Recursive Least Squares Identification with Variable-Direction Forgetting via Oblique Projection Decomposition

Kun Zhu, Chengpu Yu, Yiming Wan

***Abstract*—In this paper, a new recursive least squares identification algorithm with variable-direction forgetting (VDF) is proposed for multi-output systems. The objective is to enhance parameter estimation performance under non-persistent excitation. The proposed algorithm performs oblique projection decomposition of the information matrix, such that forgetting is applied only to directions where new information is received. Theoretical proofs show that even without persistent excitation, the information matrix remains lower and upper bounded, and the estimation error variance converges to be within a finite bound. Moreover, detailed analysis is made to compare with a recently reported VDF algorithm that exploits eigenvalue decomposition (VDF-ED). It is revealed that under non-persistent excitation, part of the forgotten subspace in the VDF-ED algorithm could discount old information without receiving new data, which could produce a more ill-conditioned information matrix than our proposed algorithm. Numerical simulation results demonstrate the efficacy and advantage of our proposed algorithm over this recent VDF-ED algorithm.**

***Index Terms*—Recursive least squares, non-persistent excitation, variable-direction forgetting, oblique projection.**

I. Introduction

RESEARCH on system identification dates back to the 1960s, but is still very active due to its critical importance in systems and controls [1], [2]. For online parameter estimation, recursive least squares (RLS) identification is one of the most well-known methods [3]. To enhance tracking capability of time-varying parameters, exponential forgetting (EF) was initially established for RLS identification of single-output (SO) systems, which discounts old information with a constant forgetting factor [3]. Various RLS extensions with or without EF have been proposed for multiple-output (MO) systems that are ubiquitous in industrial applications [4]–[11]. The parameter errors given by the EF algorithms exponentially converge if the identification data is persistently exciting [12], [13]. However, the condition of persistent excitation cannot be always satisfied in practice. With non-persistent excitation, the EF algorithm discounts old data without receiving sufficient new information. As a result, the undesirable estimator windup phenomenon occurs, i.e., the RLS gain grows unbounded, and the obtained estimates become highly sensitive to noise.

The above limitation of EF in the absence of persistent excitation is attributed to discounting old information uniformly over time and in the parameter space. To cope with this issue, various modified forgetting strategies have been reported in the literature, which can be classified into two categories: variable-rate forgetting (VRF) and variable-direction forgetting (VDF). The category of VRF algorithms adjusts a variable forgetting factor to discount old information non-uniformly over time. For example, the forgetting factor is updated according to the prediction error [14], [15] by minimizing the mean square error [16] or in accordance with Bayesian decision-making [17]. Convergence and consistency of a general VRF algorithm was recently investigated in [18]. However, data excitation in practice is not uniformly distributed over space, but might be restricted to certain directions of the parameter space over a period of time. In this case, the VRF algorithms still gradually lose information in the non-excited directions, which would lead to ill-conditioned matrix inversion and increased estimation errors [19]. This problem is addressed by the VDF algorithms in [19]–[21]. Specifically, forgetting is applied only to directions that are excited by the online data. By doing so, estimator windup does not occur under non-persistent excitation, because information in the non-excited subspace is retained.

The VRF and VDF algorithms were initially proposed for SO systems. Considering MO systems, the VRF algorithm is still applicable since it simply applies uniform forgetting to the entire parameter space [18], [22]. However, the extension of VDF algorithms to cope with MO systems is not straightforward, since the forgotten subspace varies with the online data. As the latest progress in this line of research, a VDF algorithm via eigenvalue decomposition (VDF-ED) has been proposed in [23], [24] for MO systems. The basic idea is to apply forgetting to the eigendirections of the old information matrix where new information is received. Moreover, this VDF-ED algorithm is combined with a variable forgetting factor to further enhance its tracking performance [24].

In this paper, a new VDF algorithm using oblique projection decomposition (VDF-OPD) is proposed for MO systems under non-persistent excitation. Oblique projection is exploited to decompose the old information matrix into a forgotten part and a retained part. This proposed VDF-OPD algorithm has three main contributions:

i) The proposed decomposition of the information matrix has a clear geometrical interpretation based on oblique projection. It reduces to the decomposition described in [21] when the considered system has a scalar output.
ii) A detailed comparison with the recently proposed VDF-ED algorithm in [23] is provided. The forgotten subspace in the VDF-ED algorithm has a higher dimension than that in our VDF-OPD algorithm. Under non-persistent excitation, the VDF-ED algorithm produces a

This work was supported by the National Natural Science Foundation of China (61803163, 61991414, 61873301). (*Corresponding author: Yiming Wan.*)

Kun Zhu and Yiming Wan are with Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: ywan@hust.edu.cn).

Chengpu Yu is with Beijing Institute of Technology Chongqing Innovation Center, China (e-mail: yuchengpu@bit.edu.cn).

more ill-conditioned information matrix, because part of its forgotten subspace discounts old information without receiving new data.

iii) Boundedness of the information matrix and convergence of the estimation error variance of our VDF-OPD algorithm are proved under non-persistent excitation.

The rest of this paper is organized as follows. Firstly, Section II states the problem of RLS identification of MO systems under non-persistent excitation. Our proposed VDF-OPD algorithm is presented in Section III, and compared with the VDF-EM algorithm in Section IV. Then, Section V gives the convergence analysis. Finally, simulation results and concluding remarks are provided in Sections VI and VII, respectively.

Notation: The 2-norm of a vector x is denoted by $\|x\|$. For a matrix X, Range(X), Null(X), $\|X\|_2$, and $X^\dagger$ represent its range space, nullspace, induced 2-norm, and Moore-Penrose inverse, respectively. For a square matrix X, tr(X) denotes its trace, and $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ represent its minimal and maximal eigenvalues, respectively. For a symmetric matrix X, the positive definiteness and positive semi-definiteness are denoted by $X > 0$ and $X \geq 0$, respectively. Let I_n represent an identity matrix of dimension n. The vectorization operator $\mathrm{vec}(X)$ creates a column vector by stacking the columns vectors of a matrix X. For matrices X and Y, diag(X, Y) represents a block-diagonal matrix whose diagonal blocks are X and Y.

II. Problem Statement

Consider the following MO system [25]

$$A_k(z^{-1})y_k = B_k(z^{-1})u_k + v_k, \tag{1}$$

where $y_k \in \mathbb{R}^{m_y}$ denotes the measured output vector at time instant k, $u_k \in \mathbb{R}^{m_u}$ is the system input vector, and $v_k \in \mathbb{R}^{m_y}$ represents the stochastic noise vector with zero mean. With the unit backward shift operator z^{-1} (i.e., $z^{-1}y_k = y_{k-1}$), $A_k(z^{-1})$ and $B_k(z^{-1})$ are the polynomial matrices defined as

$$\begin{aligned} A_k(z^{-1}) &= I_{m_y} + A_{1,k}z^{-1} + A_{2,k}z^{-2} + \cdots + A_{n_a,k}z^{-n_a}, \\ B_k(z^{-1}) &= B_{0,k} + B_{1,k}z^{-1} + B_{2,k}z^{-2} + \cdots + B_{n_b,k}z^{-n_b}. \end{aligned}$$

which include slowly time-varying parameters in their coefficient matrices. Define

$$\begin{aligned} \Theta_k &= \begin{bmatrix} A_{1,k} & \cdots & A_{n_a,k} & B_{0,k} & \cdots & B_{n_b,k} \end{bmatrix}^\top, \\ \varphi_k &= \begin{bmatrix} -y_{k-1}{}^\top & \cdots & -y_{k-n_a}{}^\top & u_k^\top & \cdots & u_{k-n_b}^\top \end{bmatrix}^\top, \\ n_1 &= n_a m_y + (n_b+1)m_u, \end{aligned}$$

with $\Theta_k \in \mathbb{R}^{n_1 \times m_y}$, $\varphi_k \in \mathbb{R}^{n_1}$. Then, the system model (1) is written into

$$y_k = \Theta_k^\top \varphi_k + v_k. \tag{2}$$

With the property of Kronecker product [26], i.e., $\mathrm{vec}(\Theta_k^\top \varphi_k) = (\varphi_k^\top \otimes I_{m_y})\mathrm{vec}(\Theta_k^\top)$, (2) can be further expressed as

$$y_k = \Phi_k^\top \theta_k + v_k, \tag{3}$$

where the parameter vector θ_k and the regressor matrix Φ_k are defined as

$$\begin{aligned} \theta_k &= \mathrm{vec}(\Theta_k^\top) \in \mathbb{R}^n, \ n = n_1 m_y, \\ \Phi_k &= (\varphi_k^\top \otimes I_{m_y})^\top \in \mathbb{R}^{n \times m_y}. \end{aligned} \tag{4}$$

To estimate the parameter vector θ_k in (3), the standard RLS algorithm with EF is [22]

$$\hat{\theta}_k = \hat{\theta}_{k-1} + R_k^{-1}\Phi_k(y_k - \Phi_k^\top \hat{\theta}_k), \tag{5a}$$

$$R_k = \mu_k R_{k-1} + \Phi_k \Phi_k^\top, \tag{5b}$$

where $\hat{\theta}_k$ is the parameter estimate, $R_k \in \mathbb{R}^{n \times n}$ is called the information matrix, and $\mu_k \in (0, 1)$ is the forgetting factor.

