Proceedings of Machine Learning Research vol vvv:1–33, 2025

Finite Sample Identification of Partially Observed Bilinear Dynamical Systems

Yahya Sattar[⋆1] YSATTAR@CORNELL.EDU
Yassir Jedra[⋆2] JEDRA@MIT.EDU
Maryam Fazel[3] MFAZEL@UW.EDU
Sarah Dean[1] SDEAN@CORNELL.EDU
[⋆] *equal contribution*
[1] *Department of Computer Science, Cornell University*
[2] *Laboratory of Information and Decision Systems, MIT*
[3] *Department of Electrical and Computer Engineering, University of Washington*

Abstract

We consider the problem of learning a realization of a partially observed bilinear dynamical system (BLDS) from noisy input-output data. Given a single trajectory of input-output samples, we provide a finite time analysis for learning the system's Markov-like parameters, from which a balanced realization of the bilinear system can be obtained. Our bilinear system identification algorithm learns the Markov-like parameters by regressing the outputs to highly correlated, nonlinear, and heavy-tailed covariates. Moreover, the stability of a BLDS depends on the sequence of inputs used to excite the system. These properties, unique to partially observed bilinear dynamical systems, pose significant challenges to the analysis of our algorithm for learning the unknown dynamics. We address these challenges and provide high probability error bounds on our identification algorithm under a uniform stability assumption. Our analysis provides insights into system theoretic quantities that affect learning accuracy and sample complexity. Lastly, we perform numerical experiments with synthetic data to reinforce these insights.

Keywords: bilinear dynamical systems, single trajectory learning, partial observations, infinite impulse response.

1. Introduction

Learning the dynamical behavior of nonlinear systems is an important and challenging problem with applications ranging from engineering, physics, biology (Brunton et al., 2016; Strogatz, 2018; Brunton and Kutz, 2022), to language modeling, and sequence predictions (Kombrink et al., 2011; Bahdanau et al., 2014). Bilinear systems constitute a simpler yet powerful class of nonlinear systems naturally arising in a variety of domains from engineering, biology (Mohler, 1973), quantum mechanical processes (Pardalos and Yatsenko, 2010) to recommendation systems (Koren et al., 2021). Moreover, bilinear systems approximate a much broader class of nonlinear systems via Carleman linearization (Kowalski and Steeb, 1991) or Koopman canonical transform (Goswami and Paley, 2017; Bruder et al., 2021) of control-affine nonlinear systems (Svoronos et al., 1980; Lo, 1975). Therefore, learning the dynamics of BLDS from input-output data is an important and useful problem, which has attracted significant interest both in continuous-time (Juang, 2005; Sontag et al., 2009) and discrete-time (Berk Hizir et al., 2012). However, theoretical guarantees of learning BLDS from a single trajectory of noisy input-output data is lacking, with current guarantees existing only for bilinear systems with complete state observations (Sattar et al., 2022; Chatzikiriakos et al.,

2024). Our goal in this paper is to provide theoretical guarantees for learning partially observed BLDS from noisy input-output data sampled from a single trajectory. We achieve this by learning the system's *Markov-like parameters*, which uniquely identify the end-to-end behavior of the system, and can be used to recover the state-space matrices up to a similarity transform using existing algorithms (Ho and Kálmán, 1966; Sarkar et al., 2019b; Oymak and Ozay, 2021).

Our work relates to the problem of learning linear dynamical systems (LDS) from partial state-observations. In this setting, Tu et al. (2017); Oymak and Ozay (2021); Simchowitz et al. (2019); Sun et al. (2022); Djehiche and Mazhar (2022); Tsiamis and Pappas (2019); Sarkar et al. (2021); Bakshi et al. (2023) learn the system's Markov parameters or Hankel matrix (with different types of error bounds), which can then be used to recover the state-space matrices (up to a similarity transform) using classic Ho-Kalman Algorithm (Ho and Kálmán, 1966). The papers Sun et al. (2020, 2022); Fazel et al. (2013) study system identification with Hankel nuclear norm regularization. Other works have focused on learning the end-to-end behavior of partially observed LDS via gradient descent (Hardt et al., 2018) and spectral filtering (Hazan et al., 2017). The problem of learning with partial observations has been extended to learning linear dynamical systems with nonlinear (Mhammedi et al., 2020) and bilinear (Sattar et al., 2024) observations. However, to the best of our knowledge, sample complexity and non-asymptotic analysis for partially observed nonlinear dynamical systems (including BLDS) have not been considered before.

Non-asymptotic learning of (non)linear dynamical systems with complete state observations has also attracted significant attention recently. Most of the advancements in this direction are focused on linear systems (Faradonbeh et al., 2018; Dean et al., 2018; Simchowitz et al., 2018; Dean et al., 2019; Fattahi et al., 2019; Sarkar et al., 2021; Sarkar and Rakhlin, 2019; Lale et al., 2020; Jedra and Proutiere, 2020; Wagenmaker and Jamieson, 2020), where an optimal error rate is achieved by using either mixing-time (Yu, 1994) or martingale-based arguments (Abbasi-Yadkori et al., 2011). These results have been extended to switched linear dynamical systems (Sarkar et al., 2019a; Sattar et al., 2021; Du et al., 2022; Sayedana et al., 2024), as well as certain classes of nonlinear dynamical systems (Oymak, 2019; Bahmani and Romberg, 2019; Mhammedi et al., 2020; Sattar and Oymak, 2022; Jain et al., 2021; Ziemann et al., 2022; Musavi et al., 2024; Fujiwara et al., 2024). However, the problem of learning nonlinear dynamical systems from partial observations of a single trajectory is still an open problem. In this paper, we take a step towards addressing this problem by answering the following question:

Can we learn a partially observed bilinear dynamical system from a single trajectory?

Contributions: We provide theoretical guarantees for learning partially observed bilinear dynamical systems. The main novelty in this problem comes from the fact that the hidden states evolve according to a bilinear state equation, for which the analysis tools developed for learning partially observed LDS do not work. Moreover, the stability of a BLDS explicitly depends on the input sequence, which is in stark contrast to the deterministic notion of stability in the case of LDS. These properties make the problem of learning partially observed bilinear dynamical systems challenging. In this paper we make the following major contributions towards bilinear system identification:

- **Sample complexity and error bounds:** We provide the first sample complexity analysis and finite-sample error bounds for learning a realization of a partially observed BLDS from a single trajectory of input-output data. Unlike LDS, the output of a bilinear system maps to the history of inputs via nonlinear features (obtained by the Kronecker products of past inputs) and a sequence of Markov-like parameters with exponentially increasing length. Our main result (Theorem 2)

provides $\tilde{\mathcal{O}}(1/\sqrt{T})$ error rate for learning these parameters from a single trajectory of length T. Our sample complexity lower bound $\tilde{\Omega}\left((p+1)^{L+1}\right)$ grows exponentially with the history length L (where p is the input dimension) due to the exponential increase in number of Markov-like parameters. For stable bilinear systems (defined in §2.2), this can be mitigated by choosing a smaller history of inputs (see § 5).

- **Input choice and stability:** Stability of BLDS is typically input dependent. We work with a novel notion of stability (Definition 1) that generalizes the classic notion of stability for LDS to the BLDS. We also define a notion of stability radius which governs our choice of inputs.
- **Persistence of excitation:** Of independent interest, we establish persistence of excitation (Theorem 5) for a broad class of inputs (possibly heavy-tailed) satisfying a hyper-contractivity condition. Our persistence of excitation result holds for nonlinear, correlated, and heavy-tailed covariates (i.e., the input features).
- **Numerical experiments:** Lastly, we perform experiments with synthetic data to verify our theoretical findings. Interestingly, our experiments show that exciting the system with inputs sampled uniformly at random from a sphere leads to better estimation of Markov-like parameters as compared to Gaussian inputs, further reinforcing our choice of inputs.

The rest of the paper is organized as follows: §2 sets up the problem and introduces the notion of stability. §3 provides our main result on learning Markov-like parameters of partially observed BLDS. §4 discusses our proof idea, and provides persistence of excitation result for a broader class of inputs. Lastly[1], we perform numerical experiments in §5, and conclude with a discussion of future directions in §6.

Notations: We use boldface lowercase (uppercase) letters to denote vectors (matrices). $\rho(\boldsymbol{X})$, $\|\boldsymbol{X}\|_{\text{op}}$ and $\|\boldsymbol{X}\|_F$ denote the spectral radius, spectral norm and Frobenius norm of a matrix $\boldsymbol{X}$, respectively. $\|\boldsymbol{v}\|_{\ell_2}$ denotes the Euclidean norm of a vector $\boldsymbol{v}$, and $(\boldsymbol{v})_i$ denotes its i-th element. For a positive definite matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, the Mahalanobis norm of a vector $\boldsymbol{v} \in \mathbb{R}^d$ is given by $\|\boldsymbol{v}\|_{\boldsymbol{M}} = \sqrt{\boldsymbol{v}^\top \boldsymbol{M}\boldsymbol{v}}$. For a sequence of $d \times d$ matrices $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_k$, we use the convention that $\prod_{i=1}^{k} \boldsymbol{M}_i = \boldsymbol{M}_1 \times \boldsymbol{M}_2 \times \cdots \times \boldsymbol{M}_k$. We denote by $a \vee b$, the maximum of two scalars a and b. We use $\gtrsim$ and $\lesssim$ for inequalities that hold up to an absolute constant factor. $\tilde{\mathcal{O}}(\cdot)$ and $\tilde{\Omega}(\cdot)$ are used to show the dependence on a specific quantity of interest (up to constants and logarithmic factors). Finally, $\otimes$ denotes the Kronecker product.

2. Preliminaries

2.1. Problem Formulation

Consider a partially observed bilinear dynamical system with the following state-space representation: for all $t \geq 0$,

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \boldsymbol{A}_0 \boldsymbol{x}_t + \sum_{k=1}^{p} (\boldsymbol{u}_t)_k \boldsymbol{A}_k \boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t + \boldsymbol{w}_t, \\ \boldsymbol{y}_t &= \boldsymbol{C}\boldsymbol{x}_t + \boldsymbol{D}\boldsymbol{u}_t + \boldsymbol{z}_t, \end{aligned} \tag{2.1}$$

where $\boldsymbol{x}_t \in \mathbb{R}^n$, $\boldsymbol{u}_t \in \mathbb{R}^p$, and $\boldsymbol{y}_t \in \mathbb{R}^m$, $\boldsymbol{w}_t \in \mathbb{R}^n$, and $\boldsymbol{z}_t \in \mathbb{R}^m$ represent the hidden state, input, output, process noise, and measurement noise, respectively, at time t. Without loss of generality, we consider that $x_0 = 0$. The noise processes $\{\boldsymbol{w}_t\}_{t\geq 0}$ and $\{\boldsymbol{z}_t\}_{t\geq 0}$ are assumed to be sequences of independent zero-mean, σ^2-subgaussian random vectors taking values in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively,

1. For the supplementary material which contains the appendix we refer the readers to the extended-version.

for some variance proxy parameter $\sigma > 0$. The matrices $\boldsymbol{A}_0, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_p \in \mathbb{R}^{n\times n}$, $\boldsymbol{B} \in \mathbb{R}^{n\times p}$, $\boldsymbol{C} \in \mathbb{R}^{m\times n}$, and $\boldsymbol{D} \in \mathbb{R}^{m\times p}$ represent the parameters that govern the evolution of the dynamical system and are a priori unknown.

In this work, we wish to identify the unknown parameters of the system from a single trajectory of input-output samples $\{(\boldsymbol{u}_t, \boldsymbol{y}_t)\}_{t=1}^{T}$. To that end, we focus on the task of learning the so-called *Markov-like parameters* (detailed in §3) of the system. Once learned, these Markov-like parameters can be exploited using the classic Ho-Kalman algorithm to recover the unknown matrices of the system up to some similarity transform as will be described in §3. Next, we clarify our choice of inputs and discuss the required stability assumption.

2.2. Input Choice & Stability of Bilinear Dynamical Systems

Stability of bilinear dynamical systems is typically input dependent. To see that, we can unroll the state dynamics in (2.1) to write: for all $t \geq 0$,

$$\boldsymbol{x}_{t+1} = \sum_{\ell=0}^{t} \left(\prod_{k=0}^{\ell-1} (\boldsymbol{u}_{t-k} \circ \boldsymbol{A}) \right) \boldsymbol{B}\boldsymbol{u}_{t-\ell} + \sum_{\ell=0}^{t} \left(\prod_{k=0}^{\ell-1} (\boldsymbol{u}_{t-k} \circ \boldsymbol{A}) \right) \boldsymbol{w}_{t-\ell}, \tag{2.2}$$

where we define $\boldsymbol{u}_t \circ \boldsymbol{A} := \boldsymbol{A}_0 + \sum_{k=1}^{p} (\boldsymbol{u}_t)_k \boldsymbol{A}_k$ for the ease of notation. Then, observe that the products of matrices $\prod_{k=0}^{\ell-1} (\boldsymbol{u}_{t-k} \circ \boldsymbol{A})$ may grow exponentially in norm if we consistently choose large inputs. This is precisely why stability in bilinear dynamical systems is more challenging than other classes of systems such as linear dynamical systems or switched systems. Before discussing stability, we first start by clarifying the choice of inputs we focus on.

Input choice: We consider that the inputs $\{\boldsymbol{u}_t\}_{t\geq 0}$ are sampled in an i.i.d. manner from some distribution $\mathcal{D}_{\boldsymbol{u}}$ on $\mathbb{R}^p$. For ease of exposition, we will focus on the case where inputs are sampled uniformly at random from a sphere of radius $\sqrt{p}$, that is, $\boldsymbol{u}_t \sim \mathrm{Unif}(\sqrt{p} \cdot \mathcal{S}^{p-1})$. More generally, as long as the inputs are isotropic and are bounded with high probability, our results will still hold at the expanse of longer proofs.

Traditionally, notions like Mean Square Stability (MSS) have been considered to reason about the stability behavior of bilinear systems (Kubrusly and Costa, 1985; Pardalos and Yatsenko, 2010; Sattar et al., 2022). Typically, these notions are asymptotic in nature, input dependent, and may not allow us to obtain tight guarantees. We will introduce an alternative notion of stability, akin to that considered by Monfared et al. (2023), that naturally generalizes the classical notion of stability in standard LTI systems.

Uniform stability in bilinear dynamical systems: First, let us recall that the *joint spectral radius* of a set of matrices $\mathcal{M} \subseteq \mathbb{R}^{n\times n}$ can be defined as follows:

$$\rho(\mathcal{M}) := \lim_{k\to\infty} \sup_{\boldsymbol{M}_1,\ldots,\boldsymbol{M}_k \in \mathcal{M}} \|\boldsymbol{M}_1\boldsymbol{M}_2\cdots\boldsymbol{M}_k\|_{\mathrm{op}}^{1/k}. \tag{2.3}$$

For $\rho > 0$, we define the following quantity:

$$\phi(\mathcal{M}, \rho) := \sup_{k\geq 1, \boldsymbol{M}_1,\ldots,\boldsymbol{M}_k \in \mathcal{M}} \frac{\|\boldsymbol{M}_1\boldsymbol{M}_2\cdots\boldsymbol{M}_k\|_{\mathrm{op}}}{\rho^k}. \tag{2.4}$$

The quantity $\phi(\mathcal{M}, \rho)$ is defined in similar vein to that by Mania et al. (2019) and is introduced to capture the transient behavior of a system with state transition matrices varying in $\mathcal{M}$. We note that if $\rho(\mathcal{M}) < 1$, then for any $\rho > \rho(\mathcal{M})$, the quantity $\phi(\mathcal{M}, \rho)$ is finite. Now, given a set $\mathcal{U} \subseteq \mathbb{R}^p$, we denote $\mathcal{U} \circ \boldsymbol{A} := \{\boldsymbol{A}_0 + \sum_{i=1}^p (\boldsymbol{u})_i \boldsymbol{A}_i : \boldsymbol{u} \in \mathcal{U}\}$ and introduce the following definition of stability.

Definition 1 ($(\mathcal{U}, \kappa, \rho)$-uniform-stability) *Let $\mathcal{U} \subseteq \mathbb{R}^p$, $\kappa \geq 1$, and $0 < \rho < 1$. We say that a partially observed bilinear dynamical system (as defined in (2.1)) with state-transition matrices $\boldsymbol{A} := \{\boldsymbol{A}_0, \ldots, \boldsymbol{A}_p\}$ is $(\mathcal{U}, \kappa, \rho)$-uniformly-stable, if the joint spectral radius of the set $\mathcal{U} \circ \boldsymbol{A}$ satisfies: (i) $\rho(\mathcal{U} \circ \boldsymbol{A}) < \rho < 1$; and (ii) $\phi(\mathcal{U} \circ \boldsymbol{A}, \rho) \leq \kappa$.*

Observe that if there exists a nonempty and bounded set $\mathcal{U} \subseteq \mathbb{R}^p$ such that $\rho(\mathcal{U} \circ \boldsymbol{A}) < 1$, then for any $\rho(\mathcal{U} \circ \boldsymbol{A}) < \rho < 1$, the system is $(\mathcal{U}, \kappa, \rho)$-uniformly-stable with $\kappa = \phi(\mathcal{U} \circ \boldsymbol{A}, \rho) \vee 1$. We provide detailed discussion on this claim in Appendix A. Furthermore, we note that Definition 1 naturally generalizes that of Monfared et al. (2023). Indeed, there the authors assume that there exists $\boldsymbol{u}^\star$ such that $\rho(\boldsymbol{u}^\star \circ \boldsymbol{A}) < 1$. This is equivalent to assuming that their system is $(\{\boldsymbol{u}^\star\}, \kappa, \rho)$-uniformly-stable for some $\kappa \geq 1$ and $\rho(\boldsymbol{u}^\star \circ \boldsymbol{A}) < \rho < 1$. As we shall precise shortly, we need stronger requirements on the stability of the system in comparison with Monfared et al. (2023), because we are concerned with the task of identification. This requirement stems from the need to have persistence of excitation so that estimation is possible. We are now ready to present the assumption we make on the stability of the bilinear system (2.1).

Assumption 1 (Stability) *There exists $\kappa \geq 1$ and $\rho \in (0, 1)$ such that the partially observed bilinear dynamical system (2.1) is $(\sqrt{p} \cdot \mathcal{S}^{p-1}, \kappa, \rho)$-uniformly-stable.*

In view of Assumption 1, choosing inputs uniformly at random from $\sqrt{p} \cdot \mathcal{S}^{p-1}$ guarantees stability almost surely. More generally, we can choose to sample inputs from any set $\mathcal{U}$, so long as the system is stable under such set in the sense of Definition 1. However, the quality of estimation depends on whether inputs sampled from $\mathcal{U}$ are persistently exciting or not (see §4.1).

