2018 IEEE Conference on Decision and Control (CDC)
Miami Beach, FL, USA, Dec. 17-19, 2018

A Targeted Forgetting Factor for Recursive Least Squares

Ankit Goel[1] and Dennis S Bernstein[1]

Abstract— **Recursive least squares (RLS) is widely used in signal processing, identification, and control, but is plagued by the inability to adjust quickly to changes in the unknown parameters. RLS with standard forgetting factor overcomes this problem but causes divergence due to the lack of persistency. Variable and directional forgetting factors have been proposed for overcoming this deficiency. The present paper proposes a *targeted forgetting factor* that looks directly at recent data in order to determine which directions possess new information. Targeted forgetting applies a forgetting factor directly to these directions, thereby providing a simple and effective technique for avoiding covariance divergence. Numerical examples compare targeted forgetting to standard and directional forgetting.**

I. Introduction

Recursive least squares (RLS) is a foundational algorithm in estimation and control theory [1]–[3]. In simplest terms, RLS is an iterative version of batch least squares for solving matrix-vector equations of the form $Ax = b$. RLS is often used for system identification with input-output models, where the model coefficients are the components of x and the matrix A is a time-varying regressor of input-output data [4]. The equivalence is not exact, however, since the covariance matrix in RLS must be initialized at a finite value.

RLS is closely related to the Kalman filter. In particular, for the dynamics $x_{k+1} = x_k$, which states that the unknown state x is constant, the equation $b = Ax$ can be viewed as the output equation $y = Cx$, where b plays the role of the measurement y and A plays the role of the output matrix C. The resulting Riccati equation of the Kalman filter is then precisely the RLS covariance update equation. Conversely, the Kalman filter can be derived as a consequence of RLS [3].

A useful feature of RLS that is not shared by batch least squares is the ability to employ a forgetting factor. The forgetting factor is useful in cases where the unknown parameters x change. Although batch least squares can estimate the modified parameters, the convergence is typically slow. The forgetting factor thus weights the most recent data, thereby speeding up convergence to the modified parameters.

A detrimental side effect of the forgetting factor is that, in the absence of persistency, the covariance matrix may diverge. Numerous techniques have been developed to address this problem, including bounded-noise forgetting [5], variable forgetting [6], and directional forgetting [7]–[9].

The present paper provides an alternative approach to forgetting that is robust to the lack of persistency, and still retains the tracking capabilities of standard forgetting. In particular, this approach decomposes the covariance matrix in order to determine the directions of new information entering the system through data. Forgetting is then applied only to the directions corresponding to new information. This technique, called *targeted forgetting*, therefore avoids applying the forgetting factor to directions that receive no new data, thereby avoiding divergence of the covariance matrix. Numerical examples and comparison to existing forgetting techniques demonstrate the potential advantages of this approach.

[1]Dept. Aerospace Engineering, University of Michigan, Ann Arbor, MI.

II. Recursive Least Squares

Consider the LTI SISO system

$$y(k) = G(\mathbf{q})u(k), \tag{1}$$

where $G(\mathbf{q})$ is a strictly proper nth-order rational transfer function, $\mathbf{q}$ is the forward-shift operator, u is the input to the system, and y is the measurement. The goal is to identify the coefficients of $G(\mathbf{q})$ using the measurements y and the input u. We rewrite (1) as

$$y(k) = \phi(k)^{\mathrm{T}}\theta, \tag{2}$$

where $\theta \in \mathbb{R}^{2n}$ contains the coefficients of $G(\mathbf{q})$, and

$$\phi(k) \triangleq \begin{bmatrix} u(k-1) \\ \vdots \\ u(k-n) \\ y(k-1) \\ \vdots \\ y(k-n) \end{bmatrix} \in \mathbb{R}^{2n}. \tag{3}$$

To identify θ, we minimize the cost function

$$J(k,\hat{\theta}) \triangleq \sum_{i=1}^{k} \lambda^{k-i}(y(i) - \phi(i)^{\mathrm{T}}\hat{\theta})^{\mathrm{T}}(y(i) - \phi(i)^{\mathrm{T}}\hat{\theta}) + \lambda^k \hat{\theta}^{\mathrm{T}} P(0)^{-1}\hat{\theta}, \tag{4}$$

where $\lambda \in (0,1]$ is the *forgetting factor*, and $P(0)^{-1}$ is a positive-definite matrix.

