Efficient learning of hidden state LTI state space models of unknown order

BOUALEM DJEHICHE[1,*] and OTHMANE MAZHAR[1,†]

[1]*Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden.*
E-mail: *boualem@kth.se; †othmane@kth.se*

The aim of this paper is to address two related estimation problems arising in the setup of hidden state linear time invariant (LTI) state space systems when the dimension of the hidden state is unknown. Namely, the estimation of any finite number of the system's Markov parameters and the estimation of a minimal realization for the system, both from the partial observation of a single trajectory. For both problems, we provide statistical guarantees in the form of various estimation error upper bounds, rank recovery conditions, and sample complexity estimates.

Specifically, we first show that the low rank solution of the Hankel penalized least square estimator satisfies an estimation error in S_p-norms for $p \in [1,2]$ that captures the effect of the system order better than the existing operator norm upper bound for the simple least square. We then provide a stability analysis for an estimation procedure based on a variant of the Ho-Kalman algorithm that improves both the dependence on the dimension and the least singular value of the Hankel matrix of the Markov parameters. Finally, we propose an estimation algorithm for the minimal realization that uses both the Hankel penalized least square estimator and the Ho-Kalman based estimation procedure and guarantees with high probability that we recover the correct order of the system and satisfies a new fast rate in the S_2-norm with a polynomial reduction in the dependence on the dimension and other parameters of the problem.

1. Introduction

Many control design and synthesis techniques rely on an accurate description of the system as a state space model. Deriving such accurate description is an important problem in system identification with far-reaching applications in many areas including time series analysis [24], economics [42], robotics [22] and aeronautics [3], to name a few. While in some cases it is possible to derive such models due to the simple structure of the underlying phenomena involved in the dynamical evolution [28, 12], there are many instances where such an approach is intractable because the system is too complex or some of the involved phenomena are not well understood. In those cases, one adopts a so-called black-box approach and learns the system from the input/output data generated in an experimental setting with little to no assumptions on the real system.

In recent years, there has been an increased interest in providing non-asymptotic statistical guarantees in the form of estimation error upper bounds and sample complexity estimates for data-driven estimation procedures for state space models [53, 10, 52, 18, 40, 38]. While there is a plethora of estimation procedures for learning state space models most of which are well understood in the asymptomatic regime derived e.g. in [28, 11, 4, 49], modern estimation setups present additional challenges that are not taken into account in an asymptomatic study. For example, in estimation based solution to the linear reinforcement learning problem [27] one would aim to obtain a highly accurate estimate of the dynamical system as fast as possible before moving to the control part or alternate between estimation and control in such a way to strike a trade-off between the exploration and the exploitation

arXiv:2202.01625v1 [math.ST] 3 Feb 2022

part. In these cases, asymptotic results are of limited use; the more accurate measure of estimation performance would be through the non-asymptotic estimation error and the sample complexity. Beyond the reinforcement learning use, non-asymptotic statistical guarantees are also used in conjunction with robust control techniques [6, 46], in control design using the Markov parameters [41, 19], and as theoretical guidelines for practical heuristics such as bootstrapping to establish high-probability confidence intervals [13]. Several authors provided such estimates for observed LTI state space models, and the results are essentially optimal in the sense that upper and lower bounds for both the estimation error and sample complexity match up to logarithmic terms and unknown multiplicative constants [40]. However, the situation is not as clear-cut for Hidden state LTI state space models. In a realistic setup, these systems present the additional challenge of not knowing the dimension of the state. In the absence of precise estimation lower bounds, the estimation upper bounds presented in the literature so far do not capture well the effect of the system dimension and dynamic on the non-asymptotic estimation error and sample complexity incurred by the studied estimation algorithms.

1.1. Problem statement and preliminaries

The present study aims to provide computationally effective estimation procedures satisfying sharp non-asymptotic estimation error bounds and sample complexity estimates when used for learning the parameters of hidden state LTI state space models of unknown order from the partial observation of a single trajectory of the system. We then have to deal with two ambiguities:

- The hidden dimension of the parameters is not well defined from an input/output standpoint.
- The hidden state LTI state space model parameters' are defined only up to a similarity transform.

In this section, we make these two claims more precise, introduce some necessary preliminaries from realization theory and provide a precise statement for the aim of the study.

To this end, we consider the following formulation for LTI state space models.

$$\begin{cases} x_{i+1} = A_0 x_i + B_0 u_i + w_i, \\ \quad y_i = C_0 x_i + v_i, \end{cases} \tag{1.1}$$

where $x_i \in \mathbb{R}^{d_0}$ is the hidden state variable of unknown dimension d_0 and $u_i \in \mathbb{R}^r$ are iid multivariate normal sequences $\mathcal{N}(0, \sigma_u^2 I_r)$. They excite the system to generate the output sequence $y_i \in \mathbb{R}^p$. w_i respectively v_i are the iid multivariate normal state noise sequence $\mathcal{N}(0, \sigma_w^2 I_{d_0})$ and the output noise sequence $\mathcal{N}(0, \sigma_v^2 I_p)$, respectively. These centred Gaussian random vectors can be replaced by centred subGausian centred random vectors of appropriate ψ_2 norm upper bounds and all the results remain the same. The linear dynamic is then described by the parameters (A_0, B_0, C_0) with $A_0 \in \mathcal{M}_{d_0 \times d_0}(\mathbb{R})$, $B_0 \in \mathcal{M}_{d_0 \times r}(\mathbb{R})$, and $C_0 \in \mathcal{M}_{p \times d_0}(\mathbb{R})$. If we eliminate the state variable x_i, we obtain the so-called input/output (I/O) description of the system:

$$\begin{aligned} y_n &= \sum_{i=0}^{n-1} C_0 A_0^{n-1-i} B_0 u_i + \sum_{i=10}^{n-1} C_0 A_0^{n-1-i} w_i + v_n \\ &= \sum_{l=t-2T+1}^{t-1} C_0 A_0^{t-1-l} B u_l + \sum_{l=0}^{t-2T} C_0 A_0^{t-1-l} B_0 u_l + \sum_{l=0}^{t-1} C_0 A_0^{t-1-l} w_l + v_t \\ &:= g_0 X_t + \bar{g}_0 \bar{X}_t + h W_t + v_t, \end{aligned} \tag{1.2}$$

where, $\bar{N} := N - 2T + 1$,

$$
\begin{aligned}
X_l &:= [u_{l-1}^*, u_{l-2}^*, \ldots, u_{l-2T+1}^*]^* \in \mathbb{R}^{(2T-1)r} \\
\bar{X}_l &:= [u_{l-2T}^*, u_{l-2T-1}^*, \ldots, u_0^*, 0, \ldots, 0]^* \in \mathbb{R}^{\bar{N}r} \\
W_l &:= [w_{l-1}^*, w_{l-2}^*, \ldots, w_0^*, 0, \ldots, 0]^* \in \mathbb{R}^{Nd_0},
\end{aligned}
\tag{1.3}
$$

and

$$
\begin{aligned}
g_0 &:= [C_0 B_0, C_0 A_0 B_0, \ldots, C_0 A_0^{2T-2} B_0] \in \mathcal{M}_{p \times (2T-1)r}(\mathbb{R}) \\
\bar{g}_0 &:= [C_0 A_0^{2T-1} B_0, C_0 A_0^{2T} B_0, \ldots, C_0 A_0^{N-1} B_0] \in \mathcal{M}_{p \times \bar{N}r}(\mathbb{R}) \\
h &:= [C_0, C_0 A_0, \ldots, C_0 A_0^{N-1}] \in \mathcal{M}_{p \times Nd_0}(\mathbb{R}).
\end{aligned}
\tag{1.4}
$$

Since we want to estimate the parameter g_0, the part $\bar{g}_0 \bar{X}_t + h W_t + v_t$ will play the role of a disturbance that we will refer to as the noise part. To obtain a successful estimator of g_0, this part should not grow arbitrarily large. To this end, we impose an assumption on the growth of the powers of the estimated matrix in terms of its spectral radius that we recall in the next definition.

Definition 1.1. *The spectral radius of a matrix A is defined as $\rho(A) := \min_{\lambda \in SP(A)} |\lambda|$, where $SP(A)$ is the spectrum (the set of all eigenvalues) of A.*

Assumption 1.1. *We assume that the system (1.1) is stable in the sense that the spectral radius $\rho(A_0)$ is strictly less than 1.*

By applying the Jordan decomposition to the matrix A, we readily see that there exists a positive constant ψ_A depending only on A such that for all $k \in \mathbb{N}$ we have

$$
|A^k|_{S_\infty} \leqslant \psi_A \rho(A^k).
\tag{1.5}
$$

The system identification problem in this setup would be to estimate the parameters (A_0, B_0, C_0) given that we observe a single realization of $(X_i, y_i)_{i=2T}^N$ while we do not have access to the sequence $(x_i)_i$ and in particular we do not know the dimension d_0. From (1.2) we notice that the sequence $(y_i)_i$ is related to $(u_i)_i$ in a causal fashion only through the factors $(CA^i B)_i$, commonly referred to as the Markov parameters associated with the system (A_0, B_0, C_0).

We note that for any similarity transform S, the parameters (A_0, B_0, C_0) and their transforms $(SA_0 S^{-1}, SB_0, C_0 S^{-1})$ give the same values for the Markov parameter vector g_0. This makes the problem of learning the parameters (A_0, B_0, C_0) from observations up to time N of a single trajectory $(u_i, y_i)_i$ not well defined. One can only learn a representative of the equivalence class defined by the parameters $(SA_0 S^{-1}, SB_0, C_0 S^{-1})$ for all similarity transforms S. This also makes the dimension d_0 not well defined as one can always replace the system (1.1) by the larger system

$$
\begin{cases}
\begin{bmatrix} x_{i+1} \\ z_{i+1} \end{bmatrix} = \begin{bmatrix} A_0 \\ \end{bmatrix} \begin{bmatrix} x_i \\ z_i \end{bmatrix} + \begin{bmatrix} \\ B_0 \end{bmatrix} u_i + w_i, \\
y_i = \begin{bmatrix} C_0 \end{bmatrix} \begin{bmatrix} x_i \\ z_i \end{bmatrix} + v_i.
\end{cases}
$$

Nonetheless, from the Realization Theory of linear systems, we know a representative of the equivalence class for (A_0, B_0, C_0) of minimal dimension exists.

Definition 1.2. *We refer to a representative of the equivalence class of minimal dimension as a minimal realization, and we refer to the dimension of the minimal realization as the system order.*

The system order coincides with the McMillan degree [30, 31] defined as

$$\delta(A_0, B_0, C_0) := \mathrm{rank}(Hg_0^*),$$

where $H \colon \mathcal{M}_{p \times (2T-1)r}(\mathbb{R}) \longrightarrow \mathcal{M}_{Tp \times Tr}(\mathbb{R})$ is the T order Hankel operator on g_0 defined, for any $g = [g_1, \ldots, g_{2T-1}] \in \mathcal{M}_{p \times (2T-1)r}(\mathbb{R})$, by

$$Hg^* = \begin{bmatrix} g_1 & g_2 & g_3 & & g_T \\ g_2 & g_3 & & & g_{T+1} \\ g_3 & & & & g_{T+2} \\ & & & & \\ g_T & g_{T+1} & g_{T+2} & & g_{2T-1} \end{bmatrix},$$

and where T is greater than the dimension of the matrix A_1 of some particular realization (A_1, B_1, C_1) which is not necessarily minimal. For more on Hankel operators, their properties and the role of the McMillan degree as a complexity measure for LTI models, we refer to [7, 34, 30, 31]. We also note that the McMillan degree is independent of the realization since it is defined with respect to the Markov parameters.

Hence, we deal with the ambiguity in the definition of the system dimension by adopting the following

Assumption 1.2. *We assume that the realization (A_0, B_0, C_0) is minimal in the sense that $d_0 = \delta(A_0, B_0, C_0)$.*

This assumption can be made without loss of generality since any hidden state LTI state space system has a minimal realization. We also define the T-order controllabilty matrix $\mathcal{C}$ and the T-order observability matrix $\mathcal{O}$ for the realization (A_0, B_0, C_0) by

$$\mathcal{C} = \begin{bmatrix} B_0 & A_0 B_0 & \cdots & A_0^{T-1} B_0 \end{bmatrix} \quad \text{and} \quad \mathcal{O} = \begin{bmatrix} C_0 \\ C_0 A_0 \\ \vdots \\ C_0 A_0^{T-1} \end{bmatrix} \tag{1.6}$$

and recall that the system (A_0, B_0, C_0) is a minimal realization if and only if $\mathrm{rank}(\mathcal{C}) = \mathrm{rank}(\mathcal{O}) = d_0$ and in that case, for all $T \geqslant d_0$, we have $\mathrm{rank}(Hg_0^*) = d_0$. When this occurs we say that the pair (A, B) is controllable and the pair (C, A) is observable. Therefore, we impose the following

Assumption 1.3. *We assume that $T \geqslant d_0$.*

As mentioned above, this assumption is necessary and sufficient for the Hankel matrix of the Markov parameters to capture the dimension of the minimal realization. Moreover, this assumption is even more relevant when we estimate a Hidden state LTI state space model of unknown order while given a pessimistic upper bound T on d_0, which is the high dimension estimation set up in this context.

Since a minimal realization is again defined up to a similarity transform, we introduce here the concept of a *balanced minimal realization* which is a particular minimal realization that one can compute,

given the Markov parameters. The procedure of deriving a minimal realization from the description of the Markov parameters is known as the Ho-Kalman algorithm. Suppose that we are given the Markov parameter vector g_0 and noting that $Hg_0^* = \mathcal{OC}$, the Ho-Kalman algorithm starts from the SVD decomposition of the Hankel matrix of the Markov parameters and constructs the particular minimal realization given in the following

Definition 1.3. *Assume that* $\mathrm{rank}(Hg_0^*) = d_0$ *and* $T \geqslant d_0 + 1$. *Then a minimal balanced minimal realization* $(\bar{A}, \bar{B}, \bar{C})$ *is defined through the following Ho-Kalman algorithm:*

- *Define the SVD decomposition of the Hankel matrix of the Markov parameters by*

$$Hg_0^* = U_0 \Sigma_0 V_0^*.$$

- *Take*

$$\bar{O} = U_0 \Sigma_0^{1/2} \ \ and \ \ \bar{C} = \Sigma_0^{1/2} V_0^*.$$

- *Define the minimal balanced realization as*

$$\bar{A} = \left(\bar{O}_{1:r(T-1),1:d_0} \right)^\dagger \bar{O}_{r+1:rT,1:d_0}, \ \bar{B} = \bar{O}_{1:d_0,1:r}, \ and \ \bar{C} = \bar{O}_{1:p,1:d_0}.$$

In this definition, $M^\dagger$ refers to the left pseudo inverse of a full column rank matrix M and $M_{a:b,c:d}$ refers to the sub-matrix of M composed of rows a to b and columns c to d.

We note that there are multiple variants of the Ho-Kalman algorithm described in the previous definition, but the main idea for the construction is the same for all of them.

We note as well that $\bar{O}_{1:r(T-1),1:d_0}$ has full rank since it is equal to the observability matrix up to a similarity transform. Hence, we have

$$s_{d_0}(\bar{O}_{1:r(T-1),1:d_0}) > 0,$$

where $s_k(M)$ refers to the kth singular value of the matrix M and the singular values are taken in a decreasing order. Starting from the observation that $Hg_0^* = \mathcal{OC}$ one can check that there exists a similarity transform S such that $\bar{A} = SA_0S^{-1}$, $\bar{B} = SB_0$, and $\bar{C} = C_0S^{-1}$. Thus $(\bar{A}, \bar{B}, \bar{C})$ is indeed a minimal realization and belongs to the equivalence class of $\mathcal{M} = (A_0, B_0, C_0)$.

Thus, our aim is two folds:

- to provide an estimation procedure that given the data generated from the observation of a single trajectory $(X_i, y_i)_{i=2T}^N$ up to time N outputs estimates of the Markov parameters $\hat{g} = [\hat{g}_1, \cdots, \hat{g}_{2T-2}]$ such that the following loss function

$$\mathcal{L}_p^H(\hat{g}, g_0) = |H\hat{g}^* - Hg_0^*|_{S_p}$$

is small with high probability.
- to provide an estimate $\hat{\mathcal{M}} = (\hat{A}, \hat{B}, \hat{C})$ for the system $\mathcal{M}_0 = (A_0, B_0, C_0)$ of the same dimension d_0 as some minimal realization such that the following loss function with respect to the minimal balanced realisation $\bar{\mathcal{M}} = (\bar{A}, \bar{B}, \bar{C})$

$$\mathcal{L}_p^{\mathcal{M}}(\hat{\mathcal{M}}, \bar{\mathcal{M}}) := \inf_{S: \det(S) \neq 0} |S^{-1}\hat{A}S - \bar{A}|_{S_p} + |S^{-1}\hat{\mathcal{B}} - \bar{B}|_{S_p} + |\hat{\mathcal{C}}S - \bar{C}|_{S_p}$$

is also small with high probability. Up to a multiplicative constant, this is the same as saying that the loss function $\mathcal{L}(\hat{\mathcal{M}}, \mathcal{M}_0)$ is small.

Here, $|M|_{S_p} = (\text{tr}(M^*M)^{p/2})^{1/p}$, $1 \leqslant p < \infty$, $|M|_{S_\infty} = \max_{|x|_2 \leqslant 1} |Mx|_2$ and $|x|_2^2 = \sum_{i=1}^n |x_i|^2$.

Frequently used notation

Before we review the literature related to our problem and the main contributions of the present paper, we recall some frequently used notations.

We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the underlying probability space and by $\mathbb{E}$ the corresponding expectation operator.

Here and throughout the paper c denote a positive constant whose exact value is not important for the derivation and might change from one step to another. $x \lesssim y$ is a shorthand for 'there exists a positive constant c such that $x \leqslant cy$', and $x \simeq y$ means that $x \lesssim y$ and $y \lesssim x$. The minimum (maximum) of two real numbers x and y is denoted as $\min(x, y) = x \wedge y$ ($\max(x, y) = x \vee y$).

Whenever possible, our results are provided with explicit constants to give an idea of their order. The numerical values of these constants are useful in practice but are not optimal and can be improved.

1.2. Related literature

A common estimation approach in the Hidden state LTI state space setup is the two-step approach commonly referred to as a subspace method [49, 50, 21]. In the first step of this approach, one learns the Markov parameters with a good enough precision, since unlike the true parameters (A_0, B_0, C_0) the Markov parameters are well defined, and in the second step, one uses the learned Markov parameters to provide an estimate close to a representative of the equivalence class of the true parameters. The first step is usually carried out with a regression-type estimator and the second step is carried out via some variant of the celebrated Ho-Kalman algorithm, which relies on identifying a possible realization from the output of an SVD decomposition. The popularity of the subspace approach is because it is computationally tractable, unlike the maximum likelihood approach or the predictive error method, which both results in a non-convex optimization problem [28]. Several results appeared recently in the machine learning community studying non-asymptotic properties of variants of the subspace method under various assumptions on the estimation setup. The literature on the estimation of the parameters of LTI state space models is very rich; early works in the topic date back to the nineties where [14, 35, 51, 25] provided asymptotic results. A complete overview of this vast literature falls beyond the format and the scope of the present paper. Therefore, we only mention and discuss here some recent results [33, 39, 9, 43] that provide non-asymptotic statistical guarantees for variants of the subspace method and thus are close in spirit to our work.

Remark 1.1. *We took some freedom to omit the contribution of lower order terms for some of these results. Instead, we refer to the original work for the exact statement.*

For ease of notation, we set

$$
y := \begin{bmatrix} y_{2T}^* \\ y_{2T+1}^* \\ \vdots \\ y_N^* \end{bmatrix}, \ X := \begin{bmatrix} X_{2T}^* \\ X_{2T+1}^* \\ \vdots \\ X_N^* \end{bmatrix}, \ \bar{X} := \begin{bmatrix} \bar{X}_{2T}^* \\ \bar{X}_{2T+1}^* \\ \vdots \\ \bar{X}_N^* \end{bmatrix},
$$

$$
W := \begin{bmatrix} W_{2T}^* \\ W_{2T+1}^* \\ \vdots \\ W_N^* \end{bmatrix}, \ \varepsilon = \begin{bmatrix} v_{2T}^* \\ v_{2T+1}^* \\ \vdots \\ v_N^* \end{bmatrix}.
$$

(1.7)

Thus, we can write the input/output representation (1.2) for the y vector more succinctly as follows:

$$
y = X g_0^* + \bar{X} \bar{g}_0^* + W h^* + \varepsilon.
$$

- *The context of known dimension d_0.* While this context is simpler, results in this setup are informative about what can be expected if d_0 is unknown. Oymak and Ozay [33] consider a subspace approach in this context and show that the least square estimator defined as $\hat{g}_{\mathrm{ls}} := (X^\dagger y)^*$ can effectively learn the first T Hankel parameters in the sense that with high probability and for values of N such that

$$
N \geqslant N_0 = c T q_0 \log^2(T q_0) \log^2(T N) \quad \text{with} \quad q_0 = r + p + d_0,
$$

it holds that [33, Theorem 3.1]

$$
|\hat{g}_{\mathrm{ls}} - g_0|_{S_\infty} \leqslant (\sigma_v + \sigma_e + |h|_{\mathcal{H}_\infty} \log(TN)) \sqrt{\frac{c T q_0 \log^2(T q_0)}{N}},
$$

where σ_e accounts for the variance of x_{t-T}. Under the same condition it was shown that a version of the Ho-Kalman algorithm successfully learns, up to a similarity transform, a representation of the true parameter on the same event for $N \geqslant \frac{N_0}{s_{d_0}^2(H g_0)}$ with the guarantee of [33, Theorem 5.3]

$$
\mathcal{L}_2^{\mathcal{M}}(\hat{\mathcal{M}}, \bar{\mathcal{M}}) \lesssim \frac{(\sigma_v + \sigma_e + |h|_{\mathcal{H}_\infty} \log(TN)) |H g_0|_{S_\infty} q_0 \sqrt{T \log^2(T q_0)}}{s_{d_0}^2(H g_0) \sqrt{N}}.
$$

- *The context of unknown dimension d_0.* Sarkar et al. [39] adopt a model selection approach to choose a realization of order $\hat{d}$ that is good enough. Their learning algorithm proceed in three stages:

  1. Hankel matrix estimation: the algorithm starts by solving, for all $d \in \mathcal{D}(N) = \{T \mid N \geqslant T r^2 \log^3(N r/\delta)\}$, a least square problem to get an estimated Hankel matrix of $2T + 1$ parameter $\hat{H}_T$.

