Automatica 119 (2020) 109052

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

Convergence and consistency of recursive least squares with variable-rate forgetting ☆

Adam L. Bruce *, Ankit Goel, Dennis S. Bernstein

Department of Aerospace Engineering, The University of Michigan, Ann Arbor, MI 48109, United States

ARTICLE INFO

Article history:
Received 5 August 2019
Received in revised form 27 April 2020
Accepted 7 May 2020
Available online 30 May 2020

Keywords:
System identification
Identification methods
Tracking and adaptation
Adaptive systems
Adaptive control

ABSTRACT

A recursive least squares algorithm with variable rate forgetting (VRF) is derived by minimizing a quadratic cost function. Under persistent excitation, the minimizer given by VRF is shown to converge to the true parameters. In addition, under persistent excitation and with noisy measurements, where the noise is uncorrelated with the regressor, conditions are given under which the minimizer given by VRF is a consistent estimator of the true parameters. The results are illustrated by a numerical example involving abruptly changing parameters.

1. Introduction

Recursive least squares (RLS) is one of the foundational algorithms of systems and control theory, especially for signal processing, identification, and adaptive control (Astrom, 1995; Ljung & Soderstrom, 1983). An early exposition of RLS is given in Albert and Sittler (1965).

Standard RLS employs a constant forgetting factor λ, which enhances the importance of recent data over older data. Although λ can be set by the user, the performance of RLS is often extremely sensitive to the chosen value. Consequently, choosing a suitable value of λ is typically a trial and error process.

To remedy this problem, various techniques have been proposed to automatically vary the forgetting factor in response to the fit error. In particular, Fortescue, Kershenbaum, and Ydstie (1981) reports a method for sequentially updating the forgetting factor to conserve the amount information used in the estimate, and Paleologu, Benesty, and Silviu (2008) reports an update-based algorithm that uses noise statistics to control the forgetting factor. (Leung & So, 2005) gives a gradient-based algorithm for computing a forgetting factor that locally minimizes the mean-square error of the estimate, and Song, Lim, Baek, and Sung (2000) derives a Newton-type gradient-descent algorithm that combines sequential estimation with minimization of the mean-squared error. Finally, Park, Jun, and Kim (1991) gives a formula based on exponentiation of the squared residual.

The present paper approaches the problem of varying the forgetting factor by deriving a generalization of RLS that includes time-dependent cost scaling and regularization. This formulation involves a growing-window cost function, and thus is distinct from the formulation of Ali, Hoagg, Mossberg, and Bernstein (2016), which uses a sliding-window cost function. The growing-window cost function is advantageous since it directly generalizes traditional RLS and has the ability to weigh recent data more heavily than older data.

The first contribution of the paper is given by Theorem 1, which introduces RLS with variable-rate forgetting (VRF), a novel extension of RLS in which the role of the constant forgetting factor λ in RLS is replaced by a variable forgetting factor β_k. By setting $\beta_k = \frac{1}{\lambda}$ for all k, VRF specializes to RLS with constant-rate-forgetting (CRF). The variable-rate-forgetting extensions of RLS given in Fortescue et al. (1981), Leung and So (2005), Paleologu et al. (2008), Park et al. (1991) and Song et al. (2000) are special cases of Theorem 1 with specific choices of β_k. In addition, Theorem 1 refines the variable-rate weighting used in Ljung and Soderstrom (1983, pp. 17, 18). In particular, we factor α_k in Ljung and Soderstrom (1983, Eq. (2.12)) as $\beta_k \cdots \beta_0$, where $1/\beta_k$ serves as the instantaneous forgetting factor at step k. This formulation allows the user to specify β_k at each step based on the current residual or knowledge of system changes. The second and third contributions of this paper are given by Theorems 2, 4, and Corollary 4, which prove conditions on β_k ensuring convergence

☆ This research was partially supported by AFOSR, United States under DDDAS grant FA9550-16-1-0071 (Dynamic Data-Driven Applications Systems). The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Tianshi Chen under the direction of Editor Torsten Söderström.

* Corresponding author.
E-mail addresses: admbruce@umich.edu (A.L. Bruce), ankgoel@umich.edu (A. Goel), dsbaero@umich.edu (D.S. Bernstein).

https://doi.org/10.1016/j.automatica.2020.109052
0005-1098/

under the assumption of persistency (Theorem 2), and ensuring consistency under the assumption of persistency and that the regressor and sensor noise are uncorrelated (Theorem 4, Corollary 4). Specific examples of β_k for consistent and non-consistent algorithms are given in Corollary 5. The fourth contribution is two choices of β_k that may be useful in practice. In Section 6, we demonstrate these choices on an abruptly changing system with and without measurement noise and compare the performance of VRF and CRF for the given example.

The notation used throughout this paper is as follows. The symbols $\mathbb{S}^n$, $\mathbb{N}^n$, and $\mathbb{P}^n$ denote the sets of real $n \times n$ symmetric, positive-semidefinite, and positive-definite matrices, respectively. For all $A \in \mathbb{S}^n$, $\lambda_i(A)$ denotes the ith largest eigenvalue of A, $\lambda_{\max}(A) \triangleq \lambda_1(A)$, and $\lambda_{\min}(A) \triangleq \lambda_n(A)$. $\lfloor x \rfloor$ denotes the greatest integer less than or equal to $x \in \mathbb{R}$. The notation $(x_k)_{k\geq 0} \subset X$, where X is a set, indicates that $(x_k)_{k\geq 0}$ is a sequence in X. For all $(a_k)_{k\geq 0}, (b_k)_{k\geq 0} \subset \mathbb{R}$, the notation $a_k \sim \mathcal{O}(b_k)$ indicates that there exists $M > 0$ and $K \geq 0$ such that, for all $k \geq K$, $a_k \leq M b_k$. Finally, for all $k \geq 0$ and $N \geq 0$, we define $\xi(k, N) \triangleq \lfloor \frac{k}{N+1} \rfloor$.