The batch least squares minimizer $\theta(k)$ of (4) is given by

$$\theta(k) = -A_\theta(k)^{-1} b_\theta(k), \tag{5}$$

where

$$A_\theta(k) \triangleq \sum_{i=1}^{k} \lambda^{k-i}\phi(i)\phi(i)^{\mathrm{T}} + P(0)^{-1}, \tag{6}$$

$$b_\theta(k) \triangleq \sum_{i=1}^{k} \lambda^{k-i}\phi(i)y(i). \tag{7}$$

978-1-5386-1394-8/18/$31.00 ©2018 IEEE

The following result, which presents *recursive least squares* with standard forgetting, characterizes the minimizer of (4).

Proposition II.1. Let $\theta(0) = 0$. Then, for all $k \geq 1$, the cost function (4) has a unique global minimizer $\theta(k)$, which is given by

$$\theta(k) = \theta(k-1) - P(k)\phi(k)(y(k) - \phi(k)^{\mathrm{T}}\theta(k-1)), \tag{8}$$

$$P(k) = \frac{1}{\lambda}\big(P(k-1) - P(k-1)\phi(k)\Gamma(k)\phi(k)^{\mathrm{T}}P(k-1)\big), \tag{9}$$

where

$$\Gamma(k) \triangleq [\lambda + \phi(k)^{\mathrm{T}}P(k-1)\phi(k)]^{-1}.$$

Further,

$$P(k) = A_{\theta}(k)^{-1}. \tag{10}$$

Note that, for $\lambda = 1$, P is nonincreasing in the sense that, for all $k \geq 1$, $P(k+1) \leq P(k)$. If the input u is such that the order of excitation of $\phi(k)$ is at least $2n$, then all singular values of $P(k)$ decrease at each step. If, however, the order of excitation of $\phi(k)$ is less than $2n$, then the only singular values of $P(k)$ that decrease are those corresponding to the singular vectors that receive data with new information.

We use the following proposition to identify the singular vectors receiving new information in $\phi(k)$.

Proposition II.2. Let c be a real number, let $\omega_1, \ldots, \omega_p$ be distinct nonzero real numbers no pair of which differs by an integer multiple of 2π, and define

$$u(k) = c + \sum_{i=1}^{p} \sin \omega_i k. \tag{11}$$

For all $m \geq 2p+1$, define

$$\phi(k) \triangleq \begin{bmatrix} u(k-1) & \cdots & u(k-m) \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^m, \tag{12}$$

$$R(k) \triangleq \sum_{i=1}^{m} \phi(k-i)\phi(k-i)^{\mathrm{T}}. \tag{13}$$

Then, for all $k > m$,

$$\operatorname{rank} R(k) = 2p+1. \tag{14}$$

Equivalently, if $2p+1 < m$, then $\phi(k)$ is constrained to lie in $2p+1$ dimensional subspace of $\mathbb{R}^m$. In this case, we say that $\phi(k)$ is *exciting of order* $2p+1$. Furthermore, if $c = 0$, then, for all $k > m$,

$$\operatorname{rank} R(k) = 2p. \tag{15}$$

Finally, if $u(k) \sim \mathcal{N}(0,1)$, then, for all $k > m$ and with probability 1,