  2. Order selection: the algorithm chooses a model of size $\hat{d}$ according to the rule

$$
\hat{d} = \tilde{d} \wedge \log(N/\delta),
$$

with $\alpha(h) = \sqrt{\frac{hp+h^2r+\log(N/\delta)}{N}}$, where

$$\tilde{d} := \inf\{d \in \mathcal{D}(N); \ |\hat{H}_d - \hat{H}_l|_{S_\infty} \leqslant c(\alpha(d) + \alpha(l)) \ \forall l \geqslant d, \ l \in \mathcal{D}(N)\}.$$

3. Parameter estimation: the algorithm uses a variant of the Ho-Kalman algorithm to get a realization $(\hat{A}, \hat{B}, \hat{C})$ of dimension $\hat{d}$ from $\hat{H}_{\hat{d}}$.

Their results [39, Theorem 5.1 and Proposition 5.1] imply that, for all $T \in \mathcal{D}(N)$ and

$$N \geqslant c(r^2 T \log^2(T) \log^2(r/\delta) + T \log^2(T)),$$

the estimation step outputs an estimate for the Hankel matrix of the parameters satisfying

$$|\hat{\mathcal{H}}_T - Hg_0^*|_{S_\infty} \leqslant c\sqrt{\frac{pT^2 + rT + T\log(1/\delta)}{N}}. \tag{1.8}$$

They also show [39, Theorem 5.3] that a variant of the Ho-Kalman applied to the selected model $\hat{d}$ successfully learns the best $\hat{d}$ approximation to the minimal realization after observing $N \geqslant N_*$ sample and we have with probability at least $1 - \delta$ the following

$$\mathcal{L}_\infty^{\mathcal{M}}(\hat{\mathcal{M}}, \bar{\mathcal{M}}_{\hat{d}}) \ \leqslant \ \varepsilon\Gamma(\hat{H}, \varepsilon) \ + \ \frac{\varepsilon\hat{d}}{\sqrt{s_{\hat{d}}(\hat{H})}} \ \wedge \ \sqrt{\varepsilon\hat{d}} \ + \ \frac{\varepsilon}{\sqrt{s_{\hat{d}}(\hat{H})}} \ \wedge \ \sqrt{\varepsilon}, \tag{1.9}$$

$$\text{with} \quad \varepsilon = c\sqrt{\frac{r\hat{d} + p\hat{d}^2 + \hat{d}\log(N/\delta)}{N}}, \ N_* < \infty, \ \text{and} \ \Gamma(\hat{H}, \varepsilon) < \infty.$$

Here $\bar{\mathcal{M}}_{\hat{d}}$ is the model resulting from the use of the Ho-Kalman algorithm on the truncated SVD of Hg_0^* to the first $\hat{d}$ singular values.

• *The context of unknown dimension d_0 while allowing the partial observation of N paths* $\left((u_i^j, y_i^j)_{i=1}^{2T-1}\right)_{j=1}^N$ *of length $2T - 1$ without process noise:* This setup is different from ours as it allows multiple independent realizations and assumes that $w_i = 0$ which is the main source of difficulty in our setup, nonetheless the approaches used in this context in [9, 43] are closer to our approach as they relay on restricted or penalized least square estimators to estimate the Markov parameters. Indeed, [9] analyzes the performance of the following estimator $\hat{g}_{\text{rls}}$ in the problem of robust recovery of a superposition of distinct complex exponential functions from few random Gaussian projections.

$$\hat{g}_{\text{rls}} \in \arg\min_g \ |Hg^*|_{S_1}$$
$$\text{s.t.} \ |XKg^* - y|_2 \leqslant \delta,$$

with $K = \text{diag}(\sqrt{1}, \cdots, \sqrt{T}, \sqrt{T-1}, \cdots, \sqrt{1})$. This problem is indeed equivalent to the estimation problem of single input single output LTI state space models from multiple trajectories with weighting K for the input without process noise. They show that, with probability at least $1 - e^{-cN}$ for $N \gtrsim d_0 \log^2(T) + \varepsilon$, the following holds

$$|Kg_0^* - K\hat{g}_{\text{rls}}^*|_2 \leqslant c\frac{\delta}{\varepsilon}.$$

Inspired by this result, Sun *et al.* [43] use the following nuclear norm penalized least square estimator $\hat{g}_{\text{pls}}$ for the multiple input single output case

$$\hat{g}_{\text{pls}} \in \arg\min_{g} \quad |XK^{-1}g^* - y|_2 + \lambda|HK^{-1}g^*|_{S_1}, \tag{1.10}$$

and show that with high probability, for a choice of $\lambda = \frac{T\sigma_z}{\sigma_u}\sqrt{\frac{r}{N}}\log(T)$,

$$\mathcal{L}_\infty^{\mathcal{H}}(\hat{g}_{\text{pls}}, g_0) \lesssim \begin{cases} \frac{\sigma_z}{\sigma_u}\sqrt{\frac{rT^2}{N}}\log(T), & N \geqslant d_0^2 \wedge T, \\ \frac{\sigma_z}{\sigma_u}\sqrt{\frac{d_0 rT^2}{N}}\log(T), & d_0 \leqslant N \leqslant d_0^2 \wedge T. \end{cases} \tag{1.11}$$

1.3. Main contributions

As mentioned in Section 1.1, we consider the parametric estimation task in the setup of LTI state space model (1.1) from the observation of a single trajectory when neither the state is observed nor the system's order is known. In what follows, we present our contributions.

Remark 1.2. *While some of the results presented above are provided in terms of the norm $|\cdot|_{S_\infty}$, ours are derived for the norm $|\cdot|_{S_p}$ with $p \in [1, 2]$. Whenever it is the case, we use the norm domination relation relation $|\cdot|_{S_p} \leqslant r^{1/p}|\cdot|_{S_\infty}$ for the sake of comparison, where r is the appropriate dimension.*
From the related literature we see that up to logarithmic terms all the upper bounds are of the form $P(d_0, T, s_{d_0}^{-1}(\bar{\mathcal{O}}^+), s_{d_0}^{-1}(Hg_0^*), N^{-1})$ where P is some polynomial function of these variable. All throughout, we compare different results in the asymptotic regime where $N \to \infty$, $d \to \infty$, $T \to \infty$, $s_{d_0}(Hg_0^*) \to 0$, and $s_{d_0}(\bar{\mathcal{O}}^+) \to 0$ while the upper bound still converge to 0.

In Section 2.3, we provide non-asymptotic estimation error upper-bounds and sample complexity for the Hankel penalized regression estimator given by any particular solution of the convex optimization problem

$$\hat{g} \in \arg\min_{g \in \mathcal{M}_{p \times (2T-1)r}(\mathbb{R})} \frac{1}{N}|y - Xg^*|_{S_2}^2 + \lambda|Hg^*|_{S_1}. \tag{1.12}$$

For this estimator we provide in Theorem 2.3 estimation guarantees and sample complexity for different dimension sensitive loss functions. In particular, we show with probability at least $1 - \delta$ and for $\bar{N}$ large enough, that the p-loss function $\mathcal{L}_p^H(\hat{g}, g_0)$, for $p \in (0, 1)$, satisfies

$$\mathcal{L}_p^H(\hat{g}, g_0) \lesssim d_0^{1/p}T\sqrt{\frac{p + r + \log(T/\delta)}{\bar{N}}}.$$

The available upper bounds for these loss functions are derived in the case $p = \infty$ for the least square estimator when the dimension d_0 is known; see for instance [39, Theorem 5.1]. Since the solution of the least square estimator is not low rank, the estimate (1.8) implies

$$|\hat{\mathcal{H}}_d - Hg_0^*|_{S_p} \lesssim T^{1/p}\sqrt{\frac{pT^2 + rT + T\log(1/\delta)}{\bar{N}}}$$

with a suboptimal factor $T^{1/p}$ which is the best one would hope for from a non-low rank estimation procedure. In the same fashion, our result is an improvement of the result (1.11) with unknown order,

while observing multiple trajectories. We finally show in Proposition 2.1 how we can recover the system order efficiently using a truncated SVD procedure if a lower bound on $s_{d_0}(\bar{O}_{1:r(T-1),1:d_0})$ is known. We refer to the discussion after Theorem 2.3 for more on this issue.

In Section 2.4 we provide a robustness analysis in the norm $|\cdot|_{S_2}$ of an estimation procedure for the parameters based on a variant of the Ho-Kalman algorithm. In Theorem 2.4 we show that under some stability conditions, it is possible to recover the parameters if we reduce the error term $|\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_2}$ since

$$
\mathcal{L}_2^{\mathcal{M}}(\hat{\mathcal{M}}, \bar{\mathcal{M}}) \lesssim \frac{|\bar{A}|_{S_\infty} |\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_2}}{s_{d_0}(\bar{O}^+) s_{d_0}^{1/2}(Hg_0^*)}.
$$

This is an improvement of [45, Theorem 4] which gives

$$
\mathcal{L}_2^{\mathcal{M}}(\hat{\mathcal{M}}, \bar{\mathcal{M}}) \lesssim \frac{d_0 |Hg_0^*|_{S_\infty}^{1/2} |\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_\infty}}{s_{d_0}^2(\bar{O}^+) s_{d_0}^{1/2}(Hg_0^*)}.
$$

and of [43, Theorem 5.2] which yields

$$
\mathcal{L}_2^{\mathcal{M}}(\hat{\mathcal{M}}, \bar{\mathcal{M}}) \lesssim \frac{d_0^{1/2} |Hg_0^*|_{S_\infty} |\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_\infty}}{s_{d_0}^2(Hg_0^*)}.
$$

We refer to the discussions after Theorem 2.4 for more on this.

In Section 2.5 we provide non-asymptotic estimation guarantees for Algorithm 1 introduced in Section 2.1. The algorithm yields the estimates $\hat{\mathcal{M}} = (\hat{A}, \hat{B}, \hat{C})$ for the minimal balanced realisation $\bar{\mathcal{M}} = (\bar{A}, \bar{B}, \bar{C})$ with probability $1 - \delta$ of the same dimension as a minimal realization d_0 after observing

$$
\bar{N} \geqslant c d_0 T N_0 \vee T_0 \log \frac{1}{\delta} \vee \frac{\phi^2 d_0 T_0^2}{\xi^2} \left(N_0 \vee \log \frac{1}{\delta} \right) \vee \frac{\phi d_0^{1/2} T_0 \log(T_0)}{\xi} \left(N_0 \vee \log \frac{1}{\delta} \right)
$$

such that

$$
\mathcal{L}_2^{\mathcal{M}}(\hat{\mathcal{M}}, \bar{\mathcal{M}}) \lesssim \frac{\phi |\bar{A}|_{S_\infty} d_0^{3/2}}{s_{d_0}^2(\bar{O}^+)} \left(\sqrt{\frac{N_0}{N}} \vee \frac{\log(d_0) N_0}{N} \vee \sqrt{\frac{\log \frac{1}{\delta}}{N}} \vee \frac{\log(d_0) \log \frac{1}{\delta}}{N} \right).
$$

As mentioned in Section 1.2, to the best of our knowledge, the only available result in our setup is (1.9) obtained by Sarkar $et\ al.$ [39, Theorem 5.3]. If we disregard the spectral properties of $H\hat{g}^*$ and multiply by $\hat{d}^{1/2}$ to account for the difference of norms, the dominant term in that expression is

$$
\sqrt{\frac{r\hat{d}^4 + p\hat{d}^5 + \hat{d}^4 \log(N/\delta)}{s_{\hat{d}}(\hat{H}) N}}.
$$

Hence, our result improves the bound (1.9) since it provides an upper bound in terms of the actual dimension d_0 and not the estimated dimension $\hat{d}$. Also, it reduces the dependence on the dimension by as much as d_0.

Algorithm 1 Truncated Hankel Penalized Regression with Ho-Kalman State-Space Realization.

1: **Compute**: $\hat{A}$, $\hat{B}$, $\hat{C}$
2: **Input**: $(X_i, y_i)_{i=2T_0}^N, \lambda_0, \xi$
3: **Procedure**: Hankel penalized regression
4: $\qquad \hat{g} \in \arg \min\limits_{g \in \mathcal{M}_{p \times (2T_0-1)r}(\mathbb{R})} \frac{1}{N}|y - Xg^*|_{S_2}^2 + \lambda_0|Hg^*|_{S_1}$
5: **Return**: $H\hat{g}^*$
6: $y \leftarrow 1$
7: **Compute**: $\hat{A}$, $\hat{B}$, $\hat{C}$
8: **Input**: $(X_i, y_i)_{i=2T_0}^N, \lambda_0, \xi$
9: **Procedure**: Hankel penalized regression
10: $\qquad \hat{g} \in \arg \min\limits_{g \in \mathcal{M}_{p \times (2T_0-1)r}(\mathbb{R})} \frac{1}{N}|y - Xg^*|_{S_2}^2 + \lambda_0|Hg^*|_{S_1}$
11: **Return**: $H\hat{g}^*$
12: **Procedure**: Order estimation
13: $\qquad \check{d}_\xi = \sum\limits_{i=1}^{\mathrm{rank}(H\hat{g}^*)} \mathbf{1}\{s_i(H\hat{g}^*) \geqslant 2\xi\}$
14: **Return**: $\check{d}_\xi$
15: **Procedure**: Reduced order Hankel penalized regression
16: $\qquad \hat{g}_\xi \in \arg \min\limits_{g \in \mathcal{M}_{p \times (2\check{d}_\xi+1)r}(\mathbb{R})} \frac{1}{N}|y - Xg^*|_{S_2}^2 + \lambda_1|Hg^*|_{S_1}$
17: **Return**: $H\hat{g}_\xi^*$
18: **Procedure**: Reduced order Ho-Kalman Algorithm
19: $\qquad \hat{U}_\xi \hat{\Sigma}_\xi \hat{V}_\xi^* = \mathrm{SVD}(\hat{\mathcal{H}}_{\check{d}_\xi}) \quad \text{and} \quad \hat{\mathcal{H}}_{\check{d}_\xi} = \sum\limits_{i=1}^{\check{d}_\xi+1} s_i(H\hat{g}_\xi^*)\mathbf{1}\{s_i(H\hat{g}_\xi^*) \geqslant 2\xi\}\hat{u}_i\hat{v}_i^*$
20: $\qquad \hat{O} = \hat{U}_\xi \hat{\Sigma}_\xi^{1/2} \text{ and } \hat{C} = \hat{\Sigma}_\xi^{1/2}V_\xi^*$
21: $\qquad \hat{A} = \left(\hat{O}_{1:r(T-1),1:d_\xi}\right)^\dagger \hat{O}_{r+1:rT,1:d_\xi}, \hat{B} = \hat{O}_{1:d_\xi,1:r}, \text{ and } \hat{C} = \hat{O}_{1:p,1:d_\xi}.$
22: **Return**: $\hat{A}$, $\hat{B}$, $\hat{C}$

---

2. Main results

2.1. Algorithmic details

We start first by describing our Learning Algorithm 1. The algorithm starts with the 'Hankel penalized regression' step. In this step, it computes a penalized least square estimate for the first $2T_0 - 1$ Markov parameters by solving the optimization problem

$$\hat{g} \in \arg \min_{g \in \mathcal{M}(p \times (2T_0-1)r)(\mathbb{R})} \frac{1}{N}|y - Xg^*|_{S_2}^2 + \lambda_0|Hg^*|_{S_1}. \tag{2.1}$$

The least square part is the fitting term that ensures fidelity to the data; the penalty part ensures the simplicity of the chosen model. As described in the introduction, a good measure of the model's complexity for hidden state LTI state space models is the rank of the corresponding Hankel operator since it agrees with the system order as given in Definition 1.2. The penalty term using the nuclear norm of the Hankel operator ensures that the solution to the optimization problem described in (2.1) has a low Hankel rank as it is the convex relaxation of the rank function. Thus, we would expect that via a good choice of the free parameter λ we obtain a good enough, yet simple, model in the sense that it is close to Hg_0^* with $|H\hat{g}^* - Hg_0^*|_{S_2}$ being small and has $\mathrm{rank}(H\hat{g}^*)$ small enough.

The second step of our learning algorithm is 'Order estimation' in which we compute an estimate $\check{d}_\varepsilon$ of the true dimension. If in the last step we have made the error $|H\hat{g}^* - Hg_0^*|_{S_2}$ small compared

to $s_{d_0}(Hg_0^*)$, the smallest singular value of Hg_0^*, on the one hand, for $1 \leqslant i \leqslant d_0$ the singular values $s_i(H\hat{g}^*)$ will be close to the singular values $s_i(Hg_0^*)$ and on the other hand the singular values $s_i(H\hat{g}^*)$ for $i > d_0$ will be small so that they are well separated from the others. Thus, via an appropriate choice of ξ, we can successfully ensure that $\check{d}_\xi = d_0$ with high probability.

The third step, 'Reduced order Hankel penalized regression', is similar to the first step except that it aims at estimating the first $2\check{d}_\xi + 1$ Markov parameters instead of the $2T - 1$ parameters. For this, it solves the following Hankel penalized regression problem:

$$\hat{g}_\xi \in \arg \min_{g \in \mathcal{M}_{p \times (2\check{d}_\xi+1)r}(\mathbb{R})} \frac{1}{N}|y - Xg^*|_{S_2}^2 + \lambda_1 |Hg^*|_{S_1}. \tag{2.2}$$

This is done to obtain a more accurate estimate on these first $2\check{d}_\xi + 1$ Markov parameters, since they are the only parameters needed for our estimation procedure based on the Ho-Kalman algorithm to get an accurate estimate for the minimal balanced minimal realization $(\bar{A}, \bar{B}, \bar{C})$.

The last part of our learning algorithm, 'Reduced order Ho-Kalman Algorithm', uses the previous estimate $\hat{g}_\xi$. It starts with a truncated SVD of the Hankel matrix of the estimated $2\check{d}_\varepsilon + 1$ Markov parameters from the previous part. Doing this ensures that the rank of the truncation result $\hat{\mathcal{H}}_{d_\varepsilon}$ is the same as the order of the minimal realization with high probability and that the truncation is close enough to the true model in the sense that $|\hat{\mathcal{H}}_{\check{d}_\varepsilon} - Hg_0^*|_{S_2}$ is small. This means that the eigenvalues and eigenvectors of both $\hat{\mathcal{H}}_{\check{d}_\varepsilon}$ and Hg_0^* are close to each other. Then, it proceeds with getting estimates $\hat{A}$, $\hat{B}$, $\hat{C}$ using the Ho-Kalman Algorithm steps described in Definition 1.3.

Crucial to the success of our algorithm 1 are the three choices of the free parameters: T_0 and λ_0 in the step 'Hankel penalized regression' and ξ in the step 'Order estimation'. In section 2.5 we provide values for these free parameters to ensure the high probability of success of Algorithm 1 in both the order recovery task and the estimation task. We also discuss how reasonable the assumption of knowing each of these parameters is and provide the value for the internal variable λ_1 necessary for the 'Reduced order Hankel penalized regression' step. In the next section, we provide the probabilistic estimates instrumental to those choices.

2.2. Probabilistic results

We first show that the covariance matrix of the covariates

$$X_l := [u_{l-1}^*, u_{l-2}^*, \ldots, u_{l-2T+1}^*]^* \in \mathbb{R}^{(2T-1)r}$$

generated along the path of the input of the LTI state space model (1.1) concentrate around the identity matrix. This is the main content of the following theorem, which is an extension of [17, Thoerem 3.4] to the multidimensional case. Its proof is given in Appendix A.

Theorem 2.1. *If $X_1, \ldots, X_N$ are the time shifted covariates of an LTI hidden state space model (1.1) where the components $(u_i)_i$ are independent centred standardized multivariate Gaussian $\mathcal{N}(0, \sigma_u^2 I_r)$ or subGaussian centred random vectors of subGaussian components having the same ψ_2 norm upper bound of σ_u, X is given by (1.7). Then, with probability at least $1 - \exp(-t)$ for $t \geqslant 1$, it holds that*

$$\left| \frac{1}{N} X^* X - \sigma_u^2 I_{(2T-1)r} \right|_{S_\infty} \leqslant c\sigma_u^2 \left(\sqrt{\frac{TN_1}{N}} + \frac{TN_1}{N} + \sqrt{\frac{tT}{N}} + \frac{tT}{N} \right), \tag{2.3}$$

with $N_1 = \log(T) + r$. Under the same conditions, for $\delta \in (0, e^{-1})$, with probability $1 - \delta$ and for values of N such that

$$\bar{N} \geqslant c \left(T N_1 \vee T \log \frac{1}{\delta} \right),$$

we have, for all $g \in \mathcal{M}_{p \times (2T-1)r}(\mathbb{R})$,

$$\frac{\sigma_u^2}{2} |g|_{S_2}^2 \leqslant \frac{1}{N} |X g^*|_{S_2}^2 \leqslant \frac{3\sigma_u^2}{2} |g|_{S_2}^2.$$

Concentration results for matrices with independent covariates are obtained in [2] where it is shown that with high probability the following holds

$$\left| \sum_{i=1}^{N-1} x_i x_i^* - \sigma_u^2 I_T \right|_{S_\infty} \lesssim \sigma_u^2 \left(\frac{T}{N} + \sqrt{\frac{T}{N}} + \frac{d}{N} t + \sqrt{\frac{d}{N}} t^{1/2} \right).$$

Our result, as a multidimensional extension of [17, Theorem 3.4], shows that a similar result holds for block Toplitz matrices up to the $N_1 = \log(T) + r$ factor appearing in (2.3). Comparing with the matrix of the covariates of the hidden dynamical system (1.1), it is shown in [16, Proposition 2.1] that a re-scaled version of it does concentrate around the identity if the eigenvalues of the matrix A_0 are not on the unit circle, but would fail otherwise.