3. Learning the Markov-like Parameters

The Markov-like parameters are defined in a similar vein to the classical Markov parameters for partially observed linear systems. By unrolling the dynamics (2.1), we can represent the output $\boldsymbol{y}_t$ in terms of the past L inputs $\boldsymbol{u}_{t-L}, \ldots, \boldsymbol{u}_t$, for any $t \geq L$, as follows:

$$\boldsymbol{y}_t = \boldsymbol{C}\left(\prod_{\ell=1}^{L-1} (\boldsymbol{u}_{t-\ell} \circ \boldsymbol{A})\right)\boldsymbol{x}_{t-L} + \sum_{\ell=1}^{L-1} \boldsymbol{C}\left(\prod_{i=1}^{\ell-1} (\boldsymbol{u}_{t-i} \circ \boldsymbol{A})\right)(\boldsymbol{B}\boldsymbol{u}_{t-\ell} + \boldsymbol{w}_{t-\ell}) + \boldsymbol{D}\boldsymbol{u}_t + \boldsymbol{z}_t. \quad (3.1)$$

We can simplify the form of (3.1) by adopting a more convenient notation and expanding further some of the products that involve the inputs. First, we introduce $\boldsymbol{\epsilon_t} := \boldsymbol{C}\left(\prod_{\ell=1}^{L-1} (\boldsymbol{u}_{t-\ell} \circ \boldsymbol{A})\right)\boldsymbol{x}_{t-L}$, and define: $\bar{\boldsymbol{u}}_t^\top = \begin{bmatrix} 1 & \boldsymbol{u}_t^\top \end{bmatrix}$,

$$\tilde{\boldsymbol{u}}_t := \begin{bmatrix} \boldsymbol{u}_t \\ \boldsymbol{u}_{t-1} \\ \bar{\boldsymbol{u}}_{t-1} \otimes \boldsymbol{u}_{t-2} \\ \bar{\boldsymbol{u}}_{t-1} \otimes \bar{\boldsymbol{u}}_{t-2} \otimes \boldsymbol{u}_{t-3} \\ \vdots \\ \bar{\boldsymbol{u}}_{t-1} \otimes \bar{\boldsymbol{u}}_{t-2} \otimes \cdots \otimes \boldsymbol{u}_{t-L} \end{bmatrix}, \quad \boldsymbol{F}^\top = \begin{bmatrix} \boldsymbol{I}_m \\ \boldsymbol{C} \\ \boldsymbol{C}(\boldsymbol{u}_{t-1} \circ \boldsymbol{A}) \\ \boldsymbol{C}\prod_{\ell=1}^{2}(\boldsymbol{u}_{t-\ell} \circ \boldsymbol{A}) \\ \vdots \\ \boldsymbol{C}\prod_{\ell=1}^{L-2}(\boldsymbol{u}_{t-\ell} \circ \boldsymbol{A}) \end{bmatrix}, \quad \text{and} \quad \tilde{\boldsymbol{w}}_t := \begin{bmatrix} \boldsymbol{z}_t \\ \boldsymbol{w}_{t-1} \\ \boldsymbol{w}_{t-2} \\ \boldsymbol{w}_{t-3} \\ \vdots \\ \boldsymbol{w}_{t-L} \end{bmatrix}. \quad (3.2)$$

Moreover, let us define the matrix $\boldsymbol{G}$ as follows:

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{G}_1 & \boldsymbol{G}_2 & \cdots & \boldsymbol{G}_L \end{bmatrix} \in \mathbb{R}^{m \times d_{\tilde{u}}}, \quad \text{with} \quad d_{\tilde{u}} = (p+1)^L + p - 1, \tag{3.3}$$

and where we denote $\boldsymbol{G}_1 = \boldsymbol{CB}$, $\boldsymbol{G}_\ell = \{\boldsymbol{CA}_{i_1} \times \cdots \times \boldsymbol{A}_{i_{\ell-1}}\boldsymbol{B}\}_{i_1,\ldots,i_{\ell-1} \in \{0,\ldots,p\}} \in \mathbb{R}^{m \times p(p+1)^{\ell-1}}$, for $\ell \in \{2, \ldots, L\}$. The parameters $\{\boldsymbol{CB}, \ldots, \boldsymbol{CA}_{i_1} \times \cdots \times \boldsymbol{A}_{i_{\ell-1}}\boldsymbol{B}, \ldots\}$ is what we refer to as the *Markov-like parameters* of the system. We are finally ready to rewrite (3.1) as follows: for all $t \geq L$, we have

$$\boldsymbol{y}_t = \boldsymbol{\epsilon}_t + \boldsymbol{G}\tilde{\boldsymbol{u}}_t + \boldsymbol{F}\tilde{\boldsymbol{w}}_t. \tag{3.4}$$

With the dynamics written in the form of (3.4), it is natural to use the least squares estimation method to learn the Markov-like parameters from the observations $\{\boldsymbol{y}_t, \boldsymbol{u}_t\}_{t=1}^T$. More specifically, the Least Squares Estimator (LSE) $\hat{\boldsymbol{G}}$ of $\boldsymbol{G}$ admits a closed form and can be defined as:

$$\hat{\boldsymbol{G}} := \left(\sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{u}}_t^\intercal\right)^\dagger \left(\sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \boldsymbol{y}_t^\top\right) \in \operatorname*{argmin}_{\boldsymbol{G} \in \mathbb{R}^{m \times d_{\tilde{u}}}} \sum_{t=L}^{T} \|\boldsymbol{y}_t - \boldsymbol{G}\tilde{\boldsymbol{u}}_t\|_{\ell_2}^2. \tag{3.5}$$

Moreover, when the matrix $\sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{u}}_t^\top \succ 0$, then estimation error can be expressed as:

$$\hat{\boldsymbol{G}} - \boldsymbol{G} = \left(\sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{u}}_t^\top\right)^{-1} \left(\sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t (\boldsymbol{F}\tilde{\boldsymbol{w}}_t + \boldsymbol{\epsilon}_t)^\intercal\right), \tag{3.6}$$

Below, we present Theorem 2, our main result on the recovery of the *Markov-like parameters*:

Theorem 2 (Learning Markov-like parameters) *Let $\delta \in (0,1)$, $T \geq 0$. Suppose Assumption 1 holds, then the event:*

$$\|\hat{\boldsymbol{G}} - \boldsymbol{G}\|_{\text{op}} \leq \frac{C(\kappa^2\rho^L + \kappa)}{1-\rho} \sqrt{\frac{L(p+1)^{2(L+1)}\left(\log\left(\frac{eL}{\delta}\right) + m + nL + (p+1)^{L+1}\right)}{T-L}} \tag{3.7}$$

holds with probability at least $1-\delta$, provided that

$$T - L \gtrsim L(L+1)\gamma^L \left(\log\left(\frac{e(L+1)}{\delta}\right) + (p+1)^{L+1} \log\left(\frac{e(p+1)^{L+1}}{\delta}\right)\right), \tag{3.8}$$

with positive constants $C = \text{poly}(\sigma, \sqrt{p}\|\boldsymbol{B}\|_{\text{op}}, \|\boldsymbol{C}\|_{\text{op}})$ and $\gamma > 1$.

We discuss the analysis of the estimation error leading to Theorem 2 in §4. There, we also highlight the key challenges and steps in establishing this result. From the bound in (3.7), we see the recovery error $\|\hat{\boldsymbol{G}} - \boldsymbol{G}\|_{\text{op}}$ scales as, ignoring all other dependencies, $\tilde{\mathcal{O}}(\sqrt{(p+1)^{3(L+1)}/(T-L)})$. This contrasts with partially observed linear systems where typically we only have a polynomial dependence in L, and also reflects the difficulty in learning bilinear systems from partial observations. To recover the unknown matrices $\boldsymbol{C}, \boldsymbol{A}_0, \ldots, \boldsymbol{A}_p, \boldsymbol{B}$, we require L large enough, typically larger than $2n$ (see Remark 3).

Remark 3 (The BLDS parameter recovery.) *We remark that, for every $k \in \{0, \dots, p\}$, we can directly extract from $\hat{\boldsymbol{G}}$, estimates of the matrices $\{\boldsymbol{CB}, \boldsymbol{CA}_k\boldsymbol{B}, \dots, \boldsymbol{CA}_k^{L-1}\boldsymbol{B}\}$. To see that, observe that letting $\mathcal{I}_{k,\ell} \subset \{1, \dots, d_{\tilde{\boldsymbol{u}}}\}$ be the p indices corresponding to the p-dimensional sub-vector of $\tilde{\boldsymbol{u}}_t$, $(\prod_{i=1}^{\ell-1}(\boldsymbol{u}_{t-i})_k)\boldsymbol{u}_{t-\ell}$, then $\boldsymbol{G}_{:,\mathcal{I}_{k,\ell}} = \boldsymbol{CA}_k^{\ell-1}\boldsymbol{B}$. In other words, we can take $\hat{\boldsymbol{G}}_{:,\mathcal{I}_{\ell,k}}$ to be an estimate $\boldsymbol{CA}_k^{\ell-1}\boldsymbol{B}$. We then construct a Hankel matrix from $\hat{\boldsymbol{G}}_{:,\mathcal{I}_{\ell,k}}$. Under the condition that our estimation error is sufficiently small, each pair $(\boldsymbol{A}_k, \boldsymbol{B})$ is controllable, each pair $(\boldsymbol{A}_k, \boldsymbol{C})$ is observable, and $L \geq 2n$, we can use classic Ho-Kalman algorithm (Ho and Kálmán, 1966) to estimate $\boldsymbol{A}_0, \boldsymbol{A}_1, \dots, \boldsymbol{A}_p, \boldsymbol{B}, \boldsymbol{C}$ up to a similarity transform, with robustness guarantees (Oymak and Ozay, 2021). Lastly, note that the estimate of $\boldsymbol{D}$ is obtained as the first p columns of $\hat{\boldsymbol{G}}$.*

4. Sample Complexity Analysis

To prove Theorem 2, we start our analysis by decomposing the estimation error as follows:

$$\left\|\hat{\boldsymbol{G}} - \boldsymbol{G}\right\|_{\text{op}} \leq \underbrace{\left\|\left(\sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{u}}_t^{\mathsf{T}}\right)^{\dagger}\right\|_{\text{op}}}_{\text{Excitation}} \left(\underbrace{\left\|\sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t (\boldsymbol{F}\tilde{\boldsymbol{w}}_t)^{\mathsf{T}}\right\|_{\text{op}}}_{\text{Mutiplier Process}} + \underbrace{\left\|\sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \boldsymbol{\epsilon}_t^{\mathsf{T}}\right\|_{\text{op}}}_{\text{Truncation Bias}} \right), \tag{4.1}$$

where we use the submultiplicativity of $\|\cdot\|_{\text{op}}$ and the triangular inequality. Next, we will analyze each of three terms appearing in the decomposition above separately and obtain corresponding bounds in high probability. Once these bounds have been established, the proof concludes immediately (see details in Appendix C.1). In what follows, we focus on presenting the results regarding the analysis of the three terms. We note that the challenge in analyzing this terms lies in the presence of non-trivial dependencies and nonlinearities. As such, recent analysis tools from the non-asymptotic system identification literature (Ziemann et al., 2023) do not apply, and this is precisely what we manage to tackle.

4.1. Persistence of Excitation

We show persistence of excitation which is necessary to ensure that the LSE is a consistent estimator. More precisely, we will establish that smallest singular value of the design matrix $\tilde{\boldsymbol{U}}$ whose rows correspond to $\{\tilde{\boldsymbol{u}}_t^{\mathsf{T}}\}_{t=L}^{T}$ is bounded from below by $\tilde{\Omega}(\sqrt{T-L+1})$. One of the major reasons that makes establishing this persistence of excitation result challenging is the nonlinear dependence of $\tilde{\boldsymbol{u}}_t$ on $\boldsymbol{u}_{t-L}, \dots, \boldsymbol{u}_t$ for all $t \geq L$. We need to understand how distributional properties of the input impact the lower spectrum of $\tilde{\boldsymbol{U}}^{\mathsf{T}}\tilde{\boldsymbol{U}}$. To that end, we start by introducing the property of hypercontractivity.

Definition 4 (Hypercontractivity) *A p-dimensional random vector $\boldsymbol{u}$ is $(4, 2, \gamma)$-hypercontractive, if $\mathbb{E}[(\boldsymbol{u}^{\top}\boldsymbol{x})^4] \leq \gamma \mathbb{E}[(\boldsymbol{u}^{\top}\boldsymbol{x})^2]^2$, for all $\boldsymbol{x} \in \mathbb{R}^p$.*

The $(4, 2, \gamma)$-hypercontractivity property is satisfied by many classical distributions. Notably, a p-dimensional standard gaussian random vectors satisfies it with $\gamma = 3$, while p-dimensional random vectors sampled uniformly from $\sqrt{p} \cdot \mathcal{S}^{p-1}$ satisfies $\gamma = 3/(1+2/p)$. We refer the reader to Appendix B for a proof to these claims.

Assumption 2 (Distributional properties of the input) *$\{\boldsymbol{u}_t\}_{t\geq 0}$ is a sequence of independent zero-mean, isotropic[2], and $(4, 2, \gamma)$-hypercontractive for some $\gamma > 1$, p-dimensional random vectors with zero third moment marginals[3].*

Again, it can be verified that Assumption 2 is satisfied by inputs sampled from $\mathcal{N}(0, \boldsymbol{I}_p)$ or $\mathrm{Unif}(\sqrt{p}\cdot \mathcal{S}^{p-1})$. More importantly, Assumption 2 covers a wide range of input distributions that may even be heavy-tailed, as it only requires conditions on the first four moments of the distribution. This contrast with classical assumptions that require the input distribution to have sub-Gaussian tails, and is also consistent with the intuition that bounding the smallest singular value of random matrix requires weaker moment conditions (Koltchinskii and Mendelson, 2015). We are now ready to present our main result on the persistence of excitation:

Theorem 5 (Persistence of Excitation) *Suppose the sequence of inputs $\{\boldsymbol{u}_t\}_{t\geq 0}$ are selected as per Assumption 2, , then for all $\delta \in (0, 1)$, the event:*

$$\lambda_{\min}\left(\tilde{\boldsymbol{U}}^{\intercal}\tilde{\boldsymbol{U}}\right) \equiv \lambda_{\min}\left(\sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{u}}_t^{\intercal}\right) \geq (T - L + 1)/4. \tag{4.2}$$

holds with probability at least $1 - \delta$, provided that

$$T \gtrsim L + L(L+1)(3 \vee \gamma)^{L+1}\left(\log\left(\frac{e(L+1)}{\delta}\right) + (p+1)^{L+1}\log\left(\frac{e(p+1)^{L+1}}{\delta}\right)\right). \tag{4.3}$$

The proof of Theorem 5 is deferred to Appendix B. Interestingly, despite the presence of nonlinearities and dependencies in the covariates of $\tilde{\boldsymbol{U}}$, persistence of excitation is still guaranteed. Part of the reason is because the distributional properties presented in Assumption 2 ensure that that isotropy of the vectors $\tilde{\boldsymbol{u}}_t$ is still preserved, and their third and fourth moments are well bounded. Ihe dependence in $(p+1)^{L+1}$ in (4.3) is unavoidable because of the dimension of the vectors $\tilde{\boldsymbol{u}}_t$.

4.2. Analysis of the Truncation Bias

The analysis of the truncation bias term $\|\sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \boldsymbol{\epsilon}_t^{\top}\|_{\mathrm{op}}$ is challenging for multiple reasons. Indeed, firstly, the sequences $\{\tilde{\boldsymbol{u}}_t\}_{t\geq L}$ and $\{\boldsymbol{\epsilon}_t\}_{t\geq L}$ are non-trivially dependent, and secondly $\boldsymbol{\epsilon}_t$ is only zero-mean conditioned on future inputs $\boldsymbol{u}_{t-L}, \ldots, \boldsymbol{u}_T$ and is still dependent on $\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{t-L-1}$. It is worth noting that these sort of dependency structures does not occur when learning partially observed linear dynamical systems. Nonetheless, we establish a high probability bound on the truncation bias as presented below:

Proposition 6 *Let $\delta \in (0, 1)$ and $T \geq L$. The event:*

$$\left\|\sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \boldsymbol{\epsilon}_t^{\top}\right\|_{\mathrm{op}} \leq \frac{C_1 \kappa^2 \rho^L (p+1)^{L+1}}{1-\rho}\sqrt{(T-L)\left(\log\left(\frac{e}{\delta}\right) + n + (p+1)^{L+1}\right)}$$

holds with probability at least $1 - \delta$, with a positive constant $C_1 = \mathrm{poly}(\sigma, \sqrt{p}\|\boldsymbol{B}\|_{\mathrm{op}}, \|\boldsymbol{C}\|_{\mathrm{op}})$.

2. A p-dimensional random vector $\boldsymbol{u}$ is isotropic if for all $x \in \mathbb{R}^p$, $\mathbb{E}[(\boldsymbol{u}^{\top}x)^2] = \|x\|_{\ell_2}^2$.
3. A p-dimensional random vector $\boldsymbol{u}$ has zero-third-moment-marginals if for all $\boldsymbol{x} \in \mathbb{R}^p$, $\mathbb{E}[(\boldsymbol{u}_t^{\top}\boldsymbol{x})^3] = 0$.

The proof of Proposition 6 relies on the critical observation that the truncation bias term can be rewritten as martingale, namely as follows: for all $\theta \in \mathcal{S}^{d_{\tilde{\boldsymbol{u}}}-1}, \lambda \in \mathcal{S}^{m-1}$,

$$\theta^\top \left(\sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \boldsymbol{\epsilon}_t^\top \right) \lambda = \sum_{s=0}^{T-L-1} (\boldsymbol{B}\boldsymbol{u}_s + \boldsymbol{w}_s)^\top \boldsymbol{f}_s(\theta, \lambda, , \boldsymbol{u}_{s+1}, \ldots, \boldsymbol{u}_T), \tag{4.4}$$

where the functions $\{\boldsymbol{f}_s\}_{s\geq 0}$ are nonlinear in their arguments. Moreover, thanks to our choice of inputs and the stability Assumption 1, the terms $\boldsymbol{f}_s(\theta, \lambda, , \boldsymbol{u}_{s+1}, \ldots, \boldsymbol{u}_T)$ are bounded and do not scale with T. Thus, we use classical concentration tools to deduce our final bounds. The details of the proofs including the precise definition of $\{\boldsymbol{f}_s\}_{s\geq 0}$ are delayed until Appendix C.2.