$$\operatorname{rank} R(k) = m. \tag{16}$$

If $\lambda < 1$ and the order of excitation of $\phi(k)$ is less than m, then the singular values that do not decrease in the case $\lambda = 1$ are divided by $\lambda < 1$ and thus grow unbounded. Consequently, the condition number of $P(k)$ increases, and $P(k)$ becomes ill-conditioned. To prevent this, we use a variation of directional forgetting [9] to update $P(k)$. Instead of dividing the right-hand side of (9) by λ, we propagate $P(k)$ by using

$$R(k) \triangleq \sum_{i=1}^{m} \phi(k-i)^{\mathrm{T}}\phi(k-i), \tag{17}$$

$$\bar{P}(k) = P(k-1) - P(k-1)\phi(k-1)\Gamma(k) \cdot \phi(k-1)^{\mathrm{T}}P(k-1), \tag{18}$$

$$\Sigma(k) = U(k)^{\mathrm{T}}\bar{P}(k)U(k), \tag{19}$$

$$\bar{\Sigma}(k)(i,i) = \begin{cases} \Sigma(k)(i,i), & i < 2n - \operatorname{rank} R(k), \\ \Sigma(k)(i,i)/\lambda, & i \geq 2n - \operatorname{rank} R(k), \end{cases} \tag{20}$$

$$P(k) = U(k)\bar{\Sigma}(k)U(k)^{\mathrm{T}}, \tag{21}$$

where $\Sigma(k)$ contains the singular values of $P(k)$, and $U(k)$ contains the singular vectors of $P(k)$. Note that $\Sigma(k)$ and $\bar{\Sigma}(k)$ are diagonal matrices, and $\Sigma(k)(i,i)$ denotes the (i,i) entry of $\Sigma(k)$.

We compare the efficacy of targeted forgetting with directional forgetting [8]. The update equations for $P(k)$ using directional forgetting are

$$M(k) = \begin{cases} (1-\lambda)\dfrac{R(k-1)\phi(k)\phi(k)^{\mathrm{T}}}{\phi(k)^{\mathrm{T}}R(k-1)\phi(k)}, & \|\phi(k)\|_2 > \varepsilon, \\ 0, & \|\phi(k)\|_2 \leq \varepsilon, \end{cases} \tag{22}$$

$$R(k) = (I_{2n} - M(k))R(k-1) + \phi(k)\phi(k)^{\mathrm{T}}, \tag{23}$$

$$\bar{P}(k-1) = P(k-1)(I_{2n} - M(k))^{-1}, \tag{24}$$

$$P(k) = \bar{P}(k-1) - \frac{\bar{P}(k-1)\phi(k-1)\phi(k-1)^{\mathrm{T}}\bar{P}(k-1)}{1 + \phi(k-1)^{\mathrm{T}}\bar{P}(k-1)\phi(k-1)}, \tag{25}$$

and $R(0) = P(0)^{-1}$.

III. Numerical example

In this section, we present two numerical examples to highlight the features of targeted forgetting in comparison with standard and directional forgetting. In the first example, we compare the performance of all three algorithms when the data used for estimation suffers a drop in persistency. In the second example, we compare the tracking capabilities of all three algorithms.

Example III.1. Consider the LTI system described by the transfer functions

$$G(\mathbf{q}) = \frac{\mathbf{q} + 0.2\mathbf{q}}{\mathbf{q}^2 - 0.6\mathbf{q} + 0.08}. \tag{26}$$

Equivalently,

$$\theta = \begin{bmatrix} 1.00 & 0.20 & 0.60 & -0.08 \end{bmatrix}^{\mathrm{T}}. \tag{27}$$

The input signal u is given by

$$u(k) = \begin{cases} 1, & k \in \{10000, 60000\}. \\ 1 + \sin\dfrac{2\pi k}{10} + \sin\dfrac{2\pi k}{20} + \sin\dfrac{2\pi k}{100}, & \text{otherwise.} \end{cases} \tag{28}$$

Note that $u(k)$ is sufficiently exciting to identify θ up to $k = 10000$. For $k > 10000$, $u(k)$ lacks sufficient persistency to