2. Problem formulation

Let $\lambda \in (0, 1]$, $\theta_0 \in \mathbb{R}^n$, and $P_0 \in \mathbb{P}^n$, and, for all $k \geq 0$, let $\phi_k \in \mathbb{R}^{p\times n}$, $y_k \in \mathbb{R}^p$, $e_k \triangleq y_k - \phi_k\theta$ and define $J_k\colon \mathbb{R}^n \to [0, \infty)$ by

$$J_k(\theta) \triangleq \sum_{i=0}^{k} \lambda^{k-i}\|e_k\|^2 + \lambda^{k+1}(\theta-\theta_0)^{\mathrm{T}}P_0^{-1}(\theta-\theta_0). \tag{1}$$

Since J_k is quadratic and strictly convex, it follows that its unique global minimizer, $\theta_{k+1} \triangleq \operatorname{argmin}_{\theta\in\mathbb{R}^n} J_k(\theta)$, is the only local minimizer. This minimizer is the least squares estimate of θ given $y_0, \ldots, y_k$, and can be computed efficiently using the traditional RLS update equations (Astrom, 1995; Ljung & Soderstrom, 1983; Ul Islam & Bernstein, 2019) as given in the following proposition.

Proposition 1. *Under the notation and assumptions of the preceding paragraph, for all $k \geq 0$, define $J_k\colon \mathbb{R}^n \to [0, \infty)$ by (1). Then*

$$\theta_{k+1} = \theta_k + P_{k+1}\phi_k^{\mathrm{T}}(y_k - \phi_k\theta_k), \tag{2}$$

where

$$P_{k+1} = \frac{1}{\lambda}P_k - \frac{1}{\lambda}P_k\phi_k^{\mathrm{T}}(\lambda I_p + \phi_k P_k \phi_k^{\mathrm{T}})^{-1}\phi_k P_k. \tag{3}$$

The goal of this paper is to generalize (1) so that λ can vary as a function of k, and then to prove a result analogous to Proposition 1 for the generalized cost. In addition, we will analyze the convergence and consistency properties for the family of algorithms thus obtained.

3. RLS with variable-rate forgetting

To generalize (1), for all $k \geq 0$, let $\beta_k > 0$, define

$$\rho_k \triangleq \prod_{i=0}^{k}\beta_i, \quad \rho_{-1} \triangleq 1, \tag{4}$$

and define the cost function $J_k\colon \mathbb{R}^n \to [0, \infty)$ by

$$J_k(\theta) \triangleq \sum_{i=0}^{k}\frac{\rho_i}{\rho_k}\|e_k\|^2 + \frac{1}{\rho_k}(\theta-\theta_0)^{\mathrm{T}}P_0^{-1}(\theta-\theta_0). \tag{5}$$

Since (5) is quadratic and strictly convex, like (1), its unique global minimizer is the only local minimizer. Furthermore, since (5) can be written as

$$J_k(\theta) = \theta^{\mathrm{T}}A_k\theta - 2b_k^{\mathrm{T}}\theta + c_k, \tag{6}$$

where

$$A_k \triangleq \sum_{i=0}^{k}\frac{\rho_i}{\rho_k}\phi_i^{\mathrm{T}}\phi_i + \frac{1}{\rho_k}P_0^{-1}, \tag{7}$$

$$b_k \triangleq \sum_{i=0}^{k}\frac{\rho_i}{\rho_k}\phi_i^{\mathrm{T}}y_i + \frac{1}{\rho_k}P_0^{-1}\theta_0, \tag{8}$$

$$c_k \triangleq \sum_{i=0}^{k}\frac{\rho_i}{\rho_k}y_i^{\mathrm{T}}y_i + \frac{1}{\rho_k}\theta_0^{\mathrm{T}}P_0^{-1}\theta_0, \tag{9}$$

and since A_k is positive definite, we define the positive-definite matrix

$$P_k \triangleq A_{k-1}^{-1}, \tag{10}$$

where $A_{-1} \triangleq P_0^{-1}$. The following result, *RLS with variable-rate forgetting (VRF)*, generalizes Proposition 1 to the minimizer of (5).

Theorem 1. *Let $\theta_0 \in \mathbb{R}^n$, $P_0 \in \mathbb{P}^n$, and, for all $k \geq 0$, let $\phi_k \in \mathbb{R}^{p\times n}$, $y_k \in \mathbb{R}^p$, and $\beta_k \in (0, \infty)$. Then the minimizer θ_{k+1} of (5) is given by*

$$\theta_{k+1} = \theta_k + P_{k+1}\phi_k^{\mathrm{T}}(y_k - \phi_k\theta_k), \tag{11}$$

and

$$P_{k+1} = L_k - L_k\phi_k^{\mathrm{T}}(I_p + \phi_k L_k \phi_k^{\mathrm{T}})^{-1}\phi_k L_k, \tag{12}$$

$$L_k \triangleq \beta_k P_k. \tag{13}$$

The proof of Theorem 1 requires the following lemma.

Lemma 1. *Let $P_0 \in \mathbb{P}^n$ and, for all $k \geq 0$, let $\beta_k > 0$, define ρ_k by (4), and define P_k by (10). Then, for all $k \geq 0$,*

$$P_{k+1}^{-1} = \frac{1}{\beta_k}P_k^{-1} + \phi_k^{\mathrm{T}}\phi_k \tag{14}$$

$$= \frac{1}{\rho_k}\left(P_0^{-1} + \sum_{i=0}^{k}\rho_i\phi_i^{\mathrm{T}}\phi_i\right). \tag{15}$$

Proof. Let $k \geq 0$. It follows from (7) that $A_k = \frac{1}{\beta_k}A_{k-1} + \phi_k^{\mathrm{T}}\phi_k$, which, using (10), implies (14). Furthermore, (14) implies $P_1^{-1} = \frac{1}{\rho_0}(P_0^{-1} + \rho_0\phi_0^{\mathrm{T}}\phi_0)$, which confirms (15) for $k = 0$. Next, let $k > 0$ and suppose for induction that (15) holds for $k-1$. From (14) it follows that $P_{k+1}^{-1} = \frac{1}{\beta_k}P_k^{-1} + \phi_k^{\mathrm{T}}\phi_k = \frac{1}{\rho_k}\left(P_0^{-1} + \sum_{i=0}^{k-1}\rho_i\phi_i^{\mathrm{T}}\phi_i\right) + \frac{\rho_k}{\rho_k}\phi_k^{\mathrm{T}}\phi_k = \frac{1}{\rho_k}\left(P_0^{-1} + \sum_{i=0}^{k}\rho_i\phi_i^{\mathrm{T}}\phi_i\right)$. □

