\epsilon_1}{2}$, then Eq. 13 holds with probability of at least $1-\frac{\epsilon_1}{2}$. Here, $\kappa = \frac{1}{256U}$. With $\lambda \geq 4\sqrt{\frac{3\log(8p/\epsilon_1)}{nL}}$, apply Lemma 3.1 with $\epsilon_3 = \frac{\epsilon_1}{2}$, then Eq. 11 holds with probability of at least $1-\frac{\epsilon_1}{2}$. It follows from Proposition 1 that, $\|\hat{W}_i - W_i\|_2 \leq (\frac{3}{\kappa}\lambda\sqrt{d}) = (768U\lambda\sqrt{d})$. Take $\lambda \leq \frac{m_i}{1536U\sqrt{d}}$. For $n \geq 3(6144)^2\frac{U^2}{L}d\log(\frac{8p}{\epsilon_1})(\frac{1}{m_i})^2$, $4\sqrt{\frac{3\log(8p/\epsilon_1)}{nL}}$ is smaller than $\frac{m_i}{1536U\sqrt{d}}$. Thus, $\|\hat{W}_i - W_i\|_2 \leq \frac{m_i}{2}$ holds with a probability of at least $1-\epsilon_1$. □

Proof of Theorem 2.2. Here, we combine the results of Lemma 3.4 with $\epsilon_2 = \frac{\epsilon_1}{2}$ and Lemma 3.3 with $\epsilon_3 = \frac{\epsilon_1}{2}$. The rest of the proof is analogous to proof of Theorem 2.1.

Proof of Theorem 2.3. (a) 'restart & record': Choose $\epsilon_1 = \frac{\epsilon}{p}$. It follows from definition of m, that $\frac{1}{m} \geq \frac{1}{m_i}$ for all $i \in V$. Now $n \geq max\{\frac{1}{c}\log\frac{4c'p}{\epsilon}, (3456)^2(\frac{U}{L}+0.5)\log(2p)d, \quad 3(6144)^2\frac{U^2}{L^2}d(\log\frac{8p^2}{\epsilon})(\frac{1}{m})^2\}$ and $4\sqrt{\frac{3\log(8p^2/\epsilon)}{nL}} \leq \lambda \leq \frac{m}{1536U\sqrt{d}}$ would satisfy the conditions on n and λ specified in Theorem 2.1 for a $i \in V$. Therefore, $\|\hat{W}_i - W_i\|_2 \leq \frac{m}{2}$ holds with a probability of at least $1-\frac{\epsilon}{p}$. Using a union bound for all the $p+1$ nodes, we have $\|\Im[\hat{W}_i - W_i]\|_2 \leq \|\hat{W}_i - W_i\|_2 \leq \frac{m}{2}$ holds for all $i \in V$ with a probability of at least $1-\frac{\epsilon(p+1)}{p} \approx 1-\epsilon$ for large p.

Note that if $(ij) \in E$, then $|\Im(W_i[j])| \geq m > 0$. Similarly, for $(ij) \in E \setminus E_M$, $\Im(W_i[j]) = 0$ and for $(ij) \notin E_M$, $W_i[j] = 0$. Expanding $\|\hat{W}_i - W_i\|_2$, it can thus be shown that $\hat{E}$ derived from $\hat{E}_M$ contains only the edges in E.

(b) 'consecutive': Using Theorem 2.2 for every node $i \in V$ with $\epsilon_1 = \frac{\epsilon}{p}$, the proof is analogous to the proof of Theorem 2.3. □

The next section includes the primary proof techniques for the M-estimator lemmas in Section 3.

5 PROOFS OF M-ESTIMATOR LEMMAS FOR 'RESTART & RECORD' (I.I.D.) TRAJECTORIES

The regularized regression in Eq. 7 involves working with complex-valued random variables $X_i^r, X_{\tilde{i}}^r$ defined in Eq. 4 for a node $i \in V$. Their probability distribution

is as follows,

$$\begin{bmatrix} X_i^r \\ X_{\bar{i}}^r \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \hat{\Phi}_x), \text{ where,} \quad (14)$$

$$\hat{\Phi}_x = \begin{bmatrix} \hat{\Phi}_i & \hat{\Phi}_{i,\bar{i}}, \\ \hat{\Phi}_{\bar{i},i} & \hat{\Phi}_{\bar{i}} \end{bmatrix} = \frac{1}{N} \sum_{q=-(N-1)}^{(N-1)} (N-|q|) R_x(q) e^{-\iota f q}.$$

Thus, $X_i^r \sim \mathcal{N}(0, \hat{\Phi}_i)$ and $X_{\bar{i}}^r \sim \mathcal{N}(\mathbf{0}, \hat{\Phi}_{\bar{i}})$. The following result bounds the difference between $\hat{\Phi}_x$ and Φ_x (see Eq. 3) for a N-length trajectory, and is used in our analysis.

Lemma 5.1. *If* $N > \frac{4CU\delta^{-1}}{(1-\delta^{-1})^2}$, *then* $\|\Phi_x - \hat{\Phi}_x\|_2 \leq \frac{1}{2U}$. *Moreover,* $\|\Phi_{\bar{i}} - \hat{\Phi}_{\bar{i}}\|_2 \leq \frac{1}{2U}$, *and* $\frac{1}{2U}\|v\|_2^2 \leq v^H \hat{\Phi}_{\bar{i}} v \leq [\frac{1}{L} + \frac{1}{2U}]\|v\|_2^2, \forall v \in \mathbb{C}^p$.

The proof is provided in Section 1 of the Supplementary material. Next we prove Lemma 3.1, which gives a lower bound on λ used in the regularized Wiener filter estimator. On a high level, the proof uses the Gaussianity of the complex-valued error vector $\mathcal{E} := \mathcal{Y} - \mathcal{X}W_i$. We use it to identify the Lipschitz constant associated with rows of $\frac{1}{n}\mathcal{X}^H\mathcal{E}$, and then determine the lower bound on λ using the union bound. □

Proof of Lemma 3.1. Let $\mathcal{E} := \mathcal{Y} - \mathcal{X}W_i$. We show that $\frac{1}{n}\|\mathcal{X}^H\mathcal{E}\|_\infty$ is bounded with a high probability and choose λ greater than that bound. Separating $\mathcal{X} = \mathcal{X}_R + \iota\mathcal{X}_I$, $\mathcal{E} = \mathcal{E}_R + \iota\mathcal{E}_I$, into real and imaginary parts (specified by subscripts R and I respectively), we have

$$\frac{1}{n}\|\mathcal{X}^H\mathcal{E}\|_\infty \leq \frac{1}{n}\| \begin{pmatrix} \mathcal{X}_R^T & \mathcal{X}_I^T \end{pmatrix} \begin{pmatrix} \mathcal{E}_R \\ \mathcal{E}_I \end{pmatrix} \|_\infty + \frac{1}{n}\| \begin{pmatrix} -\mathcal{X}_I^T & \mathcal{X}_R^T \end{pmatrix} \begin{pmatrix} \mathcal{E}_R \\ \mathcal{E}_I \end{pmatrix} \|_\infty. \quad (15)$$

Let $\mathcal{E}_1 := [\mathcal{E}_R[1]\ \mathcal{E}_I[1]\ ...\ \mathcal{E}_R[n]\ \mathcal{E}_I[n]]^T$ with covariance matrix $\mathcal{C}_1$. Note that $\begin{pmatrix} \mathcal{E}_R \\ \mathcal{E}_I \end{pmatrix}. = P\mathcal{E}_1$,for some symmetric permutation matrix P, such that its covariance matrix $\mathcal{C}_2 = P\mathcal{C}_1P$. Rewriting $\begin{pmatrix} \mathcal{E}_R \\ \mathcal{E}_I \end{pmatrix} = \mathcal{C}_2^{1/2} \begin{pmatrix} \mathcal{W}_R \\ \mathcal{W}_I \end{pmatrix}$ in Eq. 15, where $\begin{pmatrix} \mathcal{W}_R \\ \mathcal{W}_I \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, I)$, we have

$$\frac{1}{n}\|\mathcal{X}^H\mathcal{E}\|_\infty \leq \frac{1}{n}\| \begin{pmatrix} \mathcal{X}_R^T & \mathcal{X}_I^T \end{pmatrix} \mathcal{C}_2^{1/2} \begin{pmatrix} \mathcal{W}_R \\ \mathcal{W}_I \end{pmatrix} \|_\infty + \frac{1}{n}\| \begin{pmatrix} -\mathcal{X}_I^T & \mathcal{X}_R^T \end{pmatrix} \mathcal{C}_2^{1/2} \begin{pmatrix} \mathcal{W}_R \\ \mathcal{W}_I \end{pmatrix} \|_\infty. \quad (16)$$