The above EF algorithm works well if the regressor sequence $\{\Phi_k\}$ is persistently exciting [12], [13], i.e., there exist $\alpha > 0$ and a positive integer s_0 such that $\sum_k^{k+s_0} \Phi_k \Phi_k^\top \geq \alpha I_n$ holds for all $k > 0$. The persistently exciting data contains rich new information to compensate for discounted old data. However, under non-persistent excitation, the old information in R_k could be discounted continuously without being fully replaced by any new information from Φ_k. As a result, some eigenvalues of R_k tend to be zero, and the corresponding gain $R_k^{-1}\Phi_k$ becomes unbounded, i.e., the undesirable estimator windup occurs. In this situation, the obtained parameter estimates become highly sensitive to noise.

To address the estimator windup under non-persistent excitation, various VDF strategies have been reported in the literature for SO systems [19]–[21]. However, these VDF algorithms consider only a regressor vector, thus cannot cope with the regressor matrix Φ_k for MO systems. In this paper, we propose the VDF-OPD algorithm for MO systems, analyze its benefit over the VDF-ED algorithm recently reported in [23], and investigate its convergence properties.

III. RLS with Variable-Direction Forgetting via Oblique Projection Decomposition

For the RLS identification, the basic idea of VDF is to apply forgetting only to directions that receive new information [21]. Following this idea, (5b) is modified by decomposing the old information matrix R_{k-1} into two disjoint parts as

$$R_{k-1} = R_{k-1}^{(1)} + R_{k-1}^{(2)}, \tag{6}$$

such that $R_{k-1}^{(1)}$ and $R_{k-1}^{(2)}$ represent the retained part and the forgotten part at time k, respectively. In this section, the above decomposition is performed via oblique projection, and the VDF-OPD algorithm is proposed for MO systems. For the sake of self-containedness, necessary preliminaries on oblique projection are given in Appendix A.

In the following derivations, we assume $\Phi_k \neq 0$ and $R_{k-1} > 0$. Note that $R_{k-1} > 0$ will be proved later in Theorem 3. For MO systems, the following requirements are imposed for the decomposition in (6):

i) $R_{k-1}^{(1)}$ is the retained part which satisfies

$$\Phi_k^\top R_{k-1}^{(1)} = 0. \tag{7}$$

This means that the retained information should reside in a subspace that is orthogonal to the range space

of the new regressor matrix Φ_k, i.e., Range $\left(R_{k-1}^{(1)}\right)$ should be orthogonal to Range(Φ_k), or equivalently, Range $\left(R_{k-1}^{(1)}\right) \subseteq$ Null$(\Phi_k^\top)$.

ii) $R_{k-1}^{(2)}$ is the forgotten part which satisfies

$$\Phi_k^\top R_{k-1}^{(2)} = \Phi_k^\top R_{k-1}, \tag{8}$$

according to (6) and (7). This means that the forgotten part $R_{k-1}^{(2)}$ and the old information matrix R_{k-1} have the same amount of correlation with Φ_k.

iii) The two decomposed parts are disjoint, i.e.,

$$\text{Range}\left(R_{k-1}^{(1)}\right) \cap \text{Range}\left(R_{k-1}^{(2)}\right) = \{0\}. \tag{9}$$

iv) Positive semi-definiteness, i.e.,

$$R_{k-1}^{(1)} \geq 0, R_{k-1}^{(2)} \geq 0. \tag{10}$$

Geometrically, the above requirements can be satisfied by applying oblique projection to R_{k-1}. Define two complementary subspaces $\mathcal{V}_{k-1}$ and $\tilde{\mathcal{V}}_{k-1}$ in $\mathbb{R}^n$, satisfying

$$\mathcal{V}_{k-1} + \tilde{\mathcal{V}}_{k-1} = \mathbb{R}^n,\ \mathcal{V}_{k-1} \cap \tilde{\mathcal{V}}_{k-1} = \{0\}. \tag{11}$$

Let $P_{\mathcal{V}_{k-1}|\tilde{\mathcal{V}}_{k-1}}$ represent the oblique projection onto the subspace $\mathcal{V}_{k-1}$ along $\tilde{\mathcal{V}}_{k-1}$. According to Lemma 1 in Appendix A, $P_{\tilde{\mathcal{V}}_{k-1}|\mathcal{V}_{k-1}} = I_n - P_{\mathcal{V}_{k-1}|\tilde{\mathcal{V}}_{k-1}}$ is the oblique compliment that projects onto $\tilde{\mathcal{V}}_{k-1}$ along $\mathcal{V}_{k-1}$. By applying the above two complementary oblique projections, the decomposition in (6) is obtained as

$$R_{k-1}^{(1)} = P_{\tilde{\mathcal{V}}_{k-1}|\mathcal{V}_{k-1}} R_{k-1},\ R_{k-1}^{(2)} = P_{\mathcal{V}_{k-1}|\tilde{\mathcal{V}}_{k-1}} R_{k-1}. \tag{12}$$

This decomposition implies

$$\text{Range}\left(R_{k-1}^{(1)}\right) = \tilde{\mathcal{V}}_{k-1} \text{ and Range}\left(R_{k-1}^{(2)}\right) = \mathcal{V}_{k-1} \tag{13}$$

since R_{k-1} is non-singular. Then, requirements (7)–(9) for the above decomposition are satisfied by setting

$$\tilde{\mathcal{V}}_{k-1} \subseteq \text{Null}(\Phi_k^\top) \tag{14}$$

according to Lemma 1 in Appendix A. As indicated by (13), $\mathcal{V}_{k-1}$ and $\tilde{\mathcal{V}}_{k-1}$ are the forgotten and retained subspaces, respectively.

It is reasonable to require that information in the entire subspace Null$(\Phi_k^\top)$ is all retained, i.e.,

$$\tilde{\mathcal{V}}_{k-1} = \text{Range}\left(R_{k-1}^{(1)}\right) = \text{Null}(\Phi_k^\top). \tag{15}$$

Otherwise, certain directions within Null$(\Phi_k^\top)$ would be included in the forgotten subspace $\mathcal{V}_{k-1}$, and old information in those directions would be discounted without being compensated by new information from Φ_k.

Being a complement subspace of $\tilde{\mathcal{V}}_{k-1}$, $\mathcal{V}_{k-1}$ is to be determined such that $R_{k-1}^{(2)} = P_{\mathcal{V}_{k-1}|\tilde{\mathcal{V}}_{k-1}} R_{k-1} \geq 0$, as required in (10). For this purpose, one solution is

$$\mathcal{V}_{k-1} = \text{Range}(R_{k-1}\Phi_k), \tag{16}$$

and the corresponding oblique projection matrix onto $\mathcal{V}_{k-1}$ along $\tilde{\mathcal{V}}_{k-1}$ is

$$P_{\mathcal{V}_{k-1}|\tilde{\mathcal{V}}_{k-1}} = R_{k-1}\Phi_k(\Phi_k^\top R_{k-1}\Phi_k)^\dagger \Phi_k^\top$$

according to Lemma 1 in Appendix A. Therefore, $R_{k-1}^{(1)}$ and $R_{k-1}^{(2)}$ in (12) are

$$\begin{aligned} R_{k-1}^{(1)} &= R_{k-1} - R_{k-1}\Phi_k(\Phi_k^\top R_{k-1}\Phi_k)^\dagger \Phi_k^\top R_{k-1}, \\ R_{k-1}^{(2)} &= R_{k-1}\Phi_k(\Phi_k^\top R_{k-1}\Phi_k)^\dagger \Phi_k^\top R_{k-1}. \end{aligned} \tag{17}$$

Theorem 1. *Both $R_{k-1}^{(1)}$ and $R_{k-1}^{(2)}$ in (17) are positive semidefinite if $R_{k-1} > 0$.*

The proof is given in Appendix B. Theorem 1 shows that the requirement (10) is achieved.