4.3. Analysis of the Multiplier Process

The analysis of the the multiplier process term $\| \sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t (\boldsymbol{F}\tilde{\boldsymbol{w}}_t)^\mathsf{T} \|_{\text{op}}$ is somewhat simpler than that of truncation bias term. The reason is because the sequences $\{\tilde{\boldsymbol{u}}_t\}_{t\geq L}$, and $\{\tilde{\boldsymbol{w}}_t\}_{t\geq L}$ are independent with zero-mean vectors. However, each of these two sequences contains dependent vectors. Below, we present a high probability bound on the multiplier process term showing that we can still control this term despite the presence of these dependencies:

Proposition 7 *Let $\delta \in (0,1)$ and $T \geq L$. The event:*

$$\left\| \sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t (\boldsymbol{F}\tilde{\boldsymbol{w}}_t)^\mathsf{T} \right\|_{\text{op}} \leq C_2 \frac{\kappa(p+1)^{L+1}}{1-\rho} \sqrt{L(T-L)\left(\log\left(\frac{eL}{\delta}\right) + m + nL + (p+1)^{L+1} \right)}$$

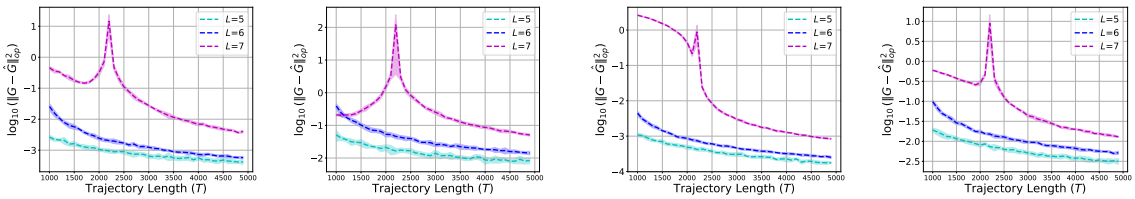
holds with probability $1-\delta$, with a positive constant $C_2 = \text{poly}(\sigma, \|\boldsymbol{C}\|_{\text{op}})$.

The key idea behind the proof of Proposition 7 is observing that each of the subsequences, $\ell \in \{1, \ldots, L\}$, $\{\tilde{\boldsymbol{w}}_{kL+\ell}\}_{k\geq 0}$, has independent vectors. Thus, we can use a blocking trick to rewrite the multiplier process as a sum of L martingales which can bound using classical concentration tools. This argument is made precise in the proof and is deferred to Appendix C.3.

5. Numerical Experiments

For our experiments, we consider a partially observed bilinear dynamical system (2.1) with n=5 hidden states, input dimension p=2, and output dimension m=2. The dynamics matrices $\boldsymbol{A}_0, \boldsymbol{A}_1, \boldsymbol{A}_2$ are constructed with i.i.d. $\mathcal{N}(0,1)$ entries, and scaled to have spectral radius $\rho(\boldsymbol{A}_0)$=ρ_0 and $\rho(\boldsymbol{A}_1)$=$\rho(\boldsymbol{A}_2)$=ρ_k, where ρ_0, ρ_u are hyper-parameters in our experiments. Similarly, $\boldsymbol{B}, \boldsymbol{C}$ and $\boldsymbol{D}$ are generated with i.i.d. $\mathcal{N}(0, 1/n)$ and $\mathcal{N}(0, 1/m)$ entries, respectively. The noise processes $\{\boldsymbol{w}_t\}_{t=0}^T$, and $\{\boldsymbol{z}_t\}_{t=0}^T$ are chosen according to i.i.d. Gaussian distribution with zero mean and variances $\sigma^2 = 0.0001$. Lastly, the control inputs are either sampled uniformly at random from the sphere $\sqrt{p} \cdot \mathcal{S}^{p-1}$ or sampled i.i.d. from a Gaussian distribution $\mathcal{N}(0, \boldsymbol{I}_p)$.

In Figure 1, we plot the estimation error $\|\boldsymbol{G} - \hat{\boldsymbol{G}}\|_{\text{op}}^2$ over different values of ρ_0, ρ_k, L and T. Each experiment is repeated 10 times and we plot the mean and one standard deviation. Figure 1*(a),(b)* correspond to estimation with inputs $\{\boldsymbol{u}_t\}_{t\geq 0} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$, whereas, Figure 1*(c),(d)* correspond to $\{\boldsymbol{u}_t\}_{t\geq 0} \overset{\text{i.i.d.}}{\sim} \text{Unif}(\sqrt{p} \cdot \mathcal{S}^{d-1})$.



(a) (ρ_0, ρ_k)=$(0.4, 0.2)$ (b) (ρ_0, ρ_k)=$(0.6, 0.4)$ (c) (ρ_0, ρ_k)=$(0.4, 0.2)$ (d) (ρ_0, ρ_k)=$(0.6, 0.4)$

Figure 1: We plot the estimation error $\|\boldsymbol{G} - \hat{\boldsymbol{G}}\|_{\text{op}}^2$ over different values of T, L, $\rho(\boldsymbol{A}_0)$, $\rho(\boldsymbol{A}_1)$, $\rho(\boldsymbol{A}_2)$ while fixing n=5, p=2 and m=2. Figure 1*(a),(b)* correspond to estimation with inputs $\{\boldsymbol{u}_t\}_{t\geq 0} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$, whereas, Figure 1*(c),(d)* correspond to $\{\boldsymbol{u}_t\}_{t\geq 0} \overset{\text{i.i.d.}}{\sim} \text{Unif}(\sqrt{p} \cdot \mathcal{S}^{d-1})$. Our plots show that the later input choice gives better estimation of $\boldsymbol{G}$ as compared to the first one. Moreover, the estimation error increases as L and ρ increases, whereas, it decreases as T increases.

Figure 1 shows that the estimation error increases with L, because the number of Markov-like parameters increases exponentially in L. More interestingly, Figure 1 suggests that choosing the inputs $\{\boldsymbol{u}_t\}_{t\geq 0} \overset{\text{i.i.d.}}{\sim} \text{Unif}(\sqrt{p} \cdot \mathcal{S}^{d-1})$ leads to more accurate estimation than $\{\boldsymbol{u}_t\}_{t\geq 0} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$ at any given T. This reinforce our theoretical guarantees as follows: Recall the hyper-contractivity parameter γ from Definition 4. In the Appendix we show that $\gamma = 3$ for Gaussian inputs, whereas, $\gamma = \frac{3}{1+2/p}$ for the uniform input. Hence, from Theorem 5, Gaussian inputs require more number of samples to guarantee persistence of excitation.

We observe double descent curves (Nakkiran et al., 2020) for L=7. This is because our regression problem is unregularized and has $(p+1)^L+p-1 = 2188$ unknown parameters, and the number of input-features ($\tilde{\boldsymbol{u}}_t$) is $T-7$. Hence, for $L = 7$, we see the peak at $T = 2195$, and the error decays smoothly after this point. Note that the peak occurs at $T - L = (p+1)^L + p - 1$ (where the number of unknown parameters become equal to the number of input-features). For $L = \{5, 6\}$, we do not see the double descent because we start at $T = 1000$ which is greater than $(p+1)^L + p - 1$.

6. Conclusion and Future Direction

We provide the first non-asymptotic learning bounds for partially observed BLDS. Given finite input-output data sampled from a single trajectory of BLDS, we learn its Markov-like parameters, provide an upper bound on the estimation error with $\tilde{\mathcal{O}}(1/\sqrt{T})$ dependence, and provide a lower bound on the number of samples required, scaling as $\tilde{\Omega}\left((p+1)^{L+1}\right)$. These parameters uniquely characterize the input-output map of a BLDS via nonlinear input features, hence, can be used to recover the state-space matrices. Our results hold under a novel notion of stability that generalizes the classic notion of stability for LDS to the BLDS.

There are several interesting future directions. First, can the exponential dependence on L be avoided? We believe this can be done by carefully designing the inputs such that the number of Markov-like parameters do not grow exponentially in L? Second, can our results be extended to account for marginally-stable BLDS? This requires stabilization of a partially observed BLDS with unknown state-space matrices which itself is an interesting future direction. Other possible directions include exploring the benefit of regularization (e.g., Hankel nuclear norm regularization) for BLDS identification, exploring gradient-based methods for learning partially observed BLDS, and adaptive control of BLDS.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Sohail Bahmani and Justin Romberg. Convex programming for estimation in nonlinear recurrent models. *arXiv preprint arXiv:1908.09915*, 2019.

Ainesh Bakshi, Allen Liu, Ankur Moitra, and Morris Yau. A new approach to learning linear dynamical systems. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 335–348, 2023.

N Berk Hizir, Minh Q Phan, Raimondo Betti, and Richard W Longman. Identification of discrete-time bilinear systems through equivalent linear models. *Nonlinear Dynamics*, 69(4):2065–2078, 2012.

Daniel Bruder, Xun Fu, and Ram Vasudevan. Advantages of bilinear koopman realizations for the modeling and control of systems with unknown dynamics. *IEEE Robotics and Automation Letters*, 6(3):4369–4376, 2021.

Steven L Brunton and J Nathan Kutz. *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press, 2022.

Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

Nicolas Chatzikiriakos, Robin Strässer, Frank Allgöwer, and Andrea Iannelli. End-to-end guarantees for indirect data-driven control of bilinear systems with finite stochastic data. *arXiv preprint arXiv:2409.18010*, 2024.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197, 2018.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *FOCM*, pages 1–47, 2019.

Boualem Djehiche and Othmane Mazhar. Efficient learning of hidden state lti state space models of unknown order. *arXiv preprint arXiv:2202.01625*, 2022.

Zhe Du, Yahya Sattar, Davoud Ataee Tarzanagh, Laura Balzano, Necmiye Ozay, and Samet Oymak. Data-driven control of markov jump systems: Sample complexity and regret bounds. In *2022 American Control Conference (ACC)*, pages 4901–4908. IEEE, 2022.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

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

M. Fazel, T. K. Pong, D. Sun, and P. Tseng. Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3): 946–977, 2013. citations: 538.

Ren Fujiwara, Yasuko Matsubara, and Yasushi Sakurai. Modeling latent non-linear dynamical system over time series. *arXiv preprint arXiv:2412.08114*, 2024.

Debdipta Goswami and Derek A Paley. Global bilinearization and controllability of control-affine nonlinear systems: A koopman spectral approach. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 6107–6112. IEEE, 2017.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *The Journal of Machine Learning Research*, 19(1):1025–1068, 2018.

Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. *Advances in Neural Information Processing Systems*, 30, 2017.

BL Ho and Rudolf E Kálmán. Effective construction of linear state-variable models from input/output functions. *at-Automatisierungstechnik*, 14(1-12):545–548, 1966.

Prateek Jain, Suhas S Kowshik, Dheeraj Nagaraj, and Praneeth Netrapalli. Near-optimal offline and streaming algorithms for learning non-linear dynamical systems. *Advances in Neural Information Processing Systems*, 34, 2021.

Yassir Jedra and Alexandre Proutiere. Finite-time identification of stable linear systems optimality of the least-squares estimator. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 996–1001. IEEE, 2020.

Jer-Nan Juang. Continuous-time bilinear system identification. *Nonlinear Dynamics*, 39(1):79–94, 2005.

Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

Stefan Kombrink, Tomas Mikolov, Martin Karafiát, and Lukás Burget. Recurrent neural network based language modeling in meeting recognition. In *Interspeech*, volume 11, pages 2877–2880, 2011.

Yehuda Koren, Steffen Rendle, and Robert Bell. Advances in collaborative filtering. *Recommender systems handbook*, pages 91–142, 2021.

Krzysztof Kowalski and W-H Steeb. *Nonlinear dynamical systems and Carleman linearization*. World Scientific, 1991.

C Kubrusly and O Costa. Mean square stability conditions for discrete stochastic bilinear systems. *IEEE Transactions on Automatic Control*, 30(11):1082–1087, 1985.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems*, 33:20876–20888, 2020.

James Ting-Ho Lo. Global bilinearization of systems with control appearing linearly. *SIAM Journal on Control*, 13(4):879–885, 1975. doi: 10.1137/0313053.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. In *NeurIPS*, 2019.

Zakaria Mhammedi, Dylan J Foster, Max Simchowitz, Dipendra Misra, Wen Sun, Akshay Krishnamurthy, Alexander Rakhlin, and John Langford. Learning the linear quadratic regulator from nonlinear observations. *Advances in Neural Information Processing Systems*, 33:14532–14543, 2020.

Ronald R. Mohler. *Bilinear Control Processes: With Applications to Engineering, Ecology, and Medicine*. Elsevier, 1973.

Morteza Nazari Monfared, Yu Kawano, Juan E Machado, Daniele Astolfi, and Michele Cucuzzella. Stabilization for a class of bilinear systems: A unified approach. *IEEE Control Systems Letters*, 7:2791–2796, 2023.

Negin Musavi, Ziyao Guo, Geir Dullerud, and Yingying Li. Identification of analytic nonlinear dynamical systems with non-asymptotic guarantees. *arXiv preprint arXiv:2411.00656*, 2024.

Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

Samet Oymak. Stochastic gradient descent learns state equations with nonlinear activations. In *Conference on Learning Theory*, pages 2551–2579, 2019.

Samet Oymak and Necmiye Ozay. Revisiting ho–kalman-based system identification: Robustness and finite-sample analysis. *IEEE Transactions on Automatic Control*, 67(4):1914–1928, 2021.

Panos M Pardalos and Vitaliy A Yatsenko. *Optimization and control of bilinear systems: theory, algorithms, and applications*, volume 11. Springer Science & Business Media, 2010.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *ICML*, pages 5610–5618. PMLR, 2019.

Tuhin Sarkar, Alexander Rakhlin, and Munther Dahleh. Nonparametric system identification of stochastic switched linear systems. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3623–3628, 2019a.

Tuhin Sarkar, Alexander Rakhlin, and Munther Dahleh. Nonparametric system identification of stochastic switched linear systems. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3623–3628. IEEE, 2019b.

Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Finite time lti system identification. *Journal of Machine Learning Research*, 22:1–61, 2021.

Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *Journal of Machine Learning Research*, 23(140):1–49, 2022.

Yahya Sattar, Zhe Du, Davoud Ataee Tarzanagh, Laura Balzano, Necmiye Ozay, and Samet Oymak. Identification and adaptive control of markov jump systems: Sample complexity and regret bounds. *arXiv preprint arXiv:2111.07018*, 2021.

Yahya Sattar, Samet Oymak, and Necmiye Ozay. Finite sample identification of bilinear dynamical systems. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 6705–6711. IEEE, 2022.

Yahya Sattar, Yassir Jedra, and Sarah Dean. Learning linear dynamics from bilinear observations. *arXiv preprint arXiv:2409.16499*, 2024.

Borna Sayedana, Mohammad Afshari, Peter E Caines, and Aditya Mahajan. Strong consistency and rate of convergence of switched least squares system identification for autonomous markov jump linear systems. *IEEE Transactions on Automatic Control*, 2024.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, pages 2714–2802. PMLR, 2019.

Eduardo D Sontag, Yuan Wang, and Alexandre Megretski. Input classes for identifiability of bilinear systems. *IEEE Transactions on Automatic Control*, 54(2):195–207, 2009.

Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

Yue Sun, Samet Oymak, and Maryam Fazel. Finite sample system identification: Optimal rates and the role of regularization. In *Learning for dynamics and control*, pages 16–25. PMLR, 2020.

Yue Sun, Samet Oymak, and Maryam Fazel. Finite sample identification of low-order lti systems via nuclear norm regularization. *IEEE Open Journal of Control Systems*, 1:237–254, 2022.

Spyros Svoronos, George Stephanopoulos, and Rutherford Aris. Bilinear approximation of general non-linear dynamic systems with linear inputs. *International Journal of Control*, 31(1):109–126, 1980.

Anastasios Tsiamis and George J Pappas. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3648–3654. IEEE, 2019.

Stephen Tu, Ross Boczar, Andrew Packard, and Benjamin Recht. Non-asymptotic analysis of robust control from coarse-grained identification. *arXiv preprint arXiv:1707.04791*, 2017.

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

Andrew Wagenmaker and Kevin Jamieson. Active learning for identification of linear dynamical systems. In *Conference on Learning Theory*, pages 3487–3582. PMLR, 2020.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Ingvar Ziemann, Anastasios Tsiamis, Bruce Lee, Yassir Jedra, Nikolai Matni, and George J Pappas. A tutorial on the non-asymptotic theory of system identification. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 8921–8939. IEEE, 2023.

Ingvar M Ziemann, Henrik Sandberg, and Nikolai Matni. Single trajectory nonparametric learning of nonlinear dynamics. In *Conference on Learning Theory*, pages 3333–3364. PMLR, 2022.

Appendix A. Stability of Bilinear Dynamical Systems

In this appendix, we present some results that concerns the stability of bilinear system in the sense of Definition 1. In Lemma 8, we present a condition under which a bilinear system satisfies uniform stability. In Lemma 9, we present a bound on $\|\boldsymbol{F}\|_{\mathrm{op}}$ which follows under our stability assumptions.

