2021 American Control Conference (ACC)
New Orleans, USA, May 25-28, 2021

Regularization-Induced Bias and Consistency in Recursive Least Squares

Brian Lai, Syed Aseem Ul Islam, and Dennis S. Bernstein

***Abstract*— Within the context of recursive least squares (RLS) parameter estimation, the goal of the present paper is to study the effect of regularization-induced bias on the transient and asymptotic accuracy of the parameter estimates. We consider this question in three stages. First, we consider regression with random data, in which case persistency is guaranteed. Next, we apply RLS to finite-impulse-response (FIR) system identification and, finally, to infinite-impulse-response (IIR) system identification. For each case, we relate the condition number of the regressor matrix to the transient response and rate of convergence of the parameter estimates.**

I. Introduction

In many parameter estimation problems, data becomes available in real time, and parameter estimates are desired at each successive step. For least-squares estimation, recursive least square (RLS) is computationally efficient relative to batch least squares due to the need to invert a matrix whose dimension is lower than the parameter dimension [1]–[4]. An additional advantage of RLS is the ability to include a forgetting factor, which weights more recent data more heavily than older data. In effect, the forgetting factor facilitates learning in response to system changes by promoting forgetting. Since system changes can occur sporadically and unexpectedly, forgetting must be adaptive and variable, see [5] and the references therein.

Since batch least squares inverts a regressor matrix over a window of data, a full-rank regressor matrix is required before the first parameter estimate is available. The invertibility of the regressor matrix depends on a persistency condition [6]. In the absence of persistency, however, a regularization term can be included, resulting in regularized batch least squares. Regularization is an essential technique in parameter estimation [7]–[10]. However, the regularization term perturbs the regressor, thus resulting in *regularization-induced bias*. Along the same lines, RLS includes a regularization term in the cost function, which provides parameter estimates from the initial time. As in the case of regularized batch least squares, the regularization term leads to bias in the parameter estimates of RLS.

For the case of constant regularization, the goal of the present paper is to study the effect of regularization-induced bias on the transient and asymptotic accuracy of the parameter estimates. Note that this bias is due solely to the regularization and thus is not due to noise, which can also result in bias. Other works have studied the bias induced by noise without regularization, [11] chapter 2, and the bias induced by noise and regularization together [12]. We consider this question in three stages. First, we consider regression with random data, in which case persistency is guaranteed. Next, we apply RLS to finite-impulse-response (FIR) system identification and, finally, to infinite-impulse-response (IIR) system identification. For each case, we relate the condition number of the regressor matrix to the transient response and rate of convergence of the parameter estimates.

Brian Lai, Syed Aseem Ul Islam, and Dennis S. Bernstein are with the Department of Aerospace Engineering, University of Michigan, Ann Arbor, MI, USA. {brianlai, aseemisl, dsbaero}@umich.edu

Notation and Terminology. Define $\mathbb{N} \triangleq \{1, 2, 3, \dots\}$ and $\mathbb{N}_0 \triangleq \{0\} \cup \mathbb{N}$. The symbols $\mathbf{S}^n$, $\mathbf{N}^n$, and $\mathbf{P}^n$ denote the sets of real $n \times n$ symmetric, symmetric positive-semidefinite, and symmetric positive-definite matrices, respectively. For $A \in \mathbf{N}^n$, $\lambda_i(A)$ denotes the ith largest eigenvalue of A, $\lambda_{\max}(A) \triangleq \lambda_1(A)$, and $\lambda_{\min}(A) \triangleq \lambda_n(A)$. Furthermore, the condition number, $\kappa(A)$, of $A \in \mathbf{P}^n$ is defined by

$$\kappa(A) \triangleq \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}. \tag{1}$$

If A is positive-semidefinite but not positive-definite, then $\kappa(A) \triangleq \infty$. $\|\cdot\|$ is the Euclidean norm.

II. Batch Least Squares

We consider the measurement process

$$y_k = \phi_k \theta, \tag{2}$$

where $k = 0, 1, 2, \dots$ is the time step, $y_k \in \mathbb{R}^p$ is the measurement at step k, the matrix $\phi_k \in \mathbb{R}^{p \times n}$ is the regressor at step k, and $\theta \in \mathbb{R}^n$ is a column vector of n unknown parameters. The objective is to use y_k and ϕ_k to estimate the components of θ. Since, in practice, y_k and ϕ_k are corrupted by noise, (2) does not hold exactly, and we thus consider least squares estimates of θ.

The batch approach to this problem is to collect a large amount of data and then apply least squares optimization to compute an estimate of θ. In particular, using data from the step window $i = 0, 1, \dots, k$, it follows from (2) that

$$Y_k = \Phi_k \theta, \tag{3}$$

where

$$Y_k \triangleq \begin{bmatrix} y_0 \\ \vdots \\ y_k \end{bmatrix}, \quad \Phi_k \triangleq \begin{bmatrix} \phi_0 \\ \vdots \\ \phi_k \end{bmatrix}. \tag{4}$$

Note that (3) has the form $Ax = b$, where A denotes Φ_k, x denotes θ, and b denotes Y_k.