Before stating the second important probabilistic estimate, we introduce the operator $H^{\dagger^*} : \mathcal{M}_{(2T-1)r \times p}(\mathbb{R}) \to \mathcal{M}_{Tp \times Tr}(\mathbb{R})$ defined, for any $h = [h_1^*, \ldots, h_{2T-1}^*]^* \in \mathcal{M}_{(2T-1)r \times p}(\mathbb{R})$, by

$$H^{\dagger^*} h = \begin{bmatrix} h_1 & \frac{1}{2} h_2 & \frac{1}{3} g_3 & & \frac{1}{T} h_T \\ \frac{1}{2} h_2 & \frac{1}{3} h_3 & & & \frac{1}{T-1} h_{T+1} \\ \frac{1}{3} h_3 & & & & \frac{1}{T-2} h_{T+2} \\ & & & & \\ \frac{1}{T} h_T & \frac{1}{T-1} g_{T+1} & \frac{1}{T-2} h_{T+2} & & h_{2T-1} \end{bmatrix}. \tag{2.4}$$

It is easy to check that $\langle h, g^* \rangle = \langle H^{\dagger^*} h, H g^* \rangle$ so that the operator $H^{\dagger^*}$ is the adjoint of the pseudo-inverse of H.

We also introduce the $\mathcal{H}_\infty$ norm for an infinite sequence of matrices $\varphi = [\varphi_0, \varphi_1, \ldots]$ given by

$$|\varphi|_{\mathcal{H}_\infty} := \sup_{x \in [0\ 1]} | \sum_{j=0}^{\infty} \varphi_j e^{i 2\pi x} |_{S_\infty}.$$

This norm relates to the notion of system norm used in control theory and turns out to be the right measure of how the hidden dynamic impacts the estimation error through the variance. The next theorem supports this claim by providing a control over the noise level induced by the term $\bar{g}_0 \bar{X}_t + h W_t + v_t$ given in (1.2).

Theorem 2.2. *Assume the random matrices X, $\bar{X}$, W, and ε as defined in (1.7) are generated by running the LTI hidden state space model (1.1) under either Gaussian or subGaussian noise condition.*

Define g, $\hat{g}$, h as in (1.2). Then, with probability at least $1 - \exp(-t)$ for $t \geqslant 1$, the following bounds for different parts of the noise term hold:

$$|H^{\dagger^*} X^* \bar{X} \bar{g}|_{S_\infty} \lesssim \sigma_u^2 |\bar{g}|_{\mathcal{H}_\infty} \left(\sqrt{\frac{N_0}{\bar{N}}} + \frac{N_0 \log(T)}{\bar{N}} + \sqrt{\frac{t}{\bar{N}}} + \frac{\log(T)t}{\bar{N}} \right). \tag{2.5}$$

with $N_0 = \log(T) + p + r$.

$$|H^{\dagger^*} X^* W h|_{S_\infty} \lesssim \sigma_u \sigma_w |h|_{\mathcal{H}_\infty} \left(\sqrt{\frac{N_2}{\bar{N}}} + \frac{N_2 \log(T)}{\bar{N}} + \sqrt{\frac{t}{\bar{N}}} + \frac{\log(T)t}{\bar{N}} \right). \tag{2.6}$$

with $N_2 = \log(T) + p$

$$|H^{\dagger^*} X^* \varepsilon|_{S_\infty} \lesssim \sigma_v^2 \left(\sqrt{\frac{N_2}{\bar{N}}} + \frac{N_2 \log(T)}{\bar{N}} + \sqrt{\frac{t}{\bar{N}}} + \frac{\log(T)t}{\bar{N}} \right). \tag{2.7}$$

2.3. Estimation guarantees for the Hankel penalized regression

This section is devoted to the analysis of the performance of the Hankel penalized regression estimator given by (1.12). This estimator plays a central role in Algorithm 1 since it is used twice. The first time it uses covariates of length $2T_0 - 1$ in (2.1) to provide a sparse estimate for estimating the true order of the system, and the second time in (2.2) where it uses $2\hat{d}_\xi + 1$ covariates to provide a more accurate estimator. To analyze the performance of this estimator, we first state a corollary to Theorem 2.2.

Corollary 2.1. *Under the same condition of Theorem 2.2, for $\delta \in (0\ e^{-1})$, there exist an absolute positive constant $c > 0$ such that for λ taken as*

$$\lambda := c\phi\sigma_u^2 \left(\sqrt{\frac{N_0}{N}} \vee \frac{\log(T)N_0}{N} \vee \frac{\sqrt{\log \frac{1}{\delta}}}{\sqrt{N}} \vee \frac{\log(T)\log \frac{1}{\delta}}{N} \right), \tag{2.8}$$

with $\phi = |\bar{g}|_{\mathcal{H}_\infty} + \frac{\sigma_w}{\sigma_u}|h|_{\mathcal{H}_\infty} + 1$, we have, with probability at least $1 - \delta$, the following upper bound:

$$\lambda \geqslant \frac{3}{N} \left(|H^{\dagger^*} X^* \bar{X} \bar{g}|_{S_\infty} + |H^{\dagger^*} X^* W h|_{S_\infty} + |H^{\dagger^*} X^* \varepsilon|_{S_\infty} \right). \tag{2.9}$$

The following theorem provides various estimation bounds and the sample complexity for the Hankel penalized regression estimator of the Markov parameters for the $|\cdot|_{S_p}$-norms with $p \in (0\ 1)$.

Theorem 2.3. *Let $(X_{2T}, y_{2T}), \ldots, (X_N, y_N)$ be the input and output values of the LTI hidden state space model (1.1) under Gaussian or subGaussian assumption for the different noise vectors. Assume that (X, y) is given by (1.7) and the estimator $\hat{g}$ given by (1.12). Define $\Delta g = \hat{g} - g_0$, where g_0 is the sequence of Markov parameters defined in (1.4). Under Assumptions 1.1 and 1.3, for the values of $\bar{N}$ such that*

$$\bar{N} \geqslant c \left(T N_1 \vee T \log \frac{1}{\delta} \right) \tag{2.10}$$

and the values of λ given by (2.8), with probability at least $1 - 2\delta$, for $\delta \in (0\ e^{-1}/2)$, the estimator $\hat{g}$ satisfies the following error bounds:

- *Slow and fast rates for the prediction error of the Markov parameters:*

$$\frac{1}{\sqrt{\bar{N}}}|X\Delta g^*|_{S_2} \leqslant \frac{5\sqrt{3}d_0^{1/2}T^{1/2}\lambda}{6\sigma_u} \wedge d_0^{1/4}\lambda^{1/2}|Hg_0^*|_{S_2}^{1/2}. \tag{2.11}$$

- *Slow and fast rates for the estimation error of the Markov parameters:*

$$|\Delta g|_{S_2} \leqslant \frac{5\sqrt{3}d_0^{1/2}T^{1/2}\lambda}{6\sigma_u^2} \wedge \frac{\sqrt{2}d_0^{1/4}\lambda^{1/2}}{\sigma_u}|Hg_0^*|_{S_2}^{1/2}. \tag{2.12}$$

- *Fast rate for the Hankel estimation spectral loss:*

$$\mathcal{L}_2^H(\hat{g}, g_0) \leqslant \frac{5\sqrt{2}}{3\sigma_u^2}d_0^{1/2}\lambda T. \tag{2.13}$$

- *Fast rate for the Hankel estimation p-loss, $p \in [1, 2]$:*

$$\mathcal{L}_p^H(\hat{g}, g_0) \leqslant \frac{20d_0^{1/p}\lambda T}{\sigma_u^2}. \tag{2.14}$$

- *Sample complexity for the spectral loss: for all $\epsilon > 0$ to obtain $\mathcal{L}_2^H(\hat{g}, g_0) \leqslant \epsilon$ we need*

$$\bar{N} \gtrsim \frac{\phi^2 d_0 T^2 N_0}{\epsilon^2} \vee \frac{\phi d_0^{1/2} T N_0 \log(T)}{\epsilon} \vee \frac{\phi^2 d_0 T^2 \log\frac{1}{\delta}}{\epsilon^2} \vee \frac{\phi d_0^{1/2} T \log(T) \log\frac{1}{\delta}}{\epsilon}. \tag{2.15}$$

Remark 2.1. *Upon inspection of the proof we notice that the result would still hold without neither Assumptions 1.1 nor 1.3. However, while all the rates hold without assumption 1.1, this assumption is necessary for these rate to converge to 0 when we observe more samples. Similarly, in the absence of Assumption 1.3 all the rates given in the theorem hold after replacing d_0 by T. Still, having $d_0 < T$ means that we are estimating less Markov parameter than necessary to be able to recover a minimal realization as was explained in Section 1.1.*

The proof of Theorem 2.3 relies on the analysis of the first-order optimality condition. This approach appeared first in [23] and in the case of matrix regression in [26] to provide oracle inequalities in the context of low-rank matrix completion. The same argument can be combined with alternative approaches, including the analysis of the zero-order optimality condition suggested in [5]. These are some of the approaches used for high dimension estimation problems. Indeed, we can cast the problem of estimating the Hankel matrix of a hidden state LTI state space model of unknown order as a high dimension matrix regression problem where we want to estimate a low rank Hankel matrix since the rank of the T Hankel matrix of the Markov parameters is the dimension of the minimal realization d_0 as long as $T \geqslant d_0$.

Existing results in the literature such as [33, Theorem 3.1] are provided for the least square estimator in terms of the $|\cdot|_{S_\infty}$-norm while the dimension is known. They do not extend to the case of $|\cdot|_{S_p}$-norm with $p \in (0\ 1)$ since, while $|\cdot|_{S_\infty}$-norm is dimension free, the least square estimator is oblivious to the rank of the estimate. Indeed, the solution of the least square estimator is not expected to be low rank and thus by simple norm domination [33, Theorem 3.1] implies,

$$|H\hat{g}_{\mathrm{ls}}^* - Hg_0^*|_{S_2} \lesssim \sqrt{\frac{T^3 q_0 \log^2(Tq_0)}{\bar{N}}}.$$

This bound misses the correct dimension scaling by a polynomial factor of $T^{1/2}$ for the $|\cdot|_{S_2}$-norm. On the other hand, Theorem 5.1 in [39] implies that

$$|\hat{\mathcal{H}}_d - Hg_0^*|_{S_2} \lesssim \sqrt{\frac{pT^3 + rT^2 + T^2\log(1/\delta)}{\bar{N}}}$$

which also misses the correct dimension scaling by a factor of $(T/d_0)^{1/2}$. For a non-low rank estimator this is the expected order as it estimates prT^2 unknowns with a variance that scales like T times the variance of $\bar{g}_0\bar{X}_t + hW_t + v_t$. We also note that the estimator $\hat{\mathcal{H}}_d$ does not preserve the Hankel structure of the matrix Hg_0^*. A low rank estimate reduces the number of the unknowns to $(p+r)d_0T$, which is consistent with our result which, after keeping only the main dimension terms, reads

$$\mathcal{L}_2^H(\hat{g}, g_0) \lesssim \sqrt{\frac{(p + r + \log(T/\delta))d_0T^2}{N}}.$$

In [43, Theorem 1] the authors study the problem of recovering the Markov parameter, while the dimension d_0 is unknown, but from the partial observation of multiple trajectories of the system and assuming that $w_i = 0$ in (1.1). To this end, they propose a penalized least square estimator for the Markov parameters as given in (1.10). While they successfully manage to control the error in the $|\cdot|_{S_\infty}$-norm, since they penalize with the transformation of the Hankel matrix $|HK^{-1}g^*|_{S_1}$, there is no reason to believe that the solution will give a low rank Hankel matrix. Their result (1.11) in the $|\cdot|_{S_2}$-norm implies that with high probability and after observing enough data we have

$$\mathcal{L}_2^{\mathcal{H}}(\hat{g}, g_0) \lesssim \begin{cases} \frac{\sigma_z}{\sigma_u}\sqrt{\frac{rT^3}{N}}\log(T) & N \geqslant d_0^2 \wedge T, \\ \frac{\sigma_z}{\sigma_u}\sqrt{\frac{d_0rT^3}{N}}\log(T) & d_0 \leqslant N \leqslant d_0^2 \wedge T, \end{cases}$$

which again does not capture well the effect of the dimension. Moreover, in this case, it also misses the effect of the dynamic captured in our case by the term $\phi = |\bar{g}|_{\mathcal{H}_\infty} + \frac{\sigma_w}{\sigma_u}|h|_{\mathcal{H}_\infty} + 1$. This is due to the fact that in their setup, we stop every realization after $2T - 1$ observation and suppose all trajectories are independent.

Proof. Set

$$\Gamma := |H^{\dagger^*}X^*\bar{X}\bar{g}^*|_{S_\infty} + |H^{\dagger^*}X^*Wh^*|_{S_\infty} + |H^{\dagger^*}X^*\varepsilon|_{S_\infty}.$$

We start by using Corollary 2.1 and Theorem 2.1 to define an event of probability $1 - 2\delta$ where we have both

$$\lambda \geqslant \frac{3\Gamma}{\bar{N}} \tag{2.16}$$

and, for all $g \in \mathcal{M}_{p\times(2T-1)r}$,

$$\frac{\sigma_u^2}{2}|g|_{S_2}^2 \leqslant \frac{1}{\bar{N}}|Xg^*|_{S_2}^2 \leqslant \frac{3\sigma_u^2}{2}|g|_{S_2}^2 \tag{2.17}$$

for values of N such that

$$\bar{N} \geqslant c\left(TN_1 \vee T\log\frac{1}{\delta}\right).$$

Since $\hat{g}$ solves the optimization problem (1.12), by Fermat's rule $0 \in \partial \operatorname{Crit}_\lambda(\hat{g})$, the subdifferential set of the criterion function. Also, by Fenchel-Rockafellar theorem (see e.g. [36]), there exists $v \in \partial|H\hat{g}^*|_{S_1}$ such that,

$$\frac{2}{N}X^*(X\hat{g}^* - y) + \lambda v = 0.$$

Using the fact that $y = Xg_0^* + \bar{X}\bar{g}_0^* + Wh_0^* + \varepsilon$ and multiplying by $\Delta g = \hat{g}^\lambda - g_0$ gives

$$|X\Delta g^*|_{S_2}^2 = \frac{\lambda\bar{N}}{2}\langle -\Delta g^*, v\rangle + \langle X^*(\bar{X}\bar{g}_0^* + Wh_0^* + \varepsilon), \Delta g\rangle.$$

By the definition of the sub-gradient we have, for all $g \in \mathcal{M}_{p\times(2T-1)r}$,

$$|Hg^*|_{S_1} \geqslant |H\hat{g}^*|_{S_1} + \langle -\Delta g^*, v\rangle.$$

Hölder's inequality yields

$$|X\Delta g^*|_{S_2}^2 \leqslant \langle X^*(\bar{X}\bar{g}^* + Wh^* + \varepsilon), \Delta g^*\rangle + \frac{\lambda\bar{N}}{2}(|Hg_0^*|_{S_1} - |H\hat{g}^*|_{S_1})$$

$$\leqslant \Gamma|H\Delta g^*|_{S_1} + \frac{\lambda\bar{N}}{2}(|Hg_0^*|_{S_1} - |H\hat{g}^*)|_{S_1} \tag{2.18}$$

$$\leqslant (\Gamma + \frac{\lambda\bar{N}}{2})|Hg_0^*|_{S_1} + (\Gamma - \frac{\lambda\bar{N}}{2})|H\hat{g}^*)|_{S_1}.$$

Since by (2.16) we have $\lambda \geqslant \frac{2\Gamma}{N}$, then it holds that

$$\frac{1}{N}|X\hat{g}^* - Xg_0^*|_{S_2}^2 \leqslant \lambda|Hg_0^*|_{S_1} \leqslant \lambda\sqrt{\operatorname{rank}(Hg_0^*)}|Hg_0^*|_{S_2}$$

which proves the slow rate in (2.11). The slow rate in (2.12) is implied by inequality (2.11), since we are in an event where the inequality (2.17) holds.

For a matrix M with a singular value decomposition $M = U\Sigma V^*$ define the projection operators $P_U^\perp := I - UU^*$, $P_V^\perp := I - V^*V$, $\mathcal{P}_M^\perp(N) := P_U^\perp N P_V^\perp$, and $\mathcal{P}_M^\perp := I - \mathcal{P}_M^\perp$. Since we have a decomposable penalty [8], we have

$$|H\hat{g}^*|_{S_1} = |Hg_0^* + H\Delta g^*|_{S_1},$$

$$= |Hg_0^* + \mathcal{P}_{Hg_0^*}^\perp(H\Delta g^*) + \mathcal{P}_{Hg_0^*}(H\Delta g_0^*)|_{S_1},$$

$$\geqslant |Hg_0^* + \mathcal{P}_{Hg^*}^\perp(H\Delta g^*)|_{S_1} - |\mathcal{P}_{Hg_0^*}(H\Delta g^*)|_{S_1},$$

$$= |Hg_0^*|_{S_1} + |\mathcal{P}_{Hg_0^*}^\perp(H\Delta g^*)|_{S_1} - |\mathcal{P}_{Hg_0^*}(H\Delta g^*)|_{S_1},$$

from which, together with (2.18), we obtain

$$|X\Delta g^*|_{S_2}^2 \leqslant \Gamma|H\Delta g^*|_{S_1} + \frac{\lambda\bar{N}}{2}(|\mathcal{P}_{Hg_0}(H\Delta g^*)|_{S_1} - |\mathcal{P}_{Hg_0^*}^\perp(H\Delta g^*)|_{S_1})$$

$$= \Gamma(|\mathcal{P}_{Hg_0}(H\Delta g^*)|_{S_1} + |\mathcal{P}_{Hg_0^*}^\perp(H\Delta g^*)|_{S_1}) + \frac{\lambda\bar{N}}{2}(|\mathcal{P}_{Hg_0^*}(H\Delta g^*)|_{S_1}$$

$$- |\mathcal{P}^{\perp}_{Hg_0^*}(H\Delta g^*)|_{S_1})$$

$$= (\Gamma - \frac{\lambda \bar{N}}{2})|\mathcal{P}^{\perp}_{Hg_0^*}(H\Delta g_*)|_{S_1} + (\Gamma + \frac{\lambda \bar{N}}{2})|\mathcal{P}_{Hg_0^*}(H\Delta g^*)|_{S_1}.$$

Again, by the particular choice of $\lambda \geqslant \frac{3\Gamma}{N}$, we have

$$0 \leqslant |X\Delta g^*|^2_{S_2} \leqslant \frac{\lambda}{6}\bar{N}(5|\mathcal{P}_{Hg_0^*}(H\Delta g^*)|_{S_1} - |\mathcal{P}^{\perp}_{Hg_0^*}(H\Delta g^*)|_{S_1}). \tag{2.19}$$

Now, by the following rank inequality

$$\mathrm{rank}(\mathcal{P}_{Hg_0^*}(H\Delta g_0^*)) = \mathrm{rank}(\mathcal{P}_{Hg_0^*}(H\hat{g}^*) + Hg_0^*) \leqslant 2\,\mathrm{rank}(Hg_0^*).$$

and the fact that we are on an event such that

$$\frac{\sigma_u^2}{2}|\Delta g^*|^2_{S_2} \leqslant \frac{1}{\bar{N}}|X\Delta g^*|^2_{S_2} \leqslant \frac{3\sigma_u^2}{2}|\Delta g^*|^2_{S_2},$$

we have

$$\frac{\sigma_u^2}{2T}|H\Delta g^*|^2_{S_2} \leqslant \frac{\sigma_u^2}{2}|\Delta g^*|^2_{S_2} \leqslant \frac{1}{\bar{N}}|X\Delta g^*|^2_{S_2}$$

$$\leqslant \frac{5\lambda}{6}\sqrt{\mathrm{rank}(\mathcal{P}_{Hg_0^*}(H\Delta g^*)|}|\mathcal{P}_{Hg_0^*}(H\Delta g^*)|_{S_2}$$

$$\leqslant \frac{5\sqrt{2}\lambda}{6}d_0^{1/2}|H\Delta g^*|_{S_2} \leqslant \frac{5\sqrt{2}\lambda}{6}(d_0 T)^{1/2}|\Delta g^*|_{S_2}$$

$$\leqslant \frac{5\sqrt{3}\lambda}{6}(d_0 T)^{1/2}\sqrt{\frac{1}{\bar{N}}}\frac{|X\Delta g^*|_{S_2}}{\sigma_u}.$$

This implies the fast rates in (2.11), (2.12), and (2.13). Also, since $\hat{g}$ satisfies (2.19), we have

$$5|\mathcal{P}_{Hg_0^*}(H\Delta g^*)|_{S_1} \leqslant |\mathcal{P}^{\perp}_{Hg_0^*}(H\Delta g^*)|_{S_1}. \tag{2.20}$$

This gives

$$|H\Delta g^*|_{S_1} \leqslant |\mathcal{P}_{Hg_0^*}(H\Delta g^*)|_{S_1} + |\mathcal{P}^{\perp}_{Hg_0^*}(H\Delta g^*)|_{S_1}$$

$$\leqslant 6|\mathcal{P}_{Hg_0^*}(H\Delta g^*)|_{S_1} \leqslant 6\sqrt{2}d_0^{1/2}|H\Delta g^*|_{S_2} \leqslant \frac{20d_0^{1/2}\lambda T}{\sigma_u^2}.$$

But, $|H\Delta g_0^*|_{S_p} = \left(\sum_{i=1}^{d} s_i^p\right)^{1/p} = |s|_p$ where s is the vector of singular values. Therefore, by the norm interpolation identity $|s|_p \leqslant (|s|_1)^{2/p-1}(|s|_2)^{2-2/p}$ for $p \in [0,1]$, we finally obtain

$$|H\Delta g^*|_{S_p} \leqslant (|H\Delta g^*|_{S_1})^{2/p-1}(|H\Delta g^*|_{S_2})^{2-2/p} \leqslant \frac{20d_0^{1/p}\lambda T}{\sigma_u^2}.$$

$\square$

Remark 2.2. *The condition* (2.10) *on the sample size is likely to be sub-optimal. One expects that the factor T should be replaced by d_0. The factor T comes from the use of the concentration result of Theorem* 2.1. *While this Theorem gives the right rate for the input covariates' concentration, the result is stronger than needed. Indeed, Theorem* 2.1 *provides us with an event in which for all $g \in \mathcal{M}_{p \times (2T-1)r}(\mathbb{R})$ we have*

$$\frac{\sigma_u^2}{2}|g|_{S_2}^2 \leqslant \frac{1}{\bar{N}}|Xg^*|_{S_2}^2 \leqslant \frac{3\sigma_u^2}{2}|g|_{S_2}^2, \tag{2.21}$$

while the proof needs such a control only on the set defined by the cone condition (2.20).