Proof of Theorem 1. Let $k \geq 0$. To prove (12), note that it follows from (13), (14), and the matrix inversion lemma that $P_{k+1} = \left(\frac{1}{\beta_k}P_k^{-1} + \phi_k^{\mathrm{T}}\phi_k\right)^{-1} = L_k - L_k\phi_k^{\mathrm{T}}\left(I_p + \phi_k L_k\phi_k^{\mathrm{T}}\right)^{-1}\phi_k L_k$. To prove (11), note that (8), (10), and (14) imply that

$$\begin{aligned}\theta_{k+1} &= P_{k+1}\left(\sum_{i=0}^{k}\frac{\rho_i}{\rho_k}\phi_i^{\mathrm{T}}y_i + \frac{1}{\rho_k}P_0^{-1}\theta_0\right)\\ &= P_{k+1}\left(\phi_k^{\mathrm{T}}y_k + \frac{\rho_{k-1}}{\rho_k}\left[\sum_{i=0}^{k-1}\frac{\rho_i}{\rho_{k-1}}\phi_i^{\mathrm{T}}y_i + \frac{1}{\rho_{k-1}}P_0^{-1}\theta_0\right]\right)\\ &= P_{k+1}\left(\phi_k^{\mathrm{T}}\phi_k + \frac{1}{\beta_k}P_k^{-1}\right)\theta_k + P_{k+1}\phi_k^{\mathrm{T}}(y_k - \phi_k\theta_k)\\ &= \theta_k + P_{k+1}\phi_k^{\mathrm{T}}(y_k - \phi_k\theta_k), \quad \square\end{aligned}$$

For all $k \geq 0$, let $\beta_k = \frac{1}{\lambda}$. Then (5) specializes to (1), and (11)–(13) specialize to (2) and (3). Theorem 1 thus includes Proposition 1 as a special case.

4. Convergence of VRF

4.1. Asymptotic convergence

Definition 1. A sequence $(S_k)_{k\geq 0} \subset \mathbb{N}^n$ is persistent if there exist $N \geq 1$ and $\alpha > 0$ such that, for all $j \geq 0$,

$$\alpha I_n \leq \sum_{i=0}^{N} S_{i+j}. \tag{16}$$

The numbers α and N are, respectively, the lower bound and persistency window of $(S_k)_{k\geq 0}$. The sequence $(\phi_k)_{k\geq 0} \subset \mathbb{R}^{n\times m}$ is persistent if $(\phi_k^{\mathrm{T}}\phi_k)_{k\geq 0}$ is persistent.

Theorem 2. *Let $(\phi_k)_{k\geq 0} \subset \mathbb{R}^{n\times m}$, be persistent, let $\theta \in \mathbb{R}^n$, and, for all $k \geq 0$, let $y_k = \phi_k\theta$. Furthermore, let $a > 1$ and, for all $k \geq 0$, let $\beta_k \geq 1$. Finally, let $\theta_0 \in \mathbb{R}^n$, let $P_0 \in \mathbb{P}^n$, and, for all $k \geq 0$, define θ_{k+1} by (11)–(13). Then $\lim_{k\to\infty} \theta_k = \theta$.*

Proof. Let $k \geq 0$ and define $\tilde{\theta}_k \triangleq \theta_k - \theta$. Using (11) and (14) it follows that $\tilde{\theta}_{k+1} = (I_n - P_{k+1}\phi_k^{\mathrm{T}}\phi_k)\tilde{\theta}_k = \frac{1}{\beta_k}P_{k+1}P_k^{-1}\tilde{\theta}_k$, thus $\tilde{\theta}_k = \frac{1}{\rho_{k-1}}P_kP_0^{-1}\tilde{\theta}_0$. From (15), it follows that

$$\begin{aligned}
\varlimsup_{k\to\infty} \|\tilde{\theta}_k\|^2 &\leq \varlimsup_{k\to\infty} \frac{\lambda_{\max}(P_k^2)}{\rho_{k-1}^2}\|P_0^{-1}\theta_0\|^2 \\
&\leq \varlimsup_{k\to\infty} \frac{\|P_0^{-1}\theta_0\|^2}{\lambda_{\max}^2\left(P_0^{-1} + \sum_{i=0}^{k-1}\rho_i\phi_i^{\mathrm{T}}\phi_i\right)} \\
&\leq \varlimsup_{k\to\infty} \frac{\|P_0^{-1}\theta_0\|^2}{[\lambda_{\max}(P_0^{-1}) + \xi(k, N+1)\alpha]^2} = 0. \quad \square
\end{aligned}$$

4.2. Convergence rate

Definition 2. Let $(S_i)_{i\geq 0} \subset \mathbb{N}^n$ be persistent with lower bound α and window N. Then the upper bound $\beta \in (0, \infty) \cup \{\infty\}$ of $(S_i)_{i\geq 0}$ is

$$\beta \triangleq \sup_{j\geq 0} \lambda_{\max}\left(\sum_{i=j}^{N} S_j\right). \tag{17}$$

Lemma 2. *Let $(S_i)_{i\geq 0} \subset \mathbb{N}^n$ be persistent with window N, lower bound α, and upper bound β, and let $(a_i)_{i\geq 0}$ be a nondecreasing sequence of nonnegative numbers. Then, for all $k \geq 0$,*

$$\alpha \ell_{\xi(k,N)-1} I_n \leq \sum_{i=0}^{k} a_i S_i \leq \beta r_{\xi(k,N)} I_n, \tag{18}$$

where $\ell_j \triangleq \sum_{i=0}^{j} a_{i(N+1)}$ and $r_j \triangleq \sum_{i=0}^{j} a_{i(N+1)+N}$.