To bound the right side of Eq. 16, we first show that either function is Lipschitz. Consider first $f(\mathcal{W}_R, \mathcal{W}_I) := \frac{1}{n} \begin{pmatrix} \mathcal{X}_R^T(j,:) & \mathcal{X}_I^T(j,:) \end{pmatrix} \mathcal{C}_2^{1/2} \begin{pmatrix} \mathcal{W}_R \\ \mathcal{W}_I \end{pmatrix}$.Then,

$$\begin{aligned} &\|f(\mathcal{W}_R, \mathcal{W}_I) - f(\mathcal{W}_R', \mathcal{W}_I')\|_2 \\ &\leq \frac{1}{n}\| \begin{pmatrix} \mathcal{X}_R^T(j,:) & \mathcal{X}_I^T(j,:) \end{pmatrix} \|_2 \|\mathcal{C}_2^{1/2}\|_2 \| \begin{pmatrix} \mathcal{W}_R - \mathcal{W}_R' \\ \mathcal{W}_I - \mathcal{W}_I' \end{pmatrix} \|_2, \\ &\leq \frac{1}{\sqrt{n}}\|P\|_2 \|\mathcal{C}_1^{1/2}\|_2 \| \begin{pmatrix} \mathcal{W}_R - \mathcal{W}_R' \\ \mathcal{W}_I - \mathcal{W}_I' \end{pmatrix} \|_2, (\because \text{using Eq. 5}) \\ &= \sqrt{\frac{3}{2nL}} \begin{pmatrix} \mathcal{W}_R - \mathcal{W}_R' \\ \mathcal{W}_I - \mathcal{W}_I' \end{pmatrix} \|_2 \end{aligned} \quad (17)$$

($\because$ Lemma 3.2 in the Supplementary material).

Thus, f is a Lipschitz function with Lipschitz constant $\sqrt{\frac{3}{2nL}}$. Using Massart (2000)'s result on concentration of Lipschitz functions, we have, for $t > 0$, $\mathbb{P}[\frac{1}{n}| \begin{pmatrix} \mathcal{X}_R^T(j,:) & \mathcal{X}_I^T(j,:) \end{pmatrix} \mathcal{C}_2^{1/2} \begin{pmatrix} \mathcal{W}_R \\ \mathcal{W}_I \end{pmatrix} | \geq t] \leq 2\exp(-\frac{t^2 nL}{3})$. Choosing $t = \sqrt{\frac{3\log(\frac{4p}{\epsilon_3})}{nL}}$ and the union bound for all $j \in \{1, \cdots, p\}$, we have $\mathbb{P}[\frac{1}{n}\| \begin{pmatrix} \mathcal{X}_R^T & \mathcal{X}_I^T \end{pmatrix} \mathcal{C}_2^{1/2} \begin{pmatrix} \mathcal{W}_R \\ \mathcal{W}_I \end{pmatrix} \|_\infty \geq \sqrt{\frac{3\log(\frac{4p}{\epsilon_3})}{nL}}] \leq \frac{\epsilon_3}{2}$. Using a similar analysis, $\mathbb{P}[\frac{1}{n}\| \begin{pmatrix} -\mathcal{X}_I^T & \mathcal{X}_R^T \end{pmatrix} \mathcal{C}_2^{1/2} \begin{pmatrix} \mathcal{W}_R \\ \mathcal{W}_I \end{pmatrix} \|_\infty \geq \sqrt{\frac{3\log(\frac{4p}{\epsilon_3})}{nL}}] \leq \frac{\epsilon_3}{2}$. Choose $\lambda \geq 4\sqrt{\frac{3\log(4p/\epsilon_3)}{nL}}$. Using the Union bound on Eq. 16, we have $\mathbb{P}[\frac{1}{n}\|\mathcal{X}^H\mathcal{E}\|_\infty \geq \lambda/2] \leq \epsilon_3$. □

Next we prove Lemma 3.2 which ensures the restricted eigenvalue property for matrix $\mathcal{X} = [X_{\bar{i}}^1, \cdots, X_{\bar{i}}^n]^T$, where $X_{\bar{i}}^r$ is computed from the r^{th} trajectory, as defined in Eq. 4. On a high level, each row in $\mathcal{X}$ can be divided into real and imaginary components, that are each Gaussian variables with known covariance matrices. The proof then follows by merging bounds on the restricted eigenvalue property of Gaussian real-valued matrices.

Proof of Lemma 3.2. Separating into real and imaginary parts, we have:

$$\frac{\|\mathcal{X}\Delta\|_2^2}{n} = \frac{\|(\mathcal{X}_R + \iota\mathcal{X}_I)(\Delta_R + \iota\Delta_I)\|_2^2}{n} = \frac{\|\mathcal{X}_1 v\|_2^2}{n} + \frac{\|\mathcal{X}_2 v\|_2^2}{n}. \quad (18)$$

where $\mathcal{X}_1 := [\mathcal{X}_R\ -\mathcal{X}_I]$, $\mathcal{X}_2 := [\mathcal{X}_I\ \mathcal{X}_R]$ and $v = (\Delta_R^T\ \Delta_I^T)^T$.

For simplicity, in this proof we drop the superscript r in X_i^r and $X_{\bar{i}}^r$. Note that the rows of $\mathcal{X}_1$ and $\mathcal{X}_2$ are i.i.d. samples of the real random vectors, $[(X_{\bar{i}})_R^T\ -(X_{\bar{i}})_I^T]^T$ and $[(X_{\bar{i}})_I^T\ (X_{\bar{i}})_R^T]^T$, respectively. To show that $\frac{1}{n}\|\mathcal{X}\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2$ holds for all $\Delta \in \mathcal{D}(W_i)$ with high probability, we prove the restricted eigenvalue property for group structured norms on both terms in Eq. 18. Let $\bar{\Sigma}$ be the covariance of random vector

$[(X_{\bar{i}})_R^T\ (X_{\bar{i}})_I^T]^T$. Then,

$$\overline{\Sigma} = \begin{bmatrix} \overline{\Sigma}_{11} & \overline{\Sigma}_{12} \\ \overline{\Sigma}_{21} & \overline{\Sigma}_{22} \end{bmatrix} = \begin{bmatrix} \mathbb{E}[(X_{\bar{i}})_R(X_{\bar{i}})_R^T] & \mathbb{E}[(X_{\bar{i}})_R(X_{\bar{i}})_I^T] \\ \mathbb{E}[(X_{\bar{i}})_I(X_{\bar{i}})_R^T] & \mathbb{E}[(X_{\bar{i}})_I(X_{\bar{i}})_I^T] \end{bmatrix}. \tag{19}$$

Thus, $[(X_{\bar{i}})_R^T\ -(X_{\bar{i}})_I^T]^T$ and $[(X_{\bar{i}})_I^T\ (X_{\bar{i}})_R^T]^T$ have means $\mathbf{0}$ and covariance Σ_1 and Σ_2, respectively, where,

$$\Sigma_1 = \begin{bmatrix} \overline{\Sigma}_{11} & -\overline{\Sigma}_{12} \\ -\overline{\Sigma}_{21} & \overline{\Sigma}_{22} \end{bmatrix},\ \Sigma_2 = \begin{bmatrix} \overline{\Sigma}_{22} & \overline{\Sigma}_{21} \\ \overline{\Sigma}_{12} & \overline{\Sigma}_{11} \end{bmatrix}. \tag{20}$$

From Eq. 18, $\|\Delta\|_1 = \|v\|_{1,2} := \sum_{j=1}^{p} \|[v[j]\ v[p+j]]^T\|_2$. Consider the following definitions: $\mathcal{M}_2 := \{c \in \mathbb{R}^{2p} |\ c[i] = 0, c[p+i] = 0,\ \text{if } (W_i)_R[i] = 0, (W_i)_I[i] = 0\}$, $\mathcal{M}_2^{\perp} := \{c \in \mathbb{R}^{2p} |\ c[i] = 0, c[p+i] = 0,\ \text{if } (W_i)_R[i] \neq 0, (W_i)_I[i] \neq 0\}$ and $\mathcal{D}_2(W_i) := \{v \in \mathbb{R}^{2p} |\ \|v_{\mathcal{M}_2^{\perp}}\|_{1,2} \leq 3\|v_{\mathcal{M}_2}\|_{1,2}\}$.

Clearly, if $\Delta \in \mathcal{D}(W_i)$, defined in Eq. 12, then $v \in \mathcal{D}_2(W_i)$ and vice versa. For $v \in \mathcal{D}_2$, the group norm $\|v\|_{1,2} \leq 4\|v_{\mathcal{M}_2}\|_{1,2}$. Using Cauchy Schwartz inequality and the definition of bounded degree d, it follows that, $\|v_{\mathcal{M}_2}\|_{1,2} \leq \sqrt{d}\|v\|_2$. Thus, $\|v\|_{1,2} \leq 4\sqrt{d}\|v\|_2$. Below is a result, derived from Negahban et al. (2012) (Section 5.1) and Ledoux and Talagrand (2013) for Gaussian random matrices.