Remark 1. *As a complement subspace of $\tilde{\mathcal{V}}_{k-1}$, the selection of $\mathcal{V}_{k-1}$ is non-unique. But not all such selections can ensure the symmetry and positive semi-definiteness of $R_{k-1}^{(2)}$. For example, a natural choice of $\mathcal{V}_{k-1}$ is $\mathcal{V}_{k-1} = Range(\Phi_k)$, then the oblique projection matrix $P_{\mathcal{V}_{k-1}|\tilde{\mathcal{V}}_{k-1}}$ becomes*

$$P_{\mathcal{V}_{k-1}|\tilde{\mathcal{V}}_{k-1}} = \Phi_k\left(\Phi_k^\top\Phi_k\right)^\dagger \Phi_k^\top.$$

However, the resulting $R_{k-1}^{(2)}$ in (12) is non-symmetric.

In order to have a well-conditioned Moore-Penrose inverse in (17), a dead zone is introduced as below for the regressor matrix Φ_k:

$$R_{k-1}^{(2)} = 0, \quad \text{if } \|\Phi_k\|_2 < \epsilon, \tag{18}$$

where ϵ is determined by the noise level in the data. If Φ_k lies in the above dead zone, Φ_k is dominated by noise and carries little new information. In this case, the VDF algorithm should not forget any old information in R_{k-1}, and the decomposition (6) is not performed.

By applying a variable forgetting factor μ_k only to $R_{k-1}^{(2)}$, the information matrix R_k is updated by

$$R_k = R_{k-1}^{(1)} + \mu_k R_{k-1}^{(2)} + \Phi_k\Phi_k^\top. \tag{19}$$

The variable forgetting factor μ_k is introduced to further improve tracking capability of the proposed VDF-OPD algorithm. Various VRF strategies, such as those found in [14]–[16], can be used to update μ_k adaptively. In this paper, μ_k is adjusted according to the prediction error

$$\boldsymbol{e}_k = y_k - \Phi_k^\top\hat{\theta}_{k-1} = \Phi_k^\top\left(\theta_k - \hat{\theta}_{k-1}\right) + v_k, \tag{20}$$

where $\hat{\theta}_{k-1}$ is the parameter estimated at time $k-1$. A large prediction error $\boldsymbol{e}_k$ implies a large parameter estimation error $\theta_k - \hat{\theta}_{k-1}$. To increase the sensitivity to the parameter variations, the forgetting factor μ_k must decrease when the prediction error $\boldsymbol{e}_k$ is large. Therefore, we use the following VRF strategy by modifying the idea in [14] for MO systems:

$$\mu_k = \max\left\{\mu_L, 1 - \frac{1}{(\eta + \boldsymbol{e}_k^\top\boldsymbol{e}_k)}\frac{\boldsymbol{e}_k^\top\boldsymbol{e}_k}{m_y + \text{tr}(\Phi_k^\top P_{k-1}\Phi_k)}\right\}, \tag{21}$$

where μ_L represents the lower bound of μ_k and η is a positive constant chosen by the user. The user-defined constant η can be viewed as a sensitivity factor: a smaller η leads to higher sensitivity of μ_k to variations of $\boldsymbol{e}_k$. As can be seen from (21), when the prediction error $\boldsymbol{e}_k$ increases, a smaller forgetting factor is used such that the parameter estimate tracks the time-varying parameters at a faster rate.

The above proposed VDF-OPD algorithm is summarized in Algorithm 1. Note that (23) is derived from (19) and (6). When the considered system (1) has only a scalar output, the regressor Φ_k defined in (4) becomes a vector, and Algorithm 1 reduces to the one proposed in [21].

Algorithm 1 Proposed VDF-OPD algorithm

Initialization: $\theta_0, \mu_L, R_0, \epsilon, \eta$
Input: Φ_k in (4) and y_k
Calculate the prediction error $\boldsymbol{e}_k$ with (20);
Adjust μ_k using (21);
if $\|\Phi_k\|_2 < \epsilon$ **then**
  $R_{k-1}^{(2)} = 0$
else
  Compute $R_{k-1}^{(2)}$ according to (17)
end if
Update R_k and $\hat{\theta}_k$:

$$\hat{\theta}_k = \hat{\theta}_{k-1} + R_k^{-1}\Phi_k(y_k - \Phi_k^\top \hat{\theta}_{k-1}), \tag{22}$$

$$R_k = R_{k-1} - (1-\mu_k)R_{k-1}^{(2)} + \Phi_k\Phi_k^\top. \tag{23}$$

IV. Comparison with variable-direction forgetting via eigenvalue decomposition

Recent progress made in the VDF-ED algorithm in [23] is applicable to MO systems, thus is closely related to our VDF-OPD algorithm. However, theoretical analysis of VDF-ED in [23] considers only the condition of persistent excitation, e.g., see Proposition 10 in [23]. Then, it is of interest to compare these two VDF algorithm under non-persistent excitation.

In the VDF-ED algorithm, the information matrix R_k is updated by [23]

$$R_k = U_{k-1}\Lambda\Sigma_{k-1}\Lambda U_{k-1}^\top + \Phi_k^\top\Phi_k, \tag{24}$$

where the orthonormal matrix U_{k-1} and the diagonal matrix Σ_{k-1} consist of eigenvectors and eigenvalues obtained from the eigenvalue decomposition $R_{k-1} = U_{k-1}\Sigma_{k-1}U_{k-1}^\top$. The diagonal matrix Λ in (24) applies forgetting to the direction of the ith column of U_{k-1} if the amount of new information along this direction is above a threshold, i.e., the diagonal entries of Λ are defined as

$$\Lambda(i,i) = \begin{cases} \sqrt{\lambda_k}, & \text{if } \|\mathrm{col}_i(\Psi_k)\| > \epsilon_{\text{th}}, \\ 1, & \text{otherwise}, \end{cases} \tag{25}$$

where

$$\Psi_k = \Phi_k^\top U_{k-1}, \tag{26}$$

$\mathrm{col}_i(\Psi_k)$ is the ith column of Ψ_k that represents the information content of the regressor matrix Φ_k along the ith column of U_{k-1}, $\lambda_k \in (0,1)$ is the forgetting factor, and ϵ_{th} is a user-defined scalar which should be larger than the noise level.

To facilitate the following analysis, according to (25), $R_{k-1} = U_{k-1}\Sigma_{k-1}U_{k-1}^\top$ is rewritten as

$$\begin{aligned} R_{k-1} &= \begin{bmatrix} U_{1,k-1} & U_{2,k-1} \end{bmatrix} \mathrm{diag}(\Sigma_{1,k-1}, \Sigma_{2,k-1}) \begin{bmatrix} U_{1,k-1}^\top \\ U_{2,k-1}^\top \end{bmatrix} \\ &= U_{1,k-1}\Sigma_{1,k-1}U_{1,k-1}^\top + U_{2,k-1}\Sigma_{2,k-1}U_{2,k-1}^\top \end{aligned} \tag{27}$$

where both $U_{1,k-1}$ and $U_{2,k-1}$ consist of columns of U_{k-1}, and satisfy

$$\left\|\mathrm{col}_i(\Phi_k^\top U_{1,k-1})\right\| \le \epsilon_{\text{th}} \text{ and } \left\|\mathrm{col}_i(\Phi_k^\top U_{2,k-1})\right\| > \epsilon_{\text{th}} \tag{28}$$

respectively. With (24) and (25), the old information in $\mathrm{Range}(U_{1,k-1})$ is retained, while the old information in $\mathrm{Range}(U_{2,k-1})$ is forgotten. Therefore, the information update in (24) and (25) can be expressed in a form similar to (19), i.e.,

$$R_k = M_{k-1}^{(1)} + \lambda_k M_{k-1}^{(2)} + \Phi_k\Phi_k^\top, \tag{29a}$$

$$M_{k-1}^{(1)} = U_{1,k-1}\Sigma_{1,k-1}U_{1,k-1}^\top, \tag{29b}$$

$$M_{k-1}^{(2)} = U_{2,k-1}\Sigma_{2,k-1}U_{2,k-1}^\top. \tag{29c}$$

Both $R_{k-1}^{(2)}$ in (19) and $M_{k-1}^{(2)}$ in (29a) are the forgotten parts in the above two forgetting algorithms.