Open Problem 2.1. *Show that for all $g \in \mathcal{M}_{p \times (2T-1)r}(\mathbb{R})$ such that $|g|_{S_2} \leqslant 1$ and* (2.20) *hold then*

$$\left| \frac{1}{\bar{N}}|Xg^*|_{S_2}^2 - \sigma_u^2 |g^*|_{S_2}^2 \right| \lesssim \sigma_u^2 \sqrt{\frac{d_0}{\bar{N}}},$$

up to logarithmic terms and lower order terms.

While the condition (2.10) on $\bar{N}$ is likely to be suboptimal, we note that it is still less restrictive than the sample complexity (2.15) which will play a major role in the analysis of Algorithm 1. As we shall see below, for this reason, the condition (2.10) will not affect the upcoming results on the estimation of the parameters $(\bar{A}, \bar{B}, \bar{C})$.

In the following proposition we show that the SVD decomposition of the Hankel matrix obtained from the Makov parameters estimate $\hat{g}$ given in (1.12) can be used to recover the system's order d_0, if given a lower bound on the smallest singular value of the true Hankel matrix of Markov parameters Hg_0^*. We also show that the fast rate for the spectral loss in (2.13) implies a fast rate for the truncation of the SVD decomposition.

To this end, we consider the SVD decomposition of the Hankel matrix of the estimated parameter $\hat{g}$ given by $H\hat{g}^* = \sum_{i=1}^{\text{rank}(H\hat{g}^*)} \hat{s}_i \hat{u}_i \hat{v}_i^*$ and define the truncation dimension $\check{d}_\xi$ and the truncated SVD matrix $\hat{\mathcal{H}}_{\check{d}_\xi}$ of the estimated Hankel matrix as:

$$\check{d}_\xi := \sum_{i=1}^{\text{rank}(H\hat{g}^*)} \mathbf{1}\{\hat{s}_i \geqslant 2\xi\} \quad \text{and} \quad \hat{\mathcal{H}}_{\check{d}_\xi} := \sum_{i=1}^{\text{rank}(H\hat{g}^*)} \mathbf{1}\{\hat{s}_i \geqslant 2\xi\}\hat{u}_i \hat{v}_i^*. \tag{2.22}$$

Proposition 2.1. *Assume the same conditions on $(X_{2T}, y_{2T}), \ldots, (X_N, y_N)$ as in Theorem* 2.3 *and suppose that*

$$s_{d_0}(Hg_0^*) \geqslant 3\xi$$

for some $\xi > 0$. Then, there exists an absolute positive constant c such that for the values of $\bar{N}$ given by

$$\bar{N} \geqslant cd_0 T N_0 \vee T \log\frac{1}{\delta} \vee \frac{\phi^2 d_0 T^2}{\xi^2}\left(N_0 \vee \log\frac{1}{\delta}\right) \vee \frac{\phi d_0^{1/2} T \log(T)}{\xi}\left(N_0 \vee \log\frac{1}{\delta}\right), \tag{2.23}$$

the dimension $\check{d}_\xi$ and then estimate $\hat{\mathcal{H}}_{\check{d}_\xi}$ defined in (2.22), *satisfy with probability at least $1 - 2\delta$ the following.*

- *Exact rank recovery:*

$$\check{d}_\xi = d_0. \tag{2.24}$$

- *Lower bound over the least singular value of the truncated estimate:*

$$s_{\check{d}_\xi}(H\hat{g}) \geqslant 2\xi. \tag{2.25}$$

- *Lower bound over the singular values after the truncated threshold:*

$$\textit{for all } d \in [\![\check{d}_\xi + 1, \mathrm{rank}(H\hat{g}^*)]\!], \qquad \hat{s}_i \leqslant \left(\sum_{i=d_0+1}^{\mathrm{rank}(H\hat{g}^*)} \hat{s}_i^2 \right)^{1/2} \leqslant \xi. \tag{2.26}$$

- *Fast rate for the truncated estimate on the 2-loss:*

$$|\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_2} \leqslant \frac{10\sqrt{2}}{3\sigma_u^2} d_0^{1/2} \lambda T. \tag{2.27}$$

Proof. By the obtained sample complexity (2.15) it follows that the condition on $\bar{N}$ in (2.23) implies that, for a large enough absolute constant c,

$$\frac{5\sqrt{2}}{3\sigma_u^2} d_0^{1/2} \lambda T \leqslant \xi$$

on the same event defined in Theorem 2.3. Therefore, in view the same Theorem and Weyl's inequality, with the same probability of at least $1 - 2\delta$, we also have

$$|s_d(H\hat{g}^*) - s_d(Hg_0^*)| \leqslant |H\Delta g^*|_{S_\infty} \leqslant |H\Delta g^*|_{S_2} \leqslant \frac{5\sqrt{2}}{3\sigma_u^2} d_0^{1/2} \lambda T \leqslant \xi.$$

Now, if we assume that $\mathrm{rank}(H\hat{g}^*) < d_0$, then $s_{\min}(Hg^*) = s_{d_0}(Hg_0^*)$ and $s_{d_0}(H\hat{g}^*) = 0$. Thus, again by Weyl's inequality, we have

$$s_{\min}(Hg_0^*) = |s_{d_0}(Hg_0^*) - s_{d_0}(H\hat{g}^*)| \leqslant |s_1(H\Delta g^*)|$$

$$\leqslant |H\Delta g^*|_{S_2} \leqslant \frac{5\sqrt{2}}{3\sigma_u^2} d_0^{1/2} \lambda T \leqslant \xi,$$

which contradicts the assumption

$$s_{\min}(Hg) \geqslant 3\xi.$$

Therefore, $d_0 \geqslant \mathrm{rank}(H\hat{g})$. This also means that

$$|s_{d_0}(Hg) - s_{d_0}(H\hat{g})| \leqslant \xi \quad \text{and} \quad s_{d_0}(H\hat{g}) \geqslant 2\xi. \tag{2.28}$$

Since we now know that $\hat{d} \geqslant d_0$, we consider the following decomposition for the SVD representation

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_{d_0} + \hat{\mathcal{H}}_{\bar{d}} = \sum_{i=1}^{d_0} \hat{s}_i \hat{u}_i \hat{v}_i^* + \sum_{i=d_0+1}^{\mathrm{rank}(H\hat{g})} \hat{s}_i \hat{u}_i \hat{v}_i^* = \begin{bmatrix} U_d & U_{\bar{d}} \end{bmatrix} \begin{bmatrix} \Sigma_d & \\ & \Sigma_{\bar{d}} \end{bmatrix} \begin{bmatrix} V_d^* \\ V_{\bar{d}}^* \end{bmatrix}.$$

Now, as the truncated SVD decomposition to rank d_0 solves the optimization problem:

$$\hat{\mathcal{H}}_{\bar{d}} \in \arg \min_{H:\ \mathrm{rank}(H) \leqslant d} |\hat{\mathcal{H}} - H|_{S_2},$$

we obtain

$$|\hat{\mathcal{H}}_{\bar{d}}|_{S_2} = \min_{H:\ \mathrm{rank}(H) \leqslant d} |\hat{\mathcal{H}} - H|_{S_2} \leqslant |H\Delta g^*|_{S_2}. \tag{2.29}$$

In particular,

$$\text{for all } d \in [\![d_0 + 1, \mathrm{rank}(H\hat{g}^*)]\!] \qquad \hat{s}_i \leqslant \left(\sum_{i=d_0+1}^{\mathrm{rank}(H\hat{g}^*)} \hat{s}_i^2 \right)^{1/2} \leqslant \xi.$$

This inequality together with (2.28) yield the following result on rank recovery:

$$\check{d}_{\xi} = \sum_{i=1}^{\mathrm{rank}(H\hat{g}^*)} \mathbf{1}\{\hat{s}_i \geqslant 2\xi\} = d.$$

It also yields (2.26) as well.

Now, since we have

$$|\hat{\mathcal{H}}_d - Hg_0^*|_{S_2} = |H\hat{g} - Hg_0^* - \hat{\mathcal{H}}_{\bar{d}}|_{S_2} \leqslant |H\Delta g|_{S_2} + |\hat{\mathcal{H}}_{\bar{d}}|_{S_2}$$

$$\leqslant 2|H\Delta g^*|_{S_2} \leqslant \frac{10\sqrt{2}}{3\sigma_u^2} d_0^{1/2} \lambda T,$$

in view of (2.29), we also have the fast rate for the SVD estimate in (2.27). $\qquad\square$

2.4. Error control for the Ho-Kalman algorithm estimates

This section provides stability results in the Hilbert-Schmidt norm for a version of an estimation procedure based on a variant of the Ho-Kalman algorithm. The variant of the Ho-Kalman algorithm in question is the one that obtains a minimal balanced realization starting from the SVD decomposition of the Hankel matrix of T Markov parameters, for $T \geqslant d_0 + 1$. Indeed, the Ho-Kalman algorithm computes, up to a similarity transform, the observability and controllability matrices are respectively

$$\bar{\mathcal{O}} = U_0 \Sigma_0^{1/2} \quad \text{and} \quad \bar{\mathcal{C}} = \Sigma_0^{1/2} V_0^*, \tag{2.30}$$

and the minimal balanced realization (see Definition 1.3) defined by

$$\bar{A} := \left(\bar{\mathcal{O}}_{1:r(T-1),1:d_0} \right)^{\dagger} \bar{\mathcal{O}}_{r+1:rT,1:d_0}, \ \ \bar{B} := \bar{\mathcal{O}}_{1:d_0,1:r}, \ \ \bar{C} := \bar{\mathcal{O}}_{1:p,1:d_0}.$$

Assuming that we have obtained an estimate $\hat{\mathcal{H}}_T$ of the Hankel matrix of order T with a rank that is higher than the dimension d_0 and of the dimension $\check{d}_{\xi}$ such that the Hilbert-Schmidt error $|\hat{\mathcal{H}}_T - Hg_0^*|_{S_2}$ is small.

The Ho-Kalman based estimation algorithm we introduce here yields an estimate of the minimal balanced realization by mimicking the Ho-Kalman algorithm described above. It starts from a truncated

SVD decomposition $\hat{\mathcal{H}}_{\check{d}_\xi} = \hat{U}_\xi \hat{\Sigma}_\xi \hat{V}_\xi^*$ of the matrix $\hat{\mathcal{H}}_T$ to the smaller estimated dimension $\check{d}_\xi$ and constructs estimates of both the observability and controllablity matrices

$$\hat{\mathcal{O}} := \hat{U}_\xi \hat{\Sigma}_\xi^{1/2}, \quad \hat{\mathcal{C}} := \hat{\Sigma}_\xi^{1/2} \hat{V}_\xi^*.$$

Thereafter, it provides an estimated minimal balanced realization as

$$\hat{A} = \left(\hat{\mathcal{O}}_{1:r(T-1),1:d_\xi} \right)^\dagger \hat{\mathcal{O}}_{r+1:rT,1:d_\xi}, \ \ \hat{B} := \hat{\mathcal{O}}_{1:d_\xi,1:r}, \ \ \hat{C} := \hat{\mathcal{O}}_{1:p,1:d_\xi}.$$

The next theorem provides error bounds for these estimates under the assumption that we have $\check{d}_\xi = d_0$,

Theorem 2.4. *Suppose that* $\check{d}_\xi = d_0$. *Set*

$$\bar{\mathcal{O}}_{1:r(T-1),1:d_0} = \bar{\mathcal{O}}^+. \tag{2.31}$$

If the following stability assumption holds

$$|\hat{\mathcal{H}}_T - H g_0^*|_{S_\infty} \wedge |\hat{\mathcal{H}}_T - H g_0^*|_{S_2} \leqslant \frac{\left(\sqrt{2}-1\right)^{1/2} s_{d_0}(\bar{\mathcal{O}}^+) s_{d_0}^{1/2}(H g_0^*)}{2\sqrt{2}}, \tag{2.32}$$

then, there exists an orthonormal matrix R such that the following holds.

- *The error on the observability and controllability matrices is controlled by the error on the truncation:*

$$|\hat{\mathcal{O}} - \bar{\mathcal{O}} R|_{S_2}^2 + |\hat{\mathcal{C}}_d - R^* \bar{\mathcal{C}}|_{S_2}^2 \leqslant \frac{2}{\sqrt{2}-1} \frac{|\hat{\mathcal{H}}_{\check{d}_\xi} - H g_0^*|_{S_2}^2}{s_{d_0}(H g_0^*)}.$$

- *The error on the C and B matrices is controlled by the error on the truncation:*

$$|\hat{C} - \bar{C} R|_{S_2}^2 + |\hat{B} - R^* \bar{B}|_{S_2}^2 \leqslant \frac{2}{\sqrt{2}-1} \frac{|\hat{\mathcal{H}}_{\check{d}_\xi} - H g_0^*|_{S_2}^2}{s_{d_0}(H g_0^*)}.$$

- *The error on the A matrix is controlled by the error on the truncation:*

$$|\hat{A} - R^* \bar{A} R|_{S_2} \leqslant \frac{2^{3/2} \left(1 + |\bar{A}|_{S_\infty}\right)}{\left(\sqrt{2}-1\right)^{1/2} s_{d_0}(\bar{\mathcal{O}}^+) s_{d_0}^{1/2}(H g_0^*)} |\hat{\mathcal{H}}_{\check{d}_\xi} - H g_0^*|_{S_2}.$$

This result provides a robustness analysis of the variant of a Ho-Kalman algorithm based estimation procedure described at the start of this section. The result is described in term of the $|\cdot|_{S_2}$-norm and shows that, under the stability condition (2.32), it is possible to recover up to an orthonormal matrix R the minimal balanced realization defined in 1.3 since we can bound the loss function $\mathcal{L}_2^{\mathcal{M}}(\hat{\mathcal{M}}, \bar{\mathcal{M}})$ in term of $|\hat{\mathcal{H}}_{\check{d}_\xi} - H g_0^*|_{S_2}$ as follows.

$$\mathcal{L}_2^{\mathcal{M}}(\hat{\mathcal{M}}, \bar{\mathcal{M}}) \leqslant \frac{2^{3/2} \left(1 + |\bar{A}|_{S_\infty}\right)}{\left(\sqrt{2}-1\right)^{1/2} s_{d_0}(\bar{\mathcal{O}}^+) s_{d_0}^{1/2}(H g_0^*)} |\hat{\mathcal{H}}_{\check{d}_\xi} - H g_0^*|_{S_2}.$$

In the next section we will use a slightly weaker version of this result to provide $\mathcal{L}_2^{\mathcal{M}}$ guarantees for Algorithm 1, namely we replace $s_{d_0}(\bar{\mathcal{O}}^+)$ with $s_{d_0}^{1/2}(Hg_0^*)$ both in the robustness condition (2.32) and in the error control of various estimates. This can be done since as argued in the proof $s_{d_0}(\bar{\mathcal{O}}^+) \leqslant s_{d_0}^{1/2}(Hg_0^*)$ and it is done so to only assume the knowledge of a lower bound $s_{d_0}^{1/2}(\bar{\mathcal{O}}^+)$. Otherwise we could work with the original statement by assuming the knowledge of a lower bound on $s_{d_0}(\bar{\mathcal{O}}^+)s_{d_0}^{1/2}(Hg_0^*)$. The condition (2.32) is stated with $|\hat{\mathcal{H}}_T - Hg_0^*|_{S_\infty} \wedge |\hat{\mathcal{H}}_T - Hg_0^*|_{S_2}$ which is always equal to $|\hat{\mathcal{H}}_T - Hg_0^*|_{S_\infty}$, it is done this way simply since sometimes it is easier to have a control over $|\hat{\mathcal{H}}_T - Hg_0^*|_{S_2}$ as it is the case for the Hankel penalized regression estimator in Theorem 2.3 .

The version of the Ho-Kalman based estimator studied here is the one studied in [45, Theorem 4]. Their guarantees suggested that

$$\mathcal{L}_2^{\mathcal{M}}(\hat{\mathcal{M}}, \bar{\mathcal{M}}) \leqslant c \frac{d_0 \left|Hg_0^*\right|_{S_\infty}^{1/2} |\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_2}}{s_{d_0}^2(\bar{\mathcal{O}}^+)s_{d_0}^{1/2}(Hg_0^*)}.$$

Our result improves it by replacing the factor $s_{d_0}^{-2}(\bar{\mathcal{O}}^+)s_{d_0}^{-1/2}(Hg_0^*)$ with the smaller factor $s_{d_0}^{-1}(\bar{\mathcal{O}}^+)s_{d_0}^{-1/2}(Hg_0^*)$ in the regime $s_{d_0}(Hg_0^*) \to 0$ and $s_{d_0}(\bar{\mathcal{O}}^+) \to 0$ which was introduced in remark 1.2 and removing the d_0 factor. Another estimator based on the Ho-Kalman algorithm is considered in [43, Theorem 5.2] where it is shown that

$$\mathcal{L}_2^{\mathcal{M}}(\hat{\mathcal{M}}, \bar{\mathcal{M}}) \leqslant c \frac{d_0^{1/2} \left|Hg_0^*\right|_{S_\infty} |\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_\infty}}{s_{d_0}^2(Hg_0^*)}.$$

Here, we improve the factor $\frac{d_0^{1/2}|Hg_0^*|_{S_\infty}}{s_{d_0}^2(Hg_0^*)}|H\hat{g}^* - Hg_0^*|_{S_\infty}$ by $\frac{|\bar{A}|_{S_\infty}|\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_2}}{s_{d_0}(\bar{\mathcal{O}}^+)s_{d_0}^{1/2}(Hg_0^*)}$ since $s_{d_0}(\bar{\mathcal{O}}^+)$ and $s_{d_0}^{1/2}(Hg_0^*)$ are usually comparable, as we shall see later in (2.41) where we have that $s_{d_0}^{1/2}(Hg_0^*) \geq s_{d_0}(\bar{\mathcal{O}}^+) \geq \frac{1}{\sqrt{2}}s_{d_0}^{1/2}(Hg_0^*)$.

Proof. We start by noting that

$$|\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_\infty} \leqslant |\hat{\mathcal{H}}_T - \hat{\mathcal{H}}_{\check{d}_\xi}|_{S_\infty} + |\hat{\mathcal{H}}_T - Hg_0^*|_{S_\infty}.$$

Since, by assumption, we have $\check{d}_\xi = d_0$ and the truncated SVD also minimizes the operator norm cost, we have

$$|\hat{\mathcal{H}}_T - \hat{\mathcal{H}}_{\check{d}_\xi}|_{S_\infty} = \min_{\operatorname{rank}(H)\leqslant\check{d}_\xi} |\hat{\mathcal{H}}_T - H|_{S_\infty} \leqslant |\hat{\mathcal{H}}_T - Hg_0^*|_{S_\infty}.$$

Since $\bar{\mathcal{O}}^+$ is a sub-matrix of $\bar{\mathcal{O}}$, we have

$$s_{d_0}(\bar{\mathcal{O}}^+) \leqslant s_{d_0}(\bar{\mathcal{O}}) \leqslant s_{d_0}^{1/2}(Hg_T^*), \tag{2.33}$$

where the second inequality follows from the construction in (2.30). Therefore, the condition (2.32) implies

$$|\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_\infty} \leqslant 2|\hat{\mathcal{H}}_T - Hg_0^*|_{S_\infty} \leqslant 2|\hat{\mathcal{H}}_T - Hg_0^*|_{S_2} \leqslant \frac{s_{d_0}(Hg_0^*)}{2}. \tag{2.34}$$

The result for the error on both the observability and controllability matrices will be derived as a direct consequence of the following lemma taken from [48]. A similar approach was used by [43] to analyze the performance of another variant of the Ho-Kalman algorithm.