Lemma 5.2. *For any Gaussian random matrix $\mathbb{X} \in \mathbb{R}^{n \times 2p}$ with i.i.d. $\mathcal{N}(\mathbf{0}, \Sigma)$ rows, then there are universal positive constants $c, c^{'}$ such that with probability at least $1 - c^{'}\exp(-cn)$,*

$$\frac{\|\mathbb{X}v\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Sigma^{1/2}v\|_2 - \frac{27}{\sqrt{n}}\sqrt{2\log(2p)}\rho(\Sigma)\|v\|_{1,2} \tag{21}$$

where, $\rho(\Sigma) := \max_{j \in \{1,\ldots,p\}, i \in \{1,2\}} [\mathbb{E}((w_{G_j}(i))^2)]^{1/2}$ and $w \sim \mathcal{N}(0, \Sigma),\ v \in \mathbb{R}^{2p \times 1}$.

We use Eq. 21 with $\mathbb{X} = \mathcal{X}_1$ and $\mathcal{X}_2$ to obtain a lower bound on $\frac{\|\mathcal{X}_1(\Delta_R^T\ \Delta_I^T)^T\|_2}{\sqrt{n}}$ and $\frac{\|\mathcal{X}_2(\Delta_R^T\ \Delta_I^T)^T\|_2}{\sqrt{n}}$. Note that, for $\bar{\Sigma}, \Sigma_1$ and Σ_2 defined in Eqs. 19, 20, we have $\rho(\bar{\Sigma}) = \rho(\Sigma_1) = \rho(\Sigma_2) = \max_j[\bar{\Sigma}_{jj}]^{1/2}$. Using the inequality $\sqrt{a^2+b^2} \leq a+b \leq \sqrt{2(a^2+b^2)}$ for two non-negative numbers a, b in Eq. 18, we get that the following holds with a probability of $1 - 2c^{'}\exp(-cn)$, where $c,\ c^{'}$ are universal positive constants,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\|\mathcal{X}\Delta\|_2 \\
&\geq \frac{1}{\sqrt{2}}[\frac{\|\mathcal{X}_1(\Delta_R^T\ \Delta_I^T)^T\|_2}{\sqrt{n}} + \frac{\|\mathcal{X}_2(\Delta_R^T\ \Delta_I^T)^T\|_2}{\sqrt{n}}] \\
&\geq \frac{1}{\sqrt{2}}[\frac{1}{4}(\|\Sigma_1^{1/2}v\|_2 + \|\Sigma_2^{1/2}v\|_2) - \frac{54}{\sqrt{n}}\sqrt{2\log(2p)}\rho(\bar{\Sigma})\|v\|_{1,2}] \\
&\geq \frac{1}{\sqrt{2}}[\frac{1}{4}\sqrt{v^T(\Sigma_1+\Sigma_2)v} - \frac{54}{\sqrt{n}}\sqrt{2\log(2p)}\rho(\bar{\Sigma})\|v\|_{1,2}] \\
&\geq \frac{1}{\sqrt{2}}[\frac{1}{4}\lambda_{min}((\Sigma_1+\Sigma_2)^{1/2}) - \frac{54}{\sqrt{n}}\sqrt{2\log(2p)}\rho(\bar{\Sigma})4\sqrt{d}]\|v\|_2],
\end{aligned} \tag{22}$$

Using Eq. 20, the definition of $\hat{\Phi}_{\bar{i}} := \mathbb{E}(X_{\bar{i}}(X_{\bar{i}})^H)$, for $v := [\Delta_R^T, \Delta_I^T]^T$, it follows that,

$$v^T(\Sigma_1+\Sigma_2)v = \Delta^H\hat{\Phi}_{\bar{i}}\Delta\ \ (\because \Sigma_1+\Sigma_2 = \begin{bmatrix} (\hat{\Phi}_{\bar{i}})_R & (\hat{\Phi}_{\bar{i}})_I \\ -(\hat{\Phi}_{\bar{i}})_I & (\hat{\Phi}_{\bar{i}})_R \end{bmatrix})$$

$$\Rightarrow \frac{1}{2U}\|\Delta\|_2^2 \leq v^T(\Sigma_1+\Sigma_2)v \leq [\frac{1}{L} + \frac{1}{2U}]\|\Delta\|_2^2 \tag{23}$$

($\because$ using Lemma 5.1)

Thus, $\lambda_{min}((\Sigma_1+\Sigma_2)^{1/2}) \geq \frac{1}{\sqrt{2U}}$, and $\rho(\bar{\Sigma}) \leq \rho(\Sigma_1+\Sigma_2) \leq \|(\Sigma_1+\Sigma_2)^{1/2}\|_2 = \|(\hat{\Phi}_{\bar{i}})^{1/2}\|_2 \leq \sqrt{\frac{1}{L}+\frac{1}{2U}}$. Thus, Eq. 22 is, $\frac{1}{\sqrt{n}}\|\mathcal{X}\Delta\|_2 \geq [\frac{1}{8\sqrt{U}} - \frac{54}{\sqrt{n}}\sqrt{\log(2p)}(\sqrt{\frac{1}{L}+\frac{1}{2U}})4\sqrt{d}]\|\Delta\|_2$. Choose $n \geq max\{\frac{1}{c}\log\frac{2c'}{\epsilon_2}, (3456)^2(\frac{U}{L} + 0.5)\log(2p)d\}$, then $\frac{54}{\sqrt{n}}\sqrt{\log(2p)}(\sqrt{\frac{1}{L}+\frac{1}{2U}})4\sqrt{d} \leq \frac{1}{16\sqrt{U}}$. Hence, $\frac{\|\mathcal{X}\Delta\|_2}{\sqrt{n}} \geq \frac{1}{16\sqrt{U}}\|\Delta\|_2$, and Eq. 13 holds with $\kappa = \frac{1}{256U}$ with a probability of at least $1-\epsilon_2$. □

The proofs of M-estimator conditions for consecutive (non-i.i.d.) trajectories (Lemmas 3.3 and 3.4) follow on similar lines, albeit with different concentration results, and are detailed in Section 3 of the supplementary material.

6 NUMERICAL RESULTS

We demonstrate the numerical implementation of recovering topology on a Desktop PC with Intel Xeon E5-1620 Processor (8x 3.7 GHz) and 32 GB RAM. We considered a two-dimensional square grid G with $p+1$ nodes and generate samples for Eq. 1 with exogenous input $e(k) = 5s_p[w(k) - 0.3w(k-1)]$, where $w(k),\ w(k-1)$ are sampled from a standard Normal distribution, $s_p \in \mathbb{R}^{(p+1)\times(p+1)}$ is a diagonal matrix containing constants; $h \in \mathbb{R}^{(p+1)\times(p+1)}$ is a weighted adjacency matrix. The scaling s_p is chosen such that $\mathcal{X}$ and $\mathcal{Y}$ in the estimator are column-normalized.

We reconstruct the topology with a probability of at least $1-\epsilon$, where $\epsilon = 0.05$. The numerical experiments are conducted in MATLAB R2020b.

For a choice of n, we generate n trajectories, either independently (restart & record) or taken as consecutive intervals of a larger trajectory (consecutive). Each trajectory is of length $N = \frac{4CU\delta^{-1}}{(1-\delta^{-1})^2}$, rounded to the nearest integer. From the trajectories, we compute the samples $\{X_i^r, X_{\bar{i}}^r\}_{r=1}^n$ for all the nodes $i \in V$ at frequency $f = \frac{2\pi}{N}$.

The Regularized Wiener Filter Estimator is solved using CVXR (Grant and Boyd, 2014), with $\lambda = 4\sqrt{\frac{3\log(8p^2/\epsilon)}{nL}}$ if trajectories are i.i.d, $\lambda =$

$4\sqrt{\frac{(3+24\sqrt{3}UC(\delta-1)^{-1})\log(8p^2/\epsilon)}{nL}}$ if trajectories are non-i.i.d. The chosen value of λ correspond to the minimum sufficient condition present in Theorem's 2.3, for the i.i.d. and consecutive settings, respectively.

After solving for $\hat{W}_i$, we construct $\hat{E} = \{(i,j)||\Im(\hat{W}_i[j])| + |\Im(\hat{W}_j[i])| \geq m\}$ (m defined in Eq. 9). The relative error in reconstructing the topology is defined as the sum of false positive and false negatives. n_{min} is the minimum value of n such that relative error is zero for 45 out of 45 random trials.

The values of n_{min} for various values of p and δ for the i.i.d. and non-i.i.d. cases are shown in Figure. 3. In Figure 3(b), the correlation strength δ^{-1} of the trajectories is high, and consequently N is large. For small δ^{-1}, the length of each trajectory can be reduced significantly for reconstructing the topology. For example, in Figure 3(c), N is much smaller. Further, n_{min} is of the order of $\approx 10^7$ rather than a more conservative estimate of $\approx 10^{16}$ as provided by the main theorems.

Numerical Comparison with prior work: We give empirical comparison with frequency-domain based gLasso-estimator in Jung et al. (2015) and unregularized regression in Talukdar et al. (2020) for a two-dimensional square grid containing 16 nodes. Note that Jung et al. (2015) does not lead to correct recovery as Conditional Independence Graph (CIG) doesn't lead to true underlying network. Unlike Talukdar et al. (2020), regularization in our algorithm gives improved exact topology recovery in low sample regime. Figure 4 shows the relative error for different values of n. The error is computed by averaging over 200 random trials for each algorithm. Further, for $\epsilon > 0$, the fraction of trials with successful topology reconstruction in our experiments is higher than $1-\epsilon$.