978-1-6654-4198-8/$31.00 ©2021 AACC

In the presence of noise corrupting the data Y and Φ, (3) may not have a solution. In this case, it is useful to replace (3) by the least-squares optimization of the cost

$$J_{R,k}(\hat{\theta}) \triangleq \sum_{i=0}^{k}(y_i - \phi_i\hat{\theta})^{\mathrm{T}}(y_i - \phi_i\hat{\theta}) + (\hat{\theta} - \theta_0)^{\mathrm{T}}R(\hat{\theta} - \theta_0)$$
$$= (Y_k - \Phi_k\hat{\theta})^{\mathrm{T}}(Y_k - \Phi_k\hat{\theta}) + (\hat{\theta} - \theta_0)^{\mathrm{T}}R(\hat{\theta} - \theta_0), \tag{5}$$

where $R \in \mathbf{P}^n$ and $\theta_0 \in \mathbb{R}^n$ is an initial estimate of θ. The regularization term $(\hat{\theta} - \theta_0)^{\mathrm{T}}R(\hat{\theta} - \theta_0)$ weighs the distance from the current estimate to the initial estimate and ensures that $J_{R,k}$ has a unique global minimizer. In particular, the *batch least-squares* (BLS) minimizer of (5) is given by

$$\theta_{\mathrm{BLS},R,k+1} \triangleq \operatorname*{argmin}_{\hat{\theta}\in\mathbb{R}^n} J_{R,k}(\hat{\theta}) \tag{6}$$
$$= (\Phi_k^{\mathrm{T}}\Phi_k + R)^{-1}(\Phi_k^{\mathrm{T}}Y_k + R\theta_0), \tag{7}$$

where the inverse $(\Phi_k^{\mathrm{T}}\Phi_k + R)^{-1}$ exists due to the positive-definite regularization R.

Note that $(\Phi_k^{\mathrm{T}}\Phi_k + R)^{-1}$ requires the computation of an $n \times n$ inverse, and thus the computational requirement of the inverse is of order n^3. Note also that the memory needed to store Φ_k grows with k. Furthermore, if Φ_k has full column rank, then the regularization is not needed, and thus R can be set to zero. In this case, (7) becomes

$$\theta_{\mathrm{BLS},0,k+1} = (\Phi_k^{\mathrm{T}}\Phi_k)^{-1}\Phi_k^{\mathrm{T}}Y_k, \tag{8}$$

which is the unique solution of (3).

III. Recursive Least Squares

Recursive least squares (RLS) provides a recursive update of the minimizer of (5) as measurements become available. Although RLS can be stated with a forgetting factor λ, the following result provides a statement of RLS with $\lambda = 1$.

Theorem 1. *For all $k \in \mathbb{N}_0$, let $\phi_k \in \mathbb{R}^{p\times n}$ and $y_k \in \mathbb{R}^p$. Furthermore, let $\theta_0 \in \mathbb{R}^n$, let $P_0 \in \mathbb{R}^{n\times n}$ be positive definite, and let $\lambda \in (0, 1]$. Furthermore, for all $k \in \mathbb{N}_0$, denote the minimizer of the function $J_{P_0^{-1},k}(\hat{\theta})$ by*

$$\theta_{k+1} \triangleq \operatorname*{argmin}_{\hat{\theta}\in\mathbb{R}^n} J_{P_0^{-1},k}(\hat{\theta}). \tag{9}$$

Then, for all $k \in \mathbb{N}_0$, θ_{k+1} is given by

$$P_{k+1} = P_k - P_k\phi_k^{\mathrm{T}}(I + \phi_k P_k \phi_k^{\mathrm{T}})^{-1}\phi_k P_k, \tag{10}$$
$$\theta_{k+1} = \theta_k + P_{k+1}\phi_k^{\mathrm{T}}(y_k - \phi_k\theta_k). \tag{11}$$

Note, that, for all $k \geq 1$,

$$\theta_k = \theta_{\mathrm{BLS},P_0^{-1},k}. \tag{12}$$

This implies any results for RLS and BLS are equivalent, where $R = P_0^{-1}$. However, the computational requirements of RLS are primarily determined by the inverse in (10), which is of size $p \times p$. When $p \ll n$, the computational burden of this inverse is much less demanding than the $n \times n$ inverse required by BLS. In addition, the storage requirements of RLS are of order n^2, which does not grow with k. Consequently, the computational and memory requirements of RLS are significantly less than those of BLS. For future analysis, we use the notation

$$\theta \triangleq \begin{bmatrix} \theta_{(1)} \\ \vdots \\ \theta_{(n)} \end{bmatrix}, \quad \theta_k \triangleq \begin{bmatrix} \theta_{k,(1)} \\ \vdots \\ \theta_{k,(n)} \end{bmatrix}. \tag{13}$$

IV. Regularization-Induced Bias

Note that $\theta_{\mathrm{BLS},R,k+1}$ given by (7) does not equal $\theta_{\mathrm{BLS},0,k+1}$ given by (8), and thus, the regularization R induces a bias in the estimate of θ. This bias decreases as more data are available, as demonstrated by the following example.

Example 1. For all $k \in \mathbb{N}_0$ let the components of $\theta \in \mathbb{R}^{10}$ and $\phi_k \in \mathbb{R}^{1\times 10}$ be sampled from the uniform distribution on $[-1, 1]$, and thus y_k is scalar. Let $R = P_0^{-1} = rI$, where r is a positive number and let $\theta_{0,(i)} = 0$ for $i = 1, \ldots, n$. For each value of r, 100 independent simulations are run, and, at each step k, the estimation error $\|\theta_k - \theta\|$ is averaged over the 100 simulations. Figure 1 shows the averaged estimation error with $k \in [0, 10^4]$. Note that $\|\theta_k - \theta\|$ sharply decreases around $k = 10$, and that the decrease is larger for smaller values of r. In addition, $\|\theta_k - \theta\|$ converges with the same linear log-log slope for all values of r. ◇