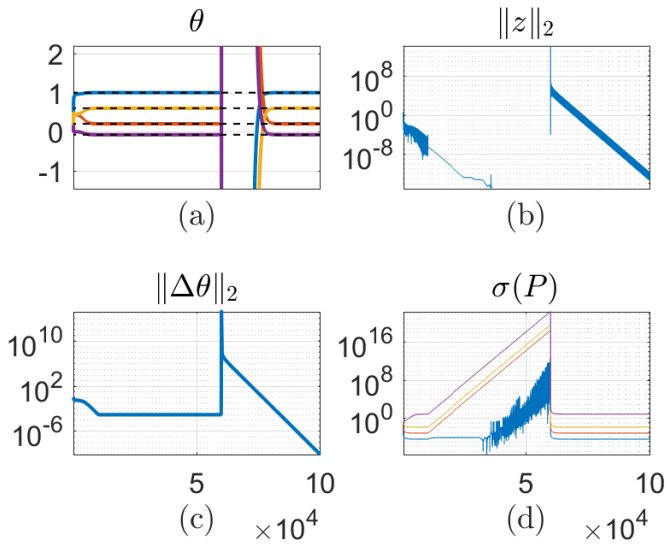

Fig. 1: RLS estimate of θ using standard forgetting. (a) shows the estimate $\theta(k)$, (b) shows the 2-norm of predicted error $z(k)$, (c) shows the 2-norm of the estimation error $\Delta\theta(k) \triangleq \theta - \theta(k)$, and (d) shows the singular values of $P(k)$. Note that the order of excitation drops between $k = 10000$ and $k = 60000$. There is no new information in three singular directions, and hence the corresponding singular values diverge. Eventually, numerical propagation of $P(k)$ becomes erroneous.

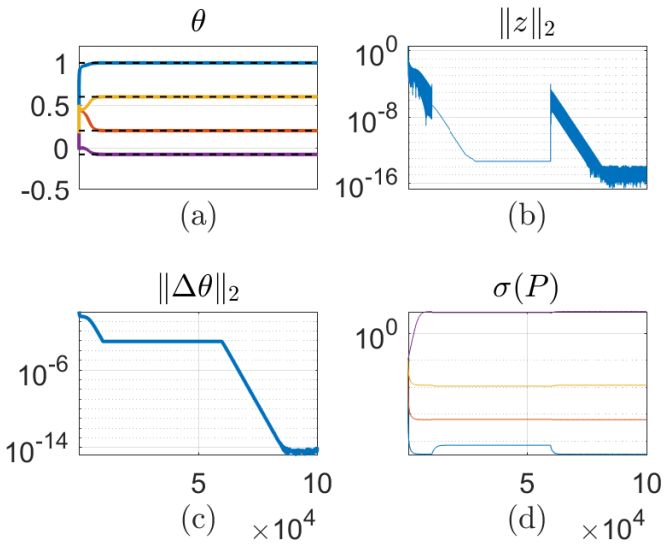

Fig. 2: RLS estimate of θ using targeted forgetting. (a) shows the estimate $\theta(k)$, (b) shows the 2-norm of predicted error $z(k)$, (c) shows the 2-norm of the estimation error $\Delta\theta(k) \triangleq \theta - \theta(k)$, and (d) shows the singular values of $P(k)$. Note that the order of excitation drops between $k = 10000$ and $k = 60000$. There is no new information in three singular directions, and hence the corresponding singular values are not divided by λ. Consequently, $P(k)$ remains well-conditioned.

estimate θ. For $k > 60000$, $u(k)$ is again sufficiently persistent to estimate θ.

We use RLS with standard, targeted, and directional forgetting to estimate θ. All algorithms are initialized with $P(0) = 0.1I_{l_\theta}$ and $\lambda = 0.999$.