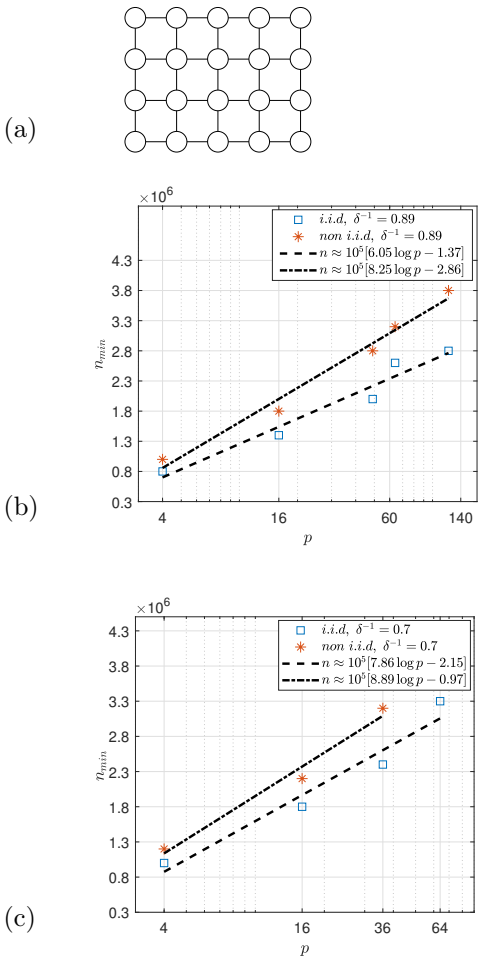

Figure 3: (a) An illustration of G as a 5×5 grid with $p = 24$. Dependence of n_{min} on $\log p$: (a) $C = 6.8$, $\delta^{-1} = 0.89$, $U = 1.55$, $L = 0.74$, $N = \frac{4CU\delta^{-1}}{(1-\delta^{-1})^2} \approx 2900$, (b) $C = 2.8$, $\delta^{-1} = 0.7$, $U = 1.3$, $L = 0.8$, $N = \frac{4CU\delta^{-1}}{(1-\delta^{-1})^2} \approx 115$. Dashed lines in (a) and (b) corresponds to least squares regression fit.

7 EXTENSIONS AND PATH FORWARD

In this article, we presented a regularized Wiener filter estimator to learn the structure of a discrete-time networked LDS. We analyzed the sample complexity of our estimator and showed that it linearly depends the logarithm of the number of nodes p in two cases, one where trajectories of nodal states are collected in independent observation windows of equal length, and another where the trajectories pertain to a single continuous observation window.

While we discuss our method for first-order discrete-time LDS, our estimator can be extended to learning related networks as highlighted next.

VAR(τ) models with correlated inputs: Lemma 2.1 and our subsequent analysis follows directly if higher-order delays (at the same node) are included in the LDS Eq. 1. L, U and the sample complexity will need to be changed accordingly.

Continuous time LDS: Considering a fixed sampling time ΔT and a time-discretization function, the continuous time LDS can be converted to a discrete-time LDS with related frequency domain-representation (Talukdar et al., 2020). The analysis will involve merging the error due to discretization with the finite sample analysis.

LDS under cyclo-stationary processes: Cyclo-stationary processes represent a generalization of WSS processes where the statistics such as mean, correlation function are periodic functions of time. As shown in

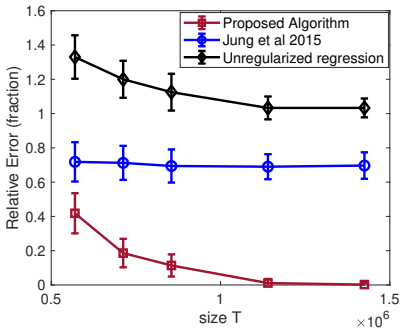

Figure 4: Reconstruction of exact topology (Proposed Algorithm) vs CIG ((Jung et al., 2015)).

Doddi et al. (2019), a lifting operation can be used to represent time-evolution of a cyclo-stationary process as a WSS process with vector-valued states. The remaining analysis of the sample complexity will be similar.

Directed graphs under correlated inputs: Note that our estimator uses properties of the inverse power spectral density (see Lemma 2.1 and discussion). Under a strict causality assumption on the linear filters (Materassi and Salapaka, 2012; Quinn et al., 2015), it has been shown that directed edges can be recovered using inverse power spectral density. The framework presented here can thus be extended to efficiently learn a family of directed graphs.

Finally, we plan to analyze the restrictions of our algorithm in learning networked LDS with spatially correlated inputs, and estimating directed networks with non-causal dependencies, where only approximate reconstruction may be possible using passive methods.

Non-linear interactions: While the theory and validating experiments are conducted for linear dynamics, we claim that the results will follow also for non-linear network dynamics that are approximately linear around an operating point. Such experiments have been described for thermal network of buildings in Talukdar et al. (2020).

8 Acknowledgements:

The authors acknowledge support from the Center for Non-Linear Studies (CNLS) and the Information Science and Technology Institute (ISTI) at Los Alamos National Laboratory.

References

Ascione, F., Bianco, N., De Masi, R. F., de'Rossi, F., and Vanoli, G. P. (2013). Simplified state space representation for evaluating thermal bridges in building: Modelling, application and validation of a methodology. *Applied Thermal Engineering*, 61(2):344–354.

Bach, F. R. and Jordan, M. I. (2004). Learning graphical models for stationary time series. *IEEE transactions on signal processing*, 52(8):2189–2199.

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

Bento, J., Ibrahimi, M., and Montanari, A. (2010). Learning networks of stochastic differential equations. *arXiv preprint arXiv:1011.0415*.

Dahlhaus, R. (2000). Graphical interaction models for multivariate time series 1. *Metrika*, 51(2):157–172.

Dankers, A., Van den Hof, P. M., Bombois, X., and Heuberger, P. S. (2015). Errors-in-variables identification in dynamic networks—consistency results for an instrumental variable approach. *Automatica*, 62:39–50.

Doddi, H., Talukdar, S., Deka, D., and Salapaka, M. (2019). Exact topology learning in a network of cyclostationary processes. In *2019 American Control Conference (ACC)*, pages 4968–4973. IEEE.

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. (2018). Finite time identification in unstable linear systems. *Automatica*, 96:342–353.

Fattahi, S., Matni, N., and Sojoudi, S. (2019). Learning sparse dynamical systems from a single sample trajectory. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2682–2689. IEEE.

Fattahi, S. and Sojoudi, S. (2018). Data-driven sparse system identification. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 462–469. IEEE.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Grant, M. and Boyd, S. (2014). Cvx: Matlab software for disciplined convex programming, version 2.1.

Inchauspe, J., Ripple, R. D., and Trück, S. (2015). The dynamics of returns on renewable energy companies: A state-space approach. *Energy Economics*, 48:325–335.

Jung, A., Hannak, G., and Goertz, N. (2015). Graphical lasso based model selection for time series. *IEEE Signal Processing Letters*, 22(10):1781–1785.

Koh, C., Wu, F.-X., Selvaraj, G., and Kusalik, A. J. (2009). Using a state-space model and location analysis to infer time-delayed regulatory networks. *EURASIP Journal on Bioinformatics and Systems Biology*, 2009:1–14.

Kroutikova, N., Hernandez-Aramburo, C. A., and Green, T. C. (2007). State-space model of grid-connected inverters under current control mode. *IET Electric Power Applications*, 1(3):329–338.

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Loh, P.-L., Wainwright, M. J., et al. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664.

Massart, P. (2000). Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 245–303.

Materassi, D. and Innocenti, G. (2010). Topological identification in networks of dynamical systems. *IEEE Transactions on Automatic Control*, 55(8):1860–1871.

Materassi, D. and Salapaka, M. V. (2012). On the problem of reconstructing an unknown topology via locality properties of the wiener filter. *IEEE transactions on automatic control*, 57(7):1765–1777.

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

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557.

Porreca, R., Drulhe, S., Jong, H. d., and Ferrari-Trecate, G. (2008). Structural identification of piecewise-linear models of genetic regulatory networks. *Journal of Computational Biology*, 15(10):1365–1380.

Quinn, C. J., Kiyavash, N., and Coleman, T. P. (2015). Directed information graphs. *IEEE Transactions on information theory*, 61(12):6887–6909.

Ravikumar, P., Raskutti, G., Wainwright, M. J., and Yu, B. (2008). Model selection in gaussian graphical models: High-dimensional consistency of l1-regularized mle. In *NIPS*, pages 1329–1336.

Sandefur, J. T. (1990). *Discrete dynamical systems: Theory and applications*. Clarendon Press.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., and Recht, B. (2018). Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR.

Songsiri, J., Dahl, J., and Vandenberghe, L. (2010). Graphical models of autoregressive processes.

Stathopoulos, A. and Karlaftis, M. G. (2003). A multivariate state space approach for urban traffic flow modeling and prediction. *Transportation Research Part C: Emerging Technologies*, 11(2):121–135.