Proof. In the case where $\beta = \infty$, the upper bound of (18) is immediate. Hence, assume $\beta < \infty$. Let $k \geq 0$. Since $(a_i)_{i\geq 0}$ is nondecreasing, for all $j \geq 0$ and $i \in \{0, \ldots, N\}$, $a_{i+j} \leq a_{N+j}$ and $a_j \leq a_{i+j}$. From (16) and (17) it follows that $\alpha a_j I_n \leq a_j \sum_{i=0}^{N} S_{i+j} \leq \sum_{i=0}^{N} a_{i+j}S_{i+j}$, and thus

$$\alpha \ell_{\xi(k,N)-1} I_n \leq \sum_{q=0}^{\xi(k,N)-1} \sum_{i=0}^{N} a_{i+q(N+1)} S_{i+q(N+1)} \leq \sum_{i=0}^{k} a_i S_i.$$

Similarly, $\sum_{i=0}^{N} a_{i+j}S_{i+j} \leq a_{N+j}\sum_{i=0}^{N} S_{i+j} \leq a_{N+j}\beta I_n$, and thus

$$\sum_{i=0}^{k} a_i S_i \leq \sum_{q=0}^{\xi(k,N)-1} a_{q(N+1)+N}\beta I_n + a_k\beta I_n \leq \beta r_{\xi(k,N)} I_n. \quad \square$$

Theorem 3. *Under the assumptions and notation of Theorem 2, $\|\tilde{\theta}_k\| \sim \mathcal{O}\left(1/\sum_{i=0}^{\xi(k,N)-1} \rho_{i(N+1)}\right)$.*

Proof. Let $M = \|P_0^{-1}\tilde{\theta}_0\|/\alpha$. From Lemma 2, it follows that, for all $k \geq 0$, $\lambda_{\min}\left(P_0^{-1}\right) + \alpha\sum_{i=0}^{\xi(k,N)-1} \rho_{i(N+1)} \leq \lambda_{\min}\left(P_0^{-1} + \sum_{i=0}^{k}\rho_i\phi_k^{\mathrm{T}}\phi_k\right)$, and therefore, for all $k \geq 0$,

$$\begin{aligned}
\|\tilde{\theta}_k\| &\leq \|P_k\|\|P_0^{-1}\tilde{\theta}_0\| \leq \frac{\|P_0^{-1}\tilde{\theta}_0\|}{\lambda_{\min}\|P_k^{-1}\|} \\
&= \frac{\|P_0^{-1}\tilde{\theta}_0\|}{\lambda_{\min}\left(P_0^{-1}\right)}\left(1 + \frac{\alpha}{\lambda_{\min}\left(P_0^{-1}\right)}\sum_{i=0}^{\xi(k,N)-1}\rho_{i(N+1)}\right)^{-1} \\
&\leq M\left(\sum_{i=0}^{\xi(k,N)-1}\rho_{i(N+1)}\right)^{-1}. \quad \square
\end{aligned}$$

The following corollary shows that Theorem 3 can be used to prove convergence rates for both RLS without forgetting and CRF.

Corollary 3. *Under the assumptions and notation of Theorem 2, assume that there exists $\gamma \in [1, \infty)$ such that, for all $k \geq 0$, $\beta_k = \gamma$. Then*

$$\|\tilde{\theta}_k\| \sim \begin{cases} \mathcal{O}(1/\xi(k, N)), & \gamma = 1, \\ \mathcal{O}(\gamma^{-(N+1)\xi(k,N)}), & \gamma > 1. \end{cases} \tag{19}$$

Proof. In the case where $\gamma = 1$, (19) is immediate from Theorem 3. Hence, suppose that $\gamma > 1$. From Theorem 3, it follows that there exist $M_0 > 0$ and $K_0 \geq 0$ such that, for all $k \geq K$, $\|\tilde{\theta}_k\| \leq M_0/\sum_{i=0}^{\xi(k,N)-1} \rho_{i(N+1)}$. Let $M = M_0\gamma^N$ and $K = \max(K_0, N+1)$. Therefore, since, for all $k \geq 0$, $\rho_{k(N+1)} = \gamma^{k(N+1)+1}$, it follows that, for all $k \geq K$,

$$\begin{aligned}
\|\tilde{\theta}_k\| &\leq \frac{M_0}{\sum_{i=0}^{\xi(k,N)-1}\gamma^{i(N+1)+1}} = \frac{M_0}{\gamma}\frac{\gamma^{N+1}-1}{\gamma^{(N+1)\xi(k,N)}-1} \\
&= \frac{M_0}{\gamma}\frac{\gamma^{N+1}-1}{\gamma^{(N+1)\xi(k,N)}-1}\frac{\gamma^{(N+1)\xi(k,N)}}{\gamma^{(N+1)\xi(k,N)}} \leq \frac{M}{\gamma^{(N+1)\xi(k,N)}}. \quad \square
\end{aligned}$$

Since generally, for all $k \geq 0$, $\beta_k \geq 1$, it follows that $1/\sum_{i=0}^{\xi(k,N)-1} \rho_{i(N+1)} \leq 1/\xi(k, N)$. Thus, this analysis suggests that, in the case where $\beta_k > 1$ for an infinite set of indices, the asymptotic convergence rate of VRF is faster than the asymptotic convergence rate of RLS without forgetting.

5. Consistency of VRF

A sequence $(X_k)_{k\geq 0}$ of vector-valued random variables on Ω is a *consistent estimator* of $\theta \in \mathbb{R}^n$ if, for all $\varepsilon > 0$,

$$\lim_{k\to\infty} \mathbb{P}(\{\omega \in \Omega\colon \|X_k(\omega) - \theta\| < \varepsilon\}) = 1. \tag{20}$$

When θ is understood, for brevity, we call such sequences *consistent*.

Theorem 4. *Let $(\phi_k)_{k\geq 0}$ be a persistently exciting sequence with window N, lower bound α, and upper bound $\beta < \infty$. Let $\theta \in \mathbb{R}^n$, $P_0 \in \mathbb{P}^n$, and $\theta_0 \sim \mathcal{N}(\theta, P_0)$. Let $(\nu_k)_{k\geq 0}$ be an $\mathbb{R}^p$-valued stationary Gaussian white-noise process with variance V and uncorrelated with θ_0, and define $y_k = \phi_k\theta + \nu_k$. Furthermore, for all $k \geq 0$, let $\beta_k \geq 1$, and define θ_{k+1} by (11)–(13). Then, for all $k \geq 0$, θ_k is a Gaussian* *random variable with mean* $\bar{\theta}$, *and*

$$\frac{\alpha\lambda_{\min}(V)}{\beta^2}\varliminf_{k\to\infty}\frac{q_{\mathrm{l},\xi(k,N)}}{s^2_{\mathrm{u},\xi(k,N)}} \le \varliminf_{k\to\infty}\lambda_{\min}(\mathrm{var}(\theta_k)) \quad (21)$$