Lemma 8 *Let $\mathcal{U}$ be a bounded and non-empty subset of $\mathbb{R}^p$ such that $\rho(\mathcal{U} \circ \boldsymbol{A}) < 1$, then for all $\rho \in (\rho(\mathcal{U} \circ \boldsymbol{A}), 1)$, there exists $\kappa \geq 1$ such that the system is $(\mathcal{U}, \kappa, \rho)$-uniformly stable.*

Proof Let $\rho \in (\rho(\mathcal{U} \circ \boldsymbol{A}), 1)$. We recall that the joint spectral radius of $\mathcal{U} \circ \boldsymbol{A}$ is defined as follows:

$$\rho(\mathcal{U} \circ \boldsymbol{A}) = \lim_{k \to \infty} \sup_{\boldsymbol{M}_1, \ldots, \boldsymbol{M}_k \in \mathcal{U} \circ \boldsymbol{A}} \|\boldsymbol{M}_1 \boldsymbol{M}_2 \cdots \boldsymbol{M}_k\|_{\mathrm{op}}^{1/k}.$$

Since by assumption the limit exists, we have by definition that

$$\forall \epsilon > 0, \;\; \exists k_0 \geq 1, \;\; \forall k \geq k_0, \quad \left| \sup_{\boldsymbol{M}_1, \ldots, \boldsymbol{M}_k \in \mathcal{U} \circ \boldsymbol{A}} \|\boldsymbol{M}_1 \boldsymbol{M}_2 \cdots \boldsymbol{M}_k\|_{\mathrm{op}}^{1/k} - \rho(\mathcal{U} \circ \boldsymbol{A}) \right| < \epsilon \rho(\mathcal{U} \circ \boldsymbol{A}).$$

Thus choosing $\epsilon > 0$ such that $\rho \geq (1 + \epsilon)\rho(\mathcal{U} \circ \boldsymbol{A})$, we can find k_0 such that

$$\forall k \geq k_0, \qquad \sup_{\boldsymbol{M}_1, \ldots, \boldsymbol{M}_k \in \mathcal{U} \circ \boldsymbol{A}} \|\boldsymbol{M}_1 \boldsymbol{M}_2 \cdots \boldsymbol{M}_k\|_{\mathrm{op}} < \rho^k.$$

Now, we can further define

$$\kappa = \max \left\{ 1, \sup_{1 \leq k \leq k_0} \sup_{\boldsymbol{M}_1, \ldots, \boldsymbol{M}_k \in \mathcal{U} \circ \boldsymbol{A}} \frac{\|\boldsymbol{M}_1 \boldsymbol{M}_2 \cdots \boldsymbol{M}_k\|_{\mathrm{op}}^{1/k}}{\rho} \right\}.$$

We note that κ is well defined because $\mathcal{U} \circ \boldsymbol{A}$ is a bounded set of matrices since $\mathcal{U}$ is bounded. Indeed, for all $\boldsymbol{M} \in \mathcal{U} \circ \boldsymbol{A}$, we have $\|\boldsymbol{M}\|_{\mathrm{op}} \leq \max\{1, \sup_{\boldsymbol{u} \in \mathcal{U}} \|\boldsymbol{u}\|_{\ell_\infty}\}(\|\boldsymbol{A}_0\|_{\mathrm{op}} + \cdots + \|\boldsymbol{A}_p\|_{\mathrm{op}})$. This concludes the proof. ■

Lemma 9 *Suppose Assumption 1 holds and let $(\boldsymbol{u}_t)_{t \geq 0}$ be a sequence of inputs taking values in $\sqrt{p} \cdot \mathcal{S}^{p-1}$, and . We have*

$$\forall t \geq 0, \qquad \left\| \prod_{\ell=0}^{t} (\boldsymbol{u}_\ell \circ \boldsymbol{A}) \right\|_{\mathrm{op}} \leq \kappa \rho^{t+1}. \tag{A.1}$$

Consequently, we have

$$\|\boldsymbol{F}\|_{\mathrm{op}} \leq 1 + \frac{\kappa \|\boldsymbol{C}\|_{\mathrm{op}}}{1 - \rho} \tag{A.2}$$

Proof The first inequality (A.1) holds immediately thanks to stability. The second inequality (A.2) is an immediate consequence of (A.1). Indeed, we have

$$\|\boldsymbol{F}\|_{\mathrm{op}} \leq 1 + \|\boldsymbol{C}\|_{\mathrm{op}} \sum_{\ell=1}^{L-1} \left\| \prod_{i=1}^{\ell-1} (\boldsymbol{u}_{t-i} \circ A) \right\|_{\mathrm{op}} \leq 1 + \|\boldsymbol{C}\|_{\mathrm{op}} \sum_{\ell=1}^{L-1} \kappa \rho^{\ell-1} \leq 1 + \frac{\kappa \|\boldsymbol{C}\|_{\mathrm{op}}}{1 - \rho}$$

■

Appendix B. Proofs for Persistence of Excitation

In this appendix, we will present our results on persistence of excitation under the following assumption unless specified otherwise.

Assumption 2 (Distributional properties of the input) *$\{\boldsymbol{u}_t\}_{t\geq 0}$ is a sequence of independent zero-mean, isotropic[4], and $(4,2,\gamma)$-hypercontractive for some $\gamma > 1$, p-dimensional random vectors with zero third moment marginals[5].*

There are many choice of input distribution that satisfy Assumption 2. Notably, if the inputs are sampled independently according to $\mathcal{N}(0, I_p)$ or $\mathrm{Unif}(\sqrt{p}\cdot\mathcal{S}^{p-1})$, then Assumption 2 holds with $\gamma = 3$ or $\gamma = \frac{3}{1+2/p}$ respectively. Before we present the proof of our main result on persistence of excitation, we present some intermediate results in the next two subsections.

B.1. Hypercontractivity to Bounded Moments (Input)

Lemma 10 *Let $\boldsymbol{Q}, \boldsymbol{R} \in \mathbb{R}^{p\times d}$, and $\gamma > 0$. Let $\boldsymbol{u}$ be an isotropic and $(4,2,\gamma)$-hypercontractive random vector taking values in $\mathbb{R}^p$. We have that*

$$\mathbb{E}[(\boldsymbol{u}^\top \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{u})(\boldsymbol{u}^\top \boldsymbol{R}\boldsymbol{R}^\top \boldsymbol{u})] \leq \gamma \|\boldsymbol{Q}\|_F^2 \|\boldsymbol{R}\|_F^2$$

Proof Note that if $\boldsymbol{u}$ is isotropic and $(4,2,\gamma)$-hypercontractive, then for any $\boldsymbol{Q}, \boldsymbol{R} \in \mathbb{R}^{p\times d}$, we have

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{u}^\top \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{u})(\boldsymbol{u}^\top \boldsymbol{R}\boldsymbol{R}^\top \boldsymbol{u})] &= \mathbb{E}\left[\left(\sum_{i=1}^d (\boldsymbol{u}^\top \boldsymbol{q}_i)^2\right)\left(\sum_{i=1}^d (\boldsymbol{u}^\top \boldsymbol{r}_i)^2\right)\right], \\
&= \mathbb{E}\left[\sum_{1\leq i,j\leq d} (\boldsymbol{u}^\top \boldsymbol{q}_i)^2 (\boldsymbol{u}^\top \boldsymbol{r}_j)^2\right], \\
&\overset{\text{(a)}}{\leq} \sum_{1\leq i,j\leq d} \sqrt{\mathbb{E}\left[(\boldsymbol{u}^\top \boldsymbol{q}_i)^4\right]}\sqrt{\mathbb{E}\left[(\boldsymbol{u}^\top \boldsymbol{r}_j)^4\right]}, \\
&\overset{\text{(b)}}{\leq} \sum_{1\leq i,j\leq d} \sqrt{\gamma}\mathbb{E}\left[(\boldsymbol{u}^\top \boldsymbol{q}_i)^2\right]\sqrt{\gamma}\mathbb{E}\left[(\boldsymbol{u}^\top \boldsymbol{r}_j)^2\right], \\
&\leq \gamma \sum_{1\leq i,j\leq d} \|\boldsymbol{q}_i\|_{\ell_2}^2 \|\boldsymbol{r}_j\|_{\ell_2}^2, \\
&\leq \gamma \|\boldsymbol{Q}\|_F^2 \|\boldsymbol{R}\|_F^2,
\end{aligned}$$

where we obtain (a) by using Cauchy-Schwarz inequality, and (b) from the $(4,2,\gamma)$-hypercontractivity assumption. This completes the proof. ■

4. A p-dimensional random vector $\boldsymbol{u}$ is isotropic if for all $\boldsymbol{x} \in \mathbb{R}^p$, $\mathbb{E}[(\boldsymbol{u}^\top \boldsymbol{x})^2] = \|\boldsymbol{x}\|_{\ell_2}^2$.
5. A p-dimensional random vector $\boldsymbol{u}$ has zero-third-moment-marginals if for all $\boldsymbol{x} \in \mathbb{R}^p$, $\mathbb{E}[(\boldsymbol{u}_t^\top \boldsymbol{x})^3] = 0$.

B.2. Hypercontractivity to Bounded Moments (Covariates)

Lemma 11 *Suppose the sequence of inputs $\{\boldsymbol{u}_t\}_{t\geq 0}$ satisfy Assumption 2. Let $\tilde{\boldsymbol{u}}_t$ be the covariates (input features) as defined in (3.2). We have that*

$$\mathbb{E}[(\tilde{\boldsymbol{u}}_t^\mathsf{T}\boldsymbol{v})^2] = 1 \quad \textit{and} \quad \mathbb{E}[(\tilde{\boldsymbol{u}}_t^\mathsf{T}\boldsymbol{v})^4] \leq L(3\vee\gamma)^{L+1},$$

for all $t = L, L+1, \ldots, T$ and all $\boldsymbol{v} \in \mathcal{S}^{d_{\tilde{u}}-1}$

Proof To begin, recall the definition of $\tilde{\boldsymbol{u}}_t$ from (3.2). For notational convenience, we define another random features of inputs $\tilde{\tilde{\boldsymbol{u}}}_t$ as follows,

$$\tilde{\boldsymbol{u}}_t := \begin{bmatrix} \boldsymbol{u}_t \\ \boldsymbol{u}_{t-1} \\ \bar{\boldsymbol{u}}_{t-1}\otimes \boldsymbol{u}_{t-2} \\ \bar{\boldsymbol{u}}_{t-1}\otimes \bar{\boldsymbol{u}}_{t-2}\otimes \boldsymbol{u}_{t-3} \\ \vdots \\ \bar{\boldsymbol{u}}_{t-1}\otimes \bar{\boldsymbol{u}}_{t-2}\otimes\cdots\otimes \boldsymbol{u}_{t-L} \end{bmatrix}, \quad \tilde{\tilde{\boldsymbol{u}}}_t := \begin{bmatrix} \bar{\boldsymbol{u}}_{t-1} \\ \bar{\boldsymbol{u}}_{t-1}\otimes \bar{\boldsymbol{u}}_{t-2} \\ \bar{\boldsymbol{u}}_{t-1}\otimes \bar{\boldsymbol{u}}_{t-2}\otimes \bar{\boldsymbol{u}}_{t-3} \\ \vdots \\ \bar{\boldsymbol{u}}_{t-1}\otimes \bar{\boldsymbol{u}}_{t-2}\otimes\cdots\otimes \bar{\boldsymbol{u}}_{t-L} \end{bmatrix}, \tag{B.1}$$

where $\bar{\boldsymbol{u}}_t = [1\ \boldsymbol{u}_t^\mathsf{T}]^\mathsf{T}$. Note that, $\tilde{\boldsymbol{u}}_t \in \mathbb{R}^{d_{\tilde{u}}}$, and $\tilde{\tilde{\boldsymbol{u}}}_t \in \mathbb{R}^{d_{\tilde{\tilde{u}}}}$,

$$d_{\tilde{u}} = p + \sum_{i=0}^{L-1} p(p+1)^i = p + p\frac{(p+1)^L - 1}{(p+1)-1} = (p+1)^L + p - 1,$$
$$\text{and} \quad d_{\tilde{\tilde{u}}} = \sum_{i=0}^{L-1} (p+1)(p+1)^i = (p+1)\frac{(p+1)^L - 1}{(p+1)-1} = \frac{(p+1)^{L+1} - (p+1)}{p}.$$

We first show that, under Assumption 2, $\tilde{\boldsymbol{u}}_t$ is isotropic as follows.

Isotropic Covariates: To begin, note that $\mathbb{E}[\tilde{\boldsymbol{u}}_t] = 0$ due to Assumption 2. Next, we will show that $\boldsymbol{\Sigma}[\tilde{\boldsymbol{u}}_t] = \mathbb{E}[\tilde{\boldsymbol{u}}_t\tilde{\boldsymbol{u}}_t^\mathsf{T}] = \boldsymbol{I}_{d_{\tilde{u}}}$ as follows: Let $[\tilde{\boldsymbol{u}}_t]_0 = \boldsymbol{u}_t$, and let $[\tilde{\boldsymbol{u}}_t]_i$ be the i-th partition of $\tilde{\boldsymbol{u}}_t$ with $p(p+1)^{i-1}$ entries for $i = 1, \ldots, L$. Then, we have that $\mathbb{E}\left[[\tilde{\boldsymbol{u}}_t]_0[\tilde{\boldsymbol{u}}_t]_0^\mathsf{T}\right] = \mathbb{E}\left[[\tilde{\boldsymbol{u}}_t]_1[\tilde{\boldsymbol{u}}_t]_1^\mathsf{T}\right] = \boldsymbol{I}_p$. Similarly, for $i = 2, \ldots, L$, we have

$$\begin{aligned} \mathbb{E}\left[[\tilde{\boldsymbol{u}}_t]_i[\tilde{\boldsymbol{u}}_t]_i^\mathsf{T}\right] &= \mathbb{E}\left[(\bar{\boldsymbol{u}}_{t-1}\otimes\cdots\otimes\boldsymbol{u}_{t-i})(\bar{\boldsymbol{u}}_{t-1}\otimes\cdots\otimes\boldsymbol{u}_{t-i})^\mathsf{T}\right], \\ &= \mathbb{E}\left[\bar{\boldsymbol{u}}_{t-1}\bar{\boldsymbol{u}}_{t-1}^\mathsf{T}\otimes\cdots\otimes\boldsymbol{u}_{t-i}\boldsymbol{u}_{t-i}^\mathsf{T}\right], \\ &= \boldsymbol{I}_{p+1}\otimes\cdots\otimes\boldsymbol{I}_p = \boldsymbol{I}_{p(p+1)^{i-1}}. \end{aligned} \tag{B.2}$$

Hence, we show that $\mathbb{E}\left[[\tilde{\boldsymbol{u}}_t]_i[\tilde{\boldsymbol{u}}_t]_i^\mathsf{T}\right] = \boldsymbol{I}_{p(p+1)^{(i-1)\vee 0}}$ for all $i = 0, \ldots, L$. Next, we will show that $\mathbb{E}\left[[\tilde{\boldsymbol{u}}_t]_i[\tilde{\boldsymbol{u}}_t]_j^\mathsf{T}\right] = 0$, for all $i, j = 0, \ldots, L$ when $i \neq j$. Because of symmetry, it suffices to consider $i < j$. In this case, we have

$$\begin{aligned} \mathbb{E}\left[[\tilde{\boldsymbol{u}}_t]_i[\tilde{\boldsymbol{u}}_t]_j^\mathsf{T}\right] &= \mathbb{E}\left[(\bar{\boldsymbol{u}}_{t-1}\otimes\cdots\otimes\boldsymbol{u}_{t-i})(\bar{\boldsymbol{u}}_{t-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})^\mathsf{T}\right], \\ &= \mathbb{E}\left[(\bar{\boldsymbol{u}}_{t-1}\otimes\cdots\otimes\boldsymbol{u}_{t-i})(\bar{\boldsymbol{u}}_{t-1}\otimes\cdots\otimes\bar{\boldsymbol{u}}_{t-i})^\mathsf{T}\right]\otimes\mathbb{E}\left[(\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})^\mathsf{T}\right], \\ &= 0, \end{aligned} \tag{B.3}$$

for all $2 \leq i < j$. Similarly, we also have $\mathbb{E}\left[[\tilde{\boldsymbol{u}}_t]_1[\tilde{\boldsymbol{u}}_t]_j^\mathsf{T}\right] = \mathbb{E}\left[\boldsymbol{u}_{t-1}(\bar{\boldsymbol{u}}_{t-1} \otimes \cdots \otimes \boldsymbol{u}_{t-j})^\mathsf{T}\right] = 0$, for all $j = 2, \ldots, L$. Lastly, $[\tilde{\boldsymbol{u}}_t]_0$ is independent of $[\tilde{\boldsymbol{u}}_t]_j$, for all $j = 1, \ldots, L$, due to Assumption 2, hence, $\mathbb{E}\left[[\tilde{\boldsymbol{u}}_t]_0[\tilde{\boldsymbol{u}}_t]_j^\mathsf{T}\right] = 0$. Combining (B.2) and (B.3), we get

$$\mathbb{E}[\tilde{\boldsymbol{u}}_t\tilde{\boldsymbol{u}}_t^\mathsf{T}] = \boldsymbol{I}_{d_{\tilde{u}}} \quad \Longrightarrow \quad \mathbb{E}[(\tilde{\boldsymbol{u}}_t^\mathsf{T}\boldsymbol{v})^2] = 1, \tag{B.4}$$

for all $\boldsymbol{v} \in \mathcal{S}^{d_{\tilde{u}}-1}$. This gives the first statement of Lemma 11. Next, we show that $\tilde{\boldsymbol{u}}_t$ also have bounded fourth moment marginals as follows:

Fourth Moment Marginals: Let $\boldsymbol{v} \in \mathcal{S}^{d_{\tilde{u}}-1}$, and consider the random variable $(\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{u}}_t)^2$. In the following we will upper bound $\mathbb{E}[(\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{u}}_t)^4]$. To begin, let $\boldsymbol{v}_i$ denote the i-th partition of $\boldsymbol{v}$ with $p(p+1)^{(i-1)\vee 0}$ entries, for $i = 0, \ldots, L$. Then, $\boldsymbol{v}$ can be represented as,

$$\boldsymbol{v} := \begin{bmatrix}\boldsymbol{v}_0^\mathsf{T} & \boldsymbol{v}_1^\mathsf{T} & \boldsymbol{v}_2^\mathsf{T} & \ldots & \boldsymbol{v}_L^\mathsf{T}\end{bmatrix}^\mathsf{T}. \tag{B.5}$$

Hence, we have

$$\begin{aligned}\mathbb{E}\left[\left(\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{u}}_t\right)^4\right] &= \mathbb{E}\left[\left(\sum_{i=0}^{L}\boldsymbol{v}_i^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_i\right)^4\right], \\ &= \mathbb{E}\left[\left(\boldsymbol{v}_0^\mathsf{T}\boldsymbol{u}_t + \boldsymbol{v}_1^\mathsf{T}\boldsymbol{u}_{t-1} + \sum_{i=2}^{L}\boldsymbol{v}_i^\mathsf{T}(\bar{\boldsymbol{u}}_{t-1} \otimes \cdots \otimes \boldsymbol{u}_{t-i})\right)^4\right].\end{aligned} \tag{B.6}$$

In the following, we will upper each term appearing on the right-hand-side (RHS) of (B.6) to get an upper bound on the fourth moment marginals of $\tilde{\boldsymbol{u}}_t$.

• **Linear Terms:** For $i \neq j \neq k \neq \ell$, consider the following expectation: We can utilize $\mathbb{E}[\boldsymbol{u}_{t-\max\{i,j,k,\ell\}}] = 0$, to show that

$$\mathbb{E}\left[(\boldsymbol{v}_i^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_i)(\boldsymbol{v}_j^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_j)(\boldsymbol{v}_k^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_k)(\boldsymbol{v}_\ell^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_\ell)\right] = 0. \tag{B.7}$$

• **Quadratic Terms I:** For $i \neq j \neq k$, consider the following expectation: When $i < \max\{j, k\}$, we can utilize $\mathbb{E}[\boldsymbol{u}_{t-\max\{j,k\}}] = 0$, to show that

$$\mathbb{E}\left[(\boldsymbol{v}_i^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_i)^2(\boldsymbol{v}_j^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_j)(\boldsymbol{v}_k^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_k)\right] = 0. \tag{B.8}$$

However, when $i > \max\{j, k\}$, the above expectation is not zero. We can upper bound these terms using Cauchy-Schwarz inequality as follows,

$$\mathbb{E}\left[(\boldsymbol{v}_i^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_i)^2(\boldsymbol{v}_j^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_j)(\boldsymbol{v}_k^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_k)\right] \leq \sqrt{\mathbb{E}\left[(\boldsymbol{v}_i^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_i)^4\right]\mathbb{E}\left[(\boldsymbol{v}_j^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_j)^2(\boldsymbol{v}_k^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_k)^2\right]}. \tag{B.9}$$

In the remaining of the proof, we will upper bound the two terms on the RHS of (B.9) individually, along-with the remaining terms on the RHS of (B.6).