Talukdar, S., Deka, D., Doddi, H., Materassi, D., Chertkov, M., and Salapaka, M. V. (2020). Physics informed topology learning in networks of linear dynamical systems. *Automatica*, 112:108705.

Talukdar, S., Prakash, M., Materassi, D., and Salapaka, M. V. (2015). Reconstruction of networks of cyclostationary processes. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 783–788. IEEE.

Tank, A., Foti, N. J., and Fox, E. B. (2015). Bayesian structure learning for stationary time series. In *UAI*.

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

Efficient and passive learning of networked dynamical systems driven by non-white exogenous inputs: Supplementary Material

In this supplementary material, we provide additional discussions of the Regularized Wiener Filter estimator (Eq. 7 in the main document). The proof of Lemma 5.1 and properties of the parameters (Eq. 14 in the article) used in the sample complexity analysis are provided in Section 1. Section 2 includes proofs on separability and errors in M-estimator in the complex domain. Section 3 contains results on norms of state covariance matrices that are used in proofs in the main paper.

1 PROOF OF LEMMA 5.1

Proof. From Eq. 9, we have $\frac{1}{U} \leq \lambda_{min}[\Phi_x]$ and $\lambda_{max}[\Phi_x] \leq \frac{1}{L}$. From Eqs. 3, 14, we have

$$\begin{aligned}
\|\Phi_x - \hat{\Phi}_x\|_2 &= \lim_{m\to\infty} \| \sum_{p=N}^{m} R_x(p)e^{-\iota fp} + \sum_{p=-m}^{-N} R_x(p)e^{-\iota fp} + \sum_{q=-(N-1)}^{N-1} \frac{|q|}{N} R_x(q)e^{-\iota fq}\|_2, \\
&\leq \lim_{m\to\infty} \sum_{p=N}^{m} \|R_x(p)\|_2 + \sum_{p=-m}^{-N} \|R_x(p)\|_2 + \sum_{q=-(N-1)}^{N-1} \frac{|q|}{N}\|R_x(q)\|_2, \\
&\leq \lim_{m\to\infty} 2\sum_{p=N}^{m} C\delta^{-p} + 2\sum_{q=1}^{N-1} C\frac{q}{N}\delta^{-q} \ (\because \text{Eq. 9}) \\
&= 2C\delta^{-N}\frac{1}{1-\delta^{-1}} + \frac{2C}{N}[\frac{\delta^{-1}(1-\delta^{-N})}{(1-\delta^{-1})^2} - \frac{N\delta^{-N}}{1-\delta^{-1}}] \\
&= \frac{2C}{N}[\frac{\delta^{-1}(1-\delta^{-N})}{(1-\delta^{-1})^2}] \leq \frac{2C\delta^{-1}}{N(1-\delta^{-1})^2}.
\end{aligned}$$

Therefore, if $N > \frac{4CU\delta^{-1}}{(1-\delta^{-1})^2}$, then $\|\Phi_x - \hat{\Phi}_x\|_2 \leq \frac{1}{2U}$. Moreover, for $v \in \mathbb{C}^{p+1}$ we have $\|(\hat{\Phi}_x)^{1/2}v\|_2^2 = v^H[\Phi_x - (\Phi_x - \hat{\Phi}_x)]v$. It thus follows that, $\|(\hat{\Phi}_x)^{1/2}v\|_2^2 \geq [\frac{1}{U} - \frac{1}{2U}]\|v\|_2^2 = \frac{1}{2U}\|v\|_2^2$, and $v^H\hat{\Phi}_x v \leq [\frac{1}{L} + \frac{1}{2U}]\|v\|_2^2$. Thus,

$$\frac{1}{2U}\|v\|_2^2 \leq v^H\hat{\Phi}_x v \leq [\frac{1}{L} + \frac{1}{2U}]\|v\|_2^2, \forall v \in \mathbb{C}^{p+1}. \tag{i}$$

Using the same approach,

$$\frac{1}{2U}\|v\|_2^2 \leq v^H\hat{\Phi}_{\bar{\imath}} v \leq [\frac{1}{L} + \frac{1}{2U}]\|v\|_2^2, \forall v \in \mathbb{C}^{p}.$$

□

2 M-ESTIMATOR IN COMPLEX-DOMAIN

Our estimator (Eq. 7) is a M-estimator in the complex domain. For completion, we present the proof of Eq. 12 and Proposition 1 by following the real-valued analysis in Negahban et al. (2012).

Eq. 12 states that if $\lambda \geq \frac{2}{n}\|\mathcal{X}^H(\mathcal{Y}-\mathcal{X}W_i)\|_\infty$ then $\hat{\Delta} := \hat{W}_i(\lambda) - W_i$ belongs to the set $\mathcal{D}(W_i) = \{\Delta \in \mathbb{C}^p \ | \|\Delta_{\mathcal{M}^\perp}\|_1 \leq 3\|\Delta_{\mathcal{M}}\|_1\}$.

To prove that, note that

$$\begin{aligned}
&\frac{1}{2n}[\|\mathcal{Y}-\mathcal{X}(W_i+\Delta)\|_2^2 - \|\mathcal{Y}-\mathcal{X}W_i\|_2^2] \\
&= \frac{1}{2n}[(\mathcal{Y}^H - (W_i+\Delta)^H\mathcal{X}^H)(\mathcal{Y}-\mathcal{X}(W_i+\Delta)) - (\mathcal{Y}^H - W_i^H\mathcal{X}^H)(\mathcal{Y}-\mathcal{X}W_i)] \\
&= \frac{1}{2n}[\Delta^H\mathcal{X}^H(\mathcal{X}W_i-\mathcal{Y}) + (\mathcal{X}W_i-\mathcal{Y})^H\mathcal{X}\Delta + \Delta^H\mathcal{X}^H\mathcal{X}\Delta] \\
&\geq \frac{1}{2n}2Re(\langle\mathcal{X}^H(\mathcal{X}W_i-\mathcal{Y}),\Delta\rangle) \geq \frac{-1}{n}|Re(\langle\mathcal{X}^H(\mathcal{X}W_i-\mathcal{Y}),\Delta\rangle)|,
\end{aligned}$$

where, $Re(x)$ denotes the real part of the complex number x. Moreover,

$$\frac{1}{n}|Re(\langle\mathcal{X}^H(\mathcal{X}W_i-\mathcal{Y}),\Delta\rangle)| \leq \frac{1}{n}|\langle\mathcal{X}^H(\mathcal{X}W_i-\mathcal{Y}),\Delta\rangle| \leq \frac{1}{n}\|\mathcal{X}^H(\mathcal{X}W_i-\mathcal{Y})\|_\infty\|\Delta\|_1 \leq \frac{\lambda}{2}\|\Delta\|_1.$$

Therefore, $\frac{1}{2n}[\|\mathcal{Y}-\mathcal{X}(W_i+\Delta)\|_2^2 - \|\mathcal{Y}-\mathcal{X}W_i\|_2^2] \geq -\frac{\lambda}{2}\|\Delta\|_1 = -\frac{\lambda}{2}(\|\Delta_{\mathcal{M}}\|_1 + \|\Delta_{\mathcal{M}^\perp}\|_1)$.

By optimality of $\hat{W}_i(\lambda) = W_i + \hat{\Delta}$ in the Regularized Wiener Filter Estimator,

$$\begin{aligned}
0 &\geq \frac{1}{2n}[\|\mathcal{Y}-\mathcal{X}(W_i+\hat{\Delta})\|_2^2 - \|\mathcal{Y}-\mathcal{X}W_i\|_2^2] + \lambda[\|W_i+\hat{\Delta}\|_1 - \|W_i\|_1] \\
&\geq -\frac{\lambda}{2}(\|\hat{\Delta}_{\mathcal{M}}\|_1 + \|\hat{\Delta}_{\mathcal{M}^\perp}\|_1) + \lambda(\|\hat{\Delta}_{\mathcal{M}^\perp}\|_1 - \|\hat{\Delta}_{\mathcal{M}}\|_1) = \frac{\lambda}{2}\|\hat{\Delta}_{\mathcal{M}^\perp}\|_1 - \frac{3\lambda}{2}\|\hat{\Delta}_{\mathcal{M}}\|_1.
\end{aligned}$$

Thus, $\hat{\Delta} \in \mathcal{D}(W_i)$. Next, we show that $\|\hat{W}_i - W_i\|_2 \leq (\frac{3}{\kappa}\lambda\sqrt{d})$, whenever Eq. 11 and Eq. 13 hold.