Lemma 2.1 (Lemma 5.14 in [48]). *Let M_1, $M_2 \in \mathcal{M}_{n_1 \times n_2}(\mathbb{R})$ be two* rank *r matrices with SVD decompositions $M_1 = U_1 \Sigma_1 V_1^*$ and $M_2 = U_2 \Sigma_2 V_2^*$. If $|M_2 - M_1|_{S_\infty} \leqslant \frac{s_r(M_1)}{2}$ then there is an orthonormal matrix R such that:*

$$|U_1 \Sigma_1^{1/2} - U_2 \Sigma_2^{1/2} R|_{S_2}^2 + |V_1 \Sigma_1^{1/2} - R^* V_1 \Sigma_1^{1/2}|_{S_2}^2 \leqslant \frac{2|M_2 - M_1|_{S_2}^2}{(\sqrt{2}-1)s_r(M_1)}.$$

Since both $\hat{\mathcal{H}}_{\breve{d}_\xi}$ and Hg_0^* are of rank d_0, we can use this lemma together with (2.34) to guarantee on the same event that there exist a matrix R such that $RR^* = I_d$ and

$$|\hat{\mathcal{O}} - \bar{\mathcal{O}}R|_{S_2}^2 + |\hat{\mathcal{C}} - R^*\bar{\mathcal{C}}|_{S_2}^2 = |\hat{U}_\xi \hat{\Sigma}_\xi^{1/2} - U_0 \Sigma_0^{1/2} R|_{S_2}^2 + |\hat{V}_\xi^* \hat{\Sigma}_\xi^{1/2} - R^* V_0^* \Sigma_0^{1/2}|_{S_2}^2$$

$$\leqslant \frac{2}{\sqrt{2}-1} \frac{|\hat{\mathcal{H}}_{\breve{d}_\xi} - Hg_0^*|_{S_2}^2}{s_{d_0}(Hg_0^*)}.$$

Since, $\bar{C}$ and $\bar{B}$ are submatrices of $\mathcal{O}$ and $\mathcal{C}$ respectively, the last inequality implies

$$|\hat{C} - \bar{C}R|_{S_2}^2 + |\hat{B} - R^*\bar{B}|_{S_2}^2 \leqslant \frac{2}{\sqrt{2}-1} \frac{|\hat{\mathcal{H}}_{\breve{d}_\xi} - Hg_0^*|_{S_2}^2}{s_{d_0}(Hg_0^*)}.$$

To derive the estimation error bound for the matrix $\bar{A}$, we recall the following notation (introduced in (2.31)),

$$\hat{\mathcal{O}}_{r+1:rT,1:d_0} := \hat{\mathcal{O}}^-, \quad \bar{\mathcal{O}}_{1:r(T-1),1:d_0} := \bar{\mathcal{O}}^+, \quad \bar{\mathcal{O}}_{r+1:rT,1:d_0} := \bar{\mathcal{O}}^-.$$

We note that

$$|\hat{A} - R^*\bar{A}R|_{S_2} = \left|\left(\hat{\mathcal{O}}^+\right)^\dagger \hat{\mathcal{O}}^- - R^*\bar{A}R\right|_{S_2}$$

$$= \left|\left(\hat{\mathcal{O}}^+\right)^\dagger \hat{\mathcal{O}}^- - \left(\hat{\mathcal{O}}^+\right)^\dagger \hat{\mathcal{O}}^+ R^*\bar{A}R\right|_{S_2} \leqslant \left|\left(\hat{\mathcal{O}}^+\right)^\dagger\right|_{S_\infty} \left|\hat{\mathcal{O}}^- - \hat{\mathcal{O}}^+ R^*\bar{A}R\right|_{S_2}$$

$$\leqslant \frac{1}{s_{d_0}(\hat{\mathcal{O}}^+)} \left(\left|\hat{\mathcal{O}}^- - \bar{\mathcal{O}}^+ RR^*\bar{A}R\right|_{S_2} + \left|\bar{\mathcal{O}}^+ RR^*\bar{A}R - \hat{\mathcal{O}}^+ R^*\bar{A}R\right|_{S_2}\right)$$

$$\leqslant \frac{1}{s_{d_0}(\hat{\mathcal{O}}^+)} \left(\left|\hat{\mathcal{O}}^- - \bar{\mathcal{O}}^- R\right|_{S_2} + \left|\bar{\mathcal{O}}^+ R - \hat{\mathcal{O}}^+\right|_{S_2} |\bar{A}|_{S_\infty}\right)$$

$$\leqslant \left(\frac{2}{\sqrt{2}-1}\right)^{1/2} \frac{\left(1 + |\bar{A}|_{S_\infty}\right)}{s_{d_0}(\hat{\mathcal{O}}^+)} \frac{|\hat{\mathcal{H}}_{\breve{d}_\xi} - Hg_0^*|_{S_2}}{s_{d_0}^{1/2}(Hg_0^*)}, \tag{2.35}$$

where in the last inequality we used the fact that both $\hat{\mathcal{O}}^- - \bar{\mathcal{O}}^- R$ and $\hat{\mathcal{O}}^+ - \bar{\mathcal{O}}^+ R$ are submatrices of $\hat{\mathcal{O}} - \bar{\mathcal{O}}R$. By Weyl's inequality we have

$$|s_{d_0}(\hat{\mathcal{O}}^+) - s_{d_0}(\bar{\mathcal{O}}^+)| = |s_{d_0}(\hat{\mathcal{O}}^+) - s_{d_0}(\bar{\mathcal{O}}^+ R)| \leqslant |\hat{\mathcal{O}}^+ - \bar{\mathcal{O}}^+ R|_{S_\infty}$$

$$\leqslant |\hat{\mathcal{O}} - \bar{\mathcal{O}}R|_{S_\infty} \leqslant |\hat{\mathcal{O}} - \bar{\mathcal{O}}R|_{S_2} \leqslant \left(\frac{2}{\sqrt{2}-1} \right)^{1/2} \frac{|\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_2}}{s_{d_0}^{1/2}(Hg_0^*)}.$$

Again, noting that

$$|\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_2} \leqslant |\hat{\mathcal{H}}_T - \hat{\mathcal{H}}_{\check{d}_\xi}|_{S_2} + |\hat{\mathcal{H}}_T - Hg_0^*|_{S_2}$$

and since the truncated SVD minimizes the Hilbert-Schmidt norm cost, we obtain

$$|\hat{\mathcal{H}}_T - \hat{\mathcal{H}}_{\check{d}_\xi}|_{S_2} = \min_{\mathrm{rank}(H)\leqslant \check{d}_\xi} |\hat{\mathcal{H}}_T - H|_{S_2} \leqslant |\hat{\mathcal{H}}_T - Hg_0^*|_{S_2}.$$

Therefore,

$$|\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_2} \leqslant 2|\hat{\mathcal{H}}_T - Hg_0^*|_{S_2} \tag{2.36}$$

and

$$|s_{d_0}(\hat{\mathcal{O}}^+) - s_{d_0}(\bar{\mathcal{O}}^+)| \leqslant \left(\frac{2}{\sqrt{2}-1} \right)^{1/2} \frac{|\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_2}}{s_{d_0}^{1/2}(Hg_0^*)}.$$

In view of the condition $\frac{|\hat{\mathcal{H}}_T - Hg_0^*|_{S_2}}{s_{d_0}^{1/2}(Hg_0^*)} \leqslant \frac{(\sqrt{2}-1)^{1/2} s_{d_0}(\bar{\mathcal{O}}^+)}{2\sqrt{2}}$ in (2.32), we have

$$s_{d_0}(\hat{\mathcal{O}}^+) \geqslant s_{d_0}(\bar{\mathcal{O}}^+) - \left(\frac{2}{\sqrt{2}-1} \right)^{1/2} \frac{|\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_2}}{s_{d_0}^{1/2}(Hg_0^*)}$$

$$\geqslant s_{d_0}(\bar{\mathcal{O}}^+) - \frac{s_{d_0}(\bar{\mathcal{O}}^+)}{2} = \frac{s_{d_0}(\bar{\mathcal{O}}^+)}{2},$$

which together with (2.35) yields

$$|\hat{A} - R^*\bar{A}R|_{S_2} \leqslant \frac{2^{3/2}\left(1 + |\bar{A}|_{S_\infty}\right)}{\left(\sqrt{2}-1\right)^{1/2} s_{d_0}(\bar{\mathcal{O}}^+) s_{d_0}^{1/2}(Hg_0^*)} |\hat{\mathcal{H}}_{\check{d}_\xi} - Hg_0^*|_{S_2}.$$

$$\square$$

2.5. Non-asymptotic guarantees for Algorithm 1

Now we are ready to derive non-asymptotic results for the complete estimation procedure described in Algorithm 1. The algorithm starts with the data obtained from the partial observation of a single trajectory $(X_i, y_i)_{i=2T}^N$ of the system and aims to obtain a possible realization $(\hat{A}, \hat{B}, \hat{C})$. To this end, we require the inputs T_0, λ_0, and ξ to satisfy the following conditions:

1. $T_0 \geqslant d_0 + 1$, a known strict upper bound for the system order which can be taken reasonably large at the expense of an additional cost in terms of the sample complexity (2.15), as it directly relates to the dimension of the unknowns in the Hankel penalized regression part of the algorithm.

2. $\lambda_0 \simeq \lambda$ as defined in (2.1). This choice requires the additional knowledge of an upper bound for $\phi\sigma_u^2$ as defined in Corollary 2.1. An upper bound on ϕ is obtained from an upper bound on the system's $\mathcal{H}_\infty$-norm and an upper bound on the variances of the involved random variables. As argued in [39], the knowledge of an upper bound on the system $\mathcal{H}_\infty$-norm is a plausible assumption. It was also shown in [47] that such upper bound could be efficiently estimated.

3. $s_{d_0}^2(\bar{\mathcal{O}}^+) \geqslant 5\xi$. This choice is made to establish a detection threshold. The assumption on the knowledge of such a threshold is also common in the literature when studying threshold based estimator for high dimension regression problems, see [32, Corollary 2], [54, Equation (8)] or [29, Assumption 3].

Remark 2.3. *Since we want to provide an estimate up to a similarity transform of a minimal realization, as explained in Section 1.1, the fact that a realisation is minimal is equivalent to the order T Observability (resp Controllability) matrix being full column (resp row) rank. This implies that the two requirements, $T \geqslant d_0$ and $s_{d_0}(\mathcal{O}) > 0$ should be satisfied. Hence, the conditions $T_0 \geqslant d_0 + 1$ and $s_{d_0}^2(\bar{\mathcal{O}}^+) \geqslant 5\xi$ strengthen those requirements to a level that permits the estimation and rank detection.*
The condition $\lambda_0 \simeq \lambda$ relates to how the system's dynamic affects the estimation error, through the variance term ϕ in (2.13). Assuming the knowledge of an upper bound on it is again strengthening this requirement to a level that permits the estimation and rank detection.

Obtaining an adaptive, entirely data-driven estimation procedure without those three additional inputs falls beyond the scope of the current paper and is left as an interesting extension for future work.

Denote $g_{0,T} = [CB, CAB, \cdots, CA^{T-1}B]$ so that $Hg_{0,T}^* \in \mathcal{M}_{rT \times pT}$. Since $s_{d_0}^{1/2}(Hg_{0,d_0+1}^*) \geqslant s_{d_0}(\bar{\mathcal{O}}) \geqslant s_{d_0}(\bar{\mathcal{O}}^+)$, then our choice of $s_{d_0}^2(\bar{\mathcal{O}}^+) \geqslant 5\xi$ also implies

$$s_{d_0}(\bar{\mathcal{O}}^+)s_{d_0}^{1/2}(Hg_{0,T}^*) \geqslant 3\xi \quad \text{and} \quad \frac{(\sqrt{2}-1)^{1/2} s_{d_0}(\bar{\mathcal{O}}^+)s_{d_0}^{1/2}(Hg_0^*)}{2\sqrt{2}} \geqslant \xi. \tag{2.37}$$

According to Theorem 2.3 with the choices of inputs above and for

$$\bar{N} \geqslant c \left(TN_1 \vee T \log\frac{1}{\delta} \right)$$

when λ_0 is taken as

$$\lambda_0 = c\phi\sigma_u^2 \left(\sqrt{\frac{N_0}{N}} \vee \frac{\log(T_0)N_0}{N} \vee \frac{\sqrt{\log\frac{1}{\delta}}}{\sqrt{\bar{N}}} \vee \frac{\log(T_0)\log\frac{1}{\delta}}{N} \right),$$

the Hankel penalized estimator $\hat{g}$ defined in (2.1) satisfies on an event $\mathcal{B}$ of probability $\mathbb{P}(\mathcal{B}) \geqslant 1 - 2\delta$ a fast rate for the Hankel estimation spectral loss

$$\mathcal{L}_2^H(\hat{g}_0, g_0) \leqslant \frac{5\sqrt{2}}{3\sigma_u^2} d_0^{1/2}\lambda_0 T_0,$$

for some absolute fixed positive constant c and ϕ as defined in Corollary 2.1. By equation (2.37) the above choice of ξ implies $s_{d_0}(Hg_{0,T_0}^*) \geqslant 3\xi$, Thus Proposition 2.1 means that on the same event, once

we have

$$\bar{N} \geqslant \bar{N}_0 = cd_0 T N_0 \vee T_0 \log \tfrac{1}{\delta} \vee \tfrac{\phi^2 d_0 T_0^2}{\xi^2} \left(N_0 \vee \log \tfrac{1}{\delta} \right)$$
$$\vee \tfrac{\phi d_0^{1/2} T_0 \log(T_0)}{\xi} \left(N_0 \vee \log \tfrac{1}{\delta} \right),$$

we can ensure exact rank recovery for Algorithm 1 in the sense that $\check{d}_\xi$ defined by $\check{d}_\xi = \sum_{i=1}^{\text{rank}(H\hat{g}^*)} \mathbf{1}\{\hat{s}_i \geqslant 2\xi\}$ satisfies $\check{d}_\xi = d_0$. Hence, the event $\mathcal{B}$ is included in the event $\{\check{d}_\xi = d_0\}$.

In a similar fashion, using the Hankel penalized regression estimator in (2.2) with $T_1 = d_0 + 1$ to get an estimate for the Hankel matrix of the Markov parameters, then on a event $\mathcal{A}$ of probability $\mathbb{P}(\mathcal{A}) \geqslant 1 - 2\delta$ and for

$$\bar{N} \geqslant c(d_0 + 1)N_1 \vee (d_0 + 1) \log \frac{1}{\delta},$$

we have a fast rate for the new Hankel estimation spectral loss:

$$\mathcal{L}_2^H(\hat{g}_1, g_0) \leqslant \frac{5\sqrt{2}}{3\sigma_u^2} d_0^{1/2} \lambda_1 (d_0 + 1)$$

with

$$\lambda_1 = c\phi \left(\sqrt{\frac{N_0}{N}} \vee \frac{\log(d_0 + 1)N_0}{N} \vee \frac{\sqrt{\log \frac{1}{\delta}}}{\sqrt{N}} \vee \frac{\log(d_0 + 1)\log \frac{1}{\delta}}{N} \right).$$

This estimated matrix is then used in the Ho-Kalman based estimation procedure to obtain estimates for the system parameters $(\hat{A}, \hat{B}, \hat{C})$. As long as $\check{d}_\xi = d_0$, Theorem 2.4 guarantees that for values of $\bar{N}$ such that

$$|\hat{\mathcal{H}}_{d_0+1} - Hg^*_{0,d_0+1}|_{S_2} \leqslant \frac{\left(\sqrt{2} - 1\right)^{1/2} s_{d_0}(\bar{\mathcal{O}}^+) s_{d_0}^{1/2}(Hg_0^*)}{2\sqrt{2}}, \tag{2.38}$$

there exists an orthonormal matrix R satisfying

- up to the same orthonormal transformation, a fast estimation rate for $\bar{C}$ and $\bar{B}$ given as

$$|\hat{C} - \bar{C}R|_{S_2}^2 + |\hat{B} - R^*\bar{B}|_{S_2}^2 \leqslant \frac{2^{1/2}}{(\sqrt{2} - 1)^{1/2}} \frac{|\hat{\mathcal{H}}_{\check{d}_\xi} - Hg^*_{0,d_0+1}|_{S_2}}{s_{d_0}(\bar{\mathcal{O}}^+)},$$

- a fast estimation rate for $\bar{A}$ given as

$$|\hat{A} - R^*\bar{A}R|_{S_2} \leqslant \frac{2^{3/2} \left(1 + |\bar{A}|_{S_\infty} \right)}{\left(\sqrt{2} - 1 \right)^{1/2} s_{d_0}^2(\bar{\mathcal{O}}^+)} |\hat{\mathcal{H}}_{\check{d}_\xi} - Hg^*_{0,d_0+1}|_{S_2}.$$

From Proposition 2.1 equation (2.27) we have the following fast rate

$$|\hat{\mathcal{H}}_{\check{d}_\xi} - Hg^*_{0,d_0+1}|_{S_2} \leqslant \frac{10\sqrt{2}}{3\sigma_u^2} (d_0 + 1)^{3/2} \lambda_1.$$

From (2.37), the condition (2.38) is satisfied as long as $|\hat{\mathcal{H}}_{d_0+1} - H g^*_{0,d_0+1}|_{S_2} \leqslant \xi$. In view of the sample complexity given in (2.15) in Theorem 2.3, this is the case if

$$\bar{N} \geqslant \bar{N}_1 = c \frac{\phi^2 d_0^3}{\xi^2} \left(N_0 \vee \log \frac{1}{\delta} \right) \vee \frac{\phi d_0^{3/2} \log(d_0)}{\xi} \left(N_0 \vee \log \frac{1}{\delta} \right).$$

For Algorithm 1 to succeed we should have both the events $\{T_1 = d_0 + 1\}$ and $\mathcal{A}$ occurring. For $\bar{N} \geqslant \bar{N}_0 \vee \bar{N}_1 = \bar{N}_0$ we obtain

$$\mathbb{P}(\{T_1 = d_0 + 1\} \cap \mathcal{A}) \geqslant \mathbb{P}(\{T_1 = \check{d}_\xi + 1\} \cap \{\check{d}_\xi = d_0\} \cap \mathcal{A})$$
$$= \mathbb{P}(\{\check{d}_\xi = d_0\} \cap \mathcal{A}) \geqslant \mathbb{P}(\mathcal{A} \cap \mathcal{B})$$
$$\geqslant 1 - 4\delta,$$

where in the equality we used the fact that Algorithm 1 always chooses $\{T_1 = \check{d}_\xi + 1\}$. In the second inequality we use the fact that our choice $\bar{N} \geqslant \bar{N}_0$ ensures $\mathcal{B} \subset \{\check{d}_\xi = d_0\}$, and in the third inequality we use the fact that both $\mathbb{P}(\mathcal{A}) \geqslant 1 - 2\delta$ and $\mathbb{P}(\mathcal{B}) \geqslant 1 - 2\delta$ and a union bound.

We summarize the results of this discussion in the following

Theorem 2.5. *Algorithm 1 succeeds with probability at least $1 - 4\delta$ for all $\delta \in (0 \ e^{-1}/4)$ after observing $\bar{N} \geqslant \bar{N}_0$ samples from a single trajectory of the system (1.1) with the particular choices T_0, λ_0, λ_1, and ξ as described above and $T_1 = \check{d}_\xi + 1$. On the event of success we have*

- *Exact order recovery $\check{d}_\xi = d_0$;*
- *There exist an orthonormal matrix R for which the estimates for C and B satisfies fast estimation rates given as*

$$|\hat{C} - \bar{C} R|_{S_2} + |\hat{B} - R^* \bar{B}|_{S_2} \leqslant \frac{20(d_0 + 1)^{3/2} \lambda_1}{3(\sqrt{2} - 1)^{1/2} s_{d_0}(\bar{\mathcal{O}}^+) \sigma_u^2}.$$

- *For the same matrix R the estimate for A also satisfies fast estimation rate given as:*

$$|\hat{A} - R^* \bar{A} R|_{S_2} \leqslant \frac{10 \left(1 + |\bar{A}|_{S_\infty} \right) (d_0 + 1)^{3/2} \lambda_1}{\left(\sqrt{2} - 1 \right)^{1/2} s_{d_0}^2(\bar{\mathcal{O}}^+) \sigma_u^2}.$$

In particular, we have the following

Corollary 2.2. *Under the same condition as Theorem 2.5, the same inputs for Algorithm 1and for $\bar{N} \geqslant \bar{N}_0$, with the same probability on the even of success, the output satisfies the following.*

- *Fast rate for the Hankel estimation spectral loss:*

$$\mathcal{L}_2^{\mathcal{M}}(\hat{\mathcal{M}}, \bar{\mathcal{M}}) \leqslant \frac{10 \left(1 + |\bar{A}|_{S_\infty} \right) (d_0 + 1)^{3/2} \lambda_1}{\left(\sqrt{2} - 1 \right)^{1/2} s_{d_0}^2(\bar{\mathcal{O}}^+) \sigma_u^2}.$$

- *Sample complexity for the spectral loss: for any $\epsilon > 0$, to obtain $\mathcal{L}_2^H(\hat{g}, g_0) \leqslant \epsilon$ we need*

$$\bar{N} \gtrsim \frac{\phi^2(1+|\bar{A}|_{S_\infty})^2 d_0^3 N_0}{s_{d_0}^4(\bar{\mathcal{O}}^+)\epsilon^2} \vee \frac{(1+|\bar{A}|_{S_\infty})d_0^{3/2}\phi N_0 \log(T)}{s_{d_0}^2(\bar{\mathcal{O}}^+)\epsilon}$$

$$\vee \frac{\phi^2(1+|\bar{A}|_{S_\infty})^2 d_0^3 \log\frac{1}{\delta}}{s_{d_0}^4(\bar{\mathcal{O}}^+)\epsilon^2} \vee \frac{(1+|\bar{A}|_{S_\infty})d_0^{3/2}\phi \log(T) \log\frac{1}{\delta}}{s_{d_0}^2(\bar{\mathcal{O}}^+)\epsilon}. \quad (2.39)$$

It is clear from the condition (2.32) in Theorem 2.4 that the condition $\bar{N} \geqslant \bar{N}_0$ could be improved by using the control in term of $|\hat{\mathcal{H}}_T - Hg_0^*|_{S_\infty}$ instead of $|\hat{\mathcal{H}}_T - Hg_0^*|_{S_2}$. This can be done using the least square estimator for which it is easier to derive estimation bounds in term of $|\hat{\mathcal{H}}_T - Hg_0^*|_{S_\infty}$ such as in [33, Theorem 3.1], which would have the effect of reducing $\bar{N}_0$ by a factor of d_0 so that it scales like T_0^2. In the high dimension regime, T_0^2 is still a big price to pay in comparison with the sample complexity necessary for the second stage. This suggest the following open problem where we only change the condition on $\bar{N}$ in Theorem 2.3,

Open Problem 2.2. *Is there an algorithm that successfully learns a minimal realization of a Hidden state LTI state space system with high probability after observing*

$$\bar{N} \gtrsim \frac{\phi^2 d_0^2}{\xi^2}\left(N_0 \vee \log\frac{1}{\delta}\right)$$

and satisfies the fast rate for the Hankel estimation spectral loss of Theorem 2.5 given as

$$\mathcal{L}_2^{\mathcal{M}}(\hat{\mathcal{M}},\bar{\mathcal{M}}) \leqslant \frac{\left(1+|\bar{A}|_{S_\infty}\right)d_0^{3/2}\lambda_1}{s_{d_0}^2(\bar{\mathcal{O}}^+)\sigma_u^2}$$

up to logarithmic terms and lower order terms.