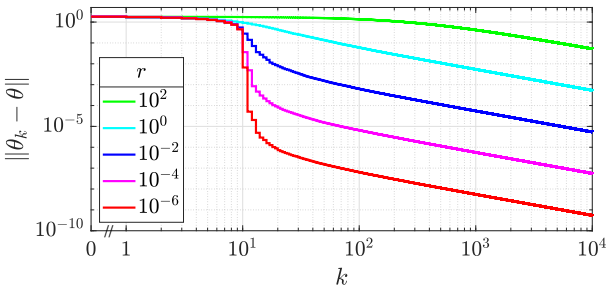

Fig. 1: Estimation error. The regularization is given by $R = rI$. For each value of r, 100 independent simulations are run where, for all $k \in \mathbb{N}_0$, each component of $\theta \in \mathbb{R}^{10}$ and $\phi_k \in \mathbb{R}^{1\times 10}$ sampled from the uniform distribution on $[-1, 1]$. The norm of the estimation error is averaged over the 100 trials for $k \in [0, 10^4]$.

V. Consistency of the Regularized Solution

The identities

$$P_{k+1}^{-1} = P_k^{-1} + \phi_k^{\mathrm{T}}\phi_k, \tag{14}$$
$$P_k^{-1} = P_0^{-1} + \sum_{i=0}^{k-1}\phi_i^{\mathrm{T}}\phi_i, \tag{15}$$
$$\theta_k - \theta = P_k P_0^{-1}(\theta_0 - \theta), \tag{16}$$
$$\theta_k = (I - P_k P_0^{-1})\theta + P_k P_0^{-1}\theta_0 \tag{17}$$

are useful. Proofs of these identities can be found in [6]. The next results follows.

Proposition 2. *Consider the notation and assumptions of Theorem 1. Then the following statements hold:*

i) For all $k \in \mathbb{N}_0$, $P_{k+1} \leq P_k$.
ii) $P_\infty \triangleq \lim_{k\to\infty} P_k$ *exists, and* $P_\infty \in \mathbf{N}^n$.
iii) $\theta_\infty \triangleq \lim_{k\to\infty} \theta_k$ *exists, and* $\theta_\infty = (I - P_\infty P_0^{-1})\theta + P_\infty P_0^{-1}\theta_0$.
iv) $\theta_\infty = \theta$ *if and only if* $P_\infty P_0^{-1}\theta = P_\infty P_0^{-1}\theta_0$.
v) If $P_\infty = 0$, *then* $\theta_\infty = \theta$.
vi) If, for all $\theta_0 \in \mathbb{R}^n$, $\theta_\infty = \theta$, *then* $P_\infty = 0$.

Proof.

i) Note (14) and that, for all $k \in \mathbb{N}_0$, $\phi_k^{\mathrm{T}}\phi_k \in \mathbf{N}^n$.
ii) Note that, for all $k \in \mathbb{N}_0$, $P_k \in \mathbf{N}^n$ and $P_{k+1} \leq P_k$.
iii) The result follows directly from (17).
iv) Substitute $\theta_\infty = \theta$ into the result of *iii)*.
v) Substitute $P_\infty = 0$ into the second equation of *iv)*.
vi) Let $\theta_0 \in \mathbb{R}^n$ and $P_\infty \neq 0$. Then, there exists $\theta \in \mathbb{R}^n$ such that $P_\infty P_0^{-1}\theta \neq P_\infty P_0^{-1}\theta_0$. So, by *iv)*, $\theta_\infty \neq \theta$.

□

A. Persistently Exciting Regressors

Definition 3. $(\phi_k)_{k=0}^\infty \subset \mathbb{R}^{p\times n}$ *is persistently exciting (PE) if there exist* $\alpha > 0$, $\beta > 0$, *and* $N \in \mathbb{N}_0$ *such that, for all* $j \in \mathbb{N}_0$,

$$\alpha I_n \leq \sum_{i=0}^{N} \phi_{i+j}^{\mathrm{T}}\phi_{i+j} \leq \beta I_n. \tag{18}$$

Proposition 4. $(\phi_k)_{k=0}^\infty \subset \mathbb{R}^{p\times n}$ *is PE if and only if*

$$C \triangleq \lim_{k\to\infty} \frac{1}{k}\Phi_k^{\mathrm{T}}\Phi_k \tag{19}$$

exists and is positive definite.

Proof. See page 64 of [11]. □

Lemma 1. *Let* $(A_k)_{k=0}^\infty \subset \mathbb{R}^{n\times n}$ *and assume that, for all* $k \in \mathbb{N}_0$, A_k *is nonsingular, and* $A \triangleq \lim_{k\to\infty} A_k$ *exists and is nonsingular. Then,* $\lim_{k\to\infty} A_k^{-1}$ *exists and* $\lim_{k\to\infty} A_k^{-1} = A^{-1}$.

Proof. This result follows from the continuity of the matrix inverse on the set of nonsingular matrices. □

Under persistent excitation, the following result describes the asymptotic rate of convergence of the regularized RLS solution.

Theorem 5. *Let* $\theta, \theta_0 \in \mathbb{R}^n$, *for all* $k \in \mathbb{N}_0$, *let* y_k *be given by* (2), *and let* P_k *and* θ_k *be given by* (10) *and* (11), *respectively. Assume that* $(\phi_k)_{k=0}^\infty$ *is PE, and define* C *by* (19). *Then,* $\lim_{k\to\infty} k(\theta_k - \theta) = C^{-1}R(\theta_0 - \theta)$.