Figure 1 shows the estimate of θ with standard forgetting. Note that, for $k < 10000$, $\theta(k)$ converges toward θ since $u(k)$ is persistent. For $10000 < k < 60000$, $\theta(k)$ converges to a minimizer of (4) that minimizes the predicted error $z(k) \triangleq y(k) - \phi(k)^{\mathrm{T}}\theta(k-1)$. In effect, the transfer function that produces $y(k)$ after $k > 10000$ is not unique. However, as shown by Figure 1, the covariance matrix diverges. Note that all but one of the singular values of P diverge since the step input $u(k) = 1$ is persistent of order one. At $k = 60000$, the estimate of θ changes abruptly since $P(k)$ is large, but eventually converges toward θ since $u(k)$ is persistent.

Figure 2 shows the estimate of θ with targeted forgetting. Unlike RLS with standard forgetting, the singular values of $P(k)$ do not diverge when $u(k)$ is not persistent, and hence the estimate $\theta(k)$ does not change abruptly.

Figure 3 shows the estimate of θ with directional forgetting. Similar to RLS with targeted forgetting, the singular values of $P(k)$ do not diverge when $u(k)$ is not persistent, and hence the estimate $\theta(k)$ does not change abruptly. However, the convergence is slow. ◇

Example III.2. Consider the abruptly changing LTI system described by the transfer functions

$$G(\mathbf{q}) = \begin{cases} \dfrac{\mathbf{q} + 0.2\mathbf{q}}{\mathbf{q}^2 - 0.6\mathbf{q} + 0.08}, & k \leq 50000, \\ \dfrac{2\mathbf{q} + 0.5\mathbf{q}}{\mathbf{q}^2 + 0.4\mathbf{q} - 0.05}, & k > 50000. \end{cases} \tag{29}$$

Equivalently,

$$\theta = \begin{cases} \begin{bmatrix} 1.00 & 0.20 & 0.60 & -0.08 \end{bmatrix}^{\mathrm{T}}, & k \leq 50000, \\ \begin{bmatrix} 2.00 & 0.50 & -0.40 & 0.05 \end{bmatrix}^{\mathrm{T}}, & k > 50000. \end{cases} \tag{30}$$

The driving signal u is given by

$$u(k) = 1 + \sin\frac{2\pi k}{10} + \sin\frac{2\pi k}{20} + \sin\frac{2\pi k}{100}. \tag{31}$$

Note that $u(k)$ is sufficiently exciting to identify θ. We use RLS with standard, targeted, and directional forgetting to estimate θ. All algorithms are initialized with $P(0) = 0.1I_{l_\theta}$, and $\lambda = 0.999$.

Figure 4 shows the estimate of θ with standard forgetting. With standard forgetting, $\theta(k)$ tracks the changing parameters despite the abrupt changes.

Figure 5 shows the estimate of θ with targetted forgetting. Note that, since the persistence of the input $u(k)$ does not decrease, targetted forgetting provides performance that is typical of standard forgetting.

Figure 6 shows the estimate of θ with directional forgetting. Note that, although the persistence of the input $u(k)$ does not decrease, directional forgetting does not recover the performance of standard forgetting. ◇