In [39, Theorem 5.3] the authors study the parameter estimation problem of a hidden state LTI state space system of unknown order where the derived results are in $|\cdot|_\infty$ norm. We have summarized their results in the Related Literature section in the introduction. The dominant term for the error bound (1.9) for their algorithm, after multiplying by $\hat{d}^{1/2}$ to get a bound in the $|\cdot|_2$ norm, is

$$\mathcal{L}_2^{\mathcal{M}}(\hat{\mathcal{M}},\bar{\mathcal{M}}) \lesssim \sqrt{\frac{r\hat{d}^4 + p\hat{d}^5 + \hat{d}^4\log(N/\delta)}{s_{\hat{d}}(\hat{H})N}}.$$

Our result improves this in a few ways. First, our result is provided in terms of the actual dimension d_0 and not the estimated dimension $\hat{d}$. Moreover, it reduces this dependence by a factor of d_0.

Regarding the burn in time $\bar{N}_0$, we have an explicit number of samples required for the result to hold $\bar{N} \geqslant \bar{N}_0$, unlike in [39, Theorem 5.3] where the result hold for $\bar{N} \geqslant N_*$ with $N_* < \infty$ (1.9). Moreover, upon inspection, a combination of [39, Proposition 13.4 and Proposition 13.7] shows that N_* depends exponentially in $\hat{d}$.

Finally, the presence of $\Gamma(\hat{H},\varepsilon) < \infty$ in (1.9) makes the result sensitive to all the singular values gapes of the matrix $\hat{H}$ and not just on the location of the smallest one. Our result, on the other hand, does not exhibit such behavior.

For the 'Reduced order Hankel penalized regression' part of Algorithm 1, let us consider a value $T_1 = \bar{d}_0 + 1$ where we take

$$\bar{d}_0 = \check{d}_\xi \vee \check{\eta} \quad \text{with} \quad \check{\eta} \geqslant \eta = \frac{\log\left(\psi_{\bar{A}}|\bar{C}|_{S_\infty}^2 / s_{d_0}(Hg_{0,d_0})\right)}{2\log\left(1/\rho(\bar{A})\right)}, \tag{2.40}$$

which on the even of success becomes $\bar{d}_0 = d_0 \vee \check{\eta}$. Applying 2.4 we obtain the same guarantees of 2.5 except that we replace d_0 with $\bar{d}$. In this case,

$$\bar{\mathcal{O}}^+ = \begin{bmatrix} \bar{C} \\ \vdots \\ \bar{C}\bar{A}^{\bar{d}-1} \end{bmatrix}$$

which is of rank d_0, since $k_0 \geqslant 1$. Moreover,

$$s_{d_0}^2(\bar{\mathcal{O}}^+) = \inf_{|u|_2=1}\left\{|\mathcal{O}u|_{S_2}^2 - |\bar{C}\bar{A}^{\bar{d}}|_{S_2}^2\right\} \geqslant s_{d_0}^2(\mathcal{O}) - |\bar{C}|_{S_\infty}^2|\bar{A}^{\bar{d}}|_{S_\infty}^2$$

$$= s_{d_0}(Hg_{0,\bar{d}}) - |\bar{C}|_{S_\infty}^2|\bar{A}^{\bar{d}}|_{S_\infty}^2 \geqslant s_{d_0}(Hg_{0,d_0}) - \psi_{\bar{A}}|\bar{C}|_{S_\infty}^2\rho(\bar{A})^{2\bar{d}}.$$

In view of the choice made in (2.40), we have

$$s_{d_0}^2(\bar{\mathcal{O}}^+) \geqslant \frac{1}{2}s_{d_0}(Hg_{0,d_0}). \tag{2.41}$$

Thus, we have the following

Corollary 2.3. *Under the same condition as Theorem 2.5, for the same inputs, except for the additional $\check{\eta}$, by taking $T_1 = \bar{d}_0 + 1$ and $s_{d_0}(Hg_{0,d_0}) \geqslant 10\xi$ with $\bar{d}_0$ defined in (2.40), Algorithm 1 succeeds for $\bar{N} \geqslant \bar{N}_0 \vee \bar{N}_\eta$ with*

$$\bar{N}_\eta \geqslant c\frac{\phi^2\eta^3}{\xi^2}\left(N_0 \vee \log\frac{1}{\delta}\right) \vee \frac{\phi\eta^{3/2}\log(\eta)}{\xi}\left(N_0 \vee \log\frac{1}{\delta}\right).$$

Furthermore, the output $\hat{\mathcal{M}}$ satisfies with probability at least $1 - 4\delta$ the following.

- *Fast rate for the Hankel estimation spectral loss:*

$$\mathcal{L}_2^{\mathcal{M}}(\hat{\mathcal{M}}, \bar{\mathcal{M}}) \leqslant \frac{20\left(1 + |\bar{A}|_{S_\infty}\right)(\bar{d}_0 + \eta + 1)^{3/2}\lambda_1}{\left(\sqrt{2} - 1\right)^{1/2}s_{d_0}(Hg_{0,d_0})\sigma_u^2}.$$

- *Sample complexity for the spectral loss: for any $\epsilon > 0$, to obtain $\mathcal{L}_2^H(\hat{g}, g_0) \leqslant \epsilon$, we need*

$$\bar{N} \gtrsim \frac{\phi^2(1 + |\bar{A}|_{S_\infty})^2(d_0 + \eta)^3 N_0}{s_{d_0}^2(Hg_{0,d_0})\epsilon^2} \vee \frac{(1 + |\bar{A}|_{S_\infty})(d_0 + \eta)^{3/2}\phi N_0\log(T)}{s_{d_0}(Hg_{0,d_0})\epsilon}$$

$$\vee \frac{\phi^2(1 + |\bar{A}|_{S_\infty})^2(d_0 + \eta)^3\log\frac{1}{\delta}}{s_{d_0}^2(Hg_{0,d_0})\epsilon^2}$$

$$\vee \frac{(1 + |\bar{A}|_{S_\infty})(d_0 + \eta)^{3/2} \phi \log(T) \log \frac{1}{\delta}}{s_{d_0}(H g_{0,d_0})\epsilon}. \quad (2.42)$$

References

[1] ABRAMOWITZ, M. and STEGUN, I. A. (1974). *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Dover.

[2] ADAMCZAK, R., LITVAK, A. E., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2011). Sharp bounds on the rate of convergence of the empirical covariance matrix. *Comptes Rendus Mathematique* **349** 195-200.

[3] ALFRIEND, K. T., VADALI, S. R., GURFIL, P., HOW, J. P. and BREGER, L. S. (2010). *Spacecraft Formation Flying*. Butterworth-Heinemann, Oxford.

[4] BAUER, D. and JANSSON, M. (2000). Analysis of the Asymptotic Properties of the MOESP Type of Subspace Algorithms. *Automatica* **36** 497–509.

[5] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37** 1705 – 1732.

[6] BOCZAR, R., MATNI, N. and RECHT, B. (2018). Finite-Data Performance Guarantees for the Output-Feedback Control of an Unknown System. In *2018 IEEE Conference on Decision and Control (CDC)* 2994-2999.

[7] BÖTTCHER, A. and SILBERMANN, B. (2012). *Introduction to Large Truncated Toeplitz Matrices*. Springer.

[8] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

[9] CAI, J.-F., QU, X., XU, W. and YE, G.-B. (2016). Robust recovery of complex exponential signals from random Gaussian projections via low rank Hankel matrix reconstruction. *Applied and computational harmonic analysis* **41** 470-490.

[10] CAMPI, M. C. and WEYER, E. (2002). Finite sample properties of system identification methods. *IEEE Transactions on Automatic Control* **47** 1329-1334.

[11] CHIUSO, A. and PICCI, G. (2001). Asymptotic variances of subspace estimates. In *Proceedings of the 40th IEEE Conference on Decision and Control (Cat. No.01CH37228)* **4** 3910-3915 vol.4.

[12] D. VECCHIO, R. M. M. (2017). *Biomolecular feedback systems*. Princeton University Press, USA.

[13] DEAN, S., MANIA, H., MATNI, N., RECHT, B. and TU, S. (2020). On the Sample Complexity of the Linear Quadratic Regulator. *Foundations of Computational Mathematics* **20** 1615-3383.

[14] DEISTLER, M., PETERNELL, K. and SCHERRER, W. (1995). Consistency and relative efficiency of subspace methods. *Autom.* **31** 1865-1875.

[15] DIRKSEN, S. (2015). Tail bounds via generic chaining. *Electronic Journal of Probability* **20**.

[16] DJEHICHE, B., MAZHAR, O. and ROJAS, C. R. (2019). Finite impulse response models: A non-asymptotic analysis of the least squares estimator.

[17] DJEHICHE, B., MAZHAR, O. and ROJAS, C. R. (2021). Finite impulse response models: A non-asymptotic analysis of the least squares estimator. *Bernoulli* **27** 976 – 1000.

[18] SHIRANI FARADONBEH, M. K., TEWARI, A. and MICHAILIDIS, G. (2018). Finite time identification in unstable linear systems. *Automatica* **96** 342-353.

[19] FURUTA, K. and WONGSAISUWAN, M. (1995). Discrete-time LQG dynamic controller design using plant Markov parameters. *Automatica* **31** 1317-1324.

[20] GINÉ, E. and NICKL, R. (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

[21] GLOVER, K. (1984). All optimal Hankel-norm approximations of linear multivariable systems and their L_∞-error bounds. *International Journal of Control* **39** 1115-1193.

[22] GOODWIN, G. C. and SIN, K. S. (2009). *Adaptive Filtering Prediction and Control*. Dover Publications, Inc., USA.

[23] GÜLER, O. (1991). On the Convergence of the Proximal Point Algorithm for Convex Minimization. *SIAM Journal on Control and Optimization* **29** 403-419.

[24] HAMILTON, J. D. (1994). *Time Series Analysis*, 1 ed. Princeton University Press.

[25] KNUDSEN, T. (2001). Consistency analysis of subspace identification methods based on a linear regression approach. *Automatica* **37** 81-89.

[26] KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics* **39** 2302 – 2329.

[27] LALE, S., AZIZZADENESHELI, K., HASSIBI, B. and ANANDKUMAR, A. (2021). Finite-time System Identification and Adaptive Control in Autoregressive Exogenous Systems. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control* (A. JADBABAIE, J. LYGEROS, G. J. PAPPAS, P. A. nbsp;PARRILO, B. RECHT, C. J. TOMLIN and M. N. ZEILINGER, eds.). *Proceedings of Machine Learning Research* **144** 967–979.

[28] LENNART, L. (1989). System identification - Theory for the user. *Autom.* **25** 475-476.

[29] LOUNICI, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics* **2** 90 – 102.

[30] MCMILLAN, B. (1952). Introduction to formal realizability theory–I. *The Bell System Technical Journal* **31** 217–279.

[31] MCMILLAN, B. (1952). Introduction to formal realizability theory–II. *The Bell System Technical Journal* **31** 541–600.

[32] MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics* **37** 246 – 270.

[33] OYMAK, S. and OZAY, N. (2021). Revisiting Ho-Kalman based system identification: robustness and finite-sample analysis. *IEEE Transactions on Automatic Control* 1-1.

[34] PARTINGTON, J. R. (1988). *An Introduction to Hankel Operators*. Cambridge University Press.

[35] PETERNELL, K., SCHERRER, W. and DEISTLER, M. (1996). Statistical analysis of novel subspace identification methods. *Signal Processing* **52** 161-177. Subspace Methods, Part II: System Identification.

[36] PEYPOUQUET, J. (2015). *Convex Optimization in Normed Spaces: Theory, Methods and Examples*. Springer.

[37] RUDELSON, M. and VERSHYNIN, R. (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.* **18** 1-9.

[38] SARKAR, T. and RAKHLIN, A. (2018). How fast can linear dynamical systems be learned? *CoRR* **abs/1812.01251**.

[39] SARKAR, T., RAKHLIN, A. and DAHLEH, M. A. (2021). Finite Time LTI System Identification. *Journal of Machine Learning Research* **22** 1-61.

[40] SIMCHOWITZ, M., MANIA, H., TU, S., JORDAN, M. I. and RECHT, B. (2018). Learning Without Mixing: Towards A Sharp Analysis of Linear System Identification. In *Proceedings of the 31st Conference On Learning Theory* (S. BUBECK, V. PERCHET and P. RIGOLLET, eds.). *Proceedings of Machine Learning Research* **75** 439–473.

[41] SKELTON, R. E. and SHI, G. (1994). The data-based LQG control problem. In *Proceedings of 1994 33rd IEEE Conference on Decision and Control* **2** 1447-1452 vol.2.

[42] STOKEY, N. and LUCAS, R. (1996). *Recursive methods in economic dynamics*. Harvard Univ. Press.

[43] SUN, Y., OYMAK, S. and FAZEL, M. (2020). Finite Sample System Identification: Optimal Rates and the Role of Regularization. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control* (A. M. BAYEN, A. JADBABAIE, G. PAPPAS, P. A. PARRILO, B. RECHT, C. TOMLIN and M. ZEILINGER, eds.). *Proceedings of Machine Learning Research* **120** 16–25.

[44] TALAGRAND, M. (2014). *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*. Springer, Berlin, Heidelberg.

[45] TSIAMIS, A. and PAPPAS, G. J. (2019). Finite Sample Analysis of Stochastic System Identification.

[46] TU, S., BOCZAR, R., PACKARD, A. and RECHT, B. (2017). Non-Asymptotic Analysis of Robust Control from Coarse-Grained Identification.

[47] TU, S., BOCZAR, R. and RECHT, B. (2018). On the Approximation of Toeplitz Operators for Nonparametric $\mathcal{H}_\infty$-norm Estimation. In *2018 Annual American Control Conference (ACC)* 1867-1872.

[48] TU, S., BOCZAR, R., SIMCHOWITZ, M., SOLTANOLKOTABI, M. and RECHT, B. (2016). Low-rank Solutions of Linear Matrix Equations via Procrustes Flow. In *Proceedings of The 33rd International Conference on Machine Learning* (M. F. BALCAN and K. Q. WEINBERGER, eds.). *Proceedings of Machine Learning Research* **48** 964–973.

[49] VAN OVERSCHEE, P. and DE MOOR, B. (1996). *Subspace Identification for Linear Systems: Theory - Implementation - Applications*. Springer US, Boston, MA.

[50] VERHAEGEN, M. and VERDULT, V. (2007). *Filtering and System Identification: A Least Squares Approach*. Cambridge University Press.

[51] VIBERG, M., WAHLBERG, B. and OTTERSTEN, B. (1997). Analysis of state space system identification methods based on instrumental variables and subspace fitting. *Automatica* **33** 1603-1616.

[52] VIDYASAGAR, M. and KARANDIKAR, R. L. (2008). A learning theory approach to system identification and stochastic adaptive control. *Journal of Process Control* **18** 421-430. Festschrift honouring Professor Dale Seborg.

[53] WEYER, E., WILLIAMSON, R. C. and MAREELS, I. M. Y. (1999). Finite sample properties of linear model identification. *IEEE Transactions on Automatic Control* **44** 1370-1383.

[54] ZHAO, P. and YU, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research* **7** 2541-2563.

Appendix A: Proofs of the main probabilistic results

In this appendix we gather the proofs of the main results stated in Section 2.2 namely the proofs of Theorem 2.1 and of (2.5) of Theorem 2.2. The proofs of the other parts of Theorem 2.2 are similar to those of (2.5) and are given in Appendix C for completeness. Their proofs use extensively generic chaining estimates. Thus, we start by recalling few concepts from the generic chaining literature to fix some notation and refer to [44, 20] for more on the topic. Let $(\mathcal{A}, d)$ be a metric space. The distance of a point $t \in \mathcal{A}$ to a subset $\mathbb{A} \subseteq \mathcal{A}$ is defined as

$$d(t, \mathbb{A}) = \inf_{s \in \mathbb{A}} d(t, s).$$

The diameter of the set $\mathbb{A}$ is

$$\Delta(\mathbb{A}) = \sup_{(s,t) \in \mathbb{A}^2} d(t, s),$$

and the covering number $N(\mathcal{A}, d, u)$ is the smallest number of balls in $(\mathcal{A}, d)$ of radius less than u needed to cover $\mathcal{A}$ (*i.e.*, whose union includes $\mathcal{A}$). A ball of center $c \in \mathcal{A}$ and radius $r \geq 0$ with respect to a distance d or a metric $\|\cdot\|$ will be denoted $B_d(c, r)$ or $B_{\|\cdot\|}(c, r)$, respectively.

The gamma-α functional $\gamma_\alpha(\mathcal{A}, d)$ for the metric space $(\mathcal{A}, d)$ and its corresponding upper bound by the Dudley chaining integral are defined as follows.

$$\gamma_\alpha(\mathcal{A}, d) := \inf \sup_{t \in \mathcal{A}} \sum_{r=0}^{\infty} 2^{r/\alpha} d(t, \mathbb{A}_r) \lesssim \int_0^{\Delta(\mathcal{A})} (\log N(\mathcal{A}, d, u))^{1/\alpha} du, \tag{A.1}$$

where the infimum is taken over all sequences of sets $(\mathbb{A}_r)_{r \in \mathbb{N}}$ in $\mathcal{A}$ with $|\mathbb{A}_0| = 1$ and $|\mathbb{A}_r| \leqslant 2^{2^r}$ ([44]). If $d(x, y) = \|x - y\|$ for some norm $\|\cdot\|$ as it is usually the case, we also use the notation $\gamma_\alpha(\mathcal{A}, \|\cdot\|)$ for $\gamma_\alpha(\mathcal{A}, d)$.

A.1. Isometric Property for the covariates of the input

Proof of Theorem 2.1. The result is an extension of [17, Thoerem 3.4] to the multidimensional case, and in the same spirit, we start the proof with a decomposition of the operator norm $|X^*X - \mathbb{E}(X^*X)|_{S_\infty}$ into the sum of 3 terms. To that end, we start by defining, for $k \in [\![1, 2T - 1]\!]$, the following shifted matrices:

$$L_k = \begin{bmatrix} 0 \cdots 0 \ u_{2T-1} \ u_{2T} \cdots u_{\bar{N}} \ 0 \cdots 0 \end{bmatrix}^*,$$

where x_{2T-1} is at the position $r(k - 1) + 1$. Then we define the matrices

$$L = [L_1 L_2 \ldots L_{2T-1}] \qquad \text{and} \qquad S = X - L,$$

to get a decomposition $X = L + S$ where L and S are independent of each other and have a shifted diagonal structure. Thus, we have

$$X^*X = L^*L + S^*S + S^*L + L^*S.$$

Using this decomposition the operator norm of deviation of X^*X is upper bounded by

$$|X^*X - \mathbb{E}(X^*X)|_{S_\infty} \leqslant |L^*L - \sigma_u^2(N - 4T + 3)I_{(2T-1)r}|_{S_\infty}$$
$$+ \sigma_u^2(2T - 2) + 2|S^*L|_{S_\infty} + |S^*S|_{S_\infty}. \tag{A.2}$$

We thus need to derive high probability bounds for the last three terms. Below, we give the derivation for the first term. The others are treated similarly, and the contribution of the first term dominates their contribution.