Proof. Since

$$\begin{aligned}
\theta_k &= (\Phi_{k-1}^{\mathrm{T}}\Phi_{k-1} + R)^{-1}(\Phi_{k-1}^{\mathrm{T}}Y_{k-1} + R\theta_0)\\
&= (\Phi_{k-1}^{\mathrm{T}}\Phi_{k-1} + R)^{-1}(\Phi_{k-1}^{\mathrm{T}}\Phi_{k-1}\theta + R\theta - R\theta + R\theta_0)\\
&= (\Phi_{k-1}^{\mathrm{T}}\Phi_{k-1} + R)^{-1}((\Phi_{k-1}^{\mathrm{T}}\Phi_{k-1} + R)\theta + R(\theta_0 - \theta))\\
&= \theta + (\Phi_{k-1}^{\mathrm{T}}\Phi_{k-1} + R)^{-1}R(\theta_0 - \theta),
\end{aligned}$$

it follows that

$$\theta_k - \theta = (\Phi_{k-1}^{\mathrm{T}}\Phi_{k-1} + R)^{-1}R(\theta_0 - \theta). \tag{20}$$

Hence, by Lemma 1,

$$\begin{aligned}
\lim_{k\to\infty} k(\theta_k - \theta) &= \lim_{k\to\infty} (\frac{1}{k}\Phi_{k-1}^{\mathrm{T}}\Phi_{k-1} + \frac{1}{k}R)^{-1}R(\theta_0 - \theta)\\
&= C^{-1}R(\theta_0 - \theta).
\end{aligned}$$

□

Assume that $\theta_0 \neq \theta$ and that the assumptions of Theorem 5 hold. Then, $v \triangleq C^{-1}R(\theta_0 - \theta) \neq 0$. Therefore, for all $i = 1, \ldots, n$, $(\theta_{k,(i)} - \theta_{(i)}) = O(1/k)$ as $k \to \infty$. If, in addition, $v_{(i)} = 0$, then $(\theta_{k,(i)} - \theta_{(i)}) = o(1/k)$ as $k \to \infty$.

B. The Condition Number

Lemma 2. *If* $A \in \mathbf{P}^n$, *then, for all nonzero* $\alpha \in \mathbb{R}$, $\kappa(A) = \kappa(\alpha A)$.

Lemma 3. *Let* $(A_k)_{k=0}^\infty \subset \mathbf{P}^n$ *and assume that* $A \triangleq \lim_{k\to\infty} A_k$ *exists. Then* $\kappa(A_k) \to \kappa(A)$ *as* $k \to \infty$.

Proposition 6. *Under the assumptions and notation of Theorem 5,* $\lim_{k\to\infty} \kappa(\Phi_k^{\mathrm{T}}\Phi_k + R) = \kappa(C) \geq 1$.

Proof. Note that

$$\lim_{k\to\infty} \frac{1}{k}(\Phi_k^{\mathrm{T}}\Phi_k + R) = \lim_{k\to\infty} \frac{1}{k}\Phi_k^{\mathrm{T}}\Phi_k. \tag{21}$$

Next, by Lemma 2, for all $k \in \mathbb{N}_0$, $\kappa(\Phi_k^{\mathrm{T}}\Phi_k + R) = \kappa(\frac{1}{k}(\Phi_k^{\mathrm{T}}\Phi_k + R))$. Hence,

$$\lim_{k\to\infty} \kappa(\Phi_k^{\mathrm{T}}\Phi_k + R) = \lim_{k\to\infty} \kappa\left(\frac{1}{k}(\Phi_k^{\mathrm{T}}\Phi_k + R)\right). \tag{22}$$

Then, by Lemma 3,

$$\lim_{k\to\infty} \kappa(\Phi_k^{\mathrm{T}}\Phi_k + R) = \kappa\left(\lim_{k\to\infty} \frac{1}{k}(\Phi_k^{\mathrm{T}}\Phi_k + R)\right). \tag{23}$$

Finally, (21) implies

$$\lim_{k\to\infty} \kappa(\Phi_k^{\mathrm{T}}\Phi_k + R) = \kappa\left(\lim_{k\to\infty} \frac{1}{k}\Phi_k^{\mathrm{T}}\Phi_k\right) = \kappa(C) \geq 1. \ \square$$

The condition number of the regressor matrix can be used as a metric assess the convergence of the regularized RLS solution under persistent excitation, particularly when θ is unknown. Note that the convergence of the condition number of $\Phi_k^{\mathrm{T}}\Phi_k + R$ is necessary but not sufficient for the convergence of $\theta_k \to \theta$ as $k \to \infty$ under persistent excitation.

VI. Analysis of Consistency

This section presents numerical examples with diagnostics in order to investigate the properties the estimation error shown in Figure 1, namely, the sudden decrease in the estimation error and its asymptotic log-log slope. To do this, we apply RLS to (2), where the components of ϕ_k are randomly generated and y_k is given by (2).