$$\le \varlimsup_{k\to\infty}\lambda_{\max}(\mathrm{var}(\theta_k)) \le \frac{\beta\lambda_{\max}(V)}{\alpha^2}\varlimsup_{k\to\infty}\frac{q_{\mathrm{u},\xi(k,N)}}{s^2_{\mathrm{l},\xi(k,N)}}, \quad (22)$$

where, for all $j \ge 0$, $s_{\mathrm{l},j} \triangleq \sum_{i=0}^{j-1}\rho_{i(N+1)}$, $s_{\mathrm{u},j} \triangleq \sum_{i=0}^{j}\rho_{i(N+1)+N}$, $q_{\mathrm{l},j} \triangleq \sum_{i=0}^{j-1}\rho^2_{i(N+1)}$, *and* $q_{\mathrm{u},j} \triangleq \sum_{i=0}^{j}\rho^2_{i(N+1)+N}$.

Proof. With base case $\theta_0 \sim \mathcal{N}(\theta, P_0)$, suppose for induction that $\theta_k \sim \mathcal{N}(\theta, \mathrm{var}(\theta_k))$. Define $\tilde{\theta}_k \triangleq \theta_k - \theta$. From (11), it follows that $\tilde{\theta}_{k+1} = \beta_k^{-1}P_{k+1}P_k^{-1}\tilde{\theta}_k + P_{k+1}\phi_k^{\mathrm{T}}\nu_k$. Since $\theta_k \sim \mathcal{N}(\theta, \mathrm{var}(\theta_k))$, it follows from Lemma A.1 that $\tilde{\theta}_k \sim \mathcal{N}(0, \mathrm{var}(\theta_k))$. Next, define $z_k \triangleq P_k^{-1}\tilde{\theta}_k$. Since $\tilde{\theta}_k \sim \mathcal{N}(0, \mathrm{var}(\theta_k))$, it follows from Lemma A.1 that $z_k \sim \mathcal{N}(0, P_k^{-1}\mathrm{var}(\theta_k)P_k^{-1})$. Since ν_k is uncorrelated with $\nu_0, \ldots, \nu_{k-1}, \theta_0$, it follows that ν_k and z_k are also uncorrelated. Furthermore, $z_{k+1} = \beta_k^{-1}z_k + \phi_k^{\mathrm{T}}\nu_k$, and thus $[z_k\ \nu_k]^{\mathrm{T}} \sim \mathcal{N}(0_{2\times1}, \mathrm{diag}(\mathrm{var}(z_k), V))$. Therefore, Lemma A.1 implies that $z_{k+1} \sim \mathcal{N}(0, \mathrm{var}(z_{k+1}))$ and $\mathrm{var}(z_{k+1}) = \beta_k^{-2}\mathrm{var}(z_k) + \phi_k^{\mathrm{T}}V\phi_k$. Since $\theta_{k+1} = P_{k+1}z_{k+1} + \theta$, it follows from Lemma A.1 that $\theta_{k+1} \sim \mathcal{N}(\theta, P_{k+1}\mathrm{var}(z_{k+1})P_{k+1})$. Thus, for all $k \ge 0$, θ_k is a Gaussian random variable with mean θ. Since $\mathrm{var}(z_0) = P_0^{-1}P_0P_0^{-1} = P_0^{-1}$, it follows that $\mathrm{var}(z_{k+1}) = \rho_k^{-2}\left(P_0^{-1} + \sum_{i=0}^{k}\rho_i^2\phi_i^{\mathrm{T}}V\phi_i\right)$. For convenience, define $M_k \triangleq \sum_{i=0}^{k}\rho_i\phi_i^{\mathrm{T}}\phi_i$, $M_{\nu,k} \triangleq \sum_{i=0}^{k}\rho_i^2\phi_i^{\mathrm{T}}V\phi_i$, $H_{0,k} \triangleq (P_0^{-1}+M_k)^{-1}P_0^{-1}(P_0^{-1}+M_k)^{-1}$, $H_{\nu,k} \triangleq (P_0^{-1}+M_k)^{-1}M_{\nu,k}(P_0^{-1} + M_k)^{-1}$. For all $k \ge 0$, it follows from Lemma 2 that

$$\alpha s_{\mathrm{l},\xi(k,N)}I_n \le M_k \le \beta s_{\mathrm{u},\xi(k,N)}I_n, \quad (23)$$

$$\alpha\lambda_{\min}(V)q_{\mathrm{l},\xi(k,N)}I_n \le M_{\nu,k} \le \beta\lambda_{\max}(V)q_{\mathrm{u},\xi(k,N)}I_n. \quad (24)$$

Since $\beta_k \ge 1$, it follows that $q_{\ell,\xi(k,N)} \to \infty$ as $k \to \infty$, and thus $\lambda_{\max}(M_k) \to \infty$ as $k \to \infty$. From this result and Lemma A.4 it follows that $\varlimsup_{k\to\infty}\lambda_{\max}(H_{0,k}) \le \varlimsup_{k\to\infty}\lambda_{\max}(P_0^{-1})/\lambda_{\max}(M_k)^2 = 0$. Hence, $\varlimsup_{k\to\infty}\lambda_{\max}(H_{0,k}) = 0$. Noting that $\mathrm{var}(\theta_k) = H_{0,k}+H_{\nu,k}$, it follows from Lemmas A.3 and A.4, (23), and (24) that

$$\varlimsup_{k\to\infty}\lambda_{\max}(\mathrm{var}(\theta_k)) \le \varlimsup_{k\to\infty}\lambda_{\max}(H_{0,k}) + \varlimsup_{k\to\infty}\lambda_{\max}(H_{\nu,k})$$

$$= \varlimsup_{k\to\infty}\lambda_{\max}(H_{\nu,k}) \le \varlimsup_{k\to\infty}\frac{\lambda_{\max}(M_{\nu,k})}{\lambda_{\max}(P_0^{-1} + M_k)^2}$$

$$\le \varlimsup_{k\to\infty}\frac{\lambda_{\max}(M_{\nu,k})}{\lambda_{\max}(M_k)^2} \le \frac{\beta\lambda_{\max}(V)}{\alpha^2}\varlimsup_{k\to\infty}\frac{q_{\mathrm{u},\xi(k,N)}}{s^2_{\mathrm{l},\xi(k,N)}},$$