Theorem 2. *Assume $R_{k-1} > 0$. Consider the noise-free case. Set $\epsilon = 0$ in (18) and $\epsilon_{th} = 0$ in (25). Our proposed VDF-OPD algorithm differs from VDF-ED in the adopted two decompositions (17) and (29), i.e.,*

$$\mathit{rank}(\Phi_k) = \mathit{rank}\left(R_{k-1}^{(2)}\right) \le \mathit{rank}\left(M_{k-1}^{(2)}\right), \tag{30}$$

$$\mathit{Null}\left(\Phi_k^\top\right) = \mathit{Range}\left(R_{k-1}^{(1)}\right) \supseteq \mathit{Range}\left(M_{k-1}^{(1)}\right). \tag{31}$$

The proof is given in Appendix C. In the noisy case, we still have (30) and (31) if the amount of informative data in Φ_k is significantly larger than noise, and the corresponding proof follows the same idea in Appendix C but with more tedious derivations.

As indicated by (30) and (31), the retained part $R_{k-1}^{(1)}$ in our proposed VDF-OPD algorithm corresponds to the entire subspace $\mathrm{Null}(\Phi_k^\top)$. In contrast, for the VDF-ED algorithm, there exist certain scenarios that some subspace $\mathcal{S} \subset \mathrm{Null}(\Phi_k^\top)$ is not included in its retained part $M_{k-1}^{(1)}$, but added to its forgotten part $M_{k-1}^{(2)}$. Since the subspace $\mathcal{S}$ is orthogonal to the subspace spanned by Φ_k, i.e., $\mathcal{S} \subset \mathrm{Null}(\Phi_k^\top)$, the forgotten information in the subspace $\mathcal{S}$ cannot be compensated for by the new information in Φ_k. Due to forgetting in the subspace $\mathcal{S}$, the eigenvalues of R_{k-1} associated with $\mathcal{S}$ would be continuously discounted until they reach a value smaller than or equal to the threshold ϵ_{th} in (25). At this point, R_k is ill-conditioned, because some of its eigenvalues are closed to ϵ_{th} which is a small value at the noise level.

Example. Consider a noise-free ARX model $y_k = a_1 y_{k-1} + a_2 y_{k-2} + b_0 u_k$ whose output is a scalar signal. Assume that the system stays at the steady state with constant input and output signals, and Φ_* becomes a constant regressor vector $\Phi_* = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^\top$. We set $R_0 = I_3$ for the two considered forgetting algorithms. Consider a vector $\omega = \begin{bmatrix} -1 & 1 & 0 \end{bmatrix}^\top$ whose range space is a subset of $\mathrm{Null}(\Phi_*^\top)$. For our VDF-OPD algorithm, the forgotten part $R_{k-1}^{(2)}$ is of rank 1, and the retained part $R_{k-1}^{(1)}$ must include $\mathrm{Range}(\omega)$ according to (31). In contrast, it will be shown in the following that whether the VDF-ED algorithm in [23] includes $\mathrm{Range}(\omega)$ in the retained subspace depends on the orthonormal matrix U_0 in

the eigenvalue decomposition of R_0 in (24). For the VDF-ED algorithm in [23], if the orthonormal matrix U_0 in (24) is chosen to be $U_0 = I_3$, the forgotten part $M_{k-1}^{(2)}$ corresponds to the first two columns of U_0, while the retained part $M_{k-1}^{(1)}$ corresponds to the last columns of U_0. Consequently, Range(ω) lies in the forgotten subspace instead of the retained subspace, i.e., Range$(\omega) \subset$ Range $\left(M_{k-1}^{(2)}\right)$. Meanwhile, no new information is received along Range(ω) since the assumed constant regressor vector Φ_* is orthogonal to ω.

It should be also noted that the orthonormal matrix U_{k-1} in (24) is non-unique if R_{k-1} has identical eigenvalues. A different selection of eigenvectors in U_{k-1} might result in a different decomposition of R_{k-1} in (29). For instance, in the above example, if we choose

$$U_0 = \begin{bmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

the two retained subspaces in VDF-OPD and VDF-ED are identical, i.e., the range space spanned by the last two columns of U_0 given above. This issue caused by the non-unique orthonormal matrix U_{k-1} in (24) does not occur in our VDF-OPD algorithm.

V. Convergence Analysis

In this section, the convergence behavior of our proposed VDF-OPD algorithm under non-persistent excitation is investigated.

For this purpose, it is important to first analyze the boundedness of the information matrix R_k at all time instants [21]. With a lower bounded R_k, the algorithm gain $R_k^{-1}\Phi_k$ remains upper bounded, which prevents the estimator windup phenomenon. With an upper bounded R_k, the algorithm gain $R_k^{-1}\Phi_k$ does not approach zero, thus it retains its tracking capability. The following two theorems show that R_k is bounded from below and above without requiring persistent excitation. In contrast, the VDF-ED algorithm in [23] only analyzes the lower bound of R_k under persistent excitation, while the VDF algorithm in [21] is not applicable to MO systems in this paper.

Theorem 3. *Consider the recursive update of R_k in (23). With ϵ defined in (18), if $R_0 > 0$ and $\varepsilon \leq \|\Phi_k\|_2 < \infty$ for all $k > 0$, then i) $R_{k-1}^{(1)} + \mu_k R_{k-1}^{(2)} > 0$ for $\mu_k > 0$ and $k > 0$; and ii) there exists $\beta_k > 0$ such that $R_k > \beta_k I_n$ for all $k > 0$.*

Theorem 4. *With ϵ defined in (18), assume $\varepsilon \leq \|\Phi_k\|_2 < \infty$ at all $k > 0$. Then there exist a finite constant $\gamma > 0$ such that $R_k < \gamma I_n$ for all $k > 0$.*

Proofs of Theorems 3 and 4 are given in Appendices D and E, respectively.

To analyze the dynamics of parameter estimation errors, we assume θ_k in (3) to be constant, as in [19], [23]. Let θ represent the true constant parameter. Then, the estimation error is defined by

$$\tilde{\theta}_k = \theta - \hat{\theta}_k. \tag{32}$$

The following theorem shows that in the presence of noise, the estimation error variance converges to be within a finite bound.

Theorem 5. *With ϵ defined in (18), assume $\varepsilon < \|\Phi_k\|_2 \leq \infty, \forall k > 0$. Define*

$$\bar{R}_{k-1} = R_{k-1}^{(1)} + \mu_k R_{k-1}^{(2)}. \tag{33}$$

There exist $a \in (0,1)$ and $b \in (0,1)$ such that

$$\tilde{\theta}^\top \bar{R}_{k-1} R_k^{-1} \bar{R}_{k-1} \tilde{\theta} \leq a \tilde{\theta}^\top R_{k-1} \tilde{\theta}, \forall \tilde{\theta} \neq 0, \tag{34}$$

$$\Phi_k^\top R_k^{-1} \Phi_k \leq b I_{m_y} \tag{35}$$

hold for $k > 0$. Let δ represent the upper bound of the noise variance $\mathbb{E}(v_k^\top v_k)$. The expected estimation error is upper bounded as $\mathbb{E}\left\|\tilde{\theta}_k\right\|^2 \leq \frac{\zeta_k}{\beta_k}$, where β_k defined in Theorem 3 is the lower bound of R_k, $\{\zeta_k\}$ is the sequence generated by

$$\zeta_k = a\zeta_{k-1} + b\delta,\ \zeta_0 = \tilde{\theta}_0^\top R_0 \tilde{\theta}_0. \tag{36}$$

The bounding sequence $\{\zeta_k\}$ converges to $\zeta_\infty = \frac{b\delta}{1-a}$ as k goes to infinity, and monotonically decreases if $\zeta_k > \zeta_\infty$.

The proof is given in Appendix F.

Remark 2. *The convergence property holds only when the parameter is constant or its change rate is slower than the algorithm's convergence speed.*

VI. Simulation Study

In this section, we present a numerical example to show the efficacy of our proposed VDF-OPD algorithm and its advantage over the VDF-ED algorithm in [23].