Example 2. Let

$$\theta = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^{\mathrm{T}}, \tag{24}$$

and $\theta_0 = 0$. For all $k \in \mathbb{N}_0$, assume $\phi_k \in \mathbb{R}^{1\times 4}$ are i.i.d. and are randomly generated by

$$\phi_k^{\mathrm{T}} \sim \mathcal{N}\big(0, \operatorname{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2)\big), \tag{25}$$

where $\sigma_1^2 = 0.1$, $\sigma_2^2 = 1$, $\sigma_3^2 = 10$, and $\sigma_4^2 = 100$. Let $R = rI$, where $r = 10^{-5}$. Then,

$$C = \lim_{k\to\infty} \frac{1}{k}\Phi_k^{\mathrm{T}}\Phi = \operatorname{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2). \tag{26}$$

Hence, Theorem 5 implies that

$$\lim_{k\to\infty} k(\theta_k - \theta) = -r\begin{bmatrix}\sigma_1^{-2} & \sigma_2^{-2} & \sigma_3^{-2} & \sigma_4^{-2}\end{bmatrix}^{\mathrm{T}}. \tag{27}$$

Next, by Proposition 6,

$$\lim_{k\to\infty} \kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI) = \kappa(C) = \frac{\sigma_4^2}{\sigma_1^2} = 10^3. \tag{28}$$

RLS is run for 10 independent simulations and, at each step k, the estimation error $\|\theta_k - \theta\|$ is averaged over the 10 simulations. Figure 2 shows the error $|\theta_{(m),k} - \theta_{(m)}|$ for $m = 1, 2, 3, 4$ and the condition number $\kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI)$, as well as the asymptotic behaviors given by (27) and (28).

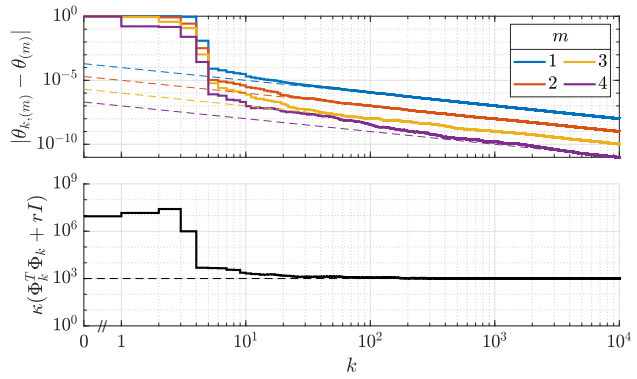

Fig. 2: Example 2: Estimation error (top) as well as condition number of $\Phi_k^{\mathrm{T}}\Phi_k + rI$ (bottom) for $k \in [0, 10^4]$, averaged over 10 independent trials. The regressor ϕ_k is sampled $\phi_k^{\mathrm{T}} \sim \mathcal{N}(0, \operatorname{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2))$, where $\sigma_1^2 = 0.1$, $\sigma_2^2 = 1$, $\sigma_3^2 = 10$, and $\sigma_4^2 = 100$. The true parameters are $\theta = [1\ 1\ 1\ 1]^{\mathrm{T}}$, and the regularization is $R = 10^{-5}I$. The dashed lines show the asymptotic behaviors of the estimation error and condition number given by (27) and (28) respectively.

For convenience, we define the *log slope* by

$$\Delta_{\log}(f_k) \triangleq \frac{\log(f_k) - \log(f_{k-1})}{\log(k) - \log(k-1)}, \tag{29}$$

which is independent of the logarithm base.

Proposition 7. *Let $\theta, \theta_k, \theta_{k-1} \in \mathbb{R}^n$, let $p \geq 1$, and assume there exists $c \in \mathbb{R}$ such that, for all $m = 1, \ldots n$,*

$$\Delta_{\log}(|\theta_{k,(m)} - \theta_{(m)}|) = c. \tag{30}$$

Then

$$\Delta_{\log}(\|\theta_k - \theta\|_p) = c. \tag{31}$$

Proof. (29) and (30) imply that, for all $m = 1, \ldots, n$,

$$\frac{|\theta_{k,(m)} - \theta_{(m)}|}{|\theta_{k-1,(m)} - \theta_{(m)}|} = q \triangleq \left(\frac{k}{k-1}\right)^c. \tag{32}$$

Next, note that,

$$\frac{\|\theta_k - \theta\|_p}{\|\theta_{k-1} - \theta\|_p} = \left(\frac{\sum_{m=1}^n |\theta_{k,(m)} - \theta_{(m)}|^p}{\sum_{m=1}^n |\theta_{k-1,(m)} - \theta_{(m)}|^p}\right)^{1/p}$$
$$= \left(\frac{\sum_{m=1}^n |\theta_{k-1,(m)} - \theta_{(m)}|^p q^p}{\sum_{m=1}^n |\theta_{k-1,(m)} - \theta_{(m)}|^p}\right)^{1/p} = (q^p)^{1/p} = q.$$

Thus, (29) yields

$$\Delta_{\log}(\|\theta_k - \theta\|_p) = \frac{\log\left(\frac{\|\theta_k-\theta\|_p}{\|\theta_{k-1}-\theta\|_p}\right)}{\log\left(\frac{k}{k-1}\right)} = \frac{c\log\left(\frac{k}{k-1}\right)}{\log\left(\frac{k}{k-1}\right)} = c.$$

□

Note that (27) implies that, for all $m = 1, \ldots, n$, $\Delta_{\log}(|\theta_{k,(m)} - \theta_{(m)}|) \to -1$ as $k \to \infty$. Hence, by Proposition 7, $\Delta_{\log}\|\theta_k - \theta\| \to -1$ as $k \to \infty$. Therefore, Figure 3 and all subsequent plots of log slope show $\Delta_{\log}\|\theta_k - \theta\|$. Note that, in Figure 3, $\Delta_{\log}\|\theta_k - \theta\|$ is also averaged over the 10 independent trials. Moreover, since we are concerned with the asymptotic log slope of the estimation error, we plot $\Delta_{\log}\|\theta_k - \theta\|$ for $9000 \leq k \leq 10000$. As shown in Figure 3, the 100-step moving average of $\Delta_{\log}\|\theta_k - \theta\|$ is approximately equal to -1. Furthermore, note that $\kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI)$ in Figure 2 approaches 10^3. The change in condition number $\kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI)$ for $k \in [9000, 10000]$ is shown in Figure 3, where

$$\Delta(\kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI)) \triangleq \kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI) - \kappa(\Phi_{k-1}^{\mathrm{T}}\Phi_{k-1} + rI). \tag{33}$$