• **Quadratic Terms II:** For $i \neq j$, consider the expectation $\mathbb{E}\left[(\boldsymbol{v}_i^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_i)^2(\boldsymbol{v}_j^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_j)^2\right]$. In order to upper bound this term, we first upper bound an intermediate term as follows: Let $[\tilde{\tilde{\boldsymbol{u}}}_t]_k$ denote

the k-th partition of $\tilde{\tilde{\boldsymbol{u}}}_t$ defined in (B.1), for $k = 1, \ldots, L$. Let $\boldsymbol{q}_k, \boldsymbol{q}'_k$ be two arbitrary vectors in $\boldsymbol{R}^{(p+1)^k}$. We will use the mathematical induction to prove the following intermediate result under Assumption 2,

$$\mathbb{E}\left[(\boldsymbol{q}_k^\mathsf{T}[\tilde{\tilde{\boldsymbol{u}}}_t]_k)^2(\boldsymbol{q}_k'^\mathsf{T}[\tilde{\tilde{\boldsymbol{u}}}_t]_k)^2\right] \leq (3 \vee \gamma)^k \|\boldsymbol{q}_k\|_{\ell_2}^2 \|\boldsymbol{q}'_k\|_{\ell_2}^2, \quad \text{for all} \quad k = 1, \ldots, L. \tag{B.10}$$

We begin the proof, by showing that, $k = 1$ obeys the induction as follows,

$$\begin{aligned}
&\mathbb{E}\left[(\boldsymbol{q}_1^\mathsf{T}[\tilde{\tilde{\boldsymbol{u}}}_t]_1)^2(\boldsymbol{q}_1'^\mathsf{T}[\tilde{\tilde{\boldsymbol{u}}}_t]_1)^2\right] \\
&\quad = \mathbb{E}\left[\left(\boldsymbol{q}_1^\mathsf{T}\bar{\boldsymbol{u}}_{t-1}\right)^2\left(\boldsymbol{q}_1'^\mathsf{T}\bar{\boldsymbol{u}}_{t-1}\right)^2\right], \\
&\quad \overset{\text{(i)}}{=} \mathbb{E}\left[\left(q_{11} + \boldsymbol{q}_{12}^\mathsf{T}\boldsymbol{u}_{t-1}\right)^2\left(q'_{11} + \boldsymbol{q}_{12}'^\mathsf{T}\boldsymbol{u}_{t-1}\right)^2\right], \\
&\quad = \mathbb{E}\left[(q_{11}^2 + (\boldsymbol{q}_{12}^\mathsf{T}\boldsymbol{u}_{t-1})^2 + 2q_{11}\boldsymbol{q}_{12}^\mathsf{T}\boldsymbol{u}_{t-1})\left(q_{11}'^2 + (\boldsymbol{q}_{12}'^\mathsf{T}\boldsymbol{u}_{t-1})^2 + 2q'_{11}\boldsymbol{q}_{12}'^\mathsf{T}\boldsymbol{u}_{t-1}\right)\right], \\
&\quad \overset{\text{(ii)}}{\leq} q_{11}^2 q_{11}'^2 + q_{11}^2\|\boldsymbol{q}'_{12}\|_{\ell_2}^2 + q_{11}'^2\|\boldsymbol{q}_{12}\|_{\ell_2}^2 + \gamma\|\boldsymbol{q}_{12}\|_{\ell_2}^2\|\boldsymbol{q}'_{12}\|_{\ell_2}^2 + 4q_{11}q'_{11}\|\boldsymbol{q}_{12}\|_{\ell_2}\|\boldsymbol{q}'_{12}\|_{\ell_2}, \\
&\quad \overset{\text{(ii)}}{\leq} q_{11}^2 q_{11}'^2 + 3q_{11}^2\|\boldsymbol{q}'_{12}\|_{\ell_2}^2 + 3q_{11}'^2\|\boldsymbol{q}_{12}\|_{\ell_2}^2 + \gamma\|\boldsymbol{q}_{12}\|_{\ell_2}^2\|\boldsymbol{q}'_{12}\|_{\ell_2}^2, \\
&\quad \leq (3 \vee \gamma)\left(q_{11}^2 + \|\boldsymbol{q}_{12}\|_{\ell_2}^2\right)\left(q_{11}'^2 + \|\boldsymbol{q}_{12}\|_{\ell_2}'^2\right) = (3 \vee \gamma)\|\boldsymbol{q}_1\|_{\ell_2}^2\|\boldsymbol{q}'_1\|_{\ell_2}^2,
\end{aligned} \tag{B.11}$$

where we obtain (i) from setting $\boldsymbol{q}_1 = [q_{11} \;\; \boldsymbol{q}_{12}^\mathsf{T}]^\mathsf{T}$, $\boldsymbol{q}'_1 = [q'_{11} \;\; \boldsymbol{q}_{12}'^\mathsf{T}]^\mathsf{T}$, (ii) from applying Lemma 10 along-with Cauchy–Schwarz inequality, and (iii) is obtained by using the identity $2ab \leq a^2 + b^2$ for $a, b \in \mathbb{R}$. Suppose we have $\mathbb{E}\left[\left(\boldsymbol{q}_{k-1}^\mathsf{T}[\tilde{\tilde{\boldsymbol{u}}}_t]_{k-1}\right)^2\left(\boldsymbol{q}_{k-1}'^\mathsf{T}[\tilde{\tilde{\boldsymbol{u}}}_t]_{k-1}\right)^2\right] \leq (3 \vee \gamma)^{k-1}\|\boldsymbol{q}_{k-1}\|_{\ell_2}^2\|\boldsymbol{q}'_{k-1}\|_{\ell_2}^2$. Then, we apply the induction as follows,

$$\begin{aligned}
\mathbb{E}\left[\left(\boldsymbol{q}_k^\mathsf{T}[\tilde{\tilde{\boldsymbol{u}}}_t]_k\right)^2\left(\boldsymbol{q}_k'^\mathsf{T}[\tilde{\tilde{\boldsymbol{u}}}_t]_k\right)^2\right] &= \mathbb{E}\left[\left(\boldsymbol{q}_k^\mathsf{T}([\tilde{\tilde{\boldsymbol{u}}}_t]_{k-1} \otimes \bar{\boldsymbol{u}}_{t-k})\right)^2\left(\boldsymbol{q}_k'^\mathsf{T}([\tilde{\tilde{\boldsymbol{u}}}_t]_{k-1} \otimes \bar{\boldsymbol{u}}_{t-k})\right)^2\right], \\
&= \mathbb{E}\left[\left(\bar{\boldsymbol{u}}_{t-k}^\mathsf{T}\boldsymbol{Q}_k[\tilde{\tilde{\boldsymbol{u}}}_t]_{k-1}\right)^2\left(\bar{\boldsymbol{u}}_{t-k}^\mathsf{T}\boldsymbol{Q}'_k[\tilde{\tilde{\boldsymbol{u}}}_t]_{k-1}\right)^2\right], \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\bar{\boldsymbol{u}}_{t-k}^\mathsf{T}\boldsymbol{Q}_k[\tilde{\tilde{\boldsymbol{u}}}_t]_{k-1}\right)^2\left(\bar{\boldsymbol{u}}_{t-k}^\mathsf{T}\boldsymbol{Q}'_k[\tilde{\tilde{\boldsymbol{u}}}_t]_{k-1}\right)^2 \mid \boldsymbol{u}_{t-k}\right]\right], \\
&\overset{\text{(i)}}{\leq} (3 \vee \gamma)^{k-1}\,\mathbb{E}\left[\|\boldsymbol{Q}_k^\mathsf{T}\bar{\boldsymbol{u}}_{t-k}\|_{\ell_2}^2\|\boldsymbol{Q}_k'^\mathsf{T}\bar{\boldsymbol{u}}_{t-k}\|_{\ell_2}^2\right], \\
&\overset{\text{(ii)}}{\leq} (3 \vee \gamma)^k\|\boldsymbol{Q}_k\|_F^2\|\boldsymbol{Q}'_k\|_F^2, \\
&= (3 \vee \gamma)^k\|\boldsymbol{q}_k\|_{\ell_2}^2\|\boldsymbol{q}'_k\|_{\ell_2}^2,
\end{aligned} \tag{B.12}$$

where $\boldsymbol{Q}_k = \mathbf{mtx}(\boldsymbol{q}_k) \in \mathbb{R}^{(p+1)\times(p+1)^{k-1}}$, $\boldsymbol{Q}'_k = \mathbf{mtx}(\boldsymbol{q}'_k) \in \mathbb{R}^{(p+1)\times(p+1)^{k-1}}$, we obtain (i) from the induction hypothesis, and (ii) follows from a similar line of reasoning as used to derive an upper bound in (B.11). This completes the proof of our intermediate result in (B.10). Next, we use (B.10) to derive an upper bound on $\mathbb{E}\left[(\boldsymbol{v}_i^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_i)^2(\boldsymbol{v}_j^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_j)^2\right]$ as follows: Due to symmetry, it is sufficient to consider $j > i$. We begin by deriving the upper bound for $j > i \geq 2$ as follows,

$$
\begin{aligned}
&\mathbb{E}\left[(\boldsymbol{v}_i^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_i)^2(\boldsymbol{v}_j^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_j)^2\right] \\
&\quad= \mathbb{E}\left[\left(\boldsymbol{v}_i^\mathsf{T}(\bar{\boldsymbol{u}}_{t-1}\otimes\cdots\otimes\boldsymbol{u}_{t-i})\right)^2\left(\boldsymbol{v}_j^\mathsf{T}(\bar{\boldsymbol{u}}_{t-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\right)^2\right], \\
&\quad= \mathbb{E}\left[\mathbb{E}\left[\left(\boldsymbol{v}_i^\mathsf{T}(\bar{\boldsymbol{u}}_{t-1}\otimes\cdots\otimes\boldsymbol{u}_{t-i})\right)^2\left(\boldsymbol{v}_j^\mathsf{T}(\bar{\boldsymbol{u}}_{t-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\right)^2 \;\middle|\; \boldsymbol{u}_{t-i},\boldsymbol{u}_{t-i-1},\ldots,\boldsymbol{u}_{t-j}\right]\right], \\
&\quad\overset{\text{(i)}}{=} \mathbb{E}\left[\mathbb{E}\left[\left(\boldsymbol{u}_{t-i}^\mathsf{T}\boldsymbol{V}_i(\bar{\boldsymbol{u}}_{t-1}\otimes\cdots\otimes\bar{\boldsymbol{u}}_{t-i+1})\right)^2\left((\bar{\boldsymbol{u}}_{t-i}^\mathsf{T}\otimes\cdots\otimes\boldsymbol{u}_{t-j}^\mathsf{T})\boldsymbol{V}_j(\bar{\boldsymbol{u}}_{t-1}\otimes\cdots\otimes\bar{\boldsymbol{u}}_{t-i+1})\right)^2\right.\right. \\
&\qquad\left.\left.\;\middle|\; \boldsymbol{u}_{t-i},\boldsymbol{u}_{t-i-1},\ldots,\boldsymbol{u}_{t-j}\right]\right], \\
&\quad= \mathbb{E}\left[\mathbb{E}\left[\left(\boldsymbol{u}_{t-i}^\mathsf{T}\boldsymbol{V}_i[\tilde{\bar{\boldsymbol{u}}}_t]_{i-1}\right)^2\left((\bar{\boldsymbol{u}}_{t-i}^\mathsf{T}\otimes\cdots\otimes\boldsymbol{u}_{t-j}^\mathsf{T})\boldsymbol{V}_j[\tilde{\bar{\boldsymbol{u}}}_t]_{i-1}\right)^2 \;\middle|\; \boldsymbol{u}_{t-i},\boldsymbol{u}_{t-i-1},\ldots,\boldsymbol{u}_{t-j}\right]\right], \\
&\quad\overset{\text{(ii)}}{\leq} (3\vee\gamma)^{i-1}\,\mathbb{E}\left[\|\boldsymbol{V}_i^\mathsf{T}\boldsymbol{u}_{t-i}\|_{\ell_2}^2\|\boldsymbol{V}_j^\mathsf{T}(\bar{\boldsymbol{u}}_{t-i}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\|_{\ell_2}^2\right], \qquad \text{(B.13)}
\end{aligned}
$$

where we obtain (i) from tower rule, and setting $\boldsymbol{V}_i = \mathbf{mtx}(\boldsymbol{v}_i) \in \mathbb{R}^{p\times(p+1)^{i-1}}$, $\boldsymbol{V}_j = \mathbf{mtx}(\boldsymbol{v}_j) \in \mathbb{R}^{p(p+1)^{j-i}\times(p+1)^{i-1}}$, and (ii) is obtained by using the intermediate result from (B.10). To proceed, observe that

$$
\bar{\boldsymbol{u}}_{t-i}\otimes\cdots\otimes\boldsymbol{u}_{t-j} = \begin{bmatrix}\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j}\\ \boldsymbol{u}_{t-i}\otimes\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j}\end{bmatrix}, \quad \boldsymbol{V}_j = \begin{bmatrix}\boldsymbol{V}_{j1}\\ \boldsymbol{V}_{j2}\end{bmatrix}, \qquad \text{(B.14)}
$$

where $\boldsymbol{V}_{j1} \in \mathbb{R}^{p(p+1)^{j-i-1}\times(p+1)^{i-1}}$, and $\boldsymbol{V}_{j2} \in \mathbb{R}^{p^2(p+1)^{j-i-1}\times(p+1)^{i-1}}$. Combining this with (B.13), we have

$$
\begin{aligned}
\mathbb{E}\left[(\boldsymbol{v}_i^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_i)^2(\boldsymbol{v}_j^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_j)^2\right] &\leq (3\vee\gamma)^{i-1}\,\mathbb{E}\left[\|\boldsymbol{V}_i^\mathsf{T}\boldsymbol{u}_{t-i}\|_{\ell_2}^2\|\boldsymbol{V}_{j1}^\mathsf{T}(\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\right. \\
&\quad\left.+\,\boldsymbol{V}_{j2}^\mathsf{T}(\boldsymbol{u}_{t-i}\otimes\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\|_{\ell_2}^2\right], \\
&\leq (3\vee\gamma)^{i-1}\,\mathbb{E}\left[\|\boldsymbol{V}_i^\mathsf{T}\boldsymbol{u}_{t-i}\|_{\ell_2}^2\left(\|\boldsymbol{V}_{j1}^\mathsf{T}(\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\|_{\ell_2}^2\right.\right. \\
&\quad+\|\boldsymbol{V}_{j2}^\mathsf{T}(\boldsymbol{u}_{t-i}\otimes\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\|_{\ell_2}^2 \\
&\quad\left.\left.+\,2(\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})^\mathsf{T}\boldsymbol{V}_{j1}\boldsymbol{V}_{j2}^\mathsf{T}(\boldsymbol{u}_{t-i}\otimes\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\right)\right], \qquad \text{(B.15)}
\end{aligned}
$$

In the following, we will upper bound each term in (B.15) separately to get an upper bound on $\mathbb{E}\left[(\boldsymbol{v}_i^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_i)^2(\boldsymbol{v}_j^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_j)^2\right]$. To begin, we have

$$
\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{V}_i^\mathsf{T}\boldsymbol{u}_{t-i}\|_{\ell_2}^2\|\boldsymbol{V}_{j1}^\mathsf{T}(\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\|_{\ell_2}^2\right] \\
&\quad= \mathbb{E}\left[\boldsymbol{u}_{t-i}^\mathsf{T}\boldsymbol{V}_i\boldsymbol{V}_i^\mathsf{T}\boldsymbol{u}_{t-i}\right]\mathbb{E}\left[(\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})^\mathsf{T}\boldsymbol{V}_{j1}\boldsymbol{V}_{j1}^\mathsf{T}(\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\right], \\
&\quad= \mathbb{E}\left[\mathbf{tr}\big(\boldsymbol{V}_i\boldsymbol{V}_i^\mathsf{T}\boldsymbol{u}_{t-i}\boldsymbol{u}_{t-i}^\mathsf{T}\big)\right]\mathbb{E}\left[\mathbf{tr}\big(\boldsymbol{V}_{j1}\boldsymbol{V}_{j1}^\mathsf{T}(\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})(\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})^\mathsf{T}\big)\right], \\
&\quad\leq \|\boldsymbol{V}_i\|_F^2\|\boldsymbol{V}_{j1}\|_F^2. \qquad \text{(B.16)}
\end{aligned}
$$

Similarly, the second term in (B.15) can be upper bounded as follows,

$$
\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{V}_i^\mathsf{T}\boldsymbol{u}_{t-i}\|_{\ell_2}^2\|\boldsymbol{V}_{j2}^\mathsf{T}(\boldsymbol{u}_{t-i}\otimes\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\|_{\ell_2}^2\right]\\
&\quad\overset{\text{(a)}}{=}\sum_{k=1}^{(p+1)^{i-1}}\mathbb{E}\left[\|\boldsymbol{V}_i^\mathsf{T}\boldsymbol{u}_{t-i}\|_{\ell_2}^2\left(\boldsymbol{v}_{j2k}^\mathsf{T}(\boldsymbol{u}_{t-i}\otimes\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\right)^2\right],\\
&\quad\overset{\text{(b)}}{=}\sum_{k=1}^{(p+1)^{i-1}}\mathbb{E}\left[\|\boldsymbol{V}_i^\mathsf{T}\boldsymbol{u}_{t-i}\|_{\ell_2}^2\left((\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})^\mathsf{T}\boldsymbol{V}_{j2k}\boldsymbol{u}_{t-i}\right)^2\right],\\
&\quad=\sum_{k=1}^{(p+1)^{i-1}}\mathbb{E}\left[\mathbb{E}\left[\boldsymbol{u}_{t-i}^\mathsf{T}\boldsymbol{V}_i\boldsymbol{V}_i^\mathsf{T}\boldsymbol{u}_{t-i}\left((\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})^\mathsf{T}\boldsymbol{V}_{j2k}\boldsymbol{u}_{t-i}\right)^2\,\middle|\,\boldsymbol{u}_{t-i-1},\ldots,\boldsymbol{u}_{t-j}\right]\right],\\
&\quad\overset{\text{(c)}}{\leq}\sum_{k=1}^{(p+1)^{i-1}}\gamma\|\boldsymbol{V}_i\|_F^2\,\mathbb{E}\left[\mathbf{tr}\left((\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})^\mathsf{T}\boldsymbol{V}_{j2k}\boldsymbol{V}_{j2k}^\mathsf{T}(\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\right)\right],\\
&\quad=\gamma\|\boldsymbol{V}_i\|_F^2\sum_{k=1}^{(p+1)^{i-1}}\|\boldsymbol{V}_{j2k}\|_F^2,\\
&\quad=\gamma\|\boldsymbol{V}_i\|_F^2\|\boldsymbol{V}_{j2}\|_F^2,
\end{aligned}
\tag{B.17}
$$

where we obtain (a) by defining $\boldsymbol{v}_{j2k}^\mathsf{T}$ to be the k-th row of $\boldsymbol{V}_{j2}^\mathsf{T}$, (b) from setting $\boldsymbol{V}_{j2k}=\mathbf{mtx}(\boldsymbol{v}_{j2k})\in\mathbb{R}^{p(p+1)^{j-i-1}\times p}$, and (c) follows from the application of Lemma 10. Lastly, note that the third term in (B.15) is zero as follows,

$$
\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{V}_i^\mathsf{T}\boldsymbol{u}_{t-i}\|_{\ell_2}^2(\boldsymbol{u}_{t-i}\otimes\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})^\mathsf{T}\boldsymbol{V}_{j2}\boldsymbol{V}_{j1}^\mathsf{T}(\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\right]\\
&\quad=\mathbb{E}\left[\|\boldsymbol{V}_i^\mathsf{T}\boldsymbol{u}_{t-i}\|_{\ell_2}^2(\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})^\mathsf{T}\mathbf{mtx}\left(\boldsymbol{V}_{j2}\boldsymbol{V}_{j1}^\mathsf{T}(\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\right)\boldsymbol{u}_{t-i}\right],\\
&\quad=\mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{V}_i^\mathsf{T}\boldsymbol{u}_{t-i}\|_{\ell_2}^2(\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})^\mathsf{T}\mathbf{mtx}\left(\boldsymbol{V}_{j2}\boldsymbol{V}_{j1}^\mathsf{T}(\bar{\boldsymbol{u}}_{t-i-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j})\right)\boldsymbol{u}_{t-i}\right.\right.\\
&\qquad\left.\left.\,\middle|\,\boldsymbol{u}_{t-i-1},\ldots,\boldsymbol{u}_{t-j}\right]\right],\\
&\quad=0.
\end{aligned}
\tag{B.18}
$$

