# Finite Sample Analysis of Stochastic System Identification

Anastasios Tsiamis and George J. Pappas

***Abstract*— In this paper, we analyze the finite sample complexity of stochastic system identification using modern tools from machine learning and statistics. An unknown discrete-time linear system evolves over time under Gaussian noise without external inputs. The objective is to recover the system parameters as well as the Kalman filter gain, given a single trajectory of output measurements over a finite horizon of length $N$. Based on a subspace identification algorithm and a finite number of $N$ output samples, we provide non-asymptotic high-probability upper bounds for the system parameter estimation errors. Our analysis uses recent results from random matrix theory, self-normalized martingales and SVD robustness, in order to show that with high probability the estimation errors decrease with a rate of $1/\sqrt{N}$ up to logarithmic terms. Our non-asymptotic bounds not only agree with classical asymptotic results, but are also valid even when the system is marginally stable.**

## I. Introduction

Identifying predictive models from data has been a fundamental problem across several fields, from classical control theory to economics and modern machine learning. System identification, in particular, has a long history of studying this problem from a control theoretic perspective [1]. Identifying linear state-space models from input-output data:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + w_k \\ y_k &= Cx_k + Du_k + v_k, \end{aligned} \tag{1}$$

has been one of its main focuses.

Most identification methods for linear systems either follow the prediction error approach [2] or the subspace method [3], [4]. The prediction error approach is usually non-convex and directly searches over the system parameters $A, B, C, D$ by minimizing a prediction error cost. The subspace approach is a convex one; first, Hankel matrices of the system are estimated, then, the parameters are realized via steps involving singular value decomposition (SVD). Methods inspired by machine learning have also been employed [5]. In this paper, we focus on the subspace identification approach–see [6] for an overview.

The asymptotic statistical properties of subspace algorithms have been well-studied in the stationary regime [7]–[13]. In [7], [8], it is shown that the identification error can decay as fast as $O(1/\sqrt{N})$ up to logarithmic factors, where $N$ is the number of data. In [12], [13] Central Limit Theorems for the identification errors are established. The aforementioned results rely on the assumption of asymptotic stability (spectral radius $\rho(A) < 1$) and hold as the number of data $N$ grows to infinity. In the non-stationary case, subspace identification for a subclass of marginally stable systems was considered in [14], where it is shown that consistency can be guaranteed asymptotically if the unit circle eigenvalues of $A$ are all equal to 1 with simple Jordan blocks.

From a machine learning perspective, finite sample analysis has been a standard tool for comparing algorithms in the non-asymptotic regime. A series of papers [15]–[18] studied the finite sample properties of system identification from a single trajectory, when the system state is fully observed ($C = I$). Finite sample results for partially observed systems ($C \neq I$), which is a more challenging problem, appeared recently in [19]–[21]. These papers provide a non-asymptotic convergence rate of $1/\sqrt{N}$ (up to logarithmic factors) for the recovery of matrices $A, B, C, D$ up to a similarity transformation. The results rely on the assumption that the system can be driven by external inputs, i.e. $B, D \neq 0$. In [20], it was shown that consistency can be achieved even for arbitrary marginally stable systems, where $\rho(A) \leq 1$. Sample complexity of prediction error methods has also been considered [22]–[26], where the main metric is prediction performance. Finite sample properties of system identification algorithms have also been used in robust and adaptive control [27], [28]. The dual problem of Kalman filtering has not been studied yet in this context; preliminary results for scalar observations appeared in [29].

In this paper, we perform the first finite sample analysis of identifying system (1) in the case $B, D = 0$, when we have no inputs, also known as *stochastic system identification* (SSI) [3]. We provide the first non-asymptotic guarantees for the estimation of matrices $A, C$ as well as the Kalman filter gain of (1). Similar to [17], [19], the analysis is based on new tools from machine learning and statistics [30]–[32]. As in [15]–[24], this paper focuses on data-independent bounds, i.e. bounds which reveal how the identification error depends on the number of data $N$, and the system's and algorithm's parameters. An alternative approach is to derive data-dependent bounds, see for example [33]. Such bounds could potentially be more tight, however it is not yet clear how they vary with the number of data $N$. In summary, the main contributions of this paper are:

- To the best of our knowledge, our paper provides the first finite sample upper bounds in the case of stochastic system identification, where we have no inputs and the system is only driven by noise. We also provide the first finite sample guarantees for the estimation error of the Kalman filter gain.
- We prove that the outputs of the system satisfy persistence of excitation in finite time with high probability.

This work was supported by the DARPA Assured Autonomy program.

The authors are with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104. Emails: {atsiamis,pappasg}@seas.upenn.edu

This result is fundamental for the analysis of many subspace identification algorithms which use outputs as regressors.

- We show that we can achieve a non-asymptotic learning rate of $O(\sqrt{1/N})$ up to logarithmic factors in the case of general marginally stable systems $\rho(A) = 1$, generalizing the asymptotic results of [14]. The learning rate is also valid in the case of repeated unit circle eigenvalues, when the system is unstable but non-explosive. For stable systems ($\rho(A) < 1$), the non-asymptotic learning rate is consistent with classical asymptotic results [7].

Due to space constraints, the proofs are omitted and can be found in the online version of the paper.

## II. Problem formulation

Consider the standard state space representation (1) with $B, D = 0$, where $x_k \in \mathbb{R}^n$ is the system state, $y_k \in \mathbb{R}^m$ is the output, $A \in \mathbb{R}^{n\times n}$ is the system matrix, $C \in \mathbb{R}^{m\times n}$ is the output matrix, $w_k \in \mathbb{R}^n$ is the process noise, and $v_k \in \mathbb{R}^m$ is the measurement noise. The noises $w_k$, $v_k$ are assumed to be i.i.d. zero mean Gaussian, with covariance matrices $Q$ and $R$ respectively, and independent of each other. The initial state $x_0$ is also assumed to be zero mean Gaussian, independent of the noises, with covariance $\Sigma_0$. Matrices $A$, $C$, $Q$, $R$, $\Sigma_0$ are initially unknown. However, the following assumption holds throughout the paper.

**Assumption 1:** The order of the system $n$ is known [1]. The spectral radius $\rho(A)$ of $A$ is $\rho(A) \leq 1$. The pair $(A, C)$ is observable, $(A, Q^{1/2})$ is controllable and $R$ is strictly positive definite. $\diamond$

The assumption $\rho(A) \leq 1$ includes marginally stable systems as well as non-explosive unstable systems with repeated unit circle roots. It is more general than the stricter condition $\rho(A) < 1$ found in previous works, see [7]–[13]. The remaining conditions in Assumption 1 are standard for the stochastic system identification problem to be well posed and the Kalman filter to converge.