The identification data is generated by the following MO system

$$y_1(k) = a_1(k)y_1(k-1) + a_2 u_1(k) + a_3(k)u_2(k) + v_1(k),$$
$$y_2(k) = b_1 y_1(k-1) + b_2 y_2(k-1) + b_3(k)u_2(k) + v_2(k).$$

whose parameters and input signals are

$$a_1(k) = -0.3 - 0.1\sin(k\pi/515),\ a_2 = 0.8,$$
$$a_3(k) = -0.2 - 0.1\cos(k/159),\ b_1 = 0.23,$$
$$b_2 = -0.67,\ b_3(k) = 0.43 - 0.1\sin(k/235),$$
$$u_1(k) = 10\sin(k\pi/140) + 10\cos(k\pi/187),$$
$$u_2(k) = 10\cos(k\pi/123).$$

This system model is equivalently written as

$$\bar{y}_k = \Phi_k^\top \theta_k + \bar{v}_k,$$

with

$$\Phi_k = \begin{bmatrix} y_1(k-1) & 0 \\ u_1(k) & 0 \\ u_2(k) & 0 \\ 0 & y_1(k-1) \\ 0 & y_2(k) \\ 0 & u_2(k) \end{bmatrix}^\top,$$

$$\theta_k = \begin{bmatrix} a_1(k) & a_2 & a_3(k) & b_1 & b_2 & b_3(k) \end{bmatrix}^\top,$$

$$\bar{y}_k = \begin{bmatrix} y_1(k) \\ y_2(k) \end{bmatrix}, \bar{v}_k = \begin{bmatrix} v_1(k) \\ v_2(k) \end{bmatrix}.$$

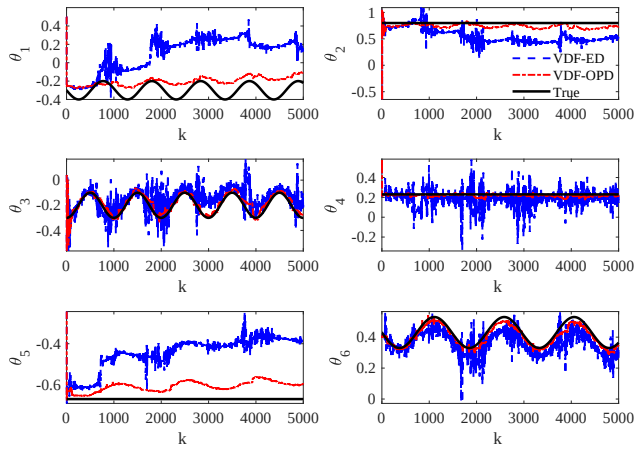


Fig. 1. The estimation results of two VDF algorithms.

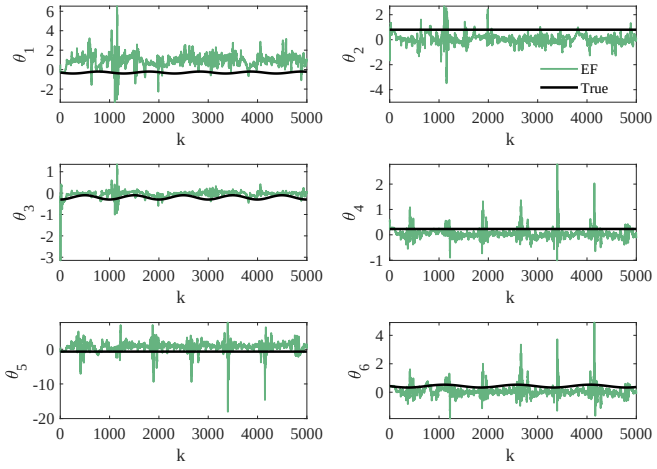


Fig. 2. The estimation results of EF algorithm.

TABLE I
RMSE OF PARAMETER ESTIMATES FROM EF, VDF-ED, AND VDF-OPD ALGORITHMS.

	EF	VDF-ED	VDF-OPD
θ_1	1.3722	0.4385	**0.1141**
θ_2	0.8438	0.2804	**0.0774**
θ_3	0.2114	0.0954	**0.0475**
θ_4	0.2616	0.0826	**0.0193**
θ_5	1.9325	0.2408	**0.0705**
θ_6	0.488	0.0746	**0.0241**

The measure noise v_k is Gaussian, with zero mean and covariance matrix $0.01I_2$.

Three RLS algorithms are implemented for comparisons: the EF algorithm, our proposed VDF-OPD algorithm, and the VDF-ED algorithm in [23]. In all three implemented algorithms, the initial guess of the parameter is $\hat{\theta}_0 = \begin{bmatrix}0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5\end{bmatrix}^\top$, and the initial information matrix is $R_0 = 10^{-3}I_6$. The constant forgetting factor μ in EF is 0.95, while the VDF-OPD and VDF-ED algorithms use the same variable forgetting factor strategy in (21) with $\eta = 10^{-2}$ and $\mu_L = 0.5$. The thresholds ϵ in (18) and ϵ_{th} in (25) are both set to 0.1.

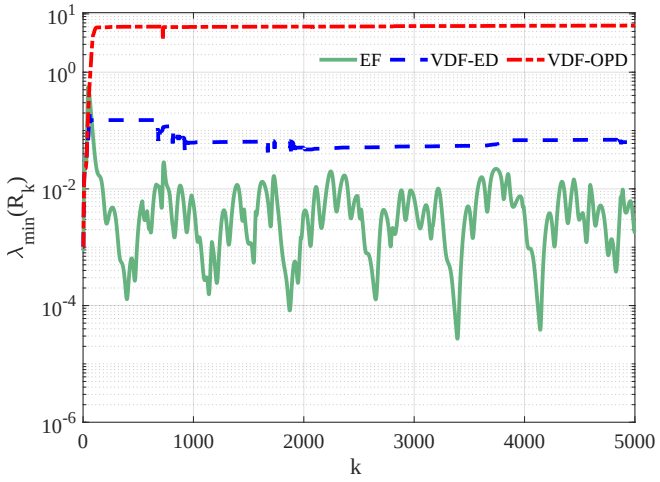


Fig. 3. $\lambda_{\min}(R_k)$ in EF, VDF-ED, and VDF-OPD algorithms.

The parameter estimates from the two VDF algorithms are depicted in Fig. 1, while those given by EF are shown in Fig. 2. The achieved estimation performance listed in Table I is evaluated by root mean square error (RMSE) of each element in θ_k, i.e.,

$$\sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\hat{\theta}_k(i) - \theta_k(i)\right)^2},$$

where $\theta_k(i)$ and $\hat{\theta}_k(i)$ represent the ith element of the true parameter and its estimate at time k, respectively.

As indicated by Fig. 2 and Table I, the parameter estimates from the EF algorithm have the largest errors, and our proposed VDF-OPD algorithm gives the smallest estimation errors. This can be explained by the evolution of the minimal eigenvalue of the information matrix R_k, i.e., $\lambda_{\min}(R_k)$, in these algorithms, as depicted in Fig. 3. For the EF algorithm, its $\lambda_{\min}(R_k)$ is significantly smaller than the other two algorithms, hence its obtained estimates are most sensitive to noise. After about time instant $k = 700$, the VDF-ED algorithm gives highly noisy estimates in Fig. 1, because its value of $\lambda_{\min}(R_k)$ decreases to around 0.1. Compared to EF and VDF-ED, our VDF-OPD algorithm gives the largest $\lambda_{\min}(R_k)$, thus is least sensitive to noise.

The robustness of VDF-OPD and VDF-ED algorithms are further compared in terms of the condition number of R_k, which is shown in Fig. 4. It can be seen that our VDF-OPD algorithm gives a much lower condition number of R_k than the VDF-ED algorithm.