Finally, combining (B.16), (B.17), and (B.18) into (B.15) we get the following upper bound,

$$
\begin{aligned}
\mathbb{E}\left[(\boldsymbol{v}_i^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_i)^2(\boldsymbol{v}_j^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_j)^2\right]&\leq(3\vee\gamma)^{i-1}\left(\|\boldsymbol{V}_i\|_F^2\|\boldsymbol{V}_{j1}\|_F^2+\gamma\|\boldsymbol{V}_i\|_F^2\|\boldsymbol{V}_{j2}\|_F^2\right),\\
&\leq(3\vee\gamma)^i\|\boldsymbol{v}_i\|_{\ell_2}^2\|\boldsymbol{v}_j\|_{\ell_2}^2,\quad\text{for all}\quad j>i\geq 2.
\end{aligned}
\tag{B.19}
$$

For $j>i=0$, it is easy to see that $\mathbb{E}\left[(\boldsymbol{v}_0^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_0)^2(\boldsymbol{v}_j^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_j)^2\right]=\mathbb{E}\left[(\boldsymbol{v}_0^\mathsf{T}\boldsymbol{u}_t)^2\right]\mathbb{E}\left[(\boldsymbol{v}_j^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_j)^2\right]=\|\boldsymbol{v}_0\|_{\ell_2}^2\|\boldsymbol{v}_j\|_{\ell_2}^2$. Finally, for $j>i=1$, we get the following upper bounds,

$$
\begin{aligned}
\mathbb{E}\left[(\boldsymbol{v}_1^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_1)^2(\boldsymbol{v}_j^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_j)^2\right]&=\mathbb{E}\left[(\boldsymbol{v}_1^\mathsf{T}\boldsymbol{u}_{t-1})^2(\boldsymbol{v}_j^\mathsf{T}(\bar{\boldsymbol{u}}_{t-1}\otimes\cdots\otimes\boldsymbol{u}_{t-j}))^2\right],\\
&=\mathbb{E}\left[\mathbb{E}\left[(\boldsymbol{v}_1^\mathsf{T}\boldsymbol{u}_{t-1})^2((\bar{\boldsymbol{u}}_{t-2}\otimes\cdots\otimes\boldsymbol{u}_{t-j})^\mathsf{T}\boldsymbol{V}_j\bar{\boldsymbol{u}}_{t-1})^2\,\middle|\,\boldsymbol{u}_{t-1}\right]\right],\\
&=\mathbb{E}\left[(\boldsymbol{v}_1^\mathsf{T}\boldsymbol{u}_{t-1})^2\|\boldsymbol{V}_j\bar{\boldsymbol{u}}_{t-1}\|_{\ell_2}^2\right],\\
&\overset{\text{(i)}}{=}\mathbb{E}\left[(\boldsymbol{v}_1^\mathsf{T}\boldsymbol{u}_{t-1})^2\|\boldsymbol{v}_{j1}+\boldsymbol{V}_{j2}\boldsymbol{u}_{t-1}\|_{\ell_2}^2\right],\\
&\overset{\text{(ii)}}{\leq}\|\boldsymbol{v}_1\|_{\ell_2}^2\|\boldsymbol{v}_{j1}\|_{\ell_2}^2+\gamma\|\boldsymbol{v}_1\|_{\ell_2}^2\|\boldsymbol{V}_{j2}\|_F^2\leq\gamma\|\boldsymbol{v}_1\|_{\ell_2}^2\|\boldsymbol{v}_j\|_{\ell_2}^2,
\end{aligned}
\tag{B.20}
$$

where we obtain (i) from setting $\boldsymbol{V}_j = \mathbf{mtx}(\boldsymbol{v}_j) := [\boldsymbol{v}_{j1} \;\; \boldsymbol{V}_{j2}] \in \mathbb{R}^{p(p+1)^{j-2} \times (p+1)}$, and (ii) follows from Lemma 10. Hence, for all $i \neq j$, we have

$$\mathbb{E}\left[(\boldsymbol{v}_i^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_i)^2 (\boldsymbol{v}_j^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_j)^2\right] \leq (3 \vee \gamma)^{\min\{i,j\}} \|\boldsymbol{v}_i\|_{\ell_2}^2 \|\boldsymbol{v}_j\|_{\ell_2}^2, \quad \text{for all} \quad i \neq j. \tag{B.21}$$

• **Cubic Terms:** For $i \neq j$, consider the following expectation: When $i < j$, we can utilize $\mathbb{E}[\boldsymbol{u}_{t-j}] = 0$, and when $i > j$, we can use $\mathbb{E}[(\boldsymbol{b}^\mathsf{T}\boldsymbol{u}_{t-i})^3] = 0$, to show that

$$\mathbb{E}\left[(\boldsymbol{v}_i^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_i)^3 (\boldsymbol{v}_j^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_j)\right] = 0. \tag{B.22}$$

• **Quartic Terms:** Consider the expectation $\mathbb{E}\left[(\boldsymbol{v}_i^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_i)^4\right]$. In order to upper bound this term, we first upper bound an intermediate term as follows: Recall the definition of $\tilde{\tilde{\boldsymbol{u}}}_t$ from (B.1), and let $[\tilde{\tilde{\boldsymbol{u}}}_t]_k$ denote its k-th partition with $(p+1)^k$ entries, for $k = 1, \ldots, L$. For any $\boldsymbol{q}_k \in \boldsymbol{R}^{(p+1)^k}$, we will use the mathematical induction to prove that

$$\mathbb{E}\left[(\boldsymbol{q}_k^\mathsf{T}[\tilde{\tilde{\boldsymbol{u}}}_t]_k)^4\right] \leq (3 \vee \gamma)^k \|\boldsymbol{q}_k\|_{\ell_2}^4, \quad \text{for all} \quad k = 1, \ldots, L. \tag{B.23}$$

We begin the proof, by showing that, $k = 1$ obeys the induction as follows,

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{q}_1^\mathsf{T}[\tilde{\tilde{\boldsymbol{u}}}_t]_1)^4\right] = \mathbb{E}\left[\left(q_{11} + \boldsymbol{q}_{12}^\mathsf{T}\boldsymbol{u}_{t-1}\right)^4\right] &= q_{11}^4 + \mathbb{E}\left[(\boldsymbol{q}_{12}^\mathsf{T}\boldsymbol{u}_{t-1})^4\right] + 6q_{11}^2\,\mathbb{E}\left[(\boldsymbol{q}_{12}^\mathsf{T}\boldsymbol{u}_{t-1})^2\right], \\
&\overset{\text{(a)}}{\leq} q_{11}^4 + \gamma\|\boldsymbol{q}_{12}\|_{\ell_2}^4 + 6q_{11}^2\|\boldsymbol{q}_{12}\|_{\ell_2}^2, \\
&\leq (3 \vee \gamma)\left(q_{11}^4 + \|\boldsymbol{q}_{12}\|_{\ell_2}^4 + 2q_{11}^2\|\boldsymbol{q}_{12}\|_{\ell_2}^2\right), \\
&= (3 \vee \gamma)\left(q_{11}^2 + \|\boldsymbol{q}_{12}\|_{\ell_2}^2\right)^2, \\
&= (3 \vee \gamma)\|\boldsymbol{q}_1\|_{\ell_2}^4,
\end{aligned} \tag{B.24}$$

where (a) follows from Lemma 10. Suppose we have $\mathbb{E}\left[(\boldsymbol{q}_{k-1}^\mathsf{T}[\tilde{\tilde{\boldsymbol{u}}}_t]_{k-1})^4\right] \leq (3 \vee \gamma)^{k-1}\|\boldsymbol{q}_{k-1}\|_{\ell_2}^4$ for any $\boldsymbol{q}_{k-1} \in \mathbb{R}^{(p+1)^{k-1}}$. Then, we apply the induction as follows,

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{q}_k^\mathsf{T}[\tilde{\tilde{\boldsymbol{u}}}_t]_k)^4\right] &= \mathbb{E}\left[\left(\boldsymbol{q}_k^\mathsf{T}([\tilde{\tilde{\boldsymbol{u}}}_t]_{k-1} \otimes \bar{\boldsymbol{u}}_{t-k})\right)^4\right], \\
&\overset{\text{(i)}}{=} \mathbb{E}\left[\mathbb{E}\left[\left(\bar{\boldsymbol{u}}_{t-k}^\mathsf{T}\boldsymbol{Q}_k[\tilde{\tilde{\boldsymbol{u}}}_t]_{k-1}\right)^4 \mid \boldsymbol{u}_{t-k}\right]\right], \\
&\overset{\text{(ii)}}{\leq} (3 \vee \gamma)^{k-1}\,\mathbb{E}\left[\|\boldsymbol{Q}_k^\mathsf{T}\bar{\boldsymbol{u}}_{t-k}\|_{\ell_2}^4\right], \\
&\overset{\text{(iii)}}{\leq} (3 \vee \gamma)^k\|\boldsymbol{q}_k\|_{\ell_2}^4,
\end{aligned} \tag{B.25}$$

where we obtain (i) from setting $\boldsymbol{Q}_k = \mathbf{mtx}(\boldsymbol{q}_k) \in \mathbb{R}^{(p+1)\times(p+1)^{k-1}}$, (ii) from the induction hypothesis, and (iii) is obtained from Lemma 10 as follows: Let $\boldsymbol{q}_{ki}^\mathsf{T}$ denote the i-th row of $\boldsymbol{Q}_k^\mathsf{T}$. Then we have,

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{Q}_k^\mathsf{T}\bar{\boldsymbol{u}}_{t-k}\|_{\ell_2}^4\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{(p+1)^{k-1}} (\boldsymbol{q}_{ki}^\mathsf{T}\bar{\boldsymbol{u}}_{t-k})^2\right)^2\right], \\
&= \mathbb{E}\left[\sum_{i=1}^{(p+1)^{k-1}} (\boldsymbol{q}_{ki}^\mathsf{T}\bar{\boldsymbol{u}}_{t-k})^4 + \sum_{i=1}^{(p+1)^{k-1}}\sum_{\substack{j=1\\ j\neq i}}^{(p+1)^{k-1}} (\boldsymbol{q}_{ki}^\mathsf{T}\bar{\boldsymbol{u}}_{t-k})^2(\boldsymbol{q}_{kj}^\mathsf{T}\bar{\boldsymbol{u}}_{t-k})^2\right], \\
&\overset{\text{(a)}}{\leq} (3\vee\gamma)\left(\sum_{i=1}^{(p+1)^{k-1}} \|\boldsymbol{q}_{ki}\|_{\ell_2}^4 + \sum_{i=1}^{(p+1)^{k-1}}\sum_{\substack{j=1\\ j\neq i}}^{(p+1)^{k-1}} \|\boldsymbol{q}_{ki}\|_{\ell_2}^2\|\boldsymbol{q}_{kj}\|_{\ell_2}^2\right), \\
&= (3\vee\gamma)\left(\sum_{i=1}^{(p+1)^{k-1}} \|\boldsymbol{q}_{ki}\|_{\ell_2}^2\right)^2, \\
&= (3\vee\gamma)\|\boldsymbol{Q}_k\|_F^4, \qquad \text{(B.26)}
\end{aligned}$$

where we obtain (a) from (B.11) and (B.24). Hence, we proved by induction that, for any $\boldsymbol{q}_k \in \boldsymbol{R}^{(p+1)^k}$, we have $\mathbb{E}\left[(\boldsymbol{q}_k^\mathsf{T}[\tilde{\tilde{\boldsymbol{u}}}_t]_k)^4\right] \leq (3\vee\gamma)^k\|\boldsymbol{q}_k\|_{\ell_2}^4$, for all $k = 1,\ldots,L$. Next, recalling the definition of $\tilde{\boldsymbol{u}}_t$ from (B.1), we use the intermediate result in (B.23) to get an upper bound on $\mathbb{E}\left[(\boldsymbol{v}_k^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_k)^4\right]$ as follows: For $k = 2,\ldots,L$, we have,

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{v}_k^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_k)^4\right] &= \mathbb{E}\left[\left(\boldsymbol{v}_k^\mathsf{T}([\tilde{\tilde{\boldsymbol{u}}}_t]_{k-1}\otimes\boldsymbol{u}_{t-k})\right)^4\right], \\
&\overset{\text{(i)}}{=} \mathbb{E}\left[\mathbb{E}\left[\left(\boldsymbol{u}_{t-k}^\mathsf{T}\boldsymbol{V}_k[\tilde{\tilde{\boldsymbol{u}}}_t]_{k-1}\right)^4 \mid \boldsymbol{u}_{t-k}\right]\right], \\
&\overset{\text{(ii)}}{\leq} (3\vee\gamma)^{k-1}\,\mathbb{E}\left[\|\boldsymbol{V}_k^\mathsf{T}\boldsymbol{u}_{t-k}\|_{\ell_2}^4\right], \\
&\overset{\text{(iii)}}{\leq} (3\vee\gamma)^k\|\boldsymbol{V}_k\|_F^4 = (3\vee\gamma)^k\|\boldsymbol{v}_k\|_{\ell_2}^4, \qquad \text{(B.27)}
\end{aligned}$$

where we obtain (i) from setting $\boldsymbol{V}_k = \mathbf{mtx}(\boldsymbol{v}_k) \in \mathbb{R}^{p\times(p+1)^{k-1}}$, (ii) is obtained from using (B.23), and (iii) follows from Lemma 10. For $k = 0, 1$, we have, $\mathbb{E}\left[(\boldsymbol{v}_0^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_0)^4\right] = \mathbb{E}\left[(\boldsymbol{v}_0^\mathsf{T}\boldsymbol{u}_t)^4\right] \leq \gamma\|\boldsymbol{v}_0\|_{\ell_2}^4$ and $\mathbb{E}\left[(\boldsymbol{v}_1^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_1)^4\right] = \mathbb{E}\left[(\boldsymbol{v}_1^\mathsf{T}\boldsymbol{u}_{t-1})^4\right] \leq \gamma\|\boldsymbol{v}_1\|_{\ell_2}^4$ using Lemma 10. Hence, we showed that,

$$\mathbb{E}\left[(\boldsymbol{v}_k^\mathsf{T}[\tilde{\boldsymbol{u}}_t]_k)^4\right] \leq (3\vee\gamma)^{\max\{1,k\}}\|\boldsymbol{v}_k\|_{\ell_2}^4, \quad \text{for all} \quad k = 0,\ldots,L. \qquad \text{(B.28)}$$

• **Finalizing the Proof:** Putting it all together, for any $\boldsymbol{v} \in \mathcal{S}^{d_{\tilde{u}}-1}$ and $\tilde{\boldsymbol{u}}_t$ in (B.1), we have

$$
\begin{aligned}
\mathbb{E}\left[\left(\boldsymbol{v}^\intercal \tilde{\boldsymbol{u}}_t\right)^4\right] &= \mathbb{E}\left[\left(\sum_{i=0}^{L} \boldsymbol{v}_i^\intercal [\tilde{\boldsymbol{u}}_t]_i\right)^4\right], \\
&= \sum_{i=0}^{L} \mathbb{E}\left[\left(\boldsymbol{v}_i^\intercal [\tilde{\boldsymbol{u}}_t]_i\right)^4\right] + 3\sum_{i=0}^{L}\sum_{\substack{j=0\\ j\neq i}}^{L} \mathbb{E}\left[\left(\boldsymbol{v}_i^\intercal [\tilde{\boldsymbol{u}}_t]_i\right)^2 \left(\boldsymbol{v}_j^\intercal [\tilde{\boldsymbol{u}}_t]_j\right)^2\right], \\
&+ 6\sum_{i=0}^{L}\sum_{\substack{j=0\\ j\neq i}}^{L}\sum_{\substack{k=0\\ k\neq i\\ k\neq j}}^{L} \mathbb{E}\left[\left(\boldsymbol{v}_i^\intercal [\tilde{\boldsymbol{u}}_t]_i\right)^2 \left(\boldsymbol{v}_j^\intercal [\tilde{\boldsymbol{u}}_t]_j\right)\left(\boldsymbol{v}_k^\intercal [\tilde{\boldsymbol{u}}_t]_k\right)\right], \\
&\leq \sum_{i=0}^{L} (3\vee\gamma)^{\max\{1,i\}} \|\boldsymbol{v}_i\|_{\ell_2}^4 + 3\sum_{i=0}^{L}\sum_{\substack{j=0\\ j\neq i}}^{L} (3\vee\gamma)^{\min\{i,j\}} \|\boldsymbol{v}_i\|_{\ell_2}^2 \|\boldsymbol{v}_j\|_{\ell_2}^2 \\
&+ 6\sum_{i=0}^{L}\sum_{\substack{j=0\\ j\neq i}}^{L}\sum_{\substack{k=0\\ k\neq i\\ k\neq j}}^{L} \sqrt{\mathbb{E}\left[\left(\boldsymbol{v}_i^\intercal [\tilde{\boldsymbol{u}}_t]_i\right)^4\right] \mathbb{E}\left[\left(\boldsymbol{v}_j^\intercal [\tilde{\boldsymbol{u}}_t]_j\right)^2 \left(\boldsymbol{v}_k^\intercal [\tilde{\boldsymbol{u}}_t]_k\right)^2\right]}, \\
&\leq (3\vee\gamma)^L \Bigg(\sum_{i=0}^{L} \|\boldsymbol{v}_i\|_{\ell_2}^4 + \sum_{i=0}^{L}\sum_{\substack{j=0\\ j\neq i}}^{L} \|\boldsymbol{v}_i\|_{\ell_2}^2 \|\boldsymbol{v}_j\|_{\ell_2}^2 \\
&+ 2\sum_{i=0}^{L}\sum_{\substack{j=0\\ j\neq i}}^{L}\sum_{\substack{k=0\\ k\neq i\\ k\neq j}}^{L} \|\boldsymbol{v}_i\|_{\ell_2}^2 \|\boldsymbol{v}_j\|_{\ell_2} \|\boldsymbol{v}_k\|_{\ell_2} \Bigg), \\
&\leq (3\vee\gamma)^L \Bigg(\sum_{i=0}^{L} \|\boldsymbol{v}_i\|_{\ell_2}^4 + \sum_{i=0}^{L}\sum_{\substack{j=0\\ j\neq i}}^{L} \|\boldsymbol{v}_i\|_{\ell_2}^2 \|\boldsymbol{v}_j\|_{\ell_2}^2 + \sum_{\substack{j=0\\ j\neq k}}^{L}\sum_{\substack{k=0\\ k\neq j}}^{L} \left(\|\boldsymbol{v}_j\|_{\ell_2}^2 + \|\boldsymbol{v}_k\|_{\ell_2}^2\right) \Bigg), \\
&\leq (3\vee\gamma)^L (1+2L) \leq L(3\vee\gamma)^{L+1}. \qquad \text{(B.29)}
\end{aligned}
$$

This completes the proof of Lemma 11. ■

B.3. Proof of Theorem 5

We are now ready to state the proof of our main result on persistence of excitation stated by Theorem 5. The proof follows a similar line of reasoning as that of Proposition 6.5 of Sattar et al. (2024). For the sake of completeness, we present the entire (modified) proof here.