The steady-state Kalman filter of system (1) is:

$$\begin{aligned} \hat{x}_{k+1} &= A\hat{x}_k + Ke_k \\ y_k &= C\hat{x}_k + e_k, \end{aligned} \tag{2}$$

where here the filter is in the predictor form:

$$\hat{x}_{k+1} = \mathbb{E}\left[x_{k+1}|y_0, \ldots, y_k\right], \hat{x}_0 = 0. \tag{3}$$

The steady-state Kalman gain $K \in \mathbb{R}^{n\times m}$ is:

$$K = APC^*(CPC^* + R)^{-1},$$

with $P$ the positive definite solution of the Riccati equation:

$$P = APA^* + Q - APC^*(CPC^* + R)^{-1}CPA^*.$$

A byproduct of Assumption 1 is that the closed-loop matrix $A - KC$ has all the eigenvalues inside the unit circle [34]. We denote the covariance matrix of the prediction $\hat{x}_k$ by:

$$\Gamma_k = \mathbb{E}\left[\hat{x}_k\hat{x}_k^*\right]. \tag{4}$$

[1]The results of Section IV do not depend on the order $n$ being known.

The innovation error sequence $e_k$ has covariance

$$\bar{R} \triangleq \mathbb{E}\left[e_k e_k^*\right] = CPC^* + R. \tag{5}$$

Since the original errors are Gaussian i.i.d., by the orthogonality principle the innovation error sequence $e_k$ is also Gaussian and i.i.d. The later property is true since we also assumed that the Kalman filter is in steady-state.

**Assumption 2:** We assume that $\Sigma_0 = P$, so that the Kalman filter (2) has converged to its steady-state. $\diamond$

Since the Kalman filter converges exponentially fast to the steady-state gain, this assumption is reasonable in many situations; it is also standard [7], [11].

In the classical stochastic subspace identification problem, the main goal is to identify the Kalman filter parameters $A, C, K$ from output samples $y_0 \ldots, y_N$, see for example Chapter 3 of [3]. The problem is ill-posed in general since the outputs are invariant under any similarity transformation $\bar{A} = S^{-1}AS$, $\bar{C} = CS$, $\bar{K} = S^{-1}K$. Thus, we can only estimate $A, C, K$ up to a similarity transformation.

In this paper, we will analyze the finite sample properties of a subspace identification algorithm, which is based on least squares.

**Problem 1 (Finite Sample Analysis of SSI):** Consider a finite number $N$ of output samples $y_0, \ldots, y_{N-1}$, which follow model (1) with $B, D = 0$, and an algorithm $\mathcal{A}$, which returns estimates $\hat{A}, \hat{C}, \hat{K}$ of the true parameters. Given a confidence level $\delta$ provide upper bounds $\epsilon_A(\delta, N)$, $\epsilon_C(\delta, N)$, $\epsilon_K(\delta, N)$ such that with probability at least $1-\delta$:

$$\begin{aligned} \left\|\hat{A} - S^{-1}AS\right\|_2 &\leq \epsilon_A(\delta, N) \\ \left\|\hat{C} - CS\right\|_2 &\leq \epsilon_C(\delta, N) \\ \left\|\hat{K} - S^{-1}K\right\|_2 &\leq \epsilon_K(\delta, N), \end{aligned} \tag{6}$$

for some invertible matrix $S$, where $\|\cdot\|_2$ denotes the spectral norm. The bounds $\epsilon$ can also depend on the model parameters $n, A, C, R, Q$ as well as the identification algorithm used. $\diamond$

## III. Subspace Identification Algorithm

The procedure of estimating the parameters $A, C, K$ is based on a least squares approach, see for example [7], [11]. It involves two stages. First, we regress future outputs to past outputs to obtain a Hankel-like matrix, which is a product of an observability and a controllability matrix. Second, we perform a balanced realization step, similar to the Ho-Kalman algorithm, to obtain estimates for $A, C, K$.

Before describing the algorithm, we need some definitions. Let $p, f$, with $p, f \geq n$ be two design parameters that define the horizons of the past and the future respectively. Assume that the total number of output samples is $\bar{N} = N+p+f-1$. Then, the future outputs $Y_k^+ \in \mathbb{R}^{mf}$ and past outputs $Y_k^- \in \mathbb{R}^{mp}$ at time $k \geq p$ are defined as follows:

$$Y_k^+ \triangleq \begin{bmatrix} y_k \\ \vdots \\ y_{k+f-1} \end{bmatrix}, \quad Y_k^- \triangleq \begin{bmatrix} y_{k-p} \\ \vdots \\ y_{k-1} \end{bmatrix}, k \geq p \tag{7}$$

By stacking the outputs for all sample sequences, over all times $p \leq k \leq N+p-1$, we form the batch outputs:

$$Y_+ \triangleq \begin{bmatrix} Y_p^+ & \dots & Y_{N+p-1}^+ \end{bmatrix},$$
$$Y_- \triangleq \begin{bmatrix} Y_p^- & \dots & Y_{N+p-1}^- \end{bmatrix},$$

The past and future noises $E_k^+, E_k^-, E_+, E_-$ are defined similarly. Finally, define the batch states:

$$\hat{X} \triangleq \begin{bmatrix} \hat{x}_0 & \dots & \hat{x}_{N-1} \end{bmatrix}.$$

The (extended) observability matrix $\mathcal{O}_k \in \mathbb{R}^{mk \times n}$ and the reversed (extended) controllability matrix $\mathcal{K}_k \in \mathbb{R}^{n \times mk}$ associated to system (2) are defined as:

$$\mathcal{O}_k \triangleq \begin{bmatrix} C^* & A^*C^* & \cdots & (A^*)^{k-1}C^* \end{bmatrix}^*, \quad (8)$$

$$\mathcal{K}_k \triangleq \begin{bmatrix} (A-KC)^{k-1}K & \dots & (A-KC)K & K \end{bmatrix} \quad (9)$$

respectively. We denote the Hankel(-like) matrix $\mathcal{O}_f\mathcal{K}_p$ by:

$$G \triangleq \mathcal{O}_f\mathcal{K}_p. \quad (10)$$

Finally, for any $s \geq 2$, define the block-Toeplitz matrix:

$$\mathcal{T}_s \triangleq \begin{bmatrix} I_m & 0 & & 0 \\ CK & I_m & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ CA^{s-2}K & CA^{s-3}K & \cdots & I_m \end{bmatrix}. \quad (11)$$