VII. CONCLUSION

In this paper, a new VDF algorithm using oblique projection decomposition is presented for MO systems under non-persistent excitation. It ensures the information matrix is lower and upper bounded, and its estimation error variance converges. In contrast, the VDF-ED algorithm in [23] discounts old information in part of its forgotten subspace where

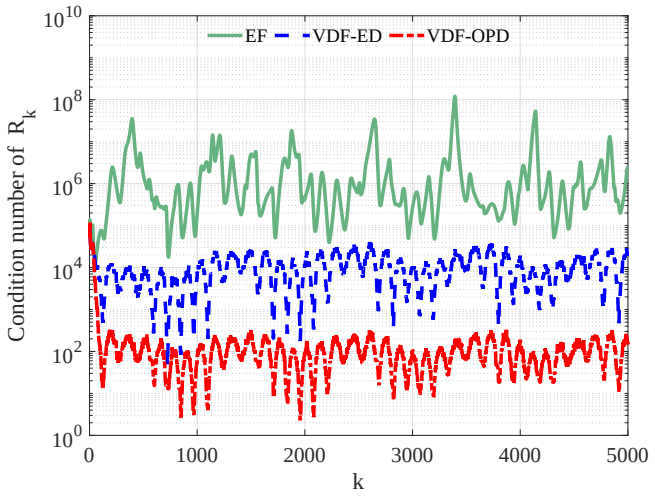


Fig. 4. Condition number of R_k in EF, VDF-ED, and VDF-OPD algorithms.

no new information is received, hence producing a more ill-conditioned information matrix under non-persistent excitation. The advantage of our proposed algorithm is illustrated by a numerical simulation example.

Appendix A
Preliminaries on Oblique Projection

Let $\mathcal{X}$ and $\mathcal{Y}$ be complementary subspaces of $\mathbb{R}^n$, i.e., $\mathcal{X}+\mathcal{Y}=\mathbb{R}^n$ and $\mathcal{X}\cap\mathcal{Y}=\{0\}$. Note that $\mathcal{X}$ and $\mathcal{Y}$ are not necessarily orthogonal. The oblique projector onto $\mathcal{X}$ along $\mathcal{Y}$ is uniquely represented by a square matrix $P_{\mathcal{X}|\mathcal{Y}}\in\mathbb{R}^{n\times n}$ that satisfies

$$P_{\mathcal{X}|\mathcal{Y}}x=x,\ P_{\mathcal{X}|\mathcal{Y}}y=0,\ P_{\mathcal{X}|\mathcal{Y}}z\in\mathcal{X}$$

for all $x\in\mathcal{X}$, $y\in\mathcal{Y}$, and $z\in\mathbb{R}^n$ [27].

Lemma 1 (Theorem 1 in [27]). *Consider two non-zero matrices $X\in\mathbb{R}^{n\times p}$ and $Y\in\mathbb{R}^{n\times q}$ satisfying $Y^\top X\neq 0$. Define two subspaces $\mathcal{X}=Range(X)$ and $\mathcal{Y}=Null(Y^\top)$. Then the oblique projection matrix $P_{\mathcal{X}|\mathcal{Y}}$ is*

$$P_{\mathcal{X}|\mathcal{Y}}=X(Y^\top X)^\dagger Y^\top. \tag{37}$$

If $\mathcal{X}+\mathcal{Y}=\mathbb{R}^n$ and $\mathcal{X}\cap\mathcal{Y}=\{0\}$, the two oblique projections $P_{\mathcal{X}|\mathcal{Y}}$ and $P_{\mathcal{Y}|\mathcal{X}}$ are complementary, i.e., $P_{\mathcal{X}|\mathcal{Y}}+P_{\mathcal{Y}|\mathcal{X}}=I_n$.

Note that the oblique projection matrix $P_{\mathcal{X}|\mathcal{Y}}$ in (37) is idempotent but can be non-symmetric. If $\mathcal{X}$ and $\mathcal{Y}$ in Lemma 1 are orthogonal complementary subspaces, $P_{\mathcal{X}|\mathcal{Y}}$ becomes an idempotent and symmetric matrix representing the orthogonal projection onto $\mathcal{X}$.

Appendix B
Proof of Theorem 1

It can be directly seen from (17) that $R_{k-1}^{(2)}\geq 0$ holds. Next, $R_{k-1}^{(1)}\geq 0$ will be proved. Since $R_{k-1}>0$, the Cholesky factorization of R_{k-1} is [28]

$$R_{k-1}=N_{k-1}N_{k-1}^\top, \tag{38}$$

where $N_{k-1}\in\mathbb{R}^{n\times n}$. Then, from (17) and (38), $R_{k-1}^{(1)}$ in (17) can be expressed as

$$\begin{aligned}R_{k-1}^{(1)}&=N_{k-1}N_{k-1}^\top-N_{k-1}D_kN_{k-1}^\top\\&=N_{k-1}(I_n-D_k)N_{k-1}^\top,\end{aligned} \tag{39}$$

with

$$D_k=N_{k-1}^\top\Phi_k(\Phi_k^\top N_{k-1}N_{k-1}^\top\Phi_k)^\dagger\Phi_k^\top N_{k-1} \tag{40}$$

being an idempotent matrix whose eigenvalues are either 0 or 1 [28]. Hence, $R_{k-1}^{(1)}\geq 0$ is proved according to (39).

Appendix C
Proof of Theorem 2

With the two complementary subspaces $\mathcal{V}_{k-1}$ in (16) and $\tilde{\mathcal{V}}_{k-1}$ in (15), the applied oblique projection in (12) results in

$$\text{rank}\left(R_{k-1}^{(2)}\right)=\text{rank}(R_{k-1}\Phi_k)=\text{rank}(\Psi_k) \tag{41}$$

due to $R_{k-1}>0$ and (26). Note that (26) can be expressed as $\text{col}_i(\Psi_k)=\Phi_k^\top\text{col}_i(U_{k-1})$, with col_i denoting the ith column of a matrix. According to (25) with $\epsilon_{\text{th}}=0$, if $\text{col}_i(\Psi_k)$ is non-zero, the associated $\text{col}_i(U_{k-1})$ should be included in the forgotten part, i.e., $U_{2,k-1}$ in (27). Otherwise, $\text{col}_i(U_{k-1})$ is included in the retained part, i.e., $U_{1,k-1}$ in (27). Hence $\text{rank}(U_{2,k-1})$ is equal to the number of non-zero columns of Ψ_k. Furthermore, since $\text{rank}(\Psi_k)$ is less than or equal to the number of non-zero columns of Ψ_k, we have $\text{rank}(\Psi_k)\leq\text{rank}(U_{2,k-1})$, thus

$$\text{rank}\left(R_{k-1}^{(2)}\right)=\text{rank}(\Psi_k)\leq\text{rank}(U_{2,k-1})=\text{rank}(M_{k-1}^{(2)})$$

holds according to (41) and (29c). This proves (30).

With $\epsilon_{\text{th}}=0$ in (28), we have $\Phi_k^\top U_{1,k-1}=0$, hence $\text{Range}(U_{1,k-1})\subseteq\text{Null}(\Phi_k^\top)$ holds. This further implies (31) according to (15) and

$$\text{Range}(U_{1,k-1})=\text{Range}\left(M_{k-1}^{(1)}\right).$$

Appendix D
Proof of Theorem 3

In the following, we prove that R_k obtained from (23) is positive if R_{k-1} is positive and Φ_k is bounded. This then leads to the proof of Theorem 3 via mathematical induction.

From (17) and (38)–(40), we have

$$R_{k-1}^{(1)}+\mu_kR_{k-1}^{(2)}=N_{k-1}\left(I_n-(1-\mu_k)D_k\right)N_{k-1}^\top. \tag{42}$$

Since D_k is an idempotent matrix, its eigenvalue decomposition can be expressed as

$$D_k=U_D\text{diag}(I_s,0)U_D^\top,$$

with $s=\text{rank}(D_k)$. Then, we have

$$\begin{aligned}I_n-(1-\mu_k)D_k&=U_DU_D^\top-(1-\mu_k)U_D\text{diag}(I_s,0)U_D^\top\\&=U_D\text{diag}(\mu_kI_s,I_{n-s})U_D^\top>0.\end{aligned} \tag{43}$$

Since R_{k-1} is assumed positive, N_{k-1} in the Cholesky factorization (38) is nonsingular. Then, it can be seen from (42) and (43) that $R_{k-1}^{(1)}+\mu_kR_{k-1}^{(2)}>0$ for $\mu_k>0$. Therefore, R_k in (23) is positive definite because $R_{k-1}-(1-\mu_k)R_{k-1}^{(2)}=R_{k-1}^{(1)}+\mu_kR_{k-1}^{(2)}>0$.