We start by relating the operator norm of $L^*L - \sigma_u^2(N - 4T + 3)I_{(2T-1)r}$ to the supremum of a multiplication process. Since the columns of L are shifted versions of each others, we have

$$L^*L = \begin{bmatrix} L_1^*L_1 & L_1^*L_2 & L_1^*L_{2T-1} \\ L_2^*L_1 & L_1^*L_1 & L_1^*L_{2T-2} \\ & & \\ L_{2T-1}^*L_1 & L_{2T-2}^*L_1 & L_1^*L_1 \end{bmatrix}.$$

Define the block Toeplitz operator $\mathcal{T} : l_{\mathbb{R}^r}(\mathbb{Z}) \to l_{\mathbb{R}^r}(\mathbb{Z})$ by the infinite diagonals of block matrices given

$$\mathcal{T}_0 = L_1^*L_1 - \sigma_u^2(N - 4T + 3)I_r, \ \mathcal{T}_l = L_1^*L_{l+1}, \ \text{and} \ \mathcal{T}_{-l} = \mathcal{T}_l^* \ \text{for} \ l \in [\![1, 2T - 2]\!].$$

The corresponding multiplication polynomial defined for $x \in [0, 1]$ is given by

$$p(x) = \sum_{l=-2T+2}^{2T-2} \mathcal{T}_l e^{2i\pi lx}.$$

Since $L^*L - \sigma_u^2(N - 4T + 3)I_{(2T-1)r}$ is a submatrix of $\mathcal{T}$, we have

$$|L^*L - \sigma_u^2(N - 4T + 3)I_{(2T-1)r}|_{S_\infty} \leqslant |\mathcal{T}|_{2\to2} = \sup_{x\in[0,1]} |p(x)|_{S_\infty}, \tag{A.3}$$

where $|\cdot|_{2\to2}$ stands for the operator norm. The last supremum can also be expressed as

$$\sup_{x\in[0\ 1]} \sup_{\substack{|v|_2=1 \\ |w|_2=1}} \Big| \sum_{j=2T-1}^{\bar{N}} (\langle u_j, v\rangle\langle u_j, w\rangle - \sigma_u^2\langle v, w\rangle)$$

$$+ \sum_{l=1}^{2T-2} \sum_{j=2T-1+l}^{\bar{N}} \langle u_j, v\rangle\langle u_{j-l}, w\rangle e^{2i\pi lx}$$

$$+ \sum_{l=1}^{2T-2} \sum_{j=2T-1+l}^{\bar{N}} \langle u_{j-l}, v\rangle\langle u_j, w\rangle e^{-2i\pi lx}\Big|.$$

Consider the block Toeplitz matrix $\mathcal{H} \in \mathcal{M}_{(N-4T+3)r\times(N-4T+3)r}(\mathbb{R})$ with block constant diagonals made of matrices $\mathcal{H}_l \in \mathcal{M}_{r\times r}(\mathbb{R})$ with (j, k) entries

$$\mathcal{H}_l(x, v, w) = e^{2i\pi lx}vw^*\mathbb{1}\{l \in [\![0, 2T - 2]\!]\} \ \text{and} \ \mathcal{H}_l(x, v, w) = \mathcal{H}_{-l}(x, v, w)^* \ \text{for} \ l < 0.$$

Taking $u = [u_{2T-1}^*, \cdots, u_{N-2T-1}^*]^*$ we obtain

$$|L^*L - \sigma_u^2(N - 4T + 3)I_{(2T-1)r}|_{S_\infty} \leqslant \sup_{x\in[0,1]} |p(x)|_{S_\infty}$$

$$= \sup_{x\in[0\ 1]} \sup_{\substack{|v|_2=1 \\ |w|_2=1}} |\langle u, \mathcal{H}(x, v, w)u\rangle|.$$

This defines a second order chaos process $\xi_{x,v,w} = \langle u, \mathcal{H}(x,v,w)u \rangle$. We control its deviation using the Hanson-Wright inequality [37] to get, for all $t > 0$ with probability at least $1 - 2e^{-ct}$,

$$
|\chi_{x_1,v_1,w_1} - \chi_{x_2,v_2,w_2}| \leqslant \sqrt{t}d_2((x_1,v_1,w_1),(x_2,v_2,w_2)) \\
+ td_\infty((x_1,v_1,w_1),(x_2,v_2,w_2)),
$$

where

$$
d_\infty((x_1,v_1,w_1),(x_2,v_2,w_2)) := |\mathcal{H}(x_1,v_1,w_1) - \mathcal{H}(x_2,v_2,w_2)|_{S_\infty}
$$

and

$$
d_2((x_1,v_1,w_1),(x_2,v_2,w_2)) := |\mathcal{H}(x_1,v_1,w_1) - \mathcal{H}(x_2,v_2,w_2)|_{S_2}.
$$

The generic chaining result in [44, Theorem 2.2.23] and [15, Theorem 3.5] provides us with the following bound for the supremum of such mixed tail process for $t \geqslant 1$:

$$
\mathbb{P}\left(\sup_{x \in [0,1]} |p(x)| \geqslant c\sigma_u^2\left(E + \sqrt{t}V + tU\right)\right) \leqslant 2\exp(-u), \tag{A.4}
$$

where

$$
E = \gamma_2([0,1] \times \mathbb{S}_2^{r-1} \times \mathbb{S}_2^{r-1}, d_2) + \gamma_1([0,1] \times \mathbb{S}_2^{r-1} \times \mathbb{S}_2^{r-1}, d_\infty),
$$
$$
V = \Delta_2([0,1] \times \mathbb{S}_2^{r-1} \times \mathbb{S}_2^{r-1}, d_2), \quad U = \Delta_\infty([0,1] \times \mathbb{S}_2^{r-1} \times \mathbb{S}_2^{r-1}, d_\infty).
$$

To conclude the proof, it suffices to estimate these three terms. We start with few inequalities to simplify the involved the norm distances

$$
d_\infty((x_1,v_1,w_1),(x_2,v_2,w_2)) := |\mathcal{H}(x_1,v_1,w_1) - \mathcal{H}(x_2,v_2,w_2)|_{S_\infty}
$$

$$
\leqslant 2\sup_{y \in [0\ 1]}\left|\sum_{l=1}^{2T-2} e^{2i\pi l(x_1+y)}u_1v_1^* - e^{2i\pi l(x_2+y)}u_2v_2^*\right|_{S_\infty} + |u_1v_1^* - u_2v_2^*|_{S_\infty}
$$

$$
\leqslant 2\sup_{y \in [0\ 1]}\left|\sum_{l=1}^{2T-2} e^{2i\pi l(x_1+y)} - e^{2i\pi l(x_2+y)}\right| + (4T-3)|u_1v_1^* - u_2v_2^*|_{S_\infty}
$$

$$
\lesssim T^2|x_1 - x_2| + T|u_1 - u_2|_2 + T|v_1 - v_2|_2,
$$

where we used Proposition B.1 and the Liptchitz property of the complex exponential in the last step. Similarly, we have

$$
d_2((x_1,v_1,w_1),(x_2,v_2,w_2)) := |\mathcal{H}(x_1,v_1,w_1) - \mathcal{H}(x_2,v_2,w_2)|_{S_2}
$$

$$
\leqslant \sqrt{N - 4T + 3}\left(2\left(\sum_{l=1}^{2T-2}\left|e^{2i\pi lx_1}u_1v_1^* - e^{2i\pi lx_2}u_2v_2^*\right|_{S_2}^2\right)^{1/2} + |u_1v_1^* - u_2v_2^*|_{S_2}\right)
$$

$$
\leqslant \sqrt{N - 4T + 3}\left(2\left|\sum_{l=1}^{2T-2}(e^{2i\pi lx_1} - e^{2i\pi lx_2})^2\right|^{1/2} + (2\sqrt{T-1}+1)|u_1v_1^* - u_2v_2^*|_{S_2}\right)
$$

$$\lesssim \sqrt{N - 4T + 3} \left(T^{3/2} |x_1 - x_2| + T^{1/2} |u_1 - u_2|_2 + T^{1/2} |v_1 - v_2|_2 \right).$$

The radii U and V become

$$U \lesssim T \text{ and } V \lesssim \sqrt{(N - 4T + 3)T}. \tag{A.5}$$

The γ_1 functional is evaluated as

$$\gamma_1 \left([0,1] \times \mathbb{S}_2^{r-1} \times \mathbb{S}_2^{r-1}, d_\infty \right) \lesssim \gamma_1 \left([0,1], T^2 |\cdot| \right) + \gamma_1 \left(\mathbb{S}_2^{r-1}, T |\cdot|_2 \right) \tag{A.6}$$

$$\int_0^T \ln N([0,1], T^2 |\cdot|, u) du + \int_0^T \ln N(\mathbb{S}_2^{r-1}, T |\cdot|_2, u) du \tag{A.7}$$

$$= \int_0^T \ln \frac{T^2}{u} du + \int_0^T \ln \left(\frac{T}{u} \right)^r du \simeq T(\ln(T) + r). \tag{A.8}$$

Similarly, we can evaluate the γ_2 functional to get

$$\gamma_2 \left([0,1] \times \mathbb{S}_2^{r-1} \times \mathbb{S}_2^{r-1}, d_2 \right) \lesssim \sqrt{T(N - 4T + 3)(\ln(T) + r)}, \tag{A.9}$$

Putting this last result together with (A.5), (A.8), (A.9) and (A.4) gives with probability at least $1 - e^{-t}$ for $t \geqslant 1$:

$$|L^* L - \sigma_u^2 (N - 4T + 3) I_{(2T-1)r}|_{S_\infty} \lesssim \sigma_u^2 \Big(T(\ln(T) + r)$$
$$+ \sqrt{T(N - 4T + 3)(\ln(T) + r)} + \sqrt{T(N - 4T + 3)} \sqrt{t} + Tt \Big).$$

We can bound the second and third term in (A.2) by modifying the argument in [17, Thoerem 3.4] the same way we did here for the first term. This give us with probability at least $1 - e^{-t}$, for all $t \geqslant 1$,

$$|S^* S|_{S_\infty} \lesssim \sigma_u^2 \Big(T(\ln(T) + r) + Tt \Big) \tag{A.10}$$

and

$$|L^* S|_{S_\infty} \lesssim \sigma_u^2 \Big(T(\ln(T) + r) + Tt \Big).$$

A straightforward union bound implies that with probability at least $1 - e^{-t}$, for all $t \geqslant 1$, we have

$$|X^* X - \mathbb{E}(X^* X)|_{S_\infty} \lesssim \sigma_u^2 \Big(T(\ln(T) + r) + \sqrt{T(N - 4T + 3)(\ln(T) + r)}$$
$$+ \sqrt{(N - 4T + 3)T} t^{1/2} + Tt \Big) \sigma_u^2 \Big).$$

This last expression directly gives the claimed result in the theorem after normalization. $\square$

Proof of (2.5) ***in Theorem*** *2.2.* Define the permuted index

$$\bar{l} = \begin{cases} l - 2T & \text{if } T + 1 \leqslant l \leqslant 2T - 1, \\ l & \text{otherwise.} \end{cases}$$

and

$$x = [u_0^*, u_1^*, \ldots, u_{N-2}^*, u_{N-1}^*]^* \in \mathbb{R}^{Nr}$$

$$U_l = [u_{2T-l}, u_{2T+1-l}, \ldots, u_{N-l}] \in \mathcal{M}_{r \times \bar{N}}(\mathbb{R}).$$

In view of the definition of $H^{\dagger^*}$ in (2.4) we have

$$H^{\dagger^*} X^* \bar{X} \bar{g}^* = \begin{bmatrix} (U_1 \bar{X} \bar{g}^*)^* & \frac{1}{2}(U_2 \bar{X} \bar{g}^*)^* & \cdots & \frac{1}{T}(U_T \bar{X} \bar{g}^*)^* \\ \frac{1}{2}(U_2 \bar{X} \bar{g}^*)^* & \frac{1}{3}(U_3 \bar{X} \bar{g}^*)^* & \cdots & \frac{1}{T-1}(U_{T+1} \bar{X} \bar{g}^*)^* \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{T}(U_T \bar{X} \bar{g}^*)^* & \frac{1}{T-1}(U_{T+1} \bar{X} \bar{g}^*)^* & \cdots & (U_{2T-1} \bar{X} \bar{g}^*)^* \end{bmatrix}.$$

So we want to find a high probability bound on the operator norm of the matrix $H^{\dagger^*} X^* \bar{X} \bar{g}$. Define the infinite block Hankel operator $\mathcal{H} : l_2(\mathbb{N}) \to l_2(\mathbb{N})$ by the $\mathcal{M}_{p \times r}(\mathbb{R})$ blocks

$$\mathcal{H}_{i,j} = \begin{cases} 1/|\bar{l}|(U_l \bar{X} \bar{g})^* & \text{for} \quad (i,j) \in \mathbb{N}^2, \text{ and } 1 \leqslant |i - j| = l \leqslant 2T - 1, \\ 0 & \text{otherwise}. \end{cases}$$

Then

$$|H^{\dagger^*} X^* \bar{X} \bar{g}^*|_{S_\infty} \leqslant |\mathcal{H}|_{2 \to 2}$$

where $|\cdot|_{2 \to 2}$ stands for the operator norm from $l_2(\mathbb{Z})$ to $l_2(\mathbb{Z})$. The corresponding multiplication polynomial defined for $u \in [0, 1]$ is given by

$$p(u) = \sum_{l=1}^{2T-1} \frac{1}{|\bar{l}|} U_l \bar{X} \bar{g}^* \exp(i 2\pi \bar{l} u).$$

where we have used a permuted Fourier basis by the mapping $l \to \bar{l}$. Thus, using Proposition B.2, we obtain

$$|H^{\dagger^*} X^* \bar{X} \bar{g}^*|_{S_\infty} \leqslant \sup_{u \in [0\ 1]} |p(u)|_{S_\infty} = \sup_{u \in [0,\ 1]} \left| \sum_{l=1}^{2T-1} \frac{e^{i 2\pi \bar{l} u}}{|\bar{l}|} U_l \bar{X} \bar{g}^* \right|_{S_\infty}$$

$$= \sup_{u \in [0\ 1]} \sup_{\substack{|v|_2 = 1 \\ |w|_2 = 1}} \left| \left\langle \sum_{l=1}^{2T-1} \frac{e^{i 2\pi \bar{l} u}}{|\bar{l}|} U_l \bar{X} \bar{g}^*, vw^* \right\rangle \right|$$

$$= \sup_{u \in [0\ 1]} \sup_{\substack{|v|_2 = 1 \\ |w|_2 = 1}} \left| \left\langle \bar{X} \bar{g}^*, \sum_{l=1}^{2T-1} \frac{e^{i 2\pi \bar{l} u}}{|\bar{l}|} U_l^* vw^* \right\rangle \right|.$$

Define, for $l \in [0\ N - 2T]$, the vectors $G_l \in \mathcal{M}_{p \times rN}(\mathbb{R})$ as

$$G_l = [C_0 A_0^{2T-1+l} B_0, C_0 A_0^{2T+l} B_0, \ldots, C_0 A_0^{2T-1} B_0, 0, \ldots, 0],$$

and for $u \in [0 \ 1]$ the matrix valued functions $w_{u,v,w,l} \in \mathcal{M}_{r \times p}(\mathbb{C})$ by $w_{u,v,w,l} = \frac{\exp(i2\pi \bar{l}u)}{|l|} vw^*$ for $l \in [1 \ 2T-1]$ and the matrix valued functions $W_{u,v,w,l} \in \mathcal{M}_{p \times Nr}(\mathbb{C})$ by

$$W_{u,v,w,k} = \left[\ 0 \ w^*_{u,v,w,2T-1} \ w^*_{u,v,w,2T-2} \cdots w^*_{u,v,w,1} \ \right],$$

with the 1st zero a the kth-position. Define G as $G = [G_0^*, \ldots, G_{N-2T}^*]^*$ and $W_{u,v,w}$ as $W_{u,v,w} = [W^*_{u,v,w,1}, \ldots, W^*_{u,v,w,N-2T-1}]^*$ satisfying

$$\left\langle \bar{X}\bar{g}^*, \sum_{l=1}^{2T-1} \frac{e^{i2\pi \bar{l}u}}{|\bar{l}|} U_l^* vw^* \right\rangle = \left\langle W_{u,v,w}x, Gx \right\rangle.$$

This gives

$$|H^{\dagger^*} X^* \bar{X}\bar{g}|_{S_\infty} \leqslant \sup_{u \in [0 \ 1]} \sup_{\substack{|v|_2=1 \\ |w|_2=1}} |\langle G^* W_{u,v,w}x, x \rangle|$$

which is the supremum of a second order chaos process defined by

$$\chi_{u,v,w} = \langle G^* W_{u,v,w}x, x \rangle.$$

To control the increment of the process we use Hanson Wright inequality [37] which yields

$$\mathbb{P}\left(|\chi_{u,v,w} - \mathbb{E}(\chi_{u,v,w})| \geqslant c\left(\sqrt{t}|G^* W_{u,v,w}|_{S_2} + t|G^* W_{u,v,w}|_{S_\infty} \right) \right) \leqslant 2\exp(-ct).$$

Since $\mathbb{E}(\chi_{u,v,w}) = 0$, we obtain a mixed tail process with probability $1 - e^{-t}$

$$\begin{aligned}
|\chi_{u_1,v_1,w_1} - \chi_{u_2,v_2,w_2}| &\leqslant c\Big(\sqrt{t}|G^* \left(W_{u_1,v_1,w_1} - W_{u_2,v_2,w_2} \right)|_{S_2} \\
&\qquad + t|G^* \left(W_{u_1,v_1,w_1} - W_{u_2,v_2,w_2} \right)|_{S_\infty} \Big) \\
&\leqslant c|G|_{S_\infty} \left(\sqrt{t}|W_{u_1,v_1,w_1} - W_{u_2,v_2,w_2}|_{S_2} \right. \\
&\qquad \left. + t|W_{u_1,v_1,w_1} - W_{u_2,v_2,w_2}|_{S_\infty} \right).
\end{aligned} \tag{A.11}$$

Consider the pseudo-distances d_2 and d_∞ defined on $[0,1] \times \mathbb{S}_2^{r-1} \times \mathbb{S}_2^{p-1}$ by

$$d_2((u_1, v_1, w_1), (u_2, v_2, w_2)) = |W_{u_1,v_1,w_1} - W_{u_2,v_2,w_2}|_{S_2}$$
$$d_\infty((u_1, v_1, w_1), (u_2, v_2, w_2)) = |W_{u_1,v_1,w_1} - W_{u_2,v_2,w_2}|_{S_\infty}.$$

The generic chaining result proved independently in [44, Theorem 2.2.23] and [15, Theorem 3.5] provides the following bound for the supremum of such mixed tail process for $t \geqslant 1$:

$$\mathbb{P}\left(\sup_{(u,v,w) \in [0,1] \times \mathbb{S}_2^{r-1} \times \mathbb{S}_2^{p-1}} |\chi_{u,v,w}| \geqslant C\sigma_u^2 |G|_{S_\infty} \left(E + \sqrt{t}\Delta_{S_2} + t\Delta_{S_\infty} \right) \right) \leqslant 2\exp(-t), \quad \text{(A.12)}$$

where

$$E = \gamma_2([0,1] \times \mathbb{S}_2^{r-1} \times \mathbb{S}_2^{p-1}, d_2) + \gamma_1([0,1] \times \mathbb{S}_2^{r-1} \times \mathbb{S}_2^{p-1}, d_\infty),$$

$$\Delta_{S_2} = \sup_{(u_1,v_1,w_1),(u_2,v_2,w_2)\in[0,1]\times\mathbb{S}_2^{r-1}\times\mathbb{S}_2^{p-1}} d_2((u_1,v_1,w_1),(u_2,v_2,w_2)),$$

$$\Delta_{S_\infty} = \sup_{(u_1,v_1,w_1),(u_2,v_2,w_2)\in[0,1]\times\mathbb{S}_2^{r-1}\times\mathbb{S}_2^{p-1}} d_\infty((u_1,v_1,w_1),(u_2,v_2,w_2)).$$

To conclude the proof, it suffices to estimate these four terms. We start with the terms which involves the distance d_∞. An estimate of d_∞ is obtained by seeing $W_{u,v,w}$ as a sub-matrix of an infinite Toeplitz matrix $\tilde{W}_{u,v,w}$ defined by,

$$\tilde{W}_{u,v,w} = \left[\mathbb{1}_{1\leqslant j-k+1\leqslant 2T-1}(w_{u,v,w}^*)_{j-k+1}\right]_{(j,k)\in\mathbb{Z}^2}$$

and the corresponding multiplication polynomial is

$$q_{u,v,w}(z) = \sum_{l=1}^{2T-1} \frac{\exp\left(i2\pi\bar{l}(u+z)\right)}{|\bar{l}|} vw^*.$$

The diameter Δ_{S_∞} becomes

$$\Delta_{S_\infty} \leqslant 2 \sup_{\substack{u\in[0\ 1] \\ |v|_2=1 \\ |w|_2=1}} \sup |\tilde{W}_{u,v,w}|_{2\to2} = 2 \sup_{\substack{u\in[0\ 1] \\ z\in[0\ 1]}} \sup_{\substack{|v|_2=1 \\ |w|_2=1}} |q(z)|_{S_2}$$

$$= 2 \sup_{\substack{u\in[0\ 1] \\ z\in[0\ 1]}} \left|\sum_{l=1}^{2T-1} \frac{\exp\left(i2\pi\bar{l}(u+z)\right)}{|\bar{l}|}\right| \leqslant \sum_{l=1}^{T-1} \frac{2}{|\bar{l}|} \lesssim \log(T).$$

Using the fact that the complex exponential is Lipschitz, we have

$$d_\infty((u_1,v_1,w_2),(u_1,v_1,w_2)) \leqslant \sum_{l=1}^{T-1} \frac{1}{|\bar{l}|}|v_1w_1^* - v_2w_2^*|_{S_2}$$

$$+ \left|\sum_{l=1}^{T-1} \frac{1}{|\bar{l}|}(e^{i2\pi l(u_1-z)} - e^{i2\pi l(u_2-z)})\right|$$

$$\lesssim \log(T)|v_1 - v_2|_2 + \log(T)|w_1 - w_2|_2 + T|u_1 - u_2|.$$

The γ_1 functional is evaluated as

$$\gamma_1([0,1] \times \mathbb{S}_2^{r-1} \times \mathbb{S}_2^{p-1}, d_{S_\infty}) \leqslant \gamma_1([0,1], T|\cdot|) + \gamma_1(\mathbb{S}_2^{r-1}, \log(T)|\cdot|_2)$$

$$+ \gamma_1(\mathbb{S}_2^{p-1}, \log(T)|\cdot|_2)$$

$$\lesssim \int_0^{\Delta_{S_\infty}} \log N([0,1], T|\cdot|, u)du + \int_0^{\Delta_{S_\infty}} \log N(\mathbb{S}_2^{r-1}, \log(T)|\cdot|_2, u)du$$

$$+ \int_0^{\Delta_{S_\infty}} \log N(\mathbb{S}_2^{p-1}, \log(T)|\cdot|_2, u)du$$

$$\lesssim \int_0^{\log(T)} \log\left(\frac{T}{u}\right) du + \int_0^{\log(T)} \log\left(\frac{\log(T)}{u}\right)^r du + \int_0^{\log(T)} \log\left(\frac{\log(T)}{u}\right)^p du$$

$$\leqslant \log^2(T) - \int_0^{\log T} \log(u) du + (p+r)\left(\log(T)\log\log(T) - \int_0^{\log T} \log(u) du\right)$$

$$\leqslant \log^2(T) + (p+r)\log(T).$$

We now turn to the terms involving the pseudo-distance $|W_{u_1,v_1,w_1} - W_{u_2,v_2,w_2}|_{S_2}$. Again the complex exponential is -Lipschitz, we have

$$d_{S_2}((u_1,v_1,w_1),(u_2,v_2,w_2))$$

$$\leqslant \sqrt{\bar{N}}\left(\left(\sum_{l=1}^{2T-1}\frac{1}{l^2}|v_1 w_1^* - v_2 w_2^*|_{S_2}^2\right)^{1/2} + \left(\sum_{l=1}^{2T-1}\frac{1}{l^2}|\exp(i2\pi l u_1) - \exp(i2\pi l u_2)|^2\right)^{1/2}\right)$$

$$\lesssim \sqrt{\bar{N}}\left(|v_1 - v_2|_2 + |w_1 - w_2|_2 + \sqrt{T}|u_1 - u_2|\right).$$