### A. Regression for Hankel Matrix Estimation

First, we establish a linear regression between the future and past outputs. From (2), for every $k$:

$$Y_k^+ = \mathcal{O}_f\hat{x}_k + \mathcal{T}_f E_k^+.$$

Meanwhile, from (2), the state prediction $\hat{x}_k$ can be expressed in terms of the past outputs:

$$\hat{x}_k = Ky_{k-1} + \cdots + (A-KC)^{p-1}Ky_{k-p} + (A-KC)^p\hat{x}_{k-p}.$$

After some algebra, we derive the linear regression:

$$Y_+ = GY_- + \mathcal{O}_f(A-KC)^p\hat{X} + \mathcal{T}_f E_+, \quad (12)$$

where the regressors $Y_-$ and the residuals $E_+$ are independent from each other column-wise. The term $\mathcal{O}_f(A-KC)^p\hat{X}$ introduces a bias due to the Kalman filter truncation, where we use only $p$ past outputs instead of all of them. Based on (12), we compute the least squares estimate

$$\hat{G} = Y_+Y_-^*(Y_-Y_-^*)^{-1}. \quad (13)$$

The Hankel matrix $G$ can be interpreted as a (truncated) Kalman filter which predicts future outputs directly from past outputs, independently of the internal state-space representation [3]. In this sense, the estimate $\hat{G}$ is a "data-driven" Kalman filter. Notice that persistence of excitation of the outputs (invertibility of $Y_-Y_-^*$) is required in order to compute the least squares estimate $\hat{G}$.

### B. Balanced Realization

This step determines a balanced realization of the state-space, which is only one of the possibly infinite state-space representations–see Section VI for comparison with other subspace methods. First, we compute a rank-$n$ factorization of the full rank matrix $\hat{G}$. Let the SVD of $\hat{G}$ be:

$$\hat{G} = \begin{bmatrix} \hat{U}_1 & \hat{U}_2 \end{bmatrix} \begin{bmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \hat{V}_1^* \\ \hat{V}_2^* \end{bmatrix}, \quad (14)$$

where $\hat{\Sigma}_1 \in \mathbb{R}^{n \times n}$ contains the $n-$largest singular values. Then, a standard realization of $\mathcal{O}_f$, $\mathcal{K}_p$ is:

$$\hat{\mathcal{O}}_f = \hat{U}_1\hat{\Sigma}_1^{1/2},\ \hat{\mathcal{K}}_p = \hat{\Sigma}_1^{1/2}\hat{V}_1^*. \quad (15)$$

This step assumed knowing the order $n$ of the system, see Assumption 1. In addition, matrix $\mathcal{K}_p$ should have full rank $n$. This is equivalent to the pair $(A, K)$ being controllable. Otherwise, $\mathcal{O}_f\mathcal{K}_p$ will have rank less than $n$ making it impossible to accurately estimate $\mathcal{O}_f$.

**Assumption 3:** The pair $(A, K)$ is controllable. ◇

The above assumption is standard–see for example [11].

Based on the estimated observability/controllability matrices, we can approximate the system parameters as follows:

$$\hat{C} = \hat{\mathcal{O}}_f(1:m,:), \quad \hat{K} = \hat{\mathcal{K}}_p(:,(p-1)m+1:pm),$$

where the notation $\hat{\mathcal{O}}_f(1:m,:)$ means we pick the first $m$ rows and all columns. The notation for $\hat{\mathcal{K}}_p$ has similar interpretation. For simplicity, define

$$\hat{\mathcal{O}}_f^u \triangleq \hat{\mathcal{O}}_f(1:m(f-1),:),$$

which includes the $m(f-1)$ "upper" rows of matrix $\hat{\mathcal{O}}_f$. Similarly, we define the lower part $\hat{\mathcal{O}}_f^l$. For matrix $A$ we exploit the structure of the extended observability matrix and solve $\hat{\mathcal{O}}_f^u\hat{A} = \hat{\mathcal{O}}_p^l$ in the least squares sense by computing

$$\hat{A} = \left(\hat{\mathcal{O}}_f^u\right)^\dagger \hat{\mathcal{O}}_p^l,$$

where $\dagger$ denotes the pseudoinverse.

The finite sample analysis of the above algorithm is divided in two parts. First, in Section IV, we provide high probability upper bounds for the error $\|G-\hat{G}\|_2$ in the regression step. Then, in Section V, we analyze the robustness of the balanced realization step.

## IV. Finite Sample Analysis of Regression

In this section, we provide the finite sample analysis of the linear regression step of the identification algorithm. We provide high-probability upper bounds for the estimation error $\|G-\hat{G}\|_2$ of the Hankel-like matrix $G$. Before we state the main result, recall the definition of the covariance matrix $\bar{R}$ in (5). We denote the past noises' weighted covariance by:

$$\Sigma_E = \mathbb{E}\left[\mathcal{T}_p E_k^-(E_k^-)^*\mathcal{T}_p^*\right] = \mathcal{T}_p\,\mathrm{diag}(\bar{R},\dots,\bar{R})\mathcal{T}_p^*. \quad (16)$$

The least singular value of the above matrix is denoted by:

$$\sigma_E \triangleq \sigma_{\min}(\Sigma_E). \quad (17)$$

Lemma 2 in the Appendix proves that $\sigma_E \geq \sigma_{\min}(R) > 0$.