Proof of Proposition 1. Let $K(\delta) := \{\Delta \in \mathbb{C}^p |\ \Delta \in \mathcal{D}(W_i^f)$ and $\|\Delta\|_2 = \delta\}$. Let $\mathsf{F}(\Delta)$ be the difference between the objective of the Regularized Wiener Filter Estimator evaluated at $W_i + \Delta$ and W_i. For a $\Delta \in K(\delta)$, the following holds:

$$\begin{aligned}
\mathsf{F}(\Delta) &= \frac{1}{2n}[2Re\langle(\mathcal{X}^H(\mathcal{X}W_i-\mathcal{Y}),\Delta\rangle + \Delta^H\mathcal{X}^H\mathcal{X}\Delta] + \lambda(\|W_i+\Delta\|_1 - \|W_i\|_1), \\
&\geq -\frac{1}{n}|Re\langle(\mathcal{X}^H(\mathcal{X}W_i-\mathcal{Y}),\Delta\rangle| + \frac{\kappa}{2}\|\Delta\|_2^2 + \lambda[\|\Delta_{\mathcal{M}^\perp}\|_1 - \|\Delta_{\mathcal{M}}\|_1], \\
&(\because \text{Using the restricted eigenvalue property}) \\
&\geq -\frac{\lambda}{2}\|\Delta\|_1 + \frac{\kappa}{2}\|\Delta\|_2^2 + \lambda[\|\Delta_{\mathcal{M}^\perp}\|_1 - \|\Delta_{\mathcal{M}}\|_1], \\
&(\because \text{Using the condition on } \lambda) \\
&= \frac{\kappa}{2}\|\Delta\|_2^2 - \frac{3}{2}\lambda\|\Delta_{\mathcal{M}}\|_1 + \frac{1}{2}\lambda\|\Delta_{\mathcal{M}^\perp}\|_1, \\
&\geq \frac{\kappa}{2}\|\Delta\|_2^2 - \frac{3}{2}\lambda\|\Delta_{\mathcal{M}}\|_1, \\
&\geq \frac{\kappa}{2}\|\Delta\|_2^2 - \frac{3}{2}\lambda\sqrt{d}\|\Delta\|_2 \ (\because \sup_{u\in\mathcal{M}\setminus\{0\}}\frac{\|u\|_1}{\|u\|_2} = \sqrt{d}\ ), \\
&= (\frac{\kappa}{2}\|\Delta\|_2 - \frac{3}{2}\lambda\sqrt{d})\|\Delta\|_2.
\end{aligned}$$

Thus, if $\|\Delta\|_2 = \delta > \frac{3}{\kappa}\lambda\sqrt{d}$, then, $\mathsf{F}(\Delta) > 0$ for all $\Delta \in K(\delta)$. Note that $\mathsf{F}(0) = 0$. Using Lemma 4 from the Supplementary material of Negahban et al. (2012) (uses convexity of $\mathsf{F}(\Delta)$), it then follows that $\|\hat{\Delta}\|_2 \leq \delta$, that is, $\|\hat{W}_i - W_i\|_2 \leq (\frac{3}{k}\lambda\sqrt{d})$. □

3 RESULTS ON COVARIANCE MATRICES

Here we present few results on norms involving state covariance and error matrices that are used in the proofs in the main document.

Lemma 3.1. *Suppose $Z := \mathcal{X}\Delta \in \mathbb{C}^n$ with non-i.i.d. rows in $\mathcal{X}$ and Δ. The real component of Z is given by $Z_R = [\mathcal{X}_R \ -\mathcal{X}_I]v$ and the imaginary component is $Z_I = [\mathcal{X}_I \ \mathcal{X}_R]v$, for $v = [\Delta_R^T, \Delta_I^T]^T$. If $N > \frac{4CU\delta^{-1}}{(1-\delta^{-1})^2}$, then $\|\mathbb{E}[Z_R Z_R^T]\|_2 + \|\mathbb{E}[Z_I Z_I^T]\|_2 \leq 2\|\Delta\|_2^2[\frac{1}{L} + \frac{1}{2U} + 4\sqrt{8}\frac{C}{\delta-1}]$.*

Proof. For $r \in \{1, \cdots, n\}$, X_i^r (r^{th} entry of $\mathcal{Y}$) and $(X_{\bar{i}}^r)^T$ (r^{th} row of $\mathcal{X}$) are computed using N consecutive samples $\{x((r-1)N), \cdots, x((r-1)N+N-1)\}$, using Eq. 4. For $r,\ c \in \{1, \cdots, n\}$ we have,

$$|\mathbb{E}[Z_R Z_R^T](r,c)| = |v^T \mathbb{E}\begin{bmatrix} [\mathcal{X}_R(r,:)]^T\mathcal{X}_R(c,:) & -[\mathcal{X}_R(r,:)]^T\mathcal{X}_I(c,:) \\ -[\mathcal{X}_I(r,:)]^T\mathcal{X}_R(c,:) & [\mathcal{X}_I(r,:)]^T\mathcal{X}_I(c,:) \end{bmatrix} v|,$$

$$\Rightarrow |\mathbb{E}[Z_R Z_R^T](r,c)| \leq \|\Delta\|_2^2 \|\mathbb{E}\begin{bmatrix} [\mathcal{X}_R(r,:)]^T\mathcal{X}_R(c,:) & -[\mathcal{X}_R(r,:)]^T\mathcal{X}_I(c,:) \\ -[\mathcal{X}_I(r,:)]^T\mathcal{X}_R(c,:) & [\mathcal{X}_I(r,:)]^T\mathcal{X}_I(c,:) \end{bmatrix}\|_2. \tag{ii}$$

$$\text{Similarly, } |\mathbb{E}[Z_I Z_I^T](r,c)| \leq \|\Delta\|_2^2 \|\mathbb{E}\begin{bmatrix} [\mathcal{X}_I(r,:)]^T\mathcal{X}_I(c,:) & [\mathcal{X}_I(r,:)]^T\mathcal{X}_R(c,:) \\ [\mathcal{X}_R(r,:)]^T\mathcal{X}_I(c,:) & [\mathcal{X}_R(r,:)]^T\mathcal{X}_R(c,:) \end{bmatrix}\|_2. \tag{iii}$$

Consider $\mathcal{Y}[1] = \frac{1}{\sqrt{N}}\sum_{t=0}^{N-1} x_i(t)e^{-\iota ft}$ and $\mathcal{Y}[2] = \frac{1}{\sqrt{N}}\sum_{s=0}^{N-1} x_i(s+N)e^{-\iota fs}$. The correlation between $\mathcal{Y}_R[1]$ and $\mathcal{Y}_R[2]$ is given by

$$\begin{aligned} |\mathbb{E}[\mathcal{Y}_R[1]\mathcal{Y}_R[2]]| &= |\mathbb{E}[\frac{1}{N}\sum_{t=0}^{N-1}\sum_{s=0}^{N-1} x_i(t)x_i(s+N)\cos(ft)\cos(fs)]| \\ &= |\frac{1}{N}\sum_{t=0}^{N-1}\sum_{s=0}^{N-1} \mathbb{E}[x_i(t)x_i(s+N)]\cos(ft)\cos(fs)| \\ &\leq \frac{1}{N}\sum_{t=0}^{N-1}\sum_{s=0}^{N-1} |R_i(t-s-N)| = \frac{1}{N}\sum_{q=-(N-1)}^{N-1}(N-|q|)|R_i(q-N)|. \end{aligned}$$

Expanding in time-domain (see Lemma 3.1's proof of the Supplementary material), it can be shown that

$$\|\mathbb{E}[\mathcal{X}_R(r,:)^T\mathcal{X}_R(c,:)]\|_2 \leq B^{rc}, \ \text{ where } \ B^{rc} := \frac{1}{N}\sum_{q=-(N-1)}^{N-1}(N-|q|)\|R_x(q+(r-c)N)\|_2. \tag{iv}$$

Similarly, $\|\mathbb{E}[\mathcal{X}_R(r,:)^T\mathcal{X}_I(c,:)]\|_2$, $\|\mathbb{E}[\mathcal{X}_I(r,:)^T\mathcal{X}_I(c,:)]\|_2$, $\|\mathbb{E}[\mathcal{Y}_R[r]\mathcal{X}_I(c,:)]\|_2$, $\|\mathbb{E}[\mathcal{Y}_R[r]\mathcal{X}_R(c,:)]\|_2$, $\|\mathbb{E}[\mathcal{Y}_R[r]\mathcal{Y}_I[c]]\|_2$ and $\|\mathbb{E}[\mathcal{Y}_R[r]\mathcal{Y}_R[c]]\|_2$ are each upper bounded by B^{rc}. From Eqs. ii, iii, it follows that, $|\mathbb{E}[Z_R Z_R^T](r,c)| \leq \|\Delta\|_2^2\sqrt{8}B^{rc}$ and $|\mathbb{E}[Z_I Z_I^T](r,c)| \leq \|\Delta\|_2^2\sqrt{8}B^{rc}$. Thus,

$$\begin{aligned} \|\mathbb{E}[Z_R Z_R^T]\|_2 + \|\mathbb{E}[Z_I Z_I^T]\|_2 &\leq \max_{r=1}^n \sum_{c=1}^n |\mathbb{E}[Z_R Z_R^T](r,c)| + \max_{r=1}^n \sum_{c=1}^n |\mathbb{E}[Z_I Z_I^T](r,c)| \\ &\leq 2\max_{r=1}^n \sum_{c=1}^n (|\mathbb{E}[Z_R Z_R^T](r,c)| + |\mathbb{E}[Z_I Z_I^T](r,c)|) \\ &= 2[v^T(\Sigma_1+\Sigma_2)v + \max_r \sum_{c=1,c\neq r}^n (|\mathbb{E}[Z_R Z_R^T](r,c)| + |\mathbb{E}[Z_I Z_I^T](r,c)|) \\ &\leq 2\|\Delta\|_2^2[(\frac{1}{L}+\frac{1}{2U}) + \max_r \sum_{c,c\neq r} 2\sqrt{8}B^{rc}]. \end{aligned} \tag{v}$$