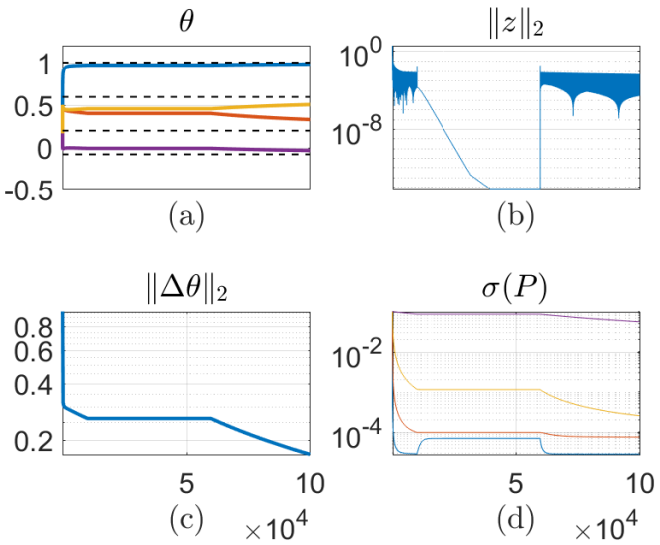

Fig. 3: RLS estimate of θ using directional forgetting. (a) shows the estimate $\theta(k)$, (b) shows the 2-norm of predicted error $z(k)$, (c) shows the 2-norm of the estimation error $\Delta\theta(k) \triangleq \theta - \theta(k)$, and (d) shows the singular values of $P(k)$. Note that the order of excitation drops between $k = 10000$ and $k = 60000$. With directional forgetting, the singular values of $P(k)$ do not increase when the order of excitation drops. Consequently, $P(k)$ remains well-conditioned. However, convergence is slow.

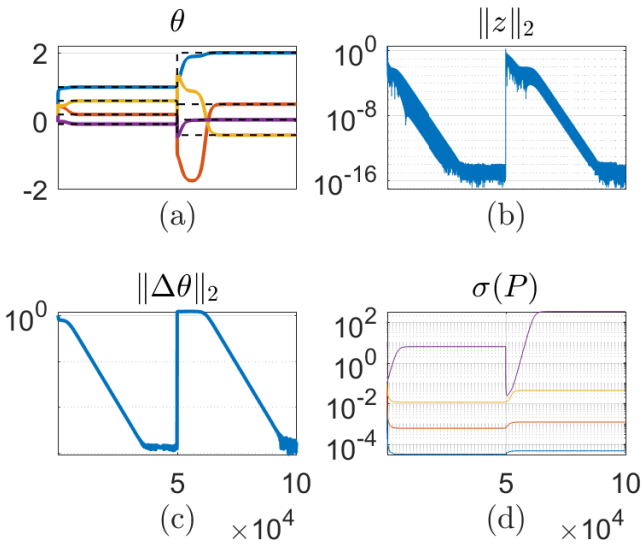

Fig. 4: RLS estimate of θ using standard forgetting. (a) shows the estimate $\theta(k)$, (b) shows the 2-norm of the predicted error $z(k)$, (c) shows the 2-norm of the estimation error $\Delta\theta(k) \triangleq \theta - \theta(k)$, and (d) shows the singular values of $P(k)$. Note that, despite the abrupt parameter change, $\theta(k)$ reconverges to the new value of θ since $u(k)$ is sufficiently persistent.

IV. Conclusions

This paper introduced a *targeted forgetting factor* for recursive least squares (RLS). Using the singular value decomposition of the covariance matrix, the targeted forgetting factor determines which singular vector possess new information and then selectively applies forgetting to those directions. Numerical results show that targeted forgetting compares favorably with directional forgetting in terms of convergence rate, but is computationally more expensive. Consequently, targeted forgetting is more suitable for low-dimensional problems. Further, targeted forgetting provides performance that is typical of standard forgetting performance for tracking changing parameters when the data is sufficiently exciting.

Future research will focus on 1) refining the choice of threshold for selectively applying targeted forgetting; 2) improving the computational efficiency of targeted forgetting; and 3) applying targeted forgetting to identification and adaptive control problems.

References

[1] K. J. Astrom and B. Wittenmark, *Adaptive Control*, 2nd ed. Addison-Wesley, 1995.

[2] L. Ljung and T. Soderstrom, *Theory and Practice of Recursive Identification*. MIT Press, 1983.

[3] A. H. Sayed, *Fundamentals of Adaptive Filtering*. Wiley, 2003.

[4] L. Ljung, *System Identification: Theory for the User*, 2nd ed., ser. Prentice Hall Information and System Sciences Series. Prentice Hall, January 1999.