**Theorem 1 (Regression Step Analysis):** Consider system (2) under the Assumptions 1, 2, 3. Let $\hat{G}$ be the estimate (13) of the subspace identification algorithm given an output trajectory $y_0, \dots, y_{N+p+f-1}$ and let $G$ be as in (10). Fix a confidence $\delta > 0$ and define:

$$\delta_N \triangleq (2(N+p-1)m)^{-\log^2(2pm)\log(2(N+p-1)m)}. \quad (18)$$

There exist $N_0, N_1, N_2$ such that if $N \geq N_0, N_1, N_2$, (see definitions (30), (34), (35) in the Appendix), then with probability at least $1 - \delta_N - 6\delta$:

$$\left\|G - \hat{G}\right\|_2 \leq \underbrace{\mathcal{C}_1 \sqrt{\frac{fmp}{N} \log \frac{5f\kappa_N}{\delta}}}_{O\left(\sqrt{p \log N / N}\right)} + \underbrace{\mathcal{C}_2 \left\|(A-KC)^p\right\|_2}_{O(\rho(A-KC)^p)}, \quad (19)$$

where

$$\kappa_N = \frac{4}{\sigma_E}\left(\left\|\mathcal{O}_p\right\|_2^2 \operatorname{tr}\Gamma_{N-1} + \operatorname{tr}\Sigma_E\right) + \delta \quad (20)$$

over-approximates the condition number of $\mathbb{E}\left[Y_- Y_-^*\right]$ and

$$\mathcal{C}_1 = 8\sqrt{\frac{\left\|\bar{R}\right\|_2}{\sigma_E}}\left\|\mathcal{T}_f\right\|_2, \quad \mathcal{C}_2 = 4\left\|\mathcal{O}_f\right\|_2 \left\|\mathcal{O}_p^\dagger\right\|_2, \quad (21)$$

are system-dependent constants. ◇

**Remark 1 (Result interpretation):** From (12), (13) the estimation error consists of two terms:

$$\hat{G} - G = \underbrace{\mathcal{T}_f E_+ Y_-^* \left(Y_- Y_-^*\right)^{-1}}_{\text{Cross term}} + \underbrace{\mathcal{O}_f \left(A - KC\right)^p \hat{X} Y_-^* \left(Y_- Y_-^*\right)^{-1}}_{\text{Kalman filter truncation bias term}}. \quad (22)$$

The first term in (19) corresponds to the cross-term error, while the second term corresponds to the Kalman filter truncation bias term. To obtain consistency for $\hat{G}$, we have to let the term $\left\|(A-KC)^p\right\|_2$ go to zero with $N$. Recall that the matrix $A - KC$ has spectral radius less than one, thus, the second term decreases exponentially with $p$. By selecting $p = c \log N$, for some $c$, we can force the Kalman truncation error term to decrease at least as fast as the first one, see for example [7]. In this sense, the dominant term is the first one, i.e. the cross-term. Notice that $f$ can be kept bounded as long as it is larger than $n$. ◇

**Remark 2 (Statistical rates):** For **marginally stable** systems or non-explosive unstable systems ($\rho(A) = 1$) and $p = c \log N$, we have $\log \kappa_N = O(\log N)$, since $\left\|\mathcal{O}_p\right\|_2, \operatorname{tr}\Gamma_N$ depend at most polynomially on $p, N$. In this case, (19) results in a rate of:

$$\left\|G - \hat{G}\right\|_2 = O\left(\frac{\log N}{\sqrt{N}}\right).$$

To the best of our knowledge, these have not been any bounds for subspace algorithms in the general case of $\rho(A) = 1$.

In the case of **asymptotically stable** systems ($\rho(A) < 1$), we have $\kappa_N = O(p)$, since $\left\|\mathcal{O}_p\right\|_2, \operatorname{tr}\Gamma_N, \left\|\mathcal{T}_p\right\|_2$ are now $O(1)$. Hence, if $p = c \log N$, we obtain a rate of:

$$\left\|G - \hat{G}\right\|_2 = O\left(\sqrt{\frac{\log N \log\log N}{N}}\right).$$

As a result, our finite sample bound (19) is consistent with the asymptotic bound in equation (14) of [7]. ◇

In the absence of inputs ($B, D = 0$), the noise both helps and obstructs identification. Larger noise leads to better excitation of the outputs, but also worsens the convergence of the least squares estimator. To see how our finite sample bounds capture that, observe that larger noise leads to bigger $\sigma_E$ but also bigger $\left\|\bar{R}\right\|_2$. This trade-off is captured by $\mathcal{C}_1$.

If $N$ is sufficiently large (condition $N \geq N_0, N_1$), the outputs are guaranteed to be persistently exciting in finite time; more details can be found in Section IV-A and the Appendix. Meanwhile, condition $N \geq N_2$ is not necessary; it just leads to a simplified expression for the bound of the Kalman filter truncation error–see Section IV-C and Appendix. The definitions of $N_0, N_1, N_2$ can be found in (30), (34), (35). Their existence is guaranteed even if $p$ varies slowly with $N$, i.e. logarithmically.

Obtaining the bound on the error $\|G - \hat{G}\|_2$ in (19) of Theorem 1 requires the following three steps:

1) Proving persistence of excitation (PE) for the past outputs, i.e. invertibility of $Y_- Y_-^*$.
2) Establishing bounds for the cross-term error in (22).
3) Establishing bounds for the the truncation term in (22).

In the following subsections, we sketch the proof steps.

### *A. Persistence of Excitation in Finite Time*

The next theorem shows that with high probability the past outputs and noises are persistently exciting in finite time. The result is of independent interest and is fundamental since many subspace algorithms use past outputs as regressors.

**Theorem 2 (Persistence of Excitation):** Consider the conditions of Theorem 1 and $N_0$, $N_1$ as in (30), (34). If $N \geq N_0, N_1$, then with probability at least $1 - \delta_N - 2\delta$ both of the following events occur:

$$\mathcal{E}_Y = \left\{Y_- Y_-^* \succeq \frac{1}{2}\mathcal{O}_p \hat{X} \hat{X}^* \mathcal{O}_p^* + \frac{1}{2}\mathcal{T}_p E_- E_-^* \mathcal{T}_p^*\right\} \quad (23)$$

$$\mathcal{E}_E = \left\{\mathcal{T}_p E_- E_-^* \mathcal{T}_p^* \succeq \frac{N}{2}\Sigma_E\right\}, \quad (24)$$

where $\succeq$ denotes comparison in the positive semidefinite cone. Hence, with probability at least $1 - \delta_N - 2\delta$ the outputs satisfy the PE condition:

$$Y_- Y_-^* \succeq \frac{N}{4}\sigma_E I_{mp},$$

where $\sigma_E > 0$ is defined in (17). ◇

A sketch of proof can be found in the Appendix. The above result implies that if the past noises satisfy a PE condition, then PE for the outputs is also guaranteed; the noises are the only way to excite the system in the absence of control inputs. The see why the outputs are persistently exciting, notice that the past output correlations satisfy:

$$Y_- Y_-^* = \mathcal{O}_p \hat{X} \hat{X}^* \mathcal{O}_p^* + \mathcal{T}_p E_- E_-^* \mathcal{T}_p^* + \mathcal{O}_p \hat{X} E_-^* \mathcal{T}_p^* + \mathcal{T}_p E_- \hat{X}^* \mathcal{O}_p^*. \quad (25)$$

We can first show PE for the noise correlations $\mathcal{T}_p E_- E_-^* \mathcal{T}_p^*$, i.e. show that the event $\mathcal{E}_E$ occurs with high probability when

$N$ is sufficiently large (condition $N \geq N_0$). This behavior is due to the fact that $\mathbb{E}\left[\mathcal{T}_p E_- E_-^* \mathcal{T}_p^*\right] = N\Sigma_E$ and the sequence $E_k^-$ is component-wise i.i.d. To prove this step, we use Lemma C.2 from [19]–see Lemma 1 in the Appendix.