In the remaining, we find an upper bound for $\max_r \sum_{c,c\neq r} B^{rc}$. From Eq. iv,

$$B^{rc} \leq \frac{1}{N}\sum_{q=-(N-1)}^{N-1}(N-|q|)C\delta^{-|q-cN+rN|}, \quad (\because \|R_x(\tau)\|_2 \leq C\delta^{-|\tau|} \text{ from Eq. 9}),$$
$$= \frac{C}{N}[\sum_{q=1}^{N-1}(N-q)(\delta^{-|-q-cN+rN|}+\delta^{-|q-cN+rN|})] + C\delta^{-|-cN+rN|},$$
$$= C\delta^{-|r-c|N}[\sum_{q=1}^{N-1}(1-\frac{q}{N})(\delta^q+\delta^{-q})+1]$$
$$= C\delta^{-|r-c|N}[1+S_a+S_b-\frac{S_c}{N}-\frac{S_d}{N}],$$
$$\text{where, } S_a=\sum_{q=1}^{N-1}\delta^q=\frac{\delta^N-\delta}{\delta-1},\ S_c=\sum_{q=1}^{N-1}q\delta^q=\frac{\delta-\delta^N}{(\delta-1)^2}+\frac{(N-1)\delta^N}{\delta-1},$$
$$S_b=\sum_{q=1}^{N-1}\delta^{-q}=\frac{1-\delta^{-(N-1)}}{\delta-1},\ S_d=\sum_{q=1}^{N-1}q\delta^{-q}=\frac{\delta^{-1}-\delta^{-N}}{(1-\delta^{-1})^2}-\frac{(N-1)\delta^{-N}}{1-\delta^{-1}}.$$

$$\text{Thus, } \max_r \sum_{c=1,c\neq r}^{n} B^{rc} \leq C[1+S_a+S_b-\frac{S_c}{N}-\frac{S_d}{N}]\sum_{c=1,c\neq r}^{n}\delta^{-|r-c|N}$$
$$\leq C[1+S_a+S_b-\frac{S_c}{N}-\frac{S_d}{N}]\sum_{c=1}^{\infty}2\delta^{-cN}$$
$$\leq C[1+S_a+S_b-\frac{S_c}{N}-\frac{S_d}{N}][\frac{2\delta^{-N}}{1-\delta^{-N}}],$$
$$= C[1+S_a+\frac{S_c}{N}(\delta^{-N}-1)][\frac{2\delta^{-N}}{1-\delta^{-N}}] \ (\because \delta^N(S_b-\frac{S_d}{N})=\frac{S_c}{N}),$$
$$\leq C(1+S_a)\frac{2\delta^{-N}}{1-\delta^{-N}} \ \ (\because \delta^{-N}-1 \text{ is negative and can be ignored}),$$
$$\leq C(1+\frac{\delta^N-\delta}{\delta-1})\frac{2\delta^{-N}}{1-\delta^{-N}} = C(\frac{\delta^N-1}{\delta-1})(\frac{2}{\delta^N-1}) = \frac{2C}{\delta-1}. \tag{vi}$$

Substituting Eq. vi in Eq. v gives $\|\mathbb{E}[Z_R Z_R^T]\|_2 + \|\mathbb{E}[Z_I Z_I^T]\|_2 \leq 2\|\Delta\|_2^2[\frac{1}{L}+\frac{1}{2U}+4\sqrt{8}\frac{C}{\delta-1}]$. □

Lemma 3.2 (covariance (restart & record))**.** *Let $\mathcal{E} := \mathcal{Y}-\mathcal{X}W_i$ with each row corresponding to an i.i.d. trajectory. Let $\mathcal{E}_1 := [\mathcal{E}_R[1]\ \mathcal{E}_I[1]\ ...\ \mathcal{E}_R[n]\ \mathcal{E}_I[n]]^T$ be the re-arranged vector of real and complex entries in $\mathcal{E}$, with covariance matrix $\mathcal{C}_1 = \mathbb{E}[\mathcal{E}_1\mathcal{E}_1^T]$. Then $\|\mathcal{C}_1\|_2 \leq \frac{3}{2L}$.*

Proof. As the n trajectories are i.i.d.,

$$\mathcal{C}_1 := diag(\mathsf{C},...,\mathsf{C}), \text{where} \begin{pmatrix} \mathcal{E}_R(j) \\ \mathcal{E}_I(j) \end{pmatrix} \sim \mathcal{N}(\mathbf{0},\mathsf{C}) \Rightarrow \|\mathcal{C}_1\|_2 = \|\mathsf{C}\|_2. \tag{vii}$$

Consider $\mathsf{E} = \Phi_i - \Phi_{i,\bar{i}}W_i - (W_i)^H\Phi_{\bar{i},i} + (W_i)^H\Phi_{\bar{i}}W_i$. Then, $(1+\|W_i\|_2^2)\frac{1}{U} \leq |\mathsf{E}| \leq (1+\|W_i\|_2^2)\frac{1}{L}$. Substituting $W_i = [\Phi_{\bar{i}}]^{-1}\Phi_{i,\bar{i}}$ in E and using the Schur complement lemma, we get,

$$|\mathsf{E}| = \Phi_i - \Phi_{i,\bar{i}}(\Phi_{\bar{i}})^{-1}\Phi_{\bar{i},i} = \frac{1}{[\Phi_x]^{-1}(i,i)}. \tag{viii}$$

From the definition of L, U in Eq. 9, we have, $\frac{1}{U} \leq |\mathsf{E}| \leq \frac{1}{L}$. Comparing the two inequality bounds of $|\mathsf{E}|$, we get

$$\frac{L}{U} \leq (1+\|W_i\|_2^2) \leq \frac{U}{L}. \tag{ix}$$

From the definition of $\mathcal{E}$, it follows that $\mathbb{E}[\mathcal{E}[j]^H\mathcal{E}[j]] = \hat{\Phi}_i - \hat{\Phi}_{i,\bar{i}}W_i - (W_i)^H\hat{\Phi}_{\bar{i},i} + (W_i)^H\hat{\Phi}_{\bar{i}}W_i$. Define $V = \hat{\Phi}_x - \Phi_x$, and from Lemma 5.1 in the article, we have $\|V\|_2 \leq \frac{1}{2U}$. Substituting $\hat{\Phi}_x = \Phi_x + V$ in $\mathbb{E}[\mathcal{E}[j]^H\mathcal{E}[j]]$, we get,

$$\begin{aligned}
&\mathbb{E}[\mathcal{E}[j]^H\mathcal{E}[j]] = \frac{1}{[\Phi_x]^{-1}(1,1)} + V_i - V_{i,\bar{i}}W_i - (W_i)^H V_{\bar{i},i} + (W_i)^H V_{\bar{i}}W_i \ (\because \text{ using Eq. viii}),\\
&\Rightarrow \mathbb{E}(\mathcal{E}_R[j]^2) + \mathbb{E}(\mathcal{E}_I[j]^2) = Tr(\mathsf{C}) \leq \frac{1}{L} + (1 + |W_i|_2^2)\|V\|_2 \leq \frac{1}{L} + \frac{U}{L}\frac{1}{2U} = \frac{3}{2L} \ (\because \text{ using Eq. ix})\\
&\Rightarrow \|\mathcal{C}_1\|_2 = \|\mathsf{C}\|_2 \leq \frac{3}{2L}. \qquad \text{(x)}
\end{aligned}$$

□

Lemma 3.3 (covariance (consecutive))**.** *Let $\mathcal{E} := \mathcal{Y} - \mathcal{X}W_i$ with non-i.i.d. trajectories per row. Let $\mathcal{E}_1 := [\mathcal{E}_R[1]\ \mathcal{E}_I[1]\ ...\ \mathcal{E}_R[n]\ \mathcal{E}_I[n]]^T$ be the re-arranged vector of real and complex entries in $\mathcal{E}$ with covariance matrix $\mathcal{C}_1 = \mathbb{E}[\mathcal{E}_1\mathcal{E}_1^T]$ Then $\|\mathcal{C}_1\|_2 \leq \frac{3}{2L} + 6\sqrt{3}\frac{U}{L}\frac{2C}{\delta-1}$.*