◇

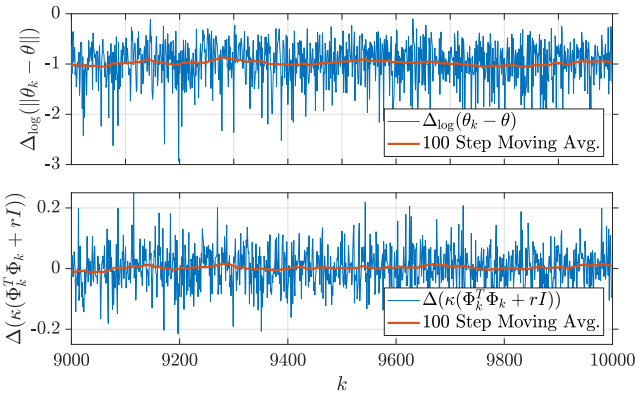

Fig. 3: Example 2: Log slope of error (top) and $\Delta(\kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI))$ (bottom) for $k \in [9000, 10000]$, averaged over 10 independent trials. The regressor ϕ_k is sampled $\phi_k^{\mathrm{T}} \sim \mathcal{N}(0, \operatorname{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2))$, where $\sigma_1^2 = 0.1$, $\sigma_2^2 = 1$, $\sigma_3^2 = 10$, and $\sigma_4^2 = 100$. The 100-step moving averages show that $\Delta_{\log}(\|\theta_k - \theta\|) \to -1$ and $\kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI) \to 10^3$ as $k \to \infty$.

VII. Application to FIR System Identification

Let a finite impulse response (FIR) be given by

$$y_k = \sum_{i=1}^{w} G_i u_{k-i}, \tag{34}$$

for $k = 0, 1, 2, \ldots$, where w is the window length, $u_k \in \mathbb{R}^q$ is the input, $y_k \in \mathbb{R}^p$ output, and $G_i \in \mathbb{R}^{p\times q}$ are the model coefficients. Note that (34) can be expressed as (2), where

$$\phi_k \triangleq \begin{bmatrix} u_{k-1}^{\mathrm{T}} & \ldots & u_{k-w}^{\mathrm{T}} \end{bmatrix} \otimes I_p \in \mathbb{R}^{p\times n}, \tag{35}$$

$$\theta \triangleq \operatorname{vec}\begin{bmatrix} G_1 & \ldots & G_w \end{bmatrix} \in \mathbb{R}^n. \tag{36}$$

To estimate θ, define

$$\theta_k \triangleq \operatorname{vec}\begin{bmatrix} G_{1,k} & \ldots & G_{w,k} \end{bmatrix} \in \mathbb{R}^n, \tag{37}$$

where $G_{i,k} \in \mathbb{R}^{p\times m}$ are the estimated model coefficients and $n = wpq$. Note that $u_k \triangleq 0$ for all $k < 0$.

Example 3. *FIR model* Consider the FIR model

$$y_k = -1.5u_{k-1} + 0.9u_{k-2} + 0.15u_{k-3} - 0.15u_{k-4}, \tag{38}$$

for all $k \in \mathbb{N}_0$, where θ is defined by (36). For all $k \in \mathbb{N}_0$, the inputs $u_k \in \mathbb{R}$ are i.i.d and are sampled from the uniform distribution on $[-1, 1]$. RLS is applied to estimate θ with regularization $R = rI$ where $r = 10^{-5}$ and $\theta_0 = 0$. Figure 4 shows the error $|\theta_{(m),k} - \theta_{(m)}|$ for $m = 1, 2, 3, 4$ as well as the condition number $\kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI)$.

Note that the error is $O(10^{-1})$ at step $k = m$, but decreases to $O(10^{-5})$ at step $k = m + 1$, for all $m = 1, 2, 3, 4$, as seen in Figure 4. Additionally, Figure 4 shows that the condition number of $\Phi_k^{\mathrm{T}}\Phi_k + rI$ drops below 10^2 at $k = 4$, the step at which $\Phi_k^{\mathrm{T}}\Phi_k$ attains full rank. Similar to the results of Example 2, Figure 5 shows that $\Delta_{\log}(\|\theta_k - \theta\|)$ approaches -1 and $\kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI) \approx 1$ asymptotically. $\diamond$

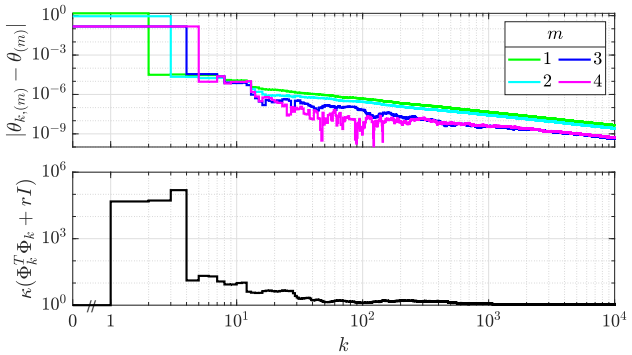

Fig. 4: Example 3: Estimation error (top) and condition number of $\Phi_k^{\mathrm{T}}\Phi_k = rI$ (bottom) for FIR system identification with steps $k \in [0, 10^4]$. For all $k \in \mathbb{N}_0$, the input u_k is sampled from the uniform distribution on $[-1, 1]$. The true parameters are $\theta = [-1.5\ 0.9\ 0.15\ \ -0.15]^{\mathrm{T}}$, and the regularization is $R = 10^{-5}I$.