Meanwhile, the cross terms $\hat{X}E^*$ are much smaller and their norm increases with a rate of at most $O(\sqrt{N})$ up to logarithmic terms. This is since $\mathbb{E}\left[\hat{X}E^*\right] = 0$ and the product $\hat{X}E^*$ has martingale structure (see Appendix and Theorem 3 below). Eventually, if the number of samples $N$ is large enough (condition $N \geq N_1$), the cross-terms will be dominated by the noise and state correlations with high probability, which establishes output PE.

### B. Cross-term error

To bound the cross-term error, we express it as a product of $E_+ Y_-^* \left(Y_- Y_-^*\right)^{-1/2}$ and $(Y_- Y_-^*)^{-1/2}$, as in [17]. The second term of the product can be bounded by applying Theorem 2. The first term is self-normalized and has martingale structure component-wise. In particular, the product $Y_- E_+^*$ is equal to:

$$Y_- E_+^* = \left[ \begin{array}{ccc} \sum_{k=p}^{N+p-1} Y_k^- e_k^* & \dots & \sum_{k=p}^{N+p-1} Y_k^- e_{k+f-1}^* \end{array} \right],$$

where every sum above is a martingale. To bound it, we apply the next theorem, which generalizes Theorem 1 in [31] and Proposition 8.2 in [17].

**Theorem 3 (Cross terms):** Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\eta_t \in \mathbb{R}^m$, $t \geq 0$ be $\mathcal{F}_t$-measurable, independent of $\mathcal{F}_{t-1}$. Suppose also that $\eta_t$ has independent components $\eta_{t,i}$ $i = 1, \dots, m$, which are $1-$sub-Gaussian:

$$\mathbb{E}\left[e^{\lambda \eta_{t,i}} | \mathcal{F}_{t-1}\right] = \mathbb{E}\left[e^{\lambda \eta_{t,i}}\right] \leq e^{\lambda^2/2}, \text{ for all } \lambda \in \mathbb{R}.$$

Let $X_t \in \mathbb{R}^d$, $t \geq 0$ be $\mathcal{F}_{t-1}-$measurable. Assume that $V$ is a $d \times d$ positive definite matrix. For any $t \geq 0$, define:

$$\bar{V}_t = V + \sum_{s=1}^{t} X_s X_s^*, \qquad S_t = \sum_{s=1}^{t} X_s H_s^*,$$

where

$$H_s^* = \left[ \begin{array}{ccc} \eta_s^* & \dots & \eta_{s+r-1}^* \end{array} \right] \in \mathbb{R}^{rm},$$

for some integer $r$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$

$$\left\| \bar{V}_t^{-1/2} S_t \right\|_2^2 \leq 8r \left( \log \frac{r 5^m}{\delta} + \frac{1}{2} \log \det \bar{V}_t V^{-1} \right). \quad \diamond$$

The above theorem along with a Markov upper bound on $Y_- Y_-^*$ (see Lemma 3 in the Appendix) are used to bound $E_+ Y_-^* \left(Y_- Y_-^*\right)^{-1/2}$.

### C. Kalman truncation error

For the Kalman truncation error term, we need to bound the term $\hat{X} Y_-^* (Y_- Y_-^*)^{-1}$, which is $O(1)$. Using the identities $\mathcal{O}_p^\dagger \mathcal{O}_p \hat{X} = \hat{X}$, and $Y_- = \mathcal{O}_p \hat{X} + \mathcal{T}_p E_-$, we derive the following equality:

$$\begin{aligned} \hat{X} Y_-^* (Y_- Y_-^*)^{-1} = \mathcal{O}_p^\dagger \big( & I_{mp} - \mathcal{T}_p E_- E_-^* \mathcal{T}_p^* (Y_- Y_-^*)^{-1} \\ & - \mathcal{T}_p E_- \hat{X}^* \mathcal{O}_p^* (Y_- Y_-^*)^{-1} \big) \end{aligned} \quad (26)$$

From Theorem 2, we obtain $\left\| \mathcal{T}_p E_- E_-^* \mathcal{T}_p^* (Y_- Y_-^*)^{-1} \right\|_2 \leq 2$. The last term in (26) can be treated like the cross-term in Section IV-B, by applying Theorems 2, 3 and Lemma 3. It decreases with a rate of $O\left(1/\sqrt{N}\right)$ up to logarithmic terms, so it is much smaller than the other terms in (26). To keep the final bound simple, we select $N_2$ such that

$$\left\| \mathcal{T}_p E_- \hat{X}^* \mathcal{O}_p^* (Y_- Y_-^*)^{-1} \right\|_2 \leq 1 \quad (27)$$

with high probability–see also (35) for the definition of $N_2$.

## V. Robustness of Balanced Realization

In this section, we analyze the robustness of the balanced realization. In particular, we upper bound the estimation errors of matrices $A, C, K$ in terms of the estimation error $\|G - \hat{G}\|_2$ obtained by Theorem 1.

Assume that we knew $G$ exactly. Then, the SVD of the true $G$, would be:

$$G = \left[ \begin{array}{cc} U_1 & U_2 \end{array} \right] \left[ \begin{array}{cc} \Sigma_1 & 0 \\ 0 & 0 \end{array} \right] \left[ \begin{array}{c} V_1^* \\ V_2^* \end{array} \right] = U_1 \Sigma_1 V_1^*,$$

for some $\Sigma_1 \in \mathbb{R}^{n \times n}$. Hence, if we knew $G$ exactly, the output of the balanced realization would be:

$$\bar{\mathcal{O}}_f = U_1 \Sigma_1^{1/2}, \, \bar{\mathcal{K}}_p = \Sigma_1^{1/2} V_1^*. \quad (28)$$

The respective matrices $\bar{C}, \bar{K}, \bar{A}$ are defined similarly, based on $\bar{\mathcal{O}}_f, \bar{\mathcal{K}}_p$, as described in Section III. The system matrices $\bar{C}, \bar{K}, \bar{A}$ are equivalent to the original matrices $C, K, A$ up to a similarity transformation $\bar{C} = CS$, $\bar{K} = S^{-1}K$, $\bar{A} = S^{-1}AS$ for some invertible $S$. For simplicity, we will quantify the estimation errors in terms of the similar $\bar{A}, \bar{C}, \bar{K}$ instead of the original $A, C, K$.