Proof. Writing real and imaginary parts of $\mathcal{E}_1$ and using inequality of matrix 2 and ∞-norms, we have,

$$\begin{aligned}
\|\mathcal{C}_1\|_2 &\leq \max\left[\max_{r=1}^{n}\sum_{c=1}^{n}\left(|\mathbb{E}[\mathcal{E}_R\mathcal{E}_R^T](r,c)| + |\mathbb{E}[\mathcal{E}_R\mathcal{E}_I^T](r,c)|\right), \max_{r=1}^{n}\sum_{c=1}^{n}\left(|\mathbb{E}[\mathcal{E}_I\mathcal{E}_R^T](r,c)| + |\mathbb{E}[\mathcal{E}_I\mathcal{E}_I^T](r,c)|\right)\right],\\
&\leq \max\begin{bmatrix}\max_{r=1}^{n}\sum_{c=1,c\neq r}^{n}\left(|\mathbb{E}[\mathcal{E}_R\mathcal{E}_R^T](r,c)| + |\mathbb{E}[\mathcal{E}_R\mathcal{E}_I^T](r,c)|\right),\\ \max_{r=1}^{n}\sum_{c=1,c\neq r}^{n}\left(|\mathbb{E}[\mathcal{E}_I\mathcal{E}_R^T](r,c)| + |\mathbb{E}[\mathcal{E}_I\mathcal{E}_I^T](r,c)|\right)\end{bmatrix} + \frac{3}{2L} \quad (\because \text{ Using Eq. x}). \qquad \text{(xi)}
\end{aligned}$$

Split $\mathcal{E} = \mathcal{Y} - \mathcal{X}W_i$ into their real and imaginary parts (subscripted by R and I respectively). Now $\mathcal{E}_R = \begin{bmatrix}\mathcal{Y}_R & -\mathcal{X}_R & \mathcal{X}_I\end{bmatrix}\begin{bmatrix}1 & (W_i)_R & (W_i)_I\end{bmatrix}^T$, and $\mathcal{E}_I = \begin{bmatrix}\mathcal{Y}_I & -\mathcal{X}_I & -\mathcal{X}_R\end{bmatrix}\begin{bmatrix}1 & (W_i)_R & (W_i)_I\end{bmatrix}^T$. For $r, c \in \{1, \cdots, n\}$, $r \neq c$, the correlation between the r^{th} and c^{th} sample is

$$\begin{aligned}
&\mathbb{E}[\mathcal{E}_R[r]\mathcal{E}_R[c]] =\\
&\begin{bmatrix}1 & (W_i^T)_R & (W_i^T)_I\end{bmatrix}\begin{bmatrix}\mathbb{E}(\mathcal{Y}_R[r]\mathcal{Y}_R[c]) & -\mathbb{E}(\mathcal{Y}_R[r]\mathcal{X}_R(c,:)) & \mathbb{E}(\mathcal{Y}_R[r]\mathcal{X}_I(c,:))\\ -\mathbb{E}(\mathcal{X}_R(r,:)^T\mathcal{Y}_R[c]) & \mathbb{E}(\mathcal{X}_R(r,:)^T\mathcal{X}_R(c,:)) & -\mathbb{E}(\mathcal{X}_R(r,:)^T\mathcal{X}_I(c,:))\\ \mathbb{E}(\mathcal{X}_I(r,:)^T\mathcal{Y}_R[c]) & -\mathbb{E}(\mathcal{X}_I(r,:)^T\mathcal{X}_R(c,:)) & \mathbb{E}(\mathcal{X}_I(r,:)^T\mathcal{X}_I(c,:))\end{bmatrix}\begin{bmatrix}1\\(W_i)_R\\(W_i)_I\end{bmatrix}.
\end{aligned}$$

From Eq. iv and Eq. ix, it follows that, $|\mathbb{E}[\mathcal{E}_R[r]\mathcal{E}_R[c]]| \leq \frac{U}{L}\sqrt{27}B^{rc}$. Similarly,

$$|\mathbb{E}[\mathcal{E}_R[r]\mathcal{E}_I[c]]| \leq \frac{U}{L}\sqrt{27}B^{rc}, |\mathbb{E}[\mathcal{E}_I[r]\mathcal{E}_R[c]]| \leq \frac{U}{L}\sqrt{27}B^{rc}, |\mathbb{E}[\mathcal{E}_I[r]\mathcal{E}_I[c]]| \leq \frac{U}{L}\sqrt{27}B^{rc}. \qquad \text{(xii)}$$

Using the inequalities Eq. xii in Eq. xi, we have $\|\mathcal{C}_1\|_2 \leq \frac{3}{2L} + 2\sqrt{27}\sum_{c=1,c\neq r}^{n}\frac{U}{L}B^{rc}$. Then, it follows from Eq. vi,

$$\|\mathcal{C}_1\|_2 \leq \frac{3}{2L} + 6\sqrt{3}\frac{U}{L}\frac{2C}{\delta-1}. \qquad \text{(xiii)}$$

□

3.1 Proofs of M-estimator conditions for consecutive trajectories

Proof of Lemma 3.3 from the article. The approach for the proof is identical to the proof of Lemma 3.1 from the article, with few changes. Define $\mathcal{E} := \mathcal{Y} - \mathcal{X}W_i$. Let $\mathcal{C}_1$ be the covariance matrix of the vector $\mathcal{E}_1 := [\mathcal{E}_R[1]\ \mathcal{E}_I[1]\ ...\ \mathcal{E}_R[n]\ \mathcal{E}_I[n]]^T$. The trajectories aren't independent and $\mathcal{C}_1$ is no more block-diagonal here. An upper bound for $\|\mathcal{C}_1\|_2$ for this case is provided in Lemma 3.3 in this supplementary material. Using that, the Lipschitz constant of $f(\mathcal{W}_R, \mathcal{W}_I)$ in Eq. 17 becomes $\sqrt{\frac{3+24\sqrt{3}UC(\delta-1)^{-1}}{2nL}}$. Following Lemma 3.1 of the article, $\lambda \geq 4\sqrt{\frac{(3+24\sqrt{3}UC(\delta-1)^{-1})\log(4p/\epsilon_3)}{nL}}$ gives the result. □

Proof of Lemma 3.4 from the article. Let $Z := \mathcal{X}\Delta \in \mathbb{C}^n$. Its real and imaginary components are $Z_R = [\mathcal{X}_R\ -\mathcal{X}_I]v$ and $Z_I = [\mathcal{X}_I\ \mathcal{X}_R]v$ where $v = [\Delta_R^T\ \Delta_I^T]^T$. We find the lower bounds on $\frac{1}{n}\|Z_R\|_2^2$, $\frac{1}{n}\|Z_I\|_2^2$ and then combine them to obtain a lower bound of $\frac{1}{n}\|Z\|_2^2$. Applying Lemma I.2 from the Supplementary material of Negahban and

Wainwright (2011) on Z_R and Z_I, we have, with individual probability at least $1-[2\exp(-\frac{n(t-\frac{2}{\sqrt{n}})^2}{2})+2\exp(-\frac{n}{2})]$ for all $t \geq \frac{2}{\sqrt{n}}$,

$$\begin{aligned}\frac{1}{n}\|Z_R\|_2^2 &\geq \frac{1}{n}Tr[\mathbb{E}(Z_R Z_R^T)] - 4t\|\mathbb{E}(Z_R Z_R^T)\|_2, \\ \frac{1}{n}\|Z_I\|_2^2 &\geq \frac{1}{n}Tr[\mathbb{E}(Z_I Z_I^T)] - 4t\|\mathbb{E}(Z_I Z_I^T)\|_2. \end{aligned} \tag{xiv}$$

Note that the diagonal values of $\mathbb{E}(Z_R Z_R^T)$ are all equal to $v^T\Sigma_1 v$, and those of $\mathbb{E}(Z_I Z_I^T)$ are equal to $v^T\Sigma_2 v$ with Σ_1, Σ_2 defined in Eq. 20. Using this with Lemma 3.1 in the Supplementary material, we get, $\frac{1}{n}\|Z\|_2^2 = \frac{1}{n}\|Z_R\|_2^2 + \frac{1}{n}\|Z_I\|_2^2 \geq$

$$\begin{aligned} & v^T(\Sigma_1+\Sigma_2)v - 8t\|\Delta\|_2^2[\frac{1}{L}+\frac{1}{2U}+4\sqrt{8}\frac{C}{\delta-1}] \\ & \geq \frac{1}{2U}\|\Delta\|_2^2 - 8t\|\Delta\|_2^2[\frac{1}{L}+\frac{1}{2U}+4\sqrt{8}\frac{C}{\delta-1}], \ (\because \text{Eq. 23}) \end{aligned} \tag{xv}$$

holds with probability of at least $1-[4\exp(-\frac{n(t-\frac{2}{\sqrt{n}})^2}{2})+4\exp(-\frac{n}{2})]$ for $t \geq \frac{2}{\sqrt{n}}$. Choose $t = \sqrt{\frac{4\log p}{n}}$($>> \frac{2}{\sqrt{n}}$ for a large n). Then for $n \geq 33^2 \log p[\frac{U}{L}+0.5+4\sqrt{8}\frac{CU}{\delta-1}]^2$, $\frac{1}{n}\|\mathcal{X}\Delta\|_2^2 \geq \frac{1}{256U}\|\Delta\|_2^2$ holds with probability at least $1-[\frac{4}{p^2}+4\exp(-\frac{n}{2})]$. Since $p \geq \sqrt{\frac{4}{\epsilon_2}}$, the statement holds whenever $n \geq 2\log(\frac{4p^2}{p^2\epsilon_2-4})$. □