Since $H_{0,k} \in \mathbb{P}^n$ and $\varliminf_{k\to\infty}\lambda_{\min}(H_{0,k}) \le \varlimsup_{k\to\infty}\lambda_{\max}(H_{0,k}) = 0$, it follows that $\varliminf_{k\to\infty}\lambda_{\min}(H_{0,k}) = 0$. Thus, from Lemmas A.3–A.5, Bernstein (2018, Fact 10.4.13), (23), and (24), it follows that

$$\frac{\alpha\lambda_{\min}(V)}{\beta^2}\varliminf_{k\to\infty}\frac{q_{\mathrm{l},\xi(k,N)}}{s^2_{\mathrm{u},\xi(k,N)}} \le \varliminf_{k\to\infty}\frac{\lambda_{\min}(M_{\nu,k})}{\lambda_{\min}(M_k)^2}$$

$$= \varliminf_{k\to\infty}\frac{\lambda_{\min}(M_{\nu,k})}{[\lambda_{\max}(P_0^{-1}) + \lambda_{\min}(M_k)]^2} \le \varliminf_{k\to\infty}\frac{\lambda_{\max}(M_{\nu,k})}{\lambda_{\min}(P_0^{-1} + M_k)^2}$$

$$\le \varliminf_{k\to\infty}[\lambda_{\min}(H_{0,k}) + \lambda_{\min}(H_{\nu,k})] = \varliminf_{k\to\infty}\lambda_{\min}(\mathrm{var}(\theta_k)). \quad \square$$

Corollary 4. *Under the notation and assumptions of Theorem 4, consider the following statements: i)* $\varlimsup_{k\to\infty}q_{\mathrm{u},\xi(k,N)}/s^2_{\mathrm{l},\xi(k,N)} = 0$, *ii)* $(\theta_k)_{k\ge0}$ *is consistent, iii),* $\varliminf_{k\to\infty}q_{\mathrm{l},\xi(k,N)}/s^2_{\mathrm{u},\xi(k,N)} = 0$.
Then (i) $\Longrightarrow$ *(ii)* $\Longrightarrow$ *(iii).*

Proof. To prove $(i) \Longrightarrow (ii)$, let $\varlimsup_{k\to\infty}q_{\mathrm{u},\xi(k,N)}/s^2_{\mathrm{l},\xi(k,N)} = 0$. Then $\lim_{k\to\infty}\lambda_{\max}(\mathrm{var}(\theta_k)) = 0$. Thus, from Lemma A.2, it follows that $(\theta_k)_{k\ge0}$ is consistent. To prove $(ii) \Longrightarrow (iii)$, suppose that $(\theta_k)_{k\ge0}$ is consistent. Then, from Lemma A.2, it follows that $\varliminf_{k\to\infty}\lambda_{\min}(\mathrm{var}(\theta_k)) = 0$, and therefore $\varliminf_{k\to\infty}q_{\mathrm{l},\xi(k,N)}/s^2_{\mathrm{u},\xi(k,N)} = 0$. $\square$

Corollary 5. *Under the notation and assumptions of Theorem 4, the following statements hold: i) assume that* $\prod_{k\ge0}\beta_k$ *is finite. Then* $(\theta_k)_{k\ge0}$ *is consistent; ii) let* $\beta_0 = 1$ *and for all* $k > 0$, *let* $\beta_k = 1 + \frac{1}{k}$. *Then* $(\theta_k)_{k\ge0}$ *is consistent; iii) let* $\gamma \in [1, \infty)$, *and, for all* $k \ge 0$, *let* $\beta_k = \gamma$. *Then* $(\theta_k)_{k\ge0}$ *is consistent if and only if* $\gamma = 1$.

Proof. To prove *i*), suppose that $\prod_{k\ge0}\beta_k = \rho$ and let $\varepsilon > 0$. Thus there exists $K > 0$ such that, for all $i \ge K$, $\rho - \varepsilon < \rho_i < \rho + \varepsilon$. Let $k_\varepsilon > 0$ be the smallest integer such that $\xi(k_\varepsilon, N)(N+1) \ge K$, and define $B_\varepsilon \triangleq \sum_{i=0}^{\xi(k_\varepsilon,N)}\rho^2_{i(N+1)+N}$ and $C_\varepsilon \triangleq \sum_{i=0}^{\xi(k_\varepsilon,N)}\rho_{i(N+1)}$. Then, for all $k > k_\varepsilon$,

$$\frac{q_{\mathrm{u},\xi(k,N)}}{s^2_{\mathrm{l},\xi(k,N)}} \le \frac{B_\varepsilon + (\rho+\varepsilon)^2(\xi(k,N) - \xi(k_\varepsilon,N) - 1)}{(C_\varepsilon + (\rho-\varepsilon)(\xi(k,N) - \xi(k_\varepsilon,N) - 1))^2}. \quad (25)$$

Since the limit superior of the left-hand side of (25) is zero, it follows that $(\theta_k)_{k\ge0}$ is consistent. To prove *ii*), for all $k \ge 0$, let $\beta_k = 1 + 1/k$. Then, for all $i \ge 0$, $\rho_i = i + 1$, and thus $q_{\mathrm{u},\xi(k,N)}$ and $s^2_{\mathrm{l},\xi(k,N)}$ are polynomials of degree three and four, respectively. Hence the limit superior is zero, and therefore $(\theta_k)_{k\ge0}$ is consistent. To prove *iii*), suppose that $\gamma = 1$. Then $\varlimsup_{k\to\infty}q_{\mathrm{u},\xi(k,N)}/s^2_{\mathrm{l},\xi(k,N)} = \varlimsup_{k\to\infty}\xi(k,N)^{-1} = 0$. Hence $(\theta_k)_{k\ge0}$ is consistent. Conversely, suppose $\gamma > 1$. Then, for all $i \ge 0$, $\rho_i = \gamma^{i+1}$, and thus

$$\varliminf_{k\to\infty}\frac{q_{\mathrm{l},\xi(k,N)}}{s^2_{\mathrm{u},\xi(k,N)}} = \frac{1}{\gamma^{2N}}\frac{(1-\gamma^{(N+1)})^2}{1-\gamma^{2(N+1)}}\varliminf_{k\to\infty}\frac{1-\gamma^{2(N+1)(\xi(k,N)+1)}}{(1-\gamma^{(N+1)(\xi(k,N)+1)})^2}$$

$$= \frac{1}{\gamma^{2N}}\frac{\gamma^{(N+1)}-1}{\gamma^{(N+1)}+1},$$

which is positive because $\gamma > 1$. Therefore, $(\theta_k)_{k\ge0}$ is not consistent. $\square$