The next result follows the steps of [19] and relies on Lemma 5.14 of [32] and Theorem 4.1 of [35]. Let $\sigma_n(\cdot)$ denote the $n-$th largest singular value.

**Theorem 4 (Realization robustness):** Consider the true Hankel-like matrix $G$ defined in (10) and the noisy estimate $\hat{G}$ defined in (13). Let $\hat{A}, \hat{C}, \hat{K}, \hat{\mathcal{O}}_f, \hat{\mathcal{K}}_p$ be the output of the balanced realization algorithm based on $\hat{G}$. Let $\bar{A}, \bar{C}, \bar{K}, \bar{\mathcal{O}}_f, \bar{\mathcal{K}}_p$ be the output of the balanced realization algorithm based on the true $G$. If $G$ has rank $n$ and the following robustness condition is satisfied:

$$\left\| \hat{G} - G \right\|_2 \leq \frac{\sigma_n(G)}{4}, \quad (29)$$

then there exists an orthonormal matrix $T \in \mathbb{R}^{n \times n}$ such that:

$$\begin{aligned} & \left\| \hat{\mathcal{O}}_f - \bar{\mathcal{O}}_f T \right\|_2 \leq 2 \sqrt{\frac{10n}{\sigma_n(G)}} \left\| G - \hat{G} \right\|_2 \\ & \left\| \hat{C} - \bar{C} T \right\|_2 \leq \left\| \hat{\mathcal{O}}_f - \bar{\mathcal{O}}_f T \right\|_2 \\ & \left\| \hat{A} - T^* \bar{A} T \right\|_2 \leq \underbrace{\frac{\sqrt{\|G\|_2} + \sigma_o}{\sigma_o^2}}_{O(1)} \left\| \hat{\mathcal{O}}_f - \bar{\mathcal{O}}_f T \right\|_2 \\ & \left\| \hat{K} - T^* \bar{K} \right\|_2 \leq 2 \sqrt{\frac{10n}{\sigma_n(G)}} \left\| G - \hat{G} \right\|_2, \end{aligned}$$

where $\sigma_o = \min\left(\sigma_n\left(\hat{\mathcal{O}}_f^u\right), \sigma_n\left(\bar{\mathcal{O}}_f^u\right)\right)$. The notation $\hat{\mathcal{O}}_f^u, \bar{\mathcal{O}}_f^u$, refers to the upper part of the respective matrix (first $(f-1)m$ rows)–see Section III-B. $\diamond$

**Remark 3:** The result states that if the error of the regression step is small enough, then the realization is robust. The singular value $\sigma_n(G)$ can be quite small. Hence, the robustness condition (29) can be restrictive in practice. However, such a condition is a fundamental limitation of the SVD procedure; it guarantees that the singular vectors related to small singular values of $G$ are separated from the singular vectors coming from the noise $G-\hat{G}$, which can be arbitrary. See also Wedin's theorem [36]. Such robustness conditions have also appeared in model reduction theory [37]. $\diamond$

The term $\frac{\sqrt{\|G\|_2}+\sigma_o}{\sigma_o^2}$ which appears in the bound of $A$ is $O(1)$. Although, the value of $\sigma_o^{-1}$ is random and depends on $\hat{\mathcal{O}}_f^h$, we could replace it by a deterministic bound. From

$$\sigma_n\left(\hat{\mathcal{O}}_f^h\right) \geq \sigma_n\left(\bar{\mathcal{O}}_f^h\right) - \left\|\hat{\mathcal{O}}_f - \bar{\mathcal{O}}_f T\right\|_2,$$

$\sigma_o$ will eventually be lower bounded by $\sigma_n\left(\bar{\mathcal{O}}_f^h\right)/2$ if the error $\|\hat{\mathcal{O}}_f - \bar{\mathcal{O}}_f T\|_2$ is small enough. The norm $\|G\|_2 \leq \|\mathcal{O}_f\|_2 \|\mathcal{K}_p\|_2$ is upper bounded for all $p$, since $A-KC$ is asymptotically stable and $f$ is fixed.

**Remark 4 (Total bounds):** The final upper bounds for the estimation of the system parameters $A, C, K$, as stated in Problem 1, can be found by combining the finite sample guarantees of the regression step (Theorem 1) with the robustness analysis of the realization step (Theorem 4). All matrix estimation errors depend linearly on the Hankel matrix estimation error $\|G-\hat{G}\|_2$. As a result, all matrix errors have the same statistical rate as the error of $G$, i.e. their estimation error decreases at least as fast as $O\left(1/\sqrt{N}\right)$ up to logarithmic factors. $\diamond$

## VI. Discussion and Future Work

One of the differences between the subspace algorithm considered in this paper and other subspace identification algorithms is the SVD step. Other algorithms perform SVD on $W_1 G W_2$ instead of $G$, where $W_1, W_2$ are full rank weighting matrices, usually data dependent [2], [3], [38]. From this point of view, the results of Section IV (upper bound for $\|G-\hat{G}\|$ in Theorem 1 and persistence of excitation in Theorem 2) are fundamental for understanding the finite sample properties of other subspace identification algorithms. Here, we studied the case $W_1 = I, W_2 = I$, which is not the standard choice [11]. It is subject of future work to explore how the choice of $W_1, W_2$ affects the realization step, especially the robustness condition of the SVD step.

## References


[1] L. Ljung, "Perspectives on System identification," *Annual Reviews in Control*, vol. 34, no. 1, pp. 1–12, 2010.

[2] ——, *System Identification: Theory for the User*. Prentice Hall, 1999.

[3] P. Van Overschee and B. De Moor, *Subspace identification for linear systems: Theory–Implementation–Applications*. Springer Science & Business Media, 2012.

[4] M. Verhaegen and V. Verdult, *Filtering and system identification: a least squares approach*. Cambridge university press, 2007.