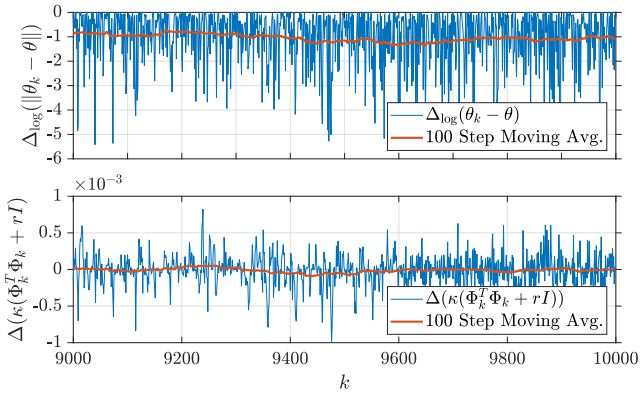

Fig. 5: Example 3: Log slope of error (top) and $\Delta(\kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI))$ (bottom) for FIR system identification with $k \in [9000, 10000]$. For all $k \in \mathbb{N}_0$, the input u_k is sampled from the uniform distribution on $[-1, 1]$. The 100-step moving averages show that $\Delta_{\log}(\|\theta_k - \theta\|)$ approaches -1 and $\kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI) \approx 1$ as $k \to \infty$.

VIII. Application to IIR System Identification

Let an infinite impulse response (IIR) be given by

$$y_k = -\sum_{i=1}^{w} F_i y_{k-i} + \sum_{i=1}^{w} G_i u_{k-i}, \tag{39}$$

for $k = 0, 1, 2, \ldots$, where w is the model window length, $u_k \in \mathbb{R}^q$ is the input, $y_k \in \mathbb{R}^p$ is the output, and $F_i \in \mathbb{R}^{p\times p}$ and $G_i \in \mathbb{R}^{p\times q}$ are the model coefficients. Note that (39) can be expressed as (2), where

$$\theta \triangleq \operatorname{vec}\begin{bmatrix} F_1 & \ldots & F_w & G_1 & \ldots & G_w \end{bmatrix} \in \mathbb{R}^n, \tag{40}$$

$$\phi_k \triangleq \begin{bmatrix} -y_{k-1}^{\mathrm{T}} \ldots -y_{k-w}^{\mathrm{T}} & u_{k-1}^{\mathrm{T}} \ldots u_{k-w}^{\mathrm{T}} \end{bmatrix} \otimes I_p \in \mathbb{R}^{p\times n}. \tag{41}$$

To estimate θ, define

$$\theta_k \triangleq \operatorname{vec}\begin{bmatrix} F_{1,k} & \ldots & F_{w,k} & G_{1,k} & \ldots & G_{w,k} \end{bmatrix} \in \mathbb{R}^n, \tag{42}$$

$$\tag{43}$$

where $F_{i,k} \in \mathbb{R}^{p\times p}$ and $G_{i,k} \in \mathbb{R}^{p\times q}$ are the estimated model coefficients and $n = wp(p+q)$. Note that $y_k \triangleq 0$ and $u_k \triangleq 0$ for all $k < 0$.

Example 4. *IIR model* Consider the IIR model

$$y_k = 1.5y_{k-1} - 0.9y_{k-2} + 0.15u_{k-1} - 0.15u_{k-2}, \tag{44}$$

for all $k \in \mathbb{N}_0$, where θ is defined by (40). For all $k \in \mathbb{N}_0$, the inputs $u_k \in \mathbb{R}$ are i.i.d. and are sampled from the uniform distribution on $[-1, 1]$. RLS is applied to estimate θ with regularization $R = rI$ where $r = 10^{-5}$ and $\theta_0 = 0$. Figure 6 shows the error $|\theta_{(m),k} - \theta_{(m)}|$ for $m = 1, 2, 3, 4$ as well as the condition number $\kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI)$.

Note that $|\theta_{k,(m)} - \theta_{(m)}|$ decreases approximately four orders of magnitude from step $k = 1$ to $k = 2$ for $m = 3$. Additionally, $|\theta_{k,(m)} - \theta_{(m)}|$ further decreases approximately three orders of magnitude from steps $k = 5$ to $k = 6$ for $m = 1, 2, 4$.

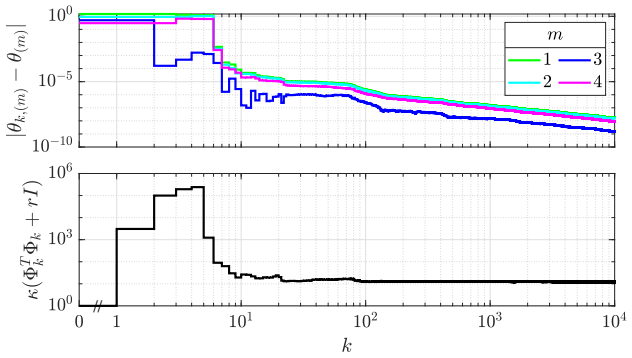

Fig. 6: Example 4: Estimation error (top) and condition number of $\Phi_k^{\mathrm{T}}\Phi_k = rI$ (bottom) for IIR system identification with steps $k \in [0, 10^4]$. For all $k \in \mathbb{N}_0$, the input u_k is sampled from the uniform distribution on $[-1, 1]$. The true parameters are $\theta = [-1.5\ 0.9\ 0.15\ \ -0.15]^{\mathrm{T}}$, and the regularization is $R = 10^{-5}I$.