Corollary 5 shows that if $\prod_{k\ge0}\beta_k$ converges, then VRF is consistent, but also that the converse is false, since $\prod_{k>0}1+\frac{1}{k} = \infty$. Furthermore, CRF is consistent if and only if $\lambda = 1$, and thus RLS with a constant, nontrivial forgetting factor is not consistent.

6. Example: abruptly changing parameters

Consider a mass–spring–damper system with $m = 5$ kg, $k = 1$ N/m, and $b = 1$ N·sec/m sampled at 1 sample/sec, and suppose that at 100 samples the parameters of the system abruptly change to $k = 10$ N/m and $b = 0.01$ N·sec/m. This process is modeled by the time-varying discrete-time transfer function

$$G_k(\mathbf{q}) = \begin{cases} \dfrac{0.4606\mathbf{q} + 0.4307}{\mathbf{q}^2 - 1.64\mathbf{q} + 0.8187}, & k < 100, \\[2ex] \dfrac{0.4218\mathbf{q} + 0.4215}{\mathbf{q}^2 - 0.3116\mathbf{q} + 0.998}, & k \ge 100, \end{cases} \quad (26)$$

where $\mathbf{q}$ is the forward shift operator. For all $k \ge 0$, let $u_k \sim \mathcal{N}(0, 1)$, and define

$$\beta_k \triangleq 1 + \eta\ \mathrm{sat}_\gamma(\|y_k - \phi_k\theta_k\|), \quad (27)$$

where $\eta, \gamma > 0$, and sat_γ is the unit-slope saturation function with saturation level γ. Fig. 1 shows the performance of VRF with $\gamma = \eta = 1$ and CRF with $\lambda = 0.99$. VRF converges to the initial system parameters and reconverges to the modified parameters in about 10 samples, illustrating Theorem 2. In contrast, while CRF converges to the initial parameters, reconvergence to the modified parameters is still not achieved at 200 samples. Next,

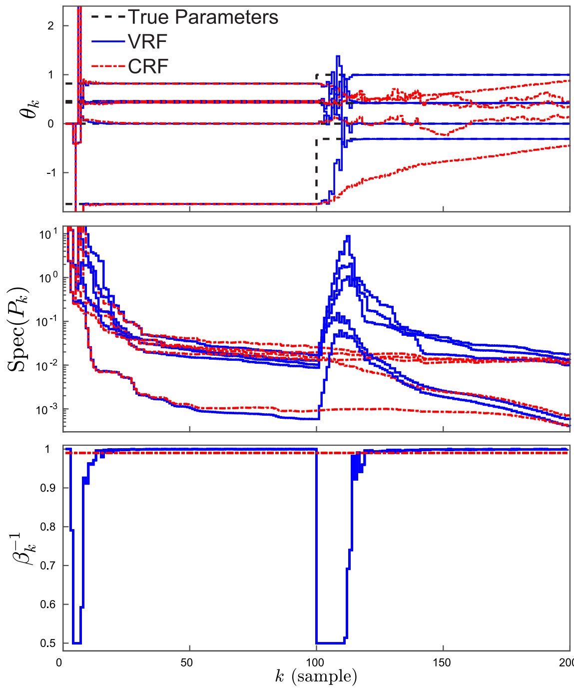

Fig. 1. The parameter estimate θ_k given by VRF with β_k defined by (27) reconverges after an abrupt change in the system as guaranteed by Theorem 2. In contrast, The parameter estimate given by CRF with $\lambda = 0.99$ requires many samples to reconverge.

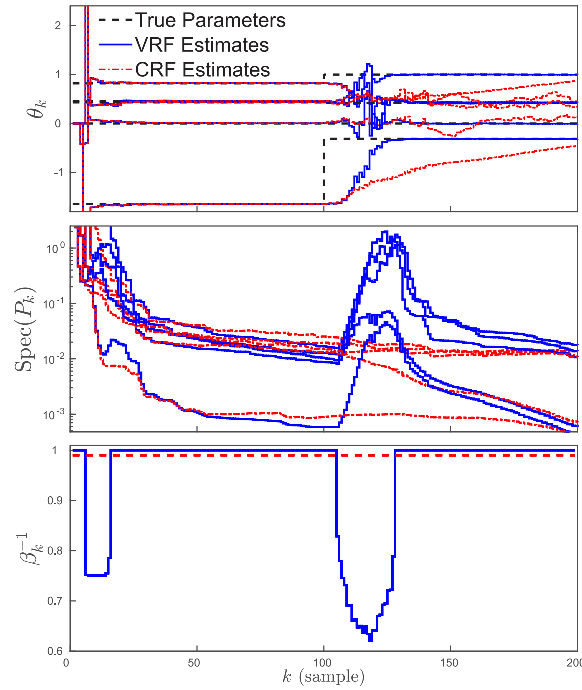

Fig. 2. The parameter estimate θ_k given by VRF with β_k defined by (28) reconverges after an abrupt change in the system with noisy measurements. In contrast, the parameter estimate given by CRF with $\lambda = 0.99$ requires many samples to reconverge.

consider the same system with the output corrupted by additive noise $v_k \sim \mathcal{N}(0, 0.05)$, and define

$$\beta_k \triangleq \begin{cases} 1 + \eta \ \mathrm{sat}_\gamma(E_\tau), & E_\tau > 1, \\ 1, & E_\tau \le 1, \end{cases} \tag{28}$$

where $\tau \in \mathbb{N}$ and $E_\tau \triangleq \left(\frac{1}{\tau}\sum_{i=k-\tau}^{k} \|y_i - \phi_i\theta_i\|^2\right)^{1/2}$. Fig. 2 shows the performance of VRF with $\eta = 1$, $\gamma = 5$, and $\tau = 10$, and CRF with $\lambda = 0.99$. VRF converges to the initial parameters and then reconverges to the new parameters in roughly 30 samples. As in the previous case, CRF converges to the initial parameters, but at 200 samples has still not reconverged to the modified parameters. ◇