[5] A. Chiuso and G. Pillonetto, "System identification: A machine learning perspective," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, no. 1, 2019.

[6] S. J. Qin, "An overview of subspace identification," *Computers & chemical engineering*, vol. 30, no. 10-12, pp. 1502–1513, 2006.

[7] M. Deistler, K. Peternell, and W. Scherrer, "Consistency and relative efficiency of subspace methods," *Automatica*, vol. 31, no. 12, pp. 1865–1875, 1995.

[8] K. Peternell, W. Scherrer, and M. Deistler, "Statistical analysis of novel subspace identification methods," *Signal Processing*, vol. 52, no. 2, pp. 161–177, 1996.

[9] M. Viberg, B. Wahlberg, and B. Ottersten, "Analysis of state space system identification methods based on instrumental variables and subspace fitting," *Automatica*, vol. 33, no. 9, pp. 1603–1616, 1997.

[10] M. Jansson and B. Wahlberg, "On consistency of subspace methods for system identification," *Automatica*, vol. 34, no. 12, pp. 1507–1519, 1998.

[11] T. Knudsen, "Consistency analysis of subspace identification methods based on a linear regression approach," *Automatica*, vol. 37, no. 1, pp. 81–89, 2001.

[12] D. Bauer, M. Deistler, and W. Scherrer, "Consistency and asymptotic normality of some subspace algorithms for systems without observed inputs," *Automatica*, vol. 35, no. 7, pp. 1243–1254, 1999.

[13] A. Chiuso and G. Picci, "The asymptotic variance of subspace estimates," *Journal of Econometrics*, vol. 118, no. 1-2, pp. 257–291, 2004.

[14] D. Bauer and M. Wagner, "Estimating cointegrated systems using subspace algorithms," *Journal of Econometrics*, vol. 111, no. 1, pp. 47–84, 2002.

[15] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, 2018.

[16] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning Without Mixing: Towards A Sharp Analysis of Linear System Identification," *arXiv preprint arXiv:1802.08334*, 2018.

[17] T. Sarkar and A. Rakhlin, "Near optimal finite time identification of arbitrary linear dynamical systems," *arXiv preprint arXiv:1812.01251*, 2018.

[18] S. Fattahi, N. Matni, and S. Sojoudi, "Learning sparse dynamical systems from a single sample trajectory," *arXiv preprint arXiv:1904.09396*, 2019.

[19] S. Oymak and N. Ozay, "Non-asymptotic Identification of LTI Systems from a Single Trajectory," *arXiv preprint arXiv:1806.05722*, 2018.

[20] M. Simchowitz, R. Boczar, and B. Recht, "Learning Linear Dynamical Systems with Semi-Parametric Least Squares," *arXiv preprint arXiv:1902.00768*, 2019.

[21] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Finite-Time System Identification for Partially Observed LTI Systems of Unknown Order," *arXiv preprint arXiv:1902.01848*, 2019.

[22] E. Weyer, R. C. Williamson, and I. M. Mareels, "Finite sample properties of linear model identification," *IEEE Transactions on Automatic Control*, vol. 44, no. 7, pp. 1370–1383, 1999.

[23] M. C. Campi and E. Weyer, "Finite sample properties of system identification methods," *IEEE Transactions on Automatic Control*, vol. 47, no. 8, pp. 1329–1334, 2002.

[24] M. Vidyasagar and R. L. Karandikar, "A learning theory approach to system identification and stochastic adaptive control," *Journal of Process Control*, vol. 18, no. 3-4, pp. 421–430, 2008.

[25] E. Hazan, H. Lee, K. Singh, C. Zhang, and Y. Zhang, "Spectral filtering for general linear dynamical systems," in *Advances in Neural Information Processing Systems*, 2018, pp. 4634–4643.

[26] M. Hardt, T. Ma, and B. Recht, "Gradient descent learns linear dynamical systems," *Journal of Machine Learning Research*, vol. 19, no. 29, pp. 1–44, 2018.

[27] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *arXiv preprint arXiv:1710.01688*, 2017.

[28] A. Rantzer, "Concentration bounds for single parameter adaptive control," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 1862–1866.

[29] M. Kozdoba, J. Marecek, T. Tchrakian, and S. Mannor, "On-line learning of linear dynamical systems: Exponential forgetting in kalman filters," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4098–4105.

[30] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018, vol. 47.
[31] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, 2011, pp. 2312–2320.
[32] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *International Conference on Machine Learning*, 2016, pp. 964–973.
[33] A. Carè, B. C. Csáji, M. C. Campi, and E. Weyer, "Finite-sample system identification: An overview and a new correlation method," *IEEE Control Systems Letters*, vol. 2, no. 1, pp. 61–66, 2018.
[34] B. Anderson and J. Moore, *Optimal Filtering*. Dover Publications, 2005.
[35] P.-Å. Wedin, "Perturbation theory for pseudo-inverses," *BIT Numerical Mathematics*, vol. 13, no. 2, pp. 217–232, 1973.
[36] ——, "Perturbation bounds in connection with singular value decomposition," *BIT Numerical Mathematics*, vol. 12, no. 1, pp. 99–111, 1972.
[37] L. Pernebo and L. Silverman, "Model reduction via balanced state space representations," *IEEE Transactions on Automatic Control*, vol. 27, no. 2, pp. 382–387, 1982.
[38] P. Van Overschee and B. De Moor, "A Unifying Theorem for Three Subspace System Identification Algorithms," *Automatica*, vol. 31, no. 12, pp. 1853–1864, 1995.
[39] F. Krahmer, S. Mendelson, and H. Rauhut, "Suprema of chaos processes and the restricted isometry property," *Communications on Pure and Applied Mathematics*, vol. 67, no. 11, pp. 1877–1904, 2014.

## APPENDIX

### A. Persistence of Excitation

The following result shows that with high probability, the past noises are persistently exciting. It follows from Lemma C.2 of [19], which in turn is based on results for random circulant matrices [39].