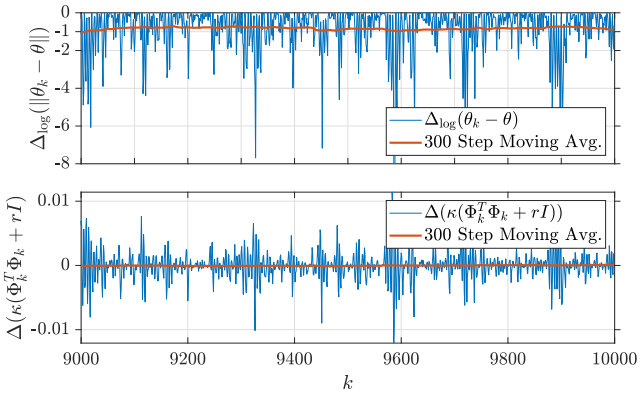

Fig. 7: Example 4: Log slope of error (top) and $\Delta(\kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI))$ (bottom) for IIR system identification with $k \in [9000, 10000]$. For all $k \in \mathbb{N}_0$, the input u_k is sampled from the uniform distribution on $[-1, 1]$. The 100-step moving averages show that $\Delta_{\log}(\|\theta_k - \theta\|)$ approaches -1 and $\kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI) \approx 10$ as $k \to \infty$.

Note that, for $k > 5$, the matrix $\Phi_k^{\mathrm{T}}\Phi_k + rI$ is ill-conditioned, with $\kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI) \approx 10$ for all $k > 100$. Yet, this ill-conditioned $(\Phi_k^{\mathrm{T}}\Phi_k + rI)$ for $k > 5$ does not affect the asymptotic behavior of $|\theta_{k,(m)} - \theta_{(m)}|$ for all values of m. In fact, Figure 7 shows that $\Delta_{\log}(\|\theta_k - \theta\|)$ approaches -1 and $\kappa(\Phi_k^{\mathrm{T}}\Phi_k + rI) \approx 10$ asymptotically.

$\diamond$

IX. Conclusions and Future Work

In optimization-based parameter estimation, regularization compensates for the lack of persistency, especially during startup. This paper examined regularization-induced bias, which refers to bias in the parameter estimates due to regularization. This paper showed that, under persistency, the parameter estimates improve at precisely the step where the regressor becomes square and nonsingular. It was also shown that the regularization-induced bias asymptotically decreases with a log slope of approximately -1. Both of these effects were connected to the condition number of the regressor.

For system identification, RLS was applied to FIR and IIR systems. For FIR systems, it was shown that the components of the estimation error decrease sequentially rather than simultaneously. However, for IIR systems, the components of the estimation error decrease less predictably. Future research will relate these trends to the condition number of the regressor.

Finally, one potential approach to overcoming regularization-induced bias is to use variable regularization [13]. By varying the regularization based on the condition number of the regressor, it may be possible to reduce the regularization-induced bias. The challenge is to vary the regularization in a computationally efficient manner without impacting the numerical stability of RLS.

References

[1] A. Albert and R. W. Sittler, "A Method for Computing Least Squares Estimators that Keep Up with the Data," *SIAM J. Contr.*, vol. 3, no. 3, pp. 384–417, 1965.

[2] K. J. Astrom, *Adaptive Control*, 2nd ed. Reading, MA: Addison-Wesley, 1995.

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

[4] S. A. U. Islam and D. S. Bernstein, "Recursive Least Squares for Real-Time Implementation," *IEEE Cont. Sys. Mag.*, vol. 39, no. 3, pp. 82–85, 2019.

[5] A. L. Bruce, A. Goel, and D. S. Bernstein, "Recursive Least Squares with Matrix Forgetting," in *Proc. Amer. Contr. Conf.*, July 2020.

[6] A. Goel, A. L. Bruce, and D. S. Bernstein, "Recursive least squares with variable-direction forgetting–compensating for the loss of persistency," *IEEE Contr. Sys. Mag.*, vol. 40, pp. 80–102, August 2020.

[7] P. C. Hansen and D. P. O'Leary, "The Use of the L-Curve in the Regularization of Discrete Ill-Posed Problems," *SIAM J. Sci. Comp.*, vol. 14, no. 6, pp. 1487–1503, 1993.

[8] G. H. Golub, P. C. Hansen, and D. P. O'Leary, "Tikhonov regularization and total least squares," *SIAM J. Matrix Anal. Appl.*, vol. 21, pp. 185–194, 1999.

[9] F. Cucker and S. Smale, "Best Choices for Regularization Parameters in Learning Theory: On the Bias-Variance Problem," *Found. Comp. Math.*, vol. 2, pp. 413–428, 2002.

[10] S. Lu, S. V. Pereverzev, and U. Tautenhahn, "Regularized Total Least Squares: Computational Aspects and Error Bounds," *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 3, pp. 918–941, 2010.

[11] K. J. Astrom and B. Wittenmark, *Adaptive Control*, 2nd ed. USA: Addison-Wesley Longman Publishing Co., Inc., 1994.

[12] A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar, "Estimation with norm regularization," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[13] A. A. Ali, J. B. Hoagg, M. Mossberg, and D. S. Bernstein, "On the stability and convergence of a sliding-window variable-regularization recursive-least-squares algorithm," *Int. J. Adapt. Control Signal Process.*, vol. 30, pp. 715–735, 2016.