[5] S. Dasgupta and Y.-F. Huang, "Asymptotically convergent modified recursive least-squares with data-dependent updating and forgetting factor for systems with bounded noise," *IEEE Trans. Infor. Th.*, vol. 33, no. 6, pp. 383–392, 1987.

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

[7] S. Bittanti, P. Bolzern, and M. Campi, "Convergence and exponential convergence of identification algorithms with directional forgetting factor," *Automatica*, vol. 26, pp. 929–932, 1990.

[8] 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.

[9] R. Kulhavỳ and M. Kárnỳ, "Tracking of slowly varying parameters by directional forgetting," *IFAC Proc. Vol.*, vol. 17, no. 2, pp. 687–692, 1984.

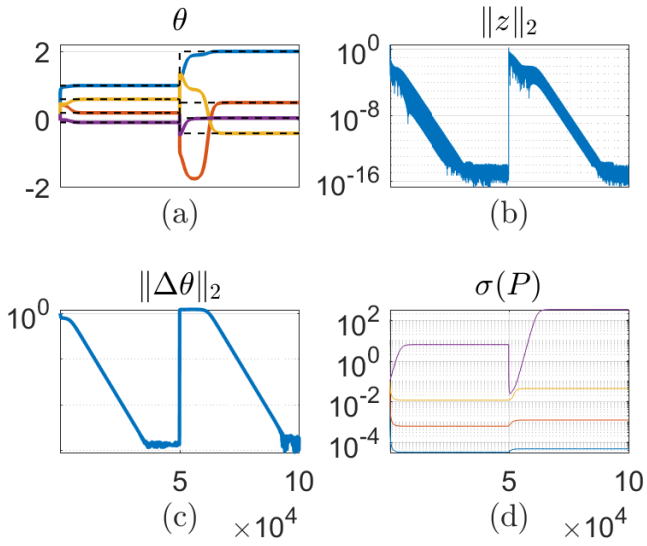

Fig. 5: RLS estimate of θ using targeted forgetting. (a) shows the estimate $\theta(k)$, (b) shows the 2-norm of the predicted error $z(k)$, (c) shows the 2-norm of the estimation error $\Delta\theta(k) \triangleq \theta - \theta(k)$, and (d) shows the singular values of $P(k)$. Note that, since the persistence of the input $u(k)$ does not decrease, targetted forgetting provides performance that is typical of standard forgetting.

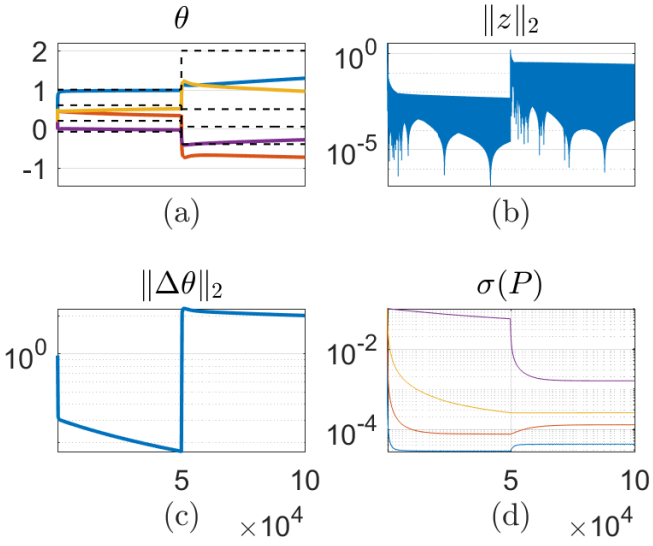

Fig. 6: RLS estimate of θ using directional forgetting. (a) shows the estimate of $\theta(k)$, (b) shows the 2-norm of the predicted error $z(k)$, (c) shows the 2-norm of the estimation error $\Delta\theta(k) \triangleq \theta - \theta(k)$, and (d) shows the singular values of $P(k)$. Note that, although the persistence of the input $u(k)$ does not decrease, directional forgetting does not provide the performance typical of standard forgetting.