**Lemma 1 (Noise PE):** Consider the conditions of Theorem 2 and the definition of $\delta_N$ in (18). There exists a universal constant $c$ (independent of system and algorithm parameters) such that if $N \geq 2cpm \log 1/\delta_N$, then with probability at least $1-\delta_N$ the event:

$$\mathcal{E}_E = \left\{ \frac{1}{N} \mathcal{T}_p E_- E_-^* \mathcal{T}_p^* \succeq \frac{1}{2} \Sigma_E \right\},$$

occurs, where $\Sigma_E$ is defined in (16). $\diamond$

From the above lemma it follows that $N$ should be large enough to guarantee PE for the noises. In particular, $N$ should be larger than $N_0$, where

$$N_0 = \min \left\{ N : \; N \geq 2cpm \log 1/\delta_N \right\}. \tag{30}$$

Such a $N_0$ exists since the term $2cpm \log 1/\delta_N$ depends logarithmically on $N$–see definition (18).

**Lemma 2:** Let $\Sigma_E$ be as in (16). Then:

$$\sigma_E \geq \sigma_{\min}(R) > 0. \qquad \diamond$$

**Lemma 3 (Markov upper bounds):** Consider system (2) and recall the definition of $\Gamma_k$, $\Sigma_E$ in (4), (16). We have the following upper bounds:

$$\mathbb{P}\left( \left\| \hat{X}\hat{X}^* \right\|_2 \geq N \frac{\operatorname{tr} \Gamma_{N-1}}{\delta} \right) \leq \delta \tag{31}$$

$$\mathbb{P}\left( \left\| Y_- Y_-^* \right\|_2 \geq N \frac{\|\mathcal{O}_p\|_2^2 \operatorname{tr} \Gamma_{N-1} + \operatorname{tr} \Sigma_E}{\delta} \right) \leq \delta. \tag{32}$$

*Sketch of proof of Theorem 2*

Some arguments are inspired from Section 9 of [17].

**Step 1: Noise PE.** Under the condition $N \geq N_0$, from Lemma 1 the event $\mathcal{E}_E$ occurs with prob. at least $1-\delta_N$.

**Step 2: Cross terms are small.** Define:

$$\bar{V}_N = N/\|\mathcal{O}_p\|_2^2 I_n + \hat{X}\hat{X}^*, \quad V_N = \hat{X}\hat{X}^*, \quad S_N = \hat{X}E_-^*.$$

By applying Theorem 3 to $\bar{V}_N, V_N, S_N$ and Lemma 3 to $\hat{X}\hat{X}^*$, we obtain that the event:

$$\mathcal{E}_{XE} = \left\{ \left\| \bar{V}_N^{-1/2} S_N \right\|_2^2 \leq \mathcal{C}_{XE} \left\| \bar{R} \right\|_2 \right\},$$

occurs with probability at least $1-2\delta$, where

$$\mathcal{C}_{XE} \triangleq 8p \left( \frac{n}{2} \log \left( \frac{\|\mathcal{O}_p\|^2 \operatorname{tr} \Gamma_{N-1}}{\delta} + 1 \right) + \log \frac{p5^m}{\delta} \right).$$

Then, if $u \in \mathbb{R}^{mp}$, $\|u\|_2 = 1$ is an arbitrary unit vector:

$$\begin{aligned} |u^* \mathcal{O}_p \hat{X} E_-^* \mathcal{T}_p^* u| &\leq \|u^* \mathcal{O}_p \bar{V}_N^{1/2} \bar{V}_N^{-1/2} S_N \mathcal{T}_p^*\|_2 \\ &\leq \sqrt{u^* \mathcal{O}_p \hat{X}\hat{X}^* \mathcal{O}_p^* u + N \frac{u^* \mathcal{O}_p \mathcal{O}_p^* u}{\|\mathcal{O}_p\|_2^2}} \sqrt{\mathcal{C}_{XE} \left\| \bar{R} \right\|_2} \|\mathcal{T}_p\|_2 \\ &\leq \sqrt{u^* \mathcal{O}_p \hat{X}\hat{X}^* \mathcal{O}_p^* u + N} \sqrt{\mathcal{C}_{XE} \left\| \bar{R} \right\|_2} \|\mathcal{T}_p\|_2 \end{aligned} \tag{33}$$

**Step 3: Output PE** Consider an arbitrary unit vector $u \in \mathbb{R}^{mp}$, $\|u\|_2 = 1$. Consider the events $\mathcal{E}_E$ and $\mathcal{E}_{XE}$ from steps 1,2. With probability $1-\delta_N-2\delta$, since $N \geq N_0$ the event $\mathcal{E}_E \cap \mathcal{E}_{XE}$ occurs. It remains to show that on $\mathcal{E}_E \cap \mathcal{E}_{XE}$ the outputs satisfy PE for sufficiently large $N$. Define

$$\alpha \triangleq \frac{1}{N} u^* \mathcal{O}_p \hat{X}\hat{X}^* \mathcal{O}_p^* u, \; \beta \triangleq \frac{1}{N} u^* \mathcal{T}_p E_- E_-^* \mathcal{T}_p^* u$$

From (25), (33) for $N \geq N_0$ on $\mathcal{E}_E \cap \mathcal{E}_{XE}$:

$$\frac{1}{N} u^* Y_- Y_-^* u \geq \alpha + \beta - \underbrace{2 \|\mathcal{T}_p\|_2 \sqrt{\frac{\mathcal{C}_{XE} \left\| \bar{R} \right\|_2}{N}}}_{\gamma_N} \sqrt{\alpha + 1}$$

with $\beta \geq \sigma_E/2$. Now let $N_1$ be such that:

$$N_1 = \min \left\{ N : \; \gamma_N \leq \min \left\{ 1, \frac{\sigma_E}{4} \right\} \right\}. \tag{34}$$

Since $\mathcal{C}_{XE}$ grows at most logarithmically with $N$, $N_1$ always exists. Now, since $N \geq N_1$ and $\beta \geq \sigma_E/2$:

$$\alpha + \beta - \gamma_N \sqrt{\alpha + 1} \geq \frac{\alpha + \beta}{2}.$$

The above inequality follows by elementary calculus. ■

### B. Definition of $N_2$.

We define:

$$N_2 = \min \left\{ N : \; 8 \sqrt{\frac{\|\bar{R}\|_2}{\sigma_E}} \|\mathcal{T}_p\|_2 \frac{\mathcal{C}_N}{\sqrt{N}} \leq 1 \right\}. \tag{35}$$

where

$$\mathcal{C}_N = \sqrt{\frac{mp^2}{2} \log \left( \frac{2 \|\mathcal{O}_p\|_2^2 \Gamma_{N-1}}{\delta \sigma_E} + 1 \right) + p \log \frac{p5^m}{\delta}}$$

Such an $N_2$ exists since $\mathcal{C}_N$ grows at most logarithmically with $N$.