Proof To begin, recall the definition of $\tilde{\boldsymbol{u}}_t$ from (B.1), and let $\tilde{\boldsymbol{U}}$ has rows $\{\tilde{\boldsymbol{u}}_t^\intercal\}_{t=L}^{T}$. From the proof of Lemma 11, we have

$$
\mathbb{E}[\tilde{\boldsymbol{U}}^\intercal \tilde{\boldsymbol{U}}] = \sum_{t=L}^{T} \mathbb{E}[\tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{u}}_t^\intercal] = \sum_{t=L}^{T} \boldsymbol{I}_{d_{\tilde{u}}} = (T-L+1)\boldsymbol{I}_{d_{\tilde{u}}}. \qquad \text{(B.30)}
$$

Next, letting $\boldsymbol{v} \in \mathcal{S}^{d_{\tilde{u}}-1}$, we consider the quantity $\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\boldsymbol{v} = \sum_{t=L}^{T}(\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{u}}_t)^2$, which can be viewed as a summation of the random process $\{(\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{u}}_t)^2\}_{t=L}^{T}$. In the following, we will derive a one-sided concentration bound for this random process.

• **Step 1) Blocking:** We begin by using blocking technique to get independent samples as follows,

$$\sum_{t=L}^{T}(\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{u}}_t)^2 = \sum_{k=0}^{L}\sum_{\tau=1}^{(T-L+1)/(L+1)}(\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{u}}_{\tau(L+1)+k-1})^2, \tag{B.31}$$

where we make the simplifying assumption that $T-L+1$ can be divided by $L+1$. (*Note that, this goes without loss of generality, and we assume it for the sake of clarity. It can be easily avoided by noting that*

$$\sum_{t=L}^{T}\tilde{\boldsymbol{u}}_t\tilde{\boldsymbol{u}}_t^\mathsf{T} \succeq \sum_{t=L}^{(L+1)\lfloor\frac{T-L+1}{L+1}\rfloor+L-1}\tilde{\boldsymbol{u}}_t\tilde{\boldsymbol{u}}_t^\mathsf{T},$$

where $\lfloor\cdot\rfloor$ denotes the floor operator. We can then analyze everything with $T_0 = (L+1)\lfloor\frac{T-L+1}{L+1}\rfloor + L - 1$, and note that $T-L \le T_0 \le T$.)

• **Step 2) Bernstein's inequality for non-negative random variables:** From Lemma 11, we have $\mathbb{E}[(\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{u}}_t)^2] = 1$ and $\mathbb{E}[(\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{u}}_t)^4] \le L(3\vee\gamma)^{L+1}$ for any $\boldsymbol{v} \in \mathcal{S}^{d_{\tilde{u}}}$. Hence, we can use one-sided Bernstein's inequality for non-negative random variables (Wainwright, 2019) to obtain a lower bound on the smallest eigenvalue of $\sum_{t=L}^{T}\tilde{\boldsymbol{u}}_t\tilde{\boldsymbol{u}}_t^\mathsf{T}$. Specifically, we have,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{\tau=1}^{(T-L+1)/(L+1)}\left((\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{u}}_{\tau(L+1)+k-1})^2 - \mathbb{E}[(\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{u}}_{\tau(L+1)+k-1})^2]\right) \le -\frac{T-L+1}{(L+1)}\eta\right) \\
&\qquad\qquad \le \exp\left(-\frac{(T-L+1)\eta^2}{2(L+1)\Xi}\right), \\
\implies &\mathbb{P}\left(\sum_{\tau=1}^{(T-L+1)/(L+1)}(\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{u}}_{\tau(L+1)+k-1})^2 \le \frac{T-L+1}{L+1}(1-\eta)\right) \le \exp\left(-\frac{(T-L+1)\eta^2}{2(L+1)\Xi}\right),
\end{aligned}$$

where we set $\Xi := L(3\vee\gamma)^{L+1}$ for notational convenience. Union bounding over $L+1$ such events, we get the following,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{k=0}^{L}\sum_{\tau=1}^{(T-L+1)/(L+1)}(\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{u}}_{\tau(L+1)+k-1})^2 \le (T-L+1)(1-\eta)\right) \\
&\qquad\qquad \le (L+1)\exp\left(-\frac{(T-L+1)\eta^2}{2(L+1)\Xi}\right).
\end{aligned} \tag{B.32}$$

This can be alternately represented by letting

$$\begin{aligned}(L+1)\exp\left(-\frac{(T-L+1)\eta^2}{2(L+1)\Xi}\right) &= \delta,\\ \iff \frac{(T-L+1)\eta^2}{2(L+1)\Xi} &= \log((L+1)/\delta),\\ \impliedby \eta &= \sqrt{\frac{L+1}{T-L+1}2\Xi\log((L+1)/\delta)}. \end{aligned} \tag{B.33}$$

Hence, we have

$$\mathbb{P}\left(\sum_{t=L}^{T}(\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{u}}_t)^2 \le (T-L+1) - \sqrt{2\Xi(L+1)(T-L+1)\log((L+1)/\delta)}\right) \le \delta. \tag{B.34}$$

• **Step 3) Covering with $\delta/(8d_{\tilde{u}})$-net:** Next, we use a covering argument as follows: Let $\mathcal{N}_\epsilon := \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{|\mathcal{N}_\epsilon|}\} \subset \mathcal{S}^{d_{\tilde{u}}-1}$ be the ϵ-net of $\mathcal{S}^{d_{\tilde{u}}-1}$ such that for any $\boldsymbol{v} \in \mathcal{S}^{d_{\tilde{u}}-1}$, there exists $\boldsymbol{v}_i \in \mathcal{N}_\epsilon$ such that $\|\boldsymbol{v}-\boldsymbol{v}_i\|_{\ell_2} \le \epsilon$. From Lemma 5.2 of Vershynin (2010), we have $|\mathcal{N}_\epsilon| \le (1+2/\epsilon)^{d_{\tilde{u}}}$.

Let us choose $\boldsymbol{v} \in \mathcal{S}^{d_{\tilde{u}}-1}$ for which $\lambda_{\min}(\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}) = \boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\boldsymbol{v}$, and choose $\boldsymbol{v}_i \in \mathcal{N}_\epsilon$ which approximates $\boldsymbol{v}$ as $\|\boldsymbol{v}-\boldsymbol{v}_i\|_{\ell_2} \le \epsilon$. By triangle inequality, we have

$$\begin{aligned}|\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\boldsymbol{v} - \boldsymbol{v}_i^\mathsf{T}\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\boldsymbol{v}_i| &= |\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}(\boldsymbol{v}-\boldsymbol{v}_i) + (\boldsymbol{v}-\boldsymbol{v}_i)^\mathsf{T}\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\boldsymbol{v}_i|,\\ &\le \|\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\|\|\boldsymbol{v}\|_{\ell_2}\|\boldsymbol{v}-\boldsymbol{v}_i\|_{\ell_2} + \|\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\|\|\boldsymbol{v}_i\|_{\ell_2}\|\boldsymbol{v}-\boldsymbol{v}_i\|_{\ell_2},\\ &\le 2\epsilon\|\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\|,\\ \implies \boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\boldsymbol{v} &\ge \boldsymbol{v}_i^\mathsf{T}\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\boldsymbol{v}_i - 2\epsilon\|\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\|,\\ \implies \lambda_{\min}(\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}) &\ge \inf_{\boldsymbol{v}_i\in\mathcal{N}_\epsilon} \boldsymbol{v}_i^\mathsf{T}\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\boldsymbol{v}_i - 2\epsilon\|\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\|. \end{aligned} \tag{B.35}$$

Hence, in order to lower bound $\lambda_{\min}(\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}})$, we also need an upper bound on $\|\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\|$. This can be done as follows: First, we have

$$\begin{aligned}\mathbb{E}[\|\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\|] = \mathbb{E}[\lambda_{\max}(\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}})] &\le \mathbb{E}[\mathbf{tr}(\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}})] = \mathbf{tr}\big(\mathbb{E}[\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}]\big),\\ &= (T-L+1)\mathbf{tr}\big(\boldsymbol{I}_{d_{\tilde{u}}}\big) = d_{\tilde{u}}(T-L+1). \end{aligned} \tag{B.36}$$

Hence, using Markov's inequality, we get

$$\mathbb{P}\left(\|\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\| > \frac{d_{\tilde{u}}(T-L+1)}{\delta}\right) \le \frac{\mathbb{E}[\|\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\|]}{d_{\tilde{u}}(T-L+1)}\delta \le \delta. \tag{B.37}$$

Let $\mathcal{E} := \{\|\tilde{\boldsymbol{U}}^\intercal\tilde{\boldsymbol{U}}\| \leq \frac{2d_{\tilde{u}}(T-L+1)}{\delta}\}$ denote the event that $\|\tilde{\boldsymbol{U}}^\intercal\tilde{\boldsymbol{U}}\|$ is bounded by the specified threshold. Then, it is straightforward to see that $\mathbb{P}(\mathcal{E}) \geq 1-\delta/2$. This further implies,

$$\begin{aligned}
&\mathbb{P}\left(\lambda_{\min}(\tilde{\boldsymbol{U}}^\intercal\tilde{\boldsymbol{U}}) < (T-L+1)(1/2-\eta)\right) \\
&\qquad \leq \mathbb{P}\left(\{\lambda_{\min}(\tilde{\boldsymbol{U}}^\intercal\tilde{\boldsymbol{U}}) < (T-L+1)(1/2-\eta)\}\bigcap\mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c), \\
&\qquad \leq \mathbb{P}\left(\{\inf_{\boldsymbol{v}_i\in\mathcal{N}_\epsilon} \boldsymbol{v}_i^\intercal\tilde{\boldsymbol{U}}^\intercal\tilde{\boldsymbol{U}}\boldsymbol{v}_i - 2\epsilon\|\tilde{\boldsymbol{U}}^\intercal\tilde{\boldsymbol{U}}\| < (T-L+1)(1/2-\eta)\}\bigcap\mathcal{E}\right) + \delta/2, \\
&\qquad \leq \mathbb{P}\left(\inf_{\boldsymbol{v}_i\in\mathcal{N}_\epsilon} \boldsymbol{v}_i^\intercal\tilde{\boldsymbol{U}}^\intercal\tilde{\boldsymbol{U}}\boldsymbol{v}_i < (T-L+1)(1/2-\eta) + 4\epsilon\frac{d_{\tilde{u}}(T-L+1)}{\delta}\right) + \delta/2, \\
&\qquad = \mathbb{P}\left(\inf_{\boldsymbol{v}_i\in\mathcal{N}_\epsilon} \boldsymbol{v}_i^\intercal\tilde{\boldsymbol{U}}^\intercal\tilde{\boldsymbol{U}}\boldsymbol{v}_i < (T-L+1)(1/2-\eta+\frac{4\epsilon d_{\tilde{u}}}{\delta})\right) + \delta/2, \\
&\qquad = \mathbb{P}\left(\inf_{\boldsymbol{v}_i\in\mathcal{N}_\epsilon} \boldsymbol{v}_i^\intercal\tilde{\boldsymbol{U}}^\intercal\tilde{\boldsymbol{U}}\boldsymbol{v}_i < (T-L+1)(1-\eta)\right) + \delta/2, \qquad \text{(B.38)}
\end{aligned}$$

where we obtained the last inequality by choosing $\epsilon = \frac{\delta}{8d_{\tilde{u}}}$. Using (B.32) with union bounding over all the elements in $\mathcal{N}_\epsilon$, we obtain,

$$\mathbb{P}\left(\inf_{\boldsymbol{v}_i\in\mathcal{N}_\epsilon} \boldsymbol{v}_i^\intercal\tilde{\boldsymbol{U}}^\intercal\tilde{\boldsymbol{U}}\boldsymbol{v}_i < (T-L+1)(1-\eta)\right) \leq |\mathcal{N}_\epsilon|(L+1)\exp\left(-\frac{(T-L+1)\eta^2}{2(L+1)\Xi}\right). \quad \text{(B.39)}$$

From Lemma 5.2 of Vershynin (2010), we have $|\mathcal{N}_\epsilon| \leq (1+2/\epsilon)^{d_{\tilde{u}}}$. Hence, we have

$$\mathbb{P}\left(\inf_{\boldsymbol{v}_i\in\mathcal{N}_\epsilon} \boldsymbol{v}_i^\intercal\tilde{\boldsymbol{U}}^\intercal\tilde{\boldsymbol{U}}\boldsymbol{v}_i < (T-L+1)(1-\eta)\right) \leq (L+1)(1+\frac{16d_{\tilde{u}}}{\delta})^{d_{\tilde{u}}}\exp\left(-\frac{(T-L+1)\eta^2}{2(L+1)\Xi}\right).$$

This can be alternately represented by letting,

$$\begin{aligned}
&(L+1)(1+\frac{16d_{\tilde{u}}}{\delta})^{d_{\tilde{u}}}\exp\left(-\frac{(T-L+1)\eta^2}{2(L+1)\Xi}\right) = \delta/2, \\
\iff &\exp\left(-\frac{(T-L+1)\eta^2}{2(L+1)\Xi}\right) = \delta/(2(L+1))(1+\frac{16d_{\tilde{u}}}{\delta})^{-d_{\tilde{u}}}, \\
\iff &\frac{(T-L+1)\eta^2}{2(L+1)\Xi} = \log(2(L+1)/\delta) + d_{\tilde{u}}\log\big(1+\frac{16d_{\tilde{u}}}{\delta}\big), \\
\impliedby &\eta = \sqrt{\frac{2(L+1)\Xi}{(T-L+1)}\big(\log\big(\frac{2(L+1)}{\delta}\big) + d_{\tilde{u}}\log\big(1+\frac{16d_{\tilde{u}}}{\delta}\big)\big)}. \qquad \text{(B.40)}
\end{aligned}$$

Plugging this back into (B.38), we have

$$\begin{aligned}
&\mathbb{P}\Bigg(\lambda_{\min}(\tilde{\boldsymbol{U}}^\intercal\tilde{\boldsymbol{U}}) < (T-L+1)/2 \\
&\qquad - \sqrt{2(L+1)(T-L+1)\Xi\big(\log\big(\frac{2(L+1)}{\delta}\big) + d_{\tilde{u}}\log\big(1+\frac{16d_{\tilde{u}}}{\delta}\big)\big)}\Bigg) \leq \delta.
\end{aligned}$$

Finally, choosing the trajectory length via

$$\begin{aligned}
&(T-L+1)/4 \geq \sqrt{2(L+1)(T-L+1)\Xi\big(\log\big(\frac{2(L+1)}{\delta}\big)+d_{\tilde{u}}\log\big(1+\frac{16d_{\tilde{u}}}{\delta}\big)\big)},\\
\iff\ &(T-L+1)/16 \geq 2(L+1)\Xi\big(\log\big(\frac{2(L+1)}{\delta}\big)+d_{\tilde{u}}\log\big(1+\frac{16d_{\tilde{u}}}{\delta}\big)\big),\\
\impliedby\ &T-L+1 \geq 32(L+1)\Xi\big(\log\big(\frac{2(L+1)}{\delta}\big)+d_{\tilde{u}}\log\big(1+\frac{16d_{\tilde{u}}}{\delta}\big)\big),
\end{aligned} \tag{B.41}$$

we obtain the following persistence of excitation result,

$$\mathbb{P}\left(\lambda_{\min}(\tilde{\boldsymbol{U}}^{\mathsf{T}}\tilde{\boldsymbol{U}}) \geq (T-L)/4\right) \geq 1-\delta. \tag{B.42}$$

This completes the proof of Theorem 5. ■

Appendix C. Proof of Theorem 2

In this appendix, we provide the proofs of Proposition 6 and Proposition 7. These propositions together with persistence of excitation presented in Theorem 5 lead immediately to Theorem 2 which concerns the main guarantee on the estimation of the *Markov-like* parameters. We present its proof in this appendix too.