APPENDIX E
PROOF OF THEOREM 4

According to (21), there exists $\bar{\mu} \in (0,1)$ such that $\mu_k \leq \bar{\mu}$. Since $\|\Phi_k\|_2 < \infty$, there exists a finite upper bound c such that $\|\Phi_k\|_2 \leq c$, which implies $\Phi_k \Phi_k^\top \leq c^2 I_n$. Then, R_k in (23) satisfies

$$R_k \leq R_0 - \sum_{i=1}^{k} Q_i \tag{44}$$

with

$$Q_i = (1-\bar{\mu}) R_{i-1}^{(2)} - c^2 I_n. \tag{45}$$

Assume $\mathrm{rank}(\Phi_k) = r_k$, where r_k may vary with Φ_k. Let the singular value decomposition of Φ_k be expressed as

$$\Phi_k = U_{\phi,k} S_{\phi,k} V_{\phi,k}^\top, \tag{46}$$

where $S_{\phi,k} \in \mathbb{R}^{r_k \times r_k}$ is a diagonal matrix whose diagonal elements are the positive singular values, $U_{\phi,k} \in \mathbb{R}^{n \times r_k}$ and $V_{\phi,k} \in \mathbb{R}^{m \times r_k}$ consist of the left-hand and right-hand singular vectors associated with $S_{\phi,k}$. With (46), $\Phi_k^\top R_{k-1} \Phi_k$ can be expressed as

$$\begin{aligned} \Phi_k^\top R_{k-1} \Phi_k &= V_{\phi,k} S_{\phi,k} U_{\phi,k}^\top R_{k-1} U_{\phi,k} S_{\phi,k} V_{\phi,k}^\top \\ &= \begin{bmatrix} V_{\phi,k} & \tilde{V}_{\phi,k} \end{bmatrix} \mathrm{diag}(\Sigma_k, 0) \begin{bmatrix} V_{\phi,k}^\top \\ \tilde{V}_{\phi,k}^\top \end{bmatrix}, \end{aligned} \tag{47}$$

where $\tilde{V}_{\phi,k} \in \mathbb{R}^{m \times (m-r_k)}$ consists of basis column vectors of the orthogonal complement of $V_{\phi,k}$, and

$$\Sigma_k = S_{\phi,k} U_{\phi,k}^\top R_{k-1} U_{\phi,k} S_{\phi,k}. \tag{48}$$

Since both R_{k-1} and S_k are positive definite, $\mathrm{Range}(U_{\phi,k})$ must be a subspace of $\mathrm{Range}(R_{k-1})$, hence $U_{\phi,k}^\top R_{k-1} U_{\phi,k}$ and Σ_k in (48) are also positive definite. Then, the Moore-Penrose inverse of $\Phi_k^\top R_{k-1} \Phi_k$ in (47) is

$$\begin{aligned} \left(\Phi_k^\top R_{k-1} \Phi_k\right)^\dagger &= V_{\phi,k} \Sigma_k^{-1} V_{\phi,k}^\top \\ &= V_{\phi,k} S_{\phi,k}^{-1} (U_{\phi,k}^\top R_{k-1} U_{\phi,k})^{-1} S_{\phi,k}^{-1} V_{\phi,k}^\top. \end{aligned} \tag{49}$$

With (46)–(49), $R_{k-1}^{(2)}$ in (17) is rewritten as

$$R_{k-1}^{(2)} = R_{k-1} U_{\phi,k} (U_{\phi,k}^\top R_{k-1} U_{\phi,k})^{-1} U_{\phi,k}^\top R_{k-1}. \tag{50}$$

From (46) and (50), Q_i in (45) can be expressed as

$$Q_i = (1-\bar{\mu}) R_{i-1} U_{\phi,i} (U_{\phi,i}^\top R_{i-1} U_{\phi,i})^{-1} U_{\phi,i}^\top R_{i-1} - c^2 I_n. \tag{51}$$

Then, using the invariance property of trace under cyclic permutations, the trace of Q_i can be rewritten as

$$\begin{aligned} \mathrm{tr}(Q_i) &= \mathrm{tr}\left[(1-\bar{\mu})(U_{\phi,i}^\top R_{i-1} U_{\phi,i})^{-1} U_{\phi,i}^\top R_{i-1}^2 U_{\phi,i} - c^2 I_n\right] \\ &= \mathrm{tr}\left[(U_{\phi,i}^\top R_{i-1} U_{\phi,i})^{-1} U_{\phi,i}^\top \Omega_{i-1} U_{\phi,i}\right] \end{aligned} \tag{52}$$

with

$$\Omega_{i-1} = R_{i-1}((1-\bar{\mu}) R_{i-1} - c^2 I_n). \tag{53}$$

Let $\lambda_{i-1,j}$ represent the jth eigenvalue of R_{i-1}. Then the jth eigenvalue of Ω_{i-1} in (52) is

$$\omega_{i-1,j} = \lambda_{i-1,j}((1-\bar{\mu})\lambda_{i-1,j} - c^2). \tag{54}$$

Next, by following the same idea in the proof of Theorem 2 in [21], we prove that it is impossible to have an unbounded R_k by contradiction. Assume that one eigenvalue $\lambda_{i-1,s}$ ($1 \leq s \leq n$) of R_{i-1} is unbounded. Then, at all time instants $q \geq i$, the eigenvalue $\lambda_{q,s}$ of R_q becomes unbounded according to the update of R_q in (23) from R_{q-1}. Hence, for all $q \geq i$, the eigenvalue $\omega_{q,s}$ in (54) is unbounded. Furthermore, each $\mathrm{tr}(Q_i)$ in (52) is dominated by the ratio

$$\frac{\omega_{q,s}}{\lambda_{q,s}} = (1-\bar{\mu})\lambda_{i-1,s} - c^2, \quad q > i$$

that is unbounded. Therefore, on the right-hand side of (44), $\mathrm{tr}\left(R_0 - \sum_{i=1}^{k} Q_i\right)$ becomes negative and unbounded, which is in contradiction with the positive definiteness of R_k proved in Theorem 3. Such a contradiction proves that R_k must be bounded from above.

APPENDIX F
PROOF OF THEOREM 5

According to Theorem 3, $\bar{R}_{k-1}$ is invertible for all $k > 0$. By applying the matrix inversion lemma to (19), we have

$$\begin{aligned} R_k^{-1} &= \bar{R}_{k-1}^{-1} - \bar{R}_{k-1}^{-1} \Phi_k \left(I_m + \Phi_k^\top \bar{R}_{k-1}^{-1} \Phi_k\right)^{-1} \Phi_k^\top \bar{R}_{k-1}^{-1}, \\ \bar{R}_{k-1} R_k^{-1} \bar{R}_{k-1} &= \bar{R}_{k-1} - \Phi_k \left(I_m + \Phi_k^\top \bar{R}_{k-1}^{-1} \Phi_k\right)^{-1} \Phi_k^\top, \end{aligned} \tag{55}$$

$$\Phi_k^\top R_k^{-1} \Phi_k = \Phi_k^\top \bar{R}_{k-1}^{-1} \Phi_k (I_m + \Phi_k^\top \bar{R}_{k-1}^{-1} \Phi_k)^{-1}. \tag{56}$$

Under the condition $\|\Phi_k\|_2 > \varepsilon$, $\bar{R}_{k-1} \leq R_{k-1}$ holds due to the adopted forgetting strategy. Hence there exists $a \in (0,1)$ such that (34) holds for all $k > 0$. According to (56), there must also exist $b \in (0,1)$ such that (35) holds for all $k \geq 0$.