Appendix. Lemmas

Lemma A.1. *Let $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. Let $X \sim \mathcal{N}(\mu, P)$ and define $Y \triangleq AX + b$. Then $Y \sim \mathcal{N}(A\mu + b, APA^{\mathrm{T}})$.*

Lemma A.2. *Let (Ω, Σ, P) be a probability space, let $\theta \in \mathbb{R}^n$, and let $(X_k\colon \Omega \to \mathbb{R}^n)_{k\ge 0}$ be a sequence of random variables such that, for all $k \ge 0$, $X_k \sim \mathcal{N}(\theta, \Sigma_k)$. Then $(X_k)_{k\ge 0}$ is a consistent estimator for θ if and only if $\lim_{k\to\infty} \Sigma_k = 0$.*

Lemma A.3. *Let $(A_k)_{k\ge 0}, (B_k)_{k\ge 0} \subset (\mathbb{N}^n)$. Then*

$$\overline{\lim_{k\to\infty}}\, \lambda_{\max}(A_k + B_k) \le \overline{\lim_{k\to\infty}}\, \lambda_{\max}(A_k) + \overline{\lim_{k\to\infty}}\, \lambda_{\max}(B_k),$$

$$\underline{\lim_{k\to\infty}}\, \lambda_{\min}(A_k + B_k) \ge \underline{\lim_{k\to\infty}}\, \lambda_{\min}(A_k) + \underline{\lim_{k\to\infty}}\, \lambda_{\min}(B_k).$$

Lemma A.4. *Let $A \in \mathbb{N}^n$ and $B \in \mathbb{P}^n$. Then, for all $i = 1, \ldots, n$,*

$$\frac{\lambda_{\min}(A)}{\lambda_i(B)^2} \le \lambda_i(B^{-1}AB^{-1}) \le \frac{\lambda_{\max}(A)}{\lambda_i(B)^2}. \tag{A.1}$$

Now assume that $A \in \mathbb{P}^n$. Then there exist $0 < b_1 \le b_2$ and $0 < a_1 \le a_2$ such that

$$a_1 I_n \le A \le a_2 I_n, \tag{A.2}$$

$$b_1 I_n \le B \le b_2 I_n. \tag{A.3}$$

Furthermore, for all a_1, a_2, b_1, b_2 satisfying (A.2), (A.3),

$$\frac{a_1}{b_2^2} I_n \le B^{-1}AB^{-1} \le \frac{a_2}{b_1^2} I_n. \tag{A.4}$$

Lemma A.5. *Let $a \in [0, \infty)$, let $(b_k)_{k\ge 0}, (c_k)_{k\ge 0} \subset [0, \infty)$, and assume that $\lim_{k\to\infty} b_k = \infty$. Then, for all $p \ge 0$,*

$$\lim_{k\to\infty} \frac{c_k}{(a + b_k)^p} = \lim_{k\to\infty} \frac{c_k}{b_k^p}. \tag{A.5}$$

References

Albert, A., & Sittler, R. W. (1965). A method for computing least squares estimators that keep up with the data. *SIAM Journal on Control*, *3*(3), 384–417.

Ali, A. A., Hoagg, J. B., Mossberg, M., & Bernstein, D. S. (2016). On the stability and convergence of a sliding-window variable-regularization recursive-least-squares algorithm. *International Journal of Adaptive Control and Signal Processing*, *30*, 715–735.

Astrom, K. J. (1995). *Adaptive control* (2nd ed.). Reading, MA: Addison-Wesley.

Bernstein, D. S. (2018). *Scalar, vector, and matrix mathematics: Theory, facts, and formulas - revised and expanded edition*. 41 William Street, Princeton, NJ, 08540: Princeton University Press.

Fortescue, T. R., Kershenbaum, L. S., & Ydstie, B. E. (1981). Implementation of self-tuning regulators with variable forgetting factors. *Automatica*, *17*(6), 831–835.

Leung, S.-H., & So, C. F. (2005). Gradient-based variable forgetting factor RLS algorithm in time-varying environments. *IEEE Transactions on Signal Processing*, *53*(8), 3141–3150.

Ljung, L., & Soderstrom, T. (1983). *Theory and practice of recursive identification*. The MIT Press.

Paleologu, C., Benesty, J., & Silviu, C. (2008). A robust variable forgetting factor recursive least-squares algorithm for system identification. *IEEE Signal Processing Letters*, *15*, 597–600.

Park, D. J., Jun, B. E., & Kim, J. H. (1991). Fast tracking RLS algorithm using novel variable forgetting factor with unity zone. *Electronics Letters*, *27*(23), 2150–2151.
Song, S., Lim, J.-S., Baek, S. J., & Sung, K.-M. (2000). Gauss Newton variable forgetting factor recursive least squares for time varying parameter tracking. *Electronics Letters*, *36*(11), 988–990.
Ul Islam, S. A., & Bernstein, D. S. (2019). Recursive least squares for real-time implementation. *IEEE Control Systems Magazine*, *39*(3), 82–85.

Adam L. Bruce received the B.A. in Physics and Applied Mathematics from the University of California, Berkeley, the M.S. in Aeronautics and Astronautics from Purdue University, and is currently a PhD Candidate at the University of Michigan, Ann Arbor. His interests are in identification, estimation, and adaptive control for nonlinear, uncertain, and complex systems.

Ankit Goel received the B.E. degree in mechanical engineering from the Delhi College of Engineering, Delhi, the M.S. and the Ph.D. degree in aerospace engineering from the University of Michigan in Ann Arbor. His interests are in data-driven estimation and control of high-dimensional complex systems such as quadcopters and scramjets.

Dennis S. Bernstein received the Sc.B. degree from Brown University and the Ph.D. degree from the University of Michigan in Ann Arbor, Michigan, where he is currently professor in the Aerospace Engineering Department. His interests are in identification, estimation, and control for aerospace applications. He is the author of Scalar, Vector, and Matrix Mathematics, published by Princeton University Press.