The radius Δ_{S_2} satisfies

$$\Delta_{S_2} = 2\sup_{u\in[0\ 1]}|W_u|_{S_2} \simeq \sqrt{\bar{N}}.$$

The γ_2 functional satisfies

$$\gamma_2([0,1]\times\mathbb{S}_2^{r-1}\times\mathbb{S}_2^{p-1}, d_2) \leqslant \gamma_2([0,1],\sqrt{\bar{N}T}|\cdot|)$$

$$+ \gamma_2(\mathbb{S}_2^{r-1},\sqrt{\bar{N}}|\cdot|_2) + \gamma_2(\mathbb{S}_2^{p-1},\sqrt{\bar{N}}|\cdot|_2)$$

$$\lesssim \int_0^{\Delta_{S_2}}\left(\log N([0,1],\sqrt{\bar{N}T}|\cdot|,u)\right)^{1/2} du + \int_0^{\Delta_{S_2}}\left(\log N(\mathbb{S}_2^{r-1},\sqrt{\bar{N}}|\cdot|_2,u)\right)^{1/2} du$$

$$+ \int_0^{\Delta_{S_2}}\left(\log N(\mathbb{S}_2^{p-1},\sqrt{\bar{N}}|\cdot|_2,u)\right)^{1/2} du$$

$$= \int_0^{\sqrt{\bar{N}}}\left(\log\frac{\sqrt{\bar{N}T}}{u}\right)^{1/2} du + \int_0^{\sqrt{\bar{N}}}\left(\log\left(\frac{3\sqrt{\bar{N}}}{u}\right)^r\right)^{1/2} du + \int_0^{\sqrt{\bar{N}}}\left(\log\left(\frac{3\sqrt{\bar{N}}}{u}\right)^p\right)^{1/2} du$$

$$= \sqrt{\bar{N}T}\int_{\sqrt{\log(T)}}^\infty t^2\exp(-t^2/2)dt + 6\sqrt{\bar{N}r}\int_{\sqrt{\log 3}}^\infty t^2\exp(-t^2/2)dt$$

$$+ 6\sqrt{\bar{N}p}\int_{\sqrt{\log 3}}^\infty t^2\exp(-t^2/2)dt$$

$$\lesssim \sqrt{\bar{N}T}\left(\frac{\sqrt{\log(T)}}{\sqrt{T}} + \frac{1}{\sqrt{T}\sqrt{\log(T)}}\right) + \sqrt{\bar{N}(p+r)} \lesssim \sqrt{\bar{N}\log(T)} + \sqrt{\bar{N}(p+r)},$$

where in the last step we did and integration by parts and used [1, Formula 7.1.13].
Putting these estimates together enables us to bound the supremum of the stochastic polynomial $\chi_{u,v,w}$ with high probability as expressed in A.12. This in turn implies that with probability at least $1 -$

$\exp(-t)$, for $t \geqslant 1$,

$$\frac{1}{\bar{N}}|H^{\dagger*}X^*\bar{X}\bar{g}|_{S_\infty} \lesssim \sigma_u^2 |G|_{S_\infty}\left(\sqrt{\frac{\log(T)+p+r}{\bar{N}}}\right.$$
$$\left. + \frac{\log^2(T)+(p+r)\log(T)}{\bar{N}} + \sqrt{\frac{t}{\bar{N}}} + \frac{\log(T)t}{\bar{N}}\right).$$

$\square$

Appendix B: Deterministic estimates

In this Appendix we provide the proofs of some deterministic inequalities that are needed especially in Appendix A.

Proposition B.1. *For (v_1, w_1) and (v_2, w_2) in $\mathbb{S}_2^{r-1} \times \mathbb{S}_2^{p-1}$ the following norm inequality holds:*

$$|v_1 w_1^* - v_2 w_2^*|_{S_\infty} \leqslant |v_1 - v_2|_2 + |w_1 - w_2|_2.$$

Proof. Take (v_1, w_1) and (v_2, w_2) both in $\mathbb{S}_2^{r-1} \times \mathbb{S}_2^{p-1}$ and note that

$$|v_1 w_1^* - v_2 w_2^*|_{S_\infty}^2 = \sup_{a \in \mathbb{S}_2^{p-1}} |v_1\langle w_1, a\rangle - v_2\langle w_2, a\rangle|_2^2$$

$$= \sup_{a \in \mathbb{S}_2^{p-1}} |v_1|_2^2 \langle w_1, a\rangle^2 + |v_2|_2^2 \langle w_2, a\rangle^2 - 2\langle v_1, v_2\rangle\langle w_1, a\rangle\langle w_2, a\rangle$$

$$= \sup_{a \in \mathbb{S}_2^{p-1}} \langle w_1, a\rangle^2 + \langle w_2, a\rangle^2 + (|v_1 - v_2|_2^2 - 2)\langle w_1, a\rangle\langle w_2, a\rangle$$

$$= \sup_{a \in \mathbb{S}_2^{p-1}} \langle w_1 - w_2, a\rangle^2 + |v_1 - v_2|_2^2\langle w_1, a\rangle\langle w_2, a\rangle$$

$$\leqslant \sup_{a \in \mathbb{S}_2^{p-1}} \langle w_1 - w_2, a\rangle^2 + |v_1 - v_2|_2^2$$

$$\leqslant |v_1 - v_2|_2^2 + |w_1 - w_2|_2^2.$$

Taking the square root we obtain

$$|v_1 w_1^* - v_2 w_2^*|_{S_\infty} \leqslant (|v_1 - v_2|_2^2 + |w_1 - w_2|_2^2)^{1/2} \leqslant |v_1 - v_2|_2 + |w_1 - w_2|_2.$$

$\square$

Proposition B.2. *Let $\mathcal{H}$ be the infinite bloc Toeplitz matrix made of $2T-1$ diagolals blocs of the matices $h_l \in \mathcal{M}_{p\times r}(\mathbb{R})$ with $l \in [\![1, 2T-1]\!]$, then its operator norm $|\mathcal{H}|_{2\to 2}$ is upper bounded by*

$$|\mathcal{H}|_{2\to 2} \leqslant \sup_{t\in[0,1]} \left(\left|\sum_{l=1}^{2T-1}\exp(i2\pi lt)h_l\right|_{S_\infty}\right).$$

Proof. Define the linear operators $\Phi : l_2(\mathbb{Z}) \to L_2(\mathbb{R}^p)$ and $\Psi : l_2(\mathbb{Z}) \to L_2(\mathbb{R}^r)$ such that $\Phi(u)(t) = \sum_{l=1}^{2T-1} u_l \exp(i2\pi lt)$ and $\Psi(v)(t) = \sum_{l=-\infty}^{+\infty} v_l \exp(i2\pi lt)$. Both are isometries since

$$|\Phi(u)|_{L_2(\mathbb{R}^p)} = \left(\int_0^1 \left| \sum_{l=1}^{2T-1} u_l \exp(i2\pi lt) \right|_2^2 dt \right)^{1/2}$$

$$= \left(\sum_{l=1}^{2T-1} \sum_{l'=1}^{2T-1} \int_0^1 \exp(i2\pi lt) \exp(-i2\pi l't) dt \, \langle u_l, u_{l'} \rangle \right)^{1/2}$$

$$= \left(\sum_{l=1}^{2T-1} \langle u_l, u_l \rangle \right)^{1/2} = |u|_2^2.$$

Ψ is the usual trigonometric isometry. Thus, for $|u|_2 = 1$ we have

$$|\mathcal{H}u|_2 = |\Phi\mathcal{H}\Psi^{-1}\Psi u|_{L_2(\mathbb{R}^p)} = |(\Phi\mathcal{H}\Psi^{-1})(\sum_{l=1}^{2T-1} u_l \exp(i2\pi lt))|_{L_2(\mathbb{R}^p)}$$

$$= \left(\int_0^1 \left| \sum_{l=1}^{2T-1} (\Phi\mathcal{H}\Psi^{-1})(u_l) \exp(i2\pi lt) \right|_2^2 dt \right)^{1/2}$$

$$= \sup_{|w|_2=1} \left(\int_0^1 \left| \sum_{l=1}^{2T-1} \sum_{l'=1}^{2T-1} \langle h_{l'}^* w, u_l \rangle \exp(i2\pi lt) \exp(i2\pi l't) \right|^2 dt \right)^{1/2}$$

$$= \sup_{|w|_2=1} \left(\int_0^1 \left| \left\langle \sum_{l=1}^{2T-1} \exp(i2\pi lt) u_l, \sum_{l'=1}^{2T-1} \exp(i2\pi l't) h_{l'}^* w \right\rangle \right|^2 dt \right)^{1/2}$$

$$\leqslant \sup_{|w|_2=1} \left(\int_0^1 \left| \sum_{l=1}^{2T-1} \exp(i2\pi lt) u_l \right|_2^2 \left| \sum_{l'=1}^{2T-1} \exp(i2\pi l't) h_{l'}^* w \right|_2^2 dt \right)^{1/2}$$

$$\leqslant \sup_{t \in [0,1]} \left| \sum_{l=1}^{2T-1} \exp(i2\pi lt) h_l \right|_{S_\infty} \left(\int_0^1 \left| \sum_{l=1}^{2T-1} \exp(i2\pi lt) u_l \right|_2^2 dt \right)^{1/2}$$

$$= \sup_{t \in [0,1]} \left(\left| \sum_{l=1}^{2T-1} \exp(i2\pi lt) h_l \right|_2 \right) |u|_2,$$

whence, the desired result

$$|\mathcal{H}|_{2\to 2} \leqslant \sup_{t\in[0,1]} \left(\left| \sum_{l=1}^{2T-1} \exp(i2\pi lt) h_l \right|_{S_\infty} \right).$$

$\square$

Appendix C: Proofs of the remaining results in Theorem 2.2

Proof of (2.6) ***Theorem 2.2.*** The proof is similar to the proof of (2.5). The difference is that X and W are independent and involve different sets of random variables. Recall the definition of the permuted index:

$$\bar{l} = \begin{cases} l - 2T & \text{if } T+1 \leqslant l \leqslant 2T-1, \\ l & \text{otherwise.} \end{cases}$$

Define

$$x = [u_0^*, \ldots, u_{N-2}^*, u_{N-1}^*]^* \in \mathbb{R}^{rN}$$

$$y = [w_0^*, \ldots, w_{N-1}^*]^* \in \mathbb{R}^{d_0 N}$$

$$z = [x^*, y^*]^* \in \mathbb{R}^{(d_0+r)N}$$

$$U_l = [u_{2T-l}, u_{2T+1-l}, \ldots, u_{N-l}] \in \mathcal{M}_{p\times\bar{N}}(\mathbb{R}).$$

From the definition of $H^{\dagger *}$ in 2.4 we have

$$H^{\dagger *} X^* W h^* = \begin{bmatrix} (U_1 W h^*)^* & \frac{1}{2}(U_2 W h^*)^* & \cdots & \frac{1}{T}(U_T W h^*)^* \\ \frac{1}{2}(U_2 W h^*)^* & \frac{1}{3}(U_3 W h)^* & \cdots & \frac{1}{T-1}(U_{T+1} W h^*)^* \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{T}(U_T W h^*)^* & \frac{1}{T-1}(U_{T+1} W h^*)^* & \cdots & (U_{2T-1} W h^*)^* \end{bmatrix}.$$

Define the infinite block Hankel operator $\mathcal{H} : l_2(\mathbb{N}) \to l_2(\mathbb{N})$ by the $\mathcal{M}_{p\times r}(\mathbb{R})$ blocks

$$\mathcal{H}_{i,j} = \begin{cases} 1/|\bar{l}|(U_l W h^*)^* & \text{for} \quad (i,j) \in \mathbb{N}^2 \text{ and } 1 \leqslant |i-j| = l \leqslant 2T-1, \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$|H^{\dagger *} X^* W h^*|_{S_\infty} \leqslant \sup_{u\in[0,\,1]} \left| \sum_{l=1}^{2T-1} \frac{e^{i2\pi \bar{l}u}}{|\bar{l}|} U_l W h^* \right|_{S_\infty}$$

$$= \sup_{u\in[0\ 1]} \sup_{\substack{|v|_2=1 \\ |w|_2=1}} \left| \left\langle W h^*, \sum_{l=1}^{2T-1} \frac{e^{i2\pi \bar{l}u}}{|\bar{l}|} U_l^* v w^* \right\rangle \right|.$$

Define, for $l \in [0 \; N - 2T]$, the vectors $G_l \in \mathcal{M}_{p \times d_0 N}(\mathbb{R})$ as

$$G_l = [C_0 A_0^{2T-1+l}, C_0 A_0^{2T+l}, \ldots, C_0, 0, \ldots, 0],$$

and for $u \in [0 \; 1]$ the matrix valued functions $w_{u,v,w,l} \in \mathcal{M}_{r \times p}(\mathbb{C})$ by $w_{u,v,w,l} = \frac{\exp(i2\pi \bar{l}u)}{|l|} vw^*$ for $l \in [1 \; 2T - 1]$ and the matrix valued functions $W_{u,v,w,l} \in \mathcal{M}_{p \times Nr}(\mathbb{C})$ by

$$W_{u,v,w,k} = \left[\begin{array}{ccccc} 0 & w_{u,v,w,2T-1}^* & w_{u,v,w,2T-2}^* & \cdots & w_{u,v,w,1}^* \end{array} \right],$$

with the 1st zero a the kth-position. Define G as $G = [G_0^*, \ldots, G_{N-2T}^*]^*$ and $W_{u,v,w}$ as $W_{u,v,w} = [W_{u,v,w,1}^*, \ldots, W_{u,v,w,N-2T-1}^*]^*$ satisfying

$$\left\langle \bar{X}\bar{g}^*, \sum_{l=1}^{2T-1} \frac{e^{i2\pi \bar{l}u}}{|l|} U_l^* vw^* \right\rangle = \left\langle W_{u,v,w}x, Gy \right\rangle.$$

This gives

$$|H^{\dagger^*} X^* Wh|_{S_\infty} \leqslant \sup_{u \in [0, \; 1]} \sup_{v \in \mathcal{S}_2^{p-1}} \left| \left\langle \begin{bmatrix} W_{u,v,w} \end{bmatrix} z, \begin{bmatrix} H \end{bmatrix} z \right\rangle \right|,$$

which is the supremum of a second order chaos process defined as follows

$$\chi_{u,v,w} = \left\langle \begin{bmatrix} H^* W_{u,v,w} \end{bmatrix} z,, z \right\rangle.$$

To control the increment of the process we use Hanson Wright inequality which gives us for $t > 0$.

$$\mathbb{P}\left(|\chi_{u,v,w} - \mathbb{E}(\chi_{u,v,w})| \geqslant \sigma_u \sigma_w \left(\sqrt{t} |H^* W_{u,v,w}|_{S_2} + t |H^* W_{u,v,w}|_{S_\infty} \right) \right) \leqslant 2\exp(-ct).$$

Since $\mathbb{E}(\chi_{u,v,w}) = 0$ we obtain a mixed tail process with probability $1 - e^{-t}$

$$|\chi_{u_1,v_1,w_1} - \chi_{u_2,v_2,w_2}| \leqslant \sigma_u \sigma_w |H|_{S_\infty} \left(\sqrt{t} |W_{u_1,v_1,w_1} - W_{u_2,v_2,w_2}|_{S_2} \right.$$
$$\left. + t |W_{u_1,v_1,w_1} - W_{u_2,v_2,w_2}|_{S_\infty} \right).$$

The rest of the proof is carried out similar to the proof of (2.5) to obtain the following bound that holds with probability at least $1 - \exp(-t)$, for $t \geqslant 1$.

$$|H^{\dagger^*} X^* Wh|_{S_\infty} \lesssim \sigma_u \sigma_w |H|_{S_\infty} \left(\sqrt{\frac{\log(T) + p}{N}} + \frac{\log^2(T) + p\log(T)}{N} + \sqrt{\frac{t}{N}} + \frac{\log(T)t}{N} \right).$$

$\square$

Proof of (2.7) ***Theorem 2.2.*** Again we follow similar steps to the proof of (2.5). We take the following definitions

$$x = [u_0^*, \ldots, u_{N-2}^*, u_{N-1}^*]^* \in \mathbb{R}^{rN}$$
$$\varepsilon = [v_{2T}^*, \ldots, v_N^*]^* \in \mathbb{R}^{p\bar{N}}$$

$$y = [x^*, \varepsilon^*]^* \in \mathbb{R}^{rN+p\bar{N}}$$

$$U_l = [u_{2T-l}, u_{2T+1-l}, \ldots, u_{N-l}] \in \mathcal{M}_{p \times \bar{N}}(\mathbb{R}).$$

From the definition of $H^{\dagger^*}$ in 2.4 we have

$$H^{\dagger^*} X^* \varepsilon = \begin{bmatrix} (U_1\varepsilon)^* & \frac{1}{2}(U_2\varepsilon)^* & \cdots & \frac{1}{T}(U_T\varepsilon)^* \\ \frac{1}{2}(U_2\varepsilon)^* & \frac{1}{3}(U_3\varepsilon)^* & \cdots & \frac{1}{T-1}(U_{T+1}\varepsilon)^* \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{T}(U_T\varepsilon)^* & \frac{1}{T-1}(U_{T+1}\varepsilon)^* & \cdots & (U_{2T-1}\varepsilon)^* \end{bmatrix}.$$

Define the infinite block Hankel operator $\mathcal{J} : l_2(\mathbb{N}) \to l_2(\mathbb{N})$ by the $\mathbb{R}^p$ blocks

$$\mathcal{J}_{i,j} = \begin{cases} 1/|\bar{l}|(U_l\varepsilon)^* & \text{for} \quad (i,j) \in \mathbb{N}^2, \text{ and } 1 \leqslant |i-j| = l \leqslant 2T-1, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$|H^{\dagger^*} X^* \varepsilon|_{S_\infty} \leqslant \sup_{u \in [0,\, 1]} \left| \sum_{l=1}^{2T-1} \frac{e^{i2\pi\bar{l}u}}{|\bar{l}|} U_l\varepsilon \right|_{S_\infty}$$

$$= \sup_{u \in [0\,1]} \sup_{\substack{|v|_2=1 \\ |w|_2=1}} \left| \left\langle \varepsilon, \sum_{l=1}^{2T-1} \frac{e^{i2\pi\bar{l}u}}{|\bar{l}|} U_l^* vw^* \right\rangle \right|$$

For $u \in [0\ 1]$ define the matrix valued functions $w_{u,v,w,l} \in \mathcal{M}_{r \times p}(\mathbb{C})$ by $w_{u,v,w,l} = \frac{\exp(i2\pi\bar{l}u)}{|l|} vw^*$ for $l \in [1\ 2T-1]$ and the matrix valued functions $W_{u,v,w,l} \in \mathcal{M}_{p \times Nr}(\mathbb{C})$ by

$$W_{u,v,w,k} = \begin{bmatrix} 0 & w_{u,v,w,2T-1}^* & w_{u,v,w,2T-2}^* & \cdots & w_{u,v,w,1}^* \end{bmatrix},$$

with the 1st zero a the kth-position. Put them together in $W_{u,v,w}$ since

$$W_{u,v,w} = [W_{u,v,w,1}^*, \ldots, W_{u,v,w,N-2T-1}^*]^*$$

which satisfy

$$\left\langle \varepsilon, \sum_{l=1}^{2T-1} \frac{e^{i2\pi\bar{l}u}}{|\bar{l}|} U_l^* vw^* \right\rangle = \langle W_{u,v,w} x, \varepsilon \rangle.$$

This gives

$$|H^{\dagger^*} X^* \varepsilon|_{S_\infty} \leqslant \sup_{u \in [0,\, 1]} \sup_{v \in \mathcal{S}_2^{p-1}} \left| \left\langle \begin{bmatrix} W_{u,v,w} \end{bmatrix} z, \begin{bmatrix} I_N \end{bmatrix} z \right\rangle \right|,$$

which is the supremum of a second order chaos process defined as follows

$$\chi_{u,v,w} = \left\langle \begin{bmatrix} W_{u,v,w} \end{bmatrix} y,, y \right\rangle.$$

To control the increment of the process we use Hanson Wright inequality which gives us for $t > 0$.

$$\mathbb{P}\left(|\chi_{u,v,w} - \mathbb{E}(\chi_{u,v,w})| \geq \sigma_v{}^2\left(\sqrt{t}|W_{u,v,w}|_{S_2} + t|W_{u,v,w}|_{S_\infty}\right)\right) \leq 2\exp\left(-ct\right).$$

Since $\mathbb{E}(\chi_{u,v,w}) = 0$ we obtain a mixed tail process with probability $1 - e^{-t}$

$$|\chi_{u_1,v_1,w_1} - \chi_{u_2,v_2,w_2}| \leq \sigma_v^2\left(\sqrt{t}|W_{u_1,v_1,w_1} - W_{u_2,v_2,w_2}|_{S_2} + t|W_{u_1,v_1,w_1} - W_{u_2,v_2,w_2}|_{S_\infty}\right).$$

The rest of the proof is carried out similarly to the proof of (2.5) to obtain the following bound that holds with probability at least $1 - \exp\left(-t\right)$ for $t \geq 1$:

$$|H^{\dagger*}X^*\varepsilon|_{S_\infty} \lesssim \sigma_v^2\left(\sqrt{\frac{\log(T) + p}{N}} + \frac{\log^2(T) + p\log(T)}{N} + \sqrt{\frac{t}{N}} + \frac{\log(T)t}{N}\right).$$

$\square$