From (19), (22), (20), (32), and (33), the parameter estimation error dynamics is expressed as

$$\tilde{\theta}_k = R_k^{-1} \left(\bar{R}_{k-1} \tilde{\theta}_{k-1} - \Phi_k v_k\right). \tag{57}$$

Then, the Lyapunov function $V_k = \tilde{\theta}_k^\top R_k \tilde{\theta}_k$ is expressed as

$$\begin{aligned} V_k &= \left(\bar{R}_{k-1}\tilde{\theta}_{k-1} - \Phi_k v_k\right)^\top R_k^{-1} \left(\bar{R}_{k-1}\tilde{\theta}_{k-1} - \Phi_k v_k\right) \\ &= \tilde{\theta}_{k-1}^\top \bar{R}_{k-1} R_k^{-1} \bar{R}_{k-1} \tilde{\theta}_{k-1} - 2 v_k^\top \Phi_k^\top R_k^{-1} \bar{R}_{k-1} \tilde{\theta}_{k-1} \\ &\quad + v_k^\top \Phi_k^\top R_k^{-1} \Phi_k v_k. \end{aligned} \tag{58}$$

Taking mathematical expectation on both sides of (58), we have $\mathbb{E}\{\Phi_k v_k\} = 0$ due to the statistical independence between v_k and Φ_k, then we derive

$$\begin{aligned} \mathbb{E}\{V_k\} &= \mathbb{E}\left\{\tilde{\theta}_{k-1}^\top \bar{R}_{k-1} R_k^{-1} \bar{R}_{k-1} \tilde{\theta}_{k-1}\right\} \\ &\quad + \mathbb{E}\left\{v_k^\top \Phi_k^\top R_k^{-1} \Phi_k v_k\right\} \\ &\leq a\mathbb{E}\left\{\tilde{\theta}_{k-1}^\top R_{k-1} \tilde{\theta}_{k-1}\right\} + b\mathbb{E}\left\{v_k^\top v_k\right\} \\ &\leq a\mathbb{E}\{V_{k-1}\} + b\delta \end{aligned} \tag{59}$$

according to (34) and (35). This implies $\mathbb{E}\{V_k\} \leq \zeta_k$, with ζ_k generated by (36). According to $R_k \geq \beta_k I_n$ in Theorem 3, $\beta_k \mathbb{E}\left\|\tilde{\theta}\right\|^2 \leq \mathbb{E}\{V_k\} \leq \zeta_k$ is derived, which further leads to $\mathbb{E}\left\|\tilde{\theta}\right\|^2 \leq \frac{\zeta_k}{\beta_k}$.

As for the bounding sequence $\{\zeta_k\}$: its convergence to ζ_∞ can be derived from (36) with $a \in (0,1)$. If $\zeta_k > \zeta_\infty$, ζ_k monotonically decreases with time because $\zeta_{k+1} - \zeta_k < -(1-a)\zeta_\infty + b\delta = 0$.

REFERENCES

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

[2] M. S. Mahmoud and M. O. Oyedeji, "Adaptive and predictive control strategies for wind turbine systems: a survey," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 2, pp. 364–378, 2019.

[3] T. Soderstrom and L. Ljung, *Theory and Practice of Recursive Identification*. The MIT Press, 1987.

[4] B. Q. Mu and H. F. Chen, "Recursive identification of multi-input multi-output errors-in-variables Hammerstein systems," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 843–849, 2014.

[5] P. Zhou, P. Dai, H. Song, and T. Chai, "Data-driven recursive subspace identification based online modelling for prediction and control of molten iron quality in blast furnace ironmaking," *IET Control Theory and Applications*, vol. 11, no. 4, pp. 2343–2351, 2017.

[6] X. Wang and F. Ding, "Convergence of the recursive identification algorithms for multivariate pseudo-linear regressive systems," *International Journal of Adaptive Control and Signal Processing*, vol. 30, no. 6, pp. 824–842, 2015.

[7] Y. Wang, F. Ding, and M. Wu, "Recursive parameter estimation algorithm for multivariate output-error systems," *Journal of the Franklin Institute*, vol. 355, no. 12, pp. 5163–5181, 2018.

[8] K. Bekiroglu, S. Srinivasan, E. Png, R. Su, and C. Lagoa, "Recursive approximation of complex behaviours with IoT-data imperfections," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 3, pp. 656–667, 2020.

[9] J. Lou, L. Jia, R. Tao, and Y. Wang, "Distributed incremental bias-compensated RLS estimation over multi-agent networks," *Science China Information Sciences*, vol. 60, no. 3, p. 032204, 2017.

[10] C. Yu, J. Chen, S. Li, and M. Verhaegen, "Identification of affinely parameterized state–space models with unknown inputs," *Automatica*, vol. 122, p. 109271, 2020.

[11] M. Dai, Y. He, and X. Yang, "Continuous-time system identification with nuclear norm minimization and GPMF-based subspace method," *IEEE/CAA Journal of Automatica Sinica*, vol. 3, no. 2, pp. 184–191, 2016.

[12] R. M. Johnstone, C. R. Johnson, R. R. Bitmead, and B. D. O. Anderson, "Exponential convergence of recursive least squares with exponential forgetting factor," *Systems and Control Letters*, vol. 2, no. 2, pp. 77–82, 1982.

[13] S. Bruggemann and R. R. Bitmead, "Exponential convergence of recursive least squares with forgetting factor for multiple-output systems," *Automatica*, vol. 124, p. 109389, 2021.

[14] D. Bertin, S. Bittanti, and P. Bolzern, "Tracking of nonstationary systems by means of different prediction error direction forgetting techniques," in *IFAC Proceedings Volumes*. Elsevier, 1987, vol. 20, no. 2, pp. 185–190.

[15] C. Paleologu, J. Benesty, and S. Ciochina, "A robust variable forgetting factor recursive least-squares algorithm for system identification," *IEEE Signal Processing Letters*, vol. 15, pp. 597–600, 2008.

[16] S. Leung and C. F. So, "Gradient-based variable forgetting factor RLS algorithm in time-varying environments," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3141–3150, 2005.

[17] J. Dokoupil, A. Voda, and P. Vaclavek, "Regularized extended estimation with stabilized exponential forgetting," *IEEE Transactions on Automatic Control*, vol. 62, no. 12, pp. 6513–6520, 2017.

[18] A. L. Bruce, A. Goel, and D. S. Bernstein, "Convergence and consistency of recursive least squares with variable-rate forgetting," *Automatica*, vol. 119, p. 109052, 2020.

[19] J. E. Parkum, N. Poulsen, and J. Holst, "Recursive forgetting algorithms," *International Journal of Control*, vol. 55, no. 1, pp. 109–128, 1992.

[20] S. Bittanti, P. Bolzern, and M. Campi, "Convergence and exponential convergence of identification algorithms with directional forgetting factor," *Automatica*, vol. 26, no. 5, pp. 929–932, 1990.

[21] L. Cao and H. Schwartz, "A directional forgetting algorithm based on the decomposition of the information matrix," *Automatica*, vol. 36, no. 11, pp. 1725–1731, 2000.

[22] S. A. U. Islam and D. S. Bernstein, "Recursive least squares for real-time implementation [Lecture Notes]," *IEEE Control Systems*, vol. 39, no. 3, pp. 82–85, 2019.

[23] A. Goel, A. L. Bruce, and D. S. Bernstein, "Recursive least squares with variable-direction forgetting: Compensating for the loss of persistency [Lecture Notes]," *IEEE Control Systems*, vol. 40, no. 4, pp. 80–102, 2020.

[24] A. L. Bruce, A. Goel, and D. S. Bernstein, "Recursive least squares with matrix forgetting," in *Proceedings of 2020 American Control Conference*, Denver, CO, USA, 2020, pp. 1406–1410.

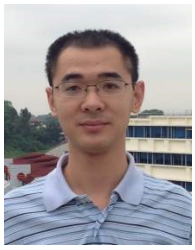
[25] F. Ding, P. X. Liu, and G. Liu, "Multiinnovation least-squares identification for system modeling," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 40, no. 3, pp. 767–778, 2009.

[26] H. V. Henderson and S. R. Searle, "The vec-permutation matrix, the vec operator and kronecker products: a review," *Linear and Multilinear Algebra*, vol. 9, no. 4, pp. 271–288, 1981.

[27] P. C. Hansen, "Oblique projections and standard-form transformations for discrete inverse problems," *Numerical Linear Algebra with Applications*, vol. 20, no. 2, pp. 250–258, 2013.

[28] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. SIAM, 2000.

Kun Zhu received his Bachelor degree in Electrical Engineering and Automation from Nanchang University in 2019. He is currently working toward a Master degree in the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. His research interests include industrial data analytics and parameter estimation.



Chengpu Yu received his B. E. and M. E. degrees from the University of Electronic Science and Technology of China, in 2006 and 2009, respectively, and a Ph.D. degree from Nanyang Technological University, Singapore, in 2014. From June 2014 to August 2017, he was a PostDoc at Delft Center for Systems and Control, The Netherlands. He is currently a professor with School of Automation, Beijing Institute of Technology, China. His research interests include system identification, distributed optimization and network system control.

Yiming Wan received his Ph.D. in Automatic Control from Tsinghua University in 2013. From 2013 to 2016, he was a postdoctoral researcher at Delft Center for Systems and Control, Delft University of Technology, The Netherlands. During 2016 to 2018, he was a Research Associate at the Massachusetts Institute of Technology, United States. Since February 2018, he has been an Associate Professor in the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. His research interests include fault-tolerant control, industrial data analytics, state and parameter estimation.