C.1. Proof of Theorem 2

The proof is an immediate consequence of Theorem 5, Proposition 6, and Proposition 7. Indeed, define the events

$$\mathcal{E}_1 \triangleq \left\{ \left\| \sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \boldsymbol{\epsilon}_t^{\top} \right\|_{\text{op}} \leq \frac{(p+1)^{L+1}\|\boldsymbol{C}\|_{\text{op}}(4p\|\boldsymbol{B}\|_{\text{op}}^2+\sigma^2)\kappa^2\rho^{L-1}}{1-\rho}\sqrt{2(T-L)\log\left(\frac{2\cdot 9^{d_{\tilde{\boldsymbol{u}}}+m}}{\delta}\right)} \right\}$$

$$\mathcal{E}_2 \triangleq \left\{ \left\| \sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t (\boldsymbol{F}\tilde{\boldsymbol{w}}_t)^{\mathsf{T}} \right\|_{\text{op}} \leq \frac{2(1+\kappa\|\boldsymbol{C}\|_{\text{op}})\sigma(p+1)^{L+1}}{1-\rho}\sqrt{2L(T-L+2)\log\left(\frac{2L9^{d_{\tilde{\boldsymbol{u}}}+m+nL}}{\delta}\right)} \right\}$$

$$\mathcal{E}_3 \triangleq \left\{ \lambda_{\min}\left(\sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{u}}_t^{\mathsf{T}}\right) \geq (T-L+1)/4. \right\}$$

Recalling the estimation error decomposition (4.1), we see that when the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ holds, then the upper bound on the estimation error presented in Theorem 2 follows. Now, we remark that by union bound, we have

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) = 1 - \mathbb{P}(\mathcal{E}_1^c \cup \mathcal{E}_2^c \cup \mathcal{E}_3^c) \geq 1 - \mathbb{P}(\mathcal{E}_1^c) - \mathbb{P}(\mathcal{E}_2^c) - \mathbb{P}(\mathcal{E}_3^c).$$

Thus, using Theorem 5, Proposition 6, and Proposition 7, we obtain $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) \leq 3\delta$, provided the condition (4.3) in Theorem 5 holds. This concludes the proof.

C.2. Proof of Proposition 6

First, using the variational form of the operator norm, we see that:

$$\left\|\sum_{t=L}^{T}\tilde{\boldsymbol{u}}_t\boldsymbol{\epsilon}_t^\top\right\|_{\text{op}} = \sup_{\theta\in\mathcal{S}^{d_{\tilde{\boldsymbol{u}}}-1},\lambda\in\mathcal{S}^{m-1}}\sum_{t=L}^{T}(\theta^\top\tilde{\boldsymbol{u}}_t)(\lambda^\top\boldsymbol{\epsilon}_t).$$

We use an $1/4$-net argument to bound the supremum. Let $\mathcal{M}$ (resp. $\mathcal{N}$) be $1/4$-nets with minimal cardinality of the sphere $\mathcal{S}^{d_{\tilde{\boldsymbol{u}}}-1}$ (resp. $\mathcal{S}^{m-1}$). Thus, using Lemma 14 we obtain: for all $r>0$:

$$\mathbb{P}\left(\left\|\sum_{t=L}^{T}\tilde{\boldsymbol{u}}_t\boldsymbol{\epsilon}_t^\top\right\|_{\text{op}} > 2r\right) \le 9^{d_{\tilde{\boldsymbol{u}}}+m}\max_{\theta\in\mathcal{M},\lambda\in\mathcal{N}}\mathbb{P}\left(\sum_{t=L}^{T}(\theta^\top\tilde{\boldsymbol{u}}_t)(\lambda^\top\boldsymbol{\epsilon}_t) > r\right). \tag{C.1}$$

It remains to bound $\sum_{t=L}^{T}(\theta^\top\tilde{\boldsymbol{u}}_t)(\lambda^\top\boldsymbol{\epsilon}_t)$ with high probability uniformly over the unit spheres. Let $\theta\in\mathcal{S}^{d_{\tilde{\boldsymbol{u}}}-1}$, and $\lambda\in\mathcal{S}^{m-1}$. We have:

$$\begin{aligned}
\sum_{t=L}^{T}(\theta^\top\tilde{\boldsymbol{u}}_t)(\lambda^\top\boldsymbol{\epsilon}_t) &= \sum_{t=L}^{T}(\theta^\top\tilde{\boldsymbol{u}}_t)(\lambda^\top\boldsymbol{C})\left(\prod_{\ell=1}^{L-1}(\boldsymbol{u}_{t-\ell}\circ\boldsymbol{A})\right)\boldsymbol{x}_{t-L}\\
&= \sum_{t=0}^{T-L}(\theta^\top\tilde{\boldsymbol{u}}_{t+L})(\lambda^\top\boldsymbol{C})\left(\prod_{\ell=1}^{L-1}(\boldsymbol{u}_{t+\ell}\circ\boldsymbol{A})\right)\boldsymbol{x}_t\\
&\overset{(a)}{=} \sum_{t=1}^{T-L}(\theta^\top\tilde{\boldsymbol{u}}_{t+L})(\lambda^\top\boldsymbol{C})\left(\prod_{\ell=1}^{L-1}(\boldsymbol{u}_{t+\ell}\circ\boldsymbol{A})\right)\sum_{s=0}^{t-1}\left(\prod_{k=s+1}^{t-1}(\boldsymbol{u}_k\circ\boldsymbol{A})\right)(\boldsymbol{B}\boldsymbol{u}_s+\boldsymbol{w}_s)\\
&= \sum_{t=1}^{T-L}\sum_{s=0}^{t-1}(\theta^\top\tilde{\boldsymbol{u}}_{t+L})(\lambda^\top\boldsymbol{C})\left(\prod_{\ell=1}^{L-1}(\boldsymbol{u}_{t+\ell}\circ\boldsymbol{A})\right)\left(\prod_{k=s+1}^{t-1}(\boldsymbol{u}_k\circ\boldsymbol{A})\right)(\boldsymbol{B}\boldsymbol{u}_s+\boldsymbol{w}_s),\\
&\overset{(b)}{=} \sum_{t=0}^{T-L-1}\sum_{s=0}^{t}\underbrace{(\theta^\top\tilde{\boldsymbol{u}}_{t+1+L})(\lambda^\top\boldsymbol{C})\left(\prod_{\ell=1}^{L-1}(\boldsymbol{u}_{t+1+\ell}\circ\boldsymbol{A})\right)}_{:=\boldsymbol{M}(\boldsymbol{u}_{t+2},\ldots,\boldsymbol{u}_{t+1+L})}\underbrace{\left(\prod_{k=s+1}^{t}(\boldsymbol{u}_k\circ\boldsymbol{A})\right)}_{:=\boldsymbol{N}(\boldsymbol{u}_{s+1},\ldots,\boldsymbol{u}_t)}(\boldsymbol{B}\boldsymbol{u}_s+\boldsymbol{w}_s)
\end{aligned}$$

where we used in (a), the dynamics (2.1) to express $\boldsymbol{x}_t$ in terms of $(\boldsymbol{u}_s,\boldsymbol{w}_s)_{s<t}$ (e.g., see (2.2)). We also introduce in (b) the quantities $\boldsymbol{M}$ and $\boldsymbol{N}$ which depends on inputs. Next, we perform next a change of indices to obtain:

$$\begin{aligned}
\sum_{t=L}^{T}(\theta^\top\tilde{\boldsymbol{u}}_t)(\lambda^\top\boldsymbol{\epsilon}_t) &= \sum_{s=0}^{T-L-1}\underbrace{\left(\sum_{t=s}^{T-L-1}\boldsymbol{M}(\boldsymbol{u}_{t+2},\ldots,\boldsymbol{u}_{t+L+1})\boldsymbol{N}(\boldsymbol{u}_{s+1},\ldots,\boldsymbol{u}_t)\right)}_{:=f_s(\boldsymbol{u}_{s+1},\ldots,\boldsymbol{u}_T)}(\boldsymbol{B}\boldsymbol{u}_s+\boldsymbol{w}_s)\\
&\overset{(c)}{=} \sum_{s=0}^{T-L-1}f_s(\boldsymbol{u}_{s+1},\ldots,\boldsymbol{u}_T)(\boldsymbol{B}\boldsymbol{u}_s+\boldsymbol{w}_s),
\end{aligned} \tag{C.2}$$

where we introduce in (c) the functions $f_s(\cdot)$. Now we clearly see that $\sum_{t=L}^{T}(\theta^\top\tilde{\boldsymbol{u}}_t)(\lambda^\top\boldsymbol{\epsilon}_t)$, written in the form (C.2) is a martingale difference. Before we use this fact, let us note that using Lemma

9 we can show that the terms involving the functions $f_s(\cdot)$ are well bounded. More specifically, we have

$$\begin{aligned}
\|f(\boldsymbol{u}_{s+1},\dots,\boldsymbol{u}_T)\|_{\ell_2} &= \left\| \sum_{t=s}^{T-L-1} \boldsymbol{M}(\boldsymbol{u}_{t+1},\dots,\boldsymbol{u}_{t+L+1})\boldsymbol{N}(\boldsymbol{u}_{s+1},\dots,\boldsymbol{u}_t) \right\|_{\ell_2} \\
&\le \sum_{t=s}^{T-L-1} \|\tilde{\boldsymbol{u}}_{t+L+1}\|_{\ell_2} \|\boldsymbol{C}\|_{\mathrm{op}} \left\| \prod_{\ell=1}^{L-1} (\boldsymbol{u}_{t+1+\ell} \circ \boldsymbol{A}) \right\|_{\mathrm{op}} \left\| \prod_{k=s+1}^{t} (\boldsymbol{u}_k \circ \boldsymbol{A}) \right\|_{\mathrm{op}} \\
&\le \sup_{t\le T} \|\tilde{\boldsymbol{u}}_t\|_{\ell_2} \|\boldsymbol{C}\|_{\mathrm{op}} \kappa^2 \rho^{L-1} \sum_{t=0}^{T-L-s-1} \rho^t \\
&\le \frac{(p+1)^{L+1}\|\boldsymbol{C}\|_{\mathrm{op}}\kappa^2\rho^{L-1}}{1-\rho}
\end{aligned}$$

where we bound $\|\tilde{\boldsymbol{u}}_t\|_{\ell_2} \le \sqrt{p^L((p+1)^{L+1}-2}$. Now, we also remark that $\boldsymbol{B}\boldsymbol{u}_s + \boldsymbol{w}_s$ is zero-mean and $(4p\|\boldsymbol{B}\|_{\mathrm{op}}^2 + \sigma^2)$-subgaussian. Hence, using Freedman's inequality (see Lemma 12), we obtain: for all $\delta \in (0,1)$, the event

$$\begin{aligned}
&\left| \sum_{s=0}^{T-L-1} f(\boldsymbol{u}_{s+1},\dots,\boldsymbol{u}_T)(\boldsymbol{B}\boldsymbol{u}_s + \boldsymbol{w}_s) \right| \\
&\qquad > \frac{(p+1)^{L+1}\|\boldsymbol{C}\|_{\mathrm{op}}(4p\|\boldsymbol{B}\|_{\mathrm{op}}^2+\sigma^2)\kappa^2\rho^{L-1}\sqrt{2(T-L)\log(2\cdot 9^{d_{\tilde{\boldsymbol{u}}}+m}/\delta)}}{1-\rho}
\end{aligned}$$

with probability at most $\delta/9^{d_{\tilde{\boldsymbol{u}}}+m}$. The conclusion follows immediately by recalling the inequality (C.1).

C.3. Proof of Proposition 7

Using the submultiplicativity of the norm and Lemma 9, we have:

$$\left\| \sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t (\boldsymbol{F}\tilde{\boldsymbol{w}}_t)^\intercal \right\|_{\mathrm{op}} \le \left\| \sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{w}}_t^\intercal \right\|_{\mathrm{op}} \|\boldsymbol{F}\|_{\mathrm{op}} \le \left(1 + \frac{\kappa\|\boldsymbol{C}\|_{\mathrm{op}}}{1-\rho}\right) \left\| \sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{w}}_t^\intercal \right\|_{\mathrm{op}} \tag{C.3}$$

We will now focus on bounding $\left\|\sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{w}}_t^\intercal\right\|_{\mathrm{op}}$ with high probability. Recalling the variational form of the operator norm, we using a $1/4$-net argument. Let $\mathcal{M}$ (resp. $\mathcal{N}$) be an $1/4$-net of $\mathcal{S}^{d_{\tilde{\boldsymbol{u}}}-1}$ (resp. $\mathcal{S}^{m+nL-1}$) with minimal cardinality. By Lemma 14, we obtain have: for all $u > 0$

$$\mathbb{P}\left(\left\| \sum_{t=L}^{T} \tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{w}}_t^\intercal \right\|_{\mathrm{op}} > 2u \right) \le 9^{d_{\tilde{\boldsymbol{u}}}+m+nl} \max_{\theta\in\mathcal{M},\lambda\in\mathcal{N}} \mathbb{P}\left(\sum_{t=L}^{T} (\theta^\intercal \tilde{\boldsymbol{u}}_t)(\lambda^\intercal \tilde{\boldsymbol{w}}_t) > u \right) \tag{C.4}$$

We observe that the sequences $\{\tilde{\boldsymbol{w}}_t\}_{t\ge L}$ and $\{\tilde{\boldsymbol{u}}_t\}_{t\ge L}$ are independent, but these are not sequences of independent random vectors. We will use a blocking trick to handle this. Let $\theta \in \mathcal{S}^{d_{\tilde{\boldsymbol{u}}}-1}, \lambda \in \mathcal{S}^{m+nL-1}$. We have:

$$\left| \sum_{t=L}^{T} (\theta^\intercal \tilde{\boldsymbol{u}}_t)(\lambda^\intercal \tilde{\boldsymbol{w}}_t) \right| = \left| \sum_{\ell=1}^{L} \sum_{s\in\mathcal{T}_\ell} (\theta^\intercal \tilde{\boldsymbol{u}}_s)(\lambda^\intercal \tilde{\boldsymbol{w}}_s) \right| \le \sum_{\ell=1}^{L} \left| \sum_{s\in\mathcal{T}_\ell} (\theta^\intercal \tilde{\boldsymbol{u}}_s)(\lambda^\intercal \tilde{\boldsymbol{w}}_s) \right| \tag{C.5}$$

where for all ℓ, $\mathcal{T}_\ell$ correspond to the indices that satisfy $(t \mod L) = \ell - 1$ and we note that $\lfloor (T-L+1)/L \rfloor \leq |\mathcal{T}_\ell| \leq \lceil (T-L+1)/L \rceil$. We note that $\{\tilde{\boldsymbol{w}}_s\}_{s\in\mathcal{T}_\ell}$ are independent for all $\ell \in [L]$. We use Lemma 12 to bound with high probability the sum $\sum_{s\in\mathcal{T}_\ell}(\theta^\mathsf{T}\tilde{\boldsymbol{u}}_s)(\lambda^\mathsf{T}\tilde{\boldsymbol{w}}_s)$ for each partition $\mathcal{T}_\ell$, then combine these bounds with a union bound over the L partitions to conclude: for all $\delta \in (0,1)$,

$$\mathbb{P}\left(\left|\sum_{t=L}^{T}(\theta^\mathsf{T}\tilde{\boldsymbol{u}}_t)(\lambda^\mathsf{T}\tilde{\boldsymbol{w}}_t)\right| > (p+1)^{L+1}\sigma\sqrt{2L(T-L+2)\log\left(\frac{2L9^{d_{\tilde{\boldsymbol{u}}}+m+nL}}{\delta}\right)}\right) \leq \frac{\delta}{9^{d_{\tilde{\boldsymbol{u}}}+m+nL}}. \tag{C.6}$$

Combining this bound with the inequality (C.4) yields

$$\mathbb{P}\left(\left\|\sum_{t=L}^{T}\tilde{\boldsymbol{u}}_t\tilde{\boldsymbol{w}}_t^\mathsf{T}\right\|_{\text{op}} \leq 2\sigma(p+1)^{L+1}\sqrt{2L(T-L+2)\log\left(\frac{2L9^{d_{\tilde{\boldsymbol{u}}}+m+nL}}{\delta}\right)}\right) \geq 1-\delta. \tag{C.7}$$

Recalling the inequality (C.3), we conclude that: for all $\delta \in (0,1)$, the event

$$\left\|\sum_{t=L}^{T}\tilde{\boldsymbol{u}}_t(\boldsymbol{F}\tilde{\boldsymbol{w}}_t)^\mathsf{T}\right\|_{\text{op}} \leq 2\left(1+\frac{\kappa\|\boldsymbol{C}\|_{\text{op}}}{1-\rho}\right)\sigma(p+1)^{L+1}\sqrt{2L(T-L+2)\log\left(\frac{2L9^{d_{\tilde{\boldsymbol{u}}}+m+nL}}{\delta}\right)} \tag{C.8}$$

holds with probability $1-\delta$.

Appendix D. Miscellaneous Lemmas & Concentration Tools

In this Appendix we present a set of lemmas and concentration inequalities that we persistently make use of in our proofs. First, we provide a version of Freedman's inequality which can also be deduced from Azuma-Hoeffding's inequality.

Lemma 12 *Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration. Let $(\boldsymbol{\eta}_t)_{t\geq 1}$ is a sequence of zero-mean, σ^2-subgaussian random vectors taking values in $\mathbb{R}^d$, such that $\boldsymbol{\eta}_t$ is $\mathcal{F}_t$-measurable for all $t \geq 1$. Let $(\boldsymbol{x}_t)$ be a sequence of random vectors taking values in $\mathbb{R}^d$ such that for all $t \geq 1$, $\boldsymbol{x}_t$ is $\mathcal{F}_{t-1}$-measurable and $\|\boldsymbol{x}_t\|_{\ell_2} \leq K$ almost surely for some $K > 0$. Then for all $\delta \in (0,1)$, $T \geq 1$,*

$$\mathbb{P}\left(\left|\sum_{t=1}^{T}\boldsymbol{x}_t^\mathsf{T}\boldsymbol{\eta}_t\right| \leq \sigma K\sqrt{2T\log(2/\delta)}\right) \geq 1-\delta$$

Next, we present an immediate generalization of Hoeffding's lemma for bounded random vectors.

Lemma 13 *Let $\boldsymbol{u}$ be a p-dimensional random vector sampled from $\text{Unif}(\sqrt{p}\cdot\mathcal{S}^{p-1})$. Then, it holds that $\boldsymbol{u}$ is zero-mean and $4p$-subgaussian, i.e., $\mathbb{E}[\exp(\theta^\top\boldsymbol{u})] \leq \exp(2\|\theta\|_{\ell_2}^2 p)$, for all $\theta \in \mathbb{R}^p$.*

Finally, we formalize the trick of a net arguments in the following lemma which is a classical argument that can be found in Vershynin (2010):

Lemma 14 (ϵ-net argument) *Let $\boldsymbol{W}$ be a $m \times n$ random matrix and $\varepsilon \in (0, 1/2)$. Let $\mathcal{M}$ (resp. $\mathcal{N}$) be an ε-net of $(\mathcal{S}^{m-1}, \|\cdot\|_{\ell_2})$ (resp. $(\mathcal{S}^{n-1}, \|\cdot\|_{\ell_2})$). For all $\rho > 0$, it holds that*

$$\mathbb{P}\left(\|\boldsymbol{W}\|_{\text{op}} > \frac{\rho}{1-2\varepsilon}\right) \leq \left(1+\frac{2}{\varepsilon}\right)^{n+m} \max_{\boldsymbol{x}\in\mathcal{M},\boldsymbol{y}\in\mathcal{N}} \mathbb{P}(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{y} > \rho).$$

We omit the proofs of these Lemmas as they are standard results within the literature.