SCIENCE CHINA
Information Sciences

July 2023, Vol. 66 170205:1–170205:2
https://doi.org/10.1007/s11432-022-3725-7

• LETTER •
Special Topic: Entanglement of Control and Optimization

On the reinforcement learning extended state observer for a class of uncertain sampled-data control systems

Guojie TANG[1,2], Wenchao XUE[1,2*], Haitao FANG[1,2] & Kun ZHANG[3,1]

[1]*LSC, NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing* 100190*, China;*
[2]*School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing* 100049*, China;*
[3]*School of Astronautics, Beihang University, Beijing* 100191*, China*

Received 19 January 2022/Revised 26 October 2022/Accepted 17 February 2023/Published online 20 June 2023

Citation Tang G J, Xue W C, Fang H T, et al. On the reinforcement learning extended state observer for a class of uncertain sampled-data control systems. Sci China Inf Sci, 2023, 66(7): 170205, https://doi.org/10.1007/s11432-022-3725-7

In the last few decades, the extended state observer (ESO) [1] has been demonstrated to be an effective tool for dealing with uncertain control systems and many modified ESOs have been proposed to get the desired estimation performance [2–4]. Nevertheless, most of these methods assume that part of the model information is already known. Therefore, a data-driven tuning law for ESO without assuming model information for disturbances and noise seems to be a significant open issue.

Reinforcement learning (RL) is a crucial part of machine learning [5] that aims to search for a policy under which the agent can maximize the cumulative rewards. Thus, the RL approach is an appealing tool for a learning-based tuning law for ESO.

Furthermore, since the output measurements are sample-data, it is straightforward to develop a discrete ESO based on the discretized system model and tune the gains of the ESO online through RL. Consequently, the reinforcement learning ESO (RLESO) will be discrete and time-varying, and the existing findings, almost all of which are for continuous-time constant gain ESO, cannot be applied. It is inferred that the discrete-time RLESO design with guaranteed stability is a critical and challenging issue to be examined.

In this study, RLESO, whose gains can be optimized online with a data-driven mechanism, is proposed. The major contributions are threefold:

(i) The framework of RLESO is proposed for a class of sampled-data control systems under unknown dynamics and stochastic noise.

(ii) The stability of RLESO is assessed and the quantitative conditions to guarantee the boundedness of the RLESO's estimation error using mean square are introduced.

(iii) It is proven that convergence to zero of the estimation error can be guaranteed if the noise's variance and the disturbance's higher-order derivatives approach zero as time goes to infinity.

Problem formulation. Consider the class of uncertain sampled-data systems with continuous dynamics,

$$\begin{cases} \dot{x}(t) = Ax(t) + B(u(t) + d_1(x(t), t)), \\ y(kh) = x_1(kh) + v(kh), \\ u(t) = u(kh), \\ t \in [kh, kh + h), \quad k \geqslant 0, \end{cases} \tag{1}$$

where t is the time, $x(t) = [x_1(t), x_2(t), \ldots, x_n(t)]^{\mathrm{T}}$ denotes the system state vector, $u(t) \in \mathbb{R}$ depicts the system input, $d_1(x(t), t) \in \mathbb{R}$ comprises both unknown dynamics and external disturbances, h is the sampling period, $k \in \mathbb{N}$ is the discrete-time index, $y(kh)$ is the system output, and $v(kh)$ is the measurement of stochastic noise at the k-th sampling time. A and B are defined as

$$A = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Specifically, this study assumes that $d_1(t)$ satisfies [6, 7] Assumption 1.

Assumption 1. For $t \in [iph, (i+1)ph), i = 0, 1, \ldots,$ we have

$$\dot{d}_1(t) = d_2(t), \ldots, \dot{d}_{m-1}(t) = d_m(t), \dot{d}_m(t) = c_i + \sigma(t), \tag{2}$$

where $\{c_i\}_{i=0}^{\infty}$ is a stochastic sequence satisfying

$$R_i^c \triangleq E[{c_i}^2] < R^c \quad \forall i, \quad E[c_i c_j] = 0 \quad \forall i \neq j. \tag{3}$$

$\{\sigma(t)\}_{t>0}$ is a stochastic process satisfying

$$R_k^\sigma \triangleq E\left[\int_{kh}^{(k+1)h} \sigma^2(s)\mathrm{d}s\right] < R^\sigma \quad \forall k, \tag{4}$$

* Corresponding author (email: wenchaoxue@amss.ac.cn)

where R^c and R^σ are positive constants. More explanations for Assumption 1 are given in Appendix B.

A popular assumption regarding measurement noise $\{v(kh)\}_{k=1}^{\infty}$ is explained below [4].

Assumption 2.

$$R_k^v \triangleq E[v(kh)^2] < R^v \quad \forall k, \tag{5}$$

where R^v is positive.

After defining $z = [x, d_1, {d_1}^{(1)}, \ldots, {d_1}^{(m-1)}]^{\mathrm{T}}$, the linear ESO is designed as (6) to estimate z,

$$\begin{bmatrix} \hat{z}((k+1)h) \\ \hat{c}((k+1)h) \end{bmatrix} = \begin{bmatrix} A_z & B_d \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{z}(kh) \\ \hat{c}(kh) \end{bmatrix} + \begin{bmatrix} B_x \\ 0 \end{bmatrix} u(kh) + h\beta_k(y(kh) - \hat{z}_1(kh)), \tag{6}$$

where $\beta_k = [\beta_{k,1}, \ldots, \beta_{k,m+n+1}]^{\mathrm{T}}$ is the gain vector. Definitions for $\hat{z}$, $\hat{c}$, A_z, B_d, and B_x are given in Appendix C.

Framework of RLESO. The value of ESO is calculated at the end of each sample period, but the RL tuning is performed for every q sample periods, that is, the gains of ESO are updated every q samples. More details on the RLESO framework are provided in Appendix D.

The three basic elements of the RL algorithm are state, action, and reward. The design procedures for these elements are presented below.

State. State $s_j = [s_{j,1}, s_{j,2}]$, where

$$\begin{cases} s_{j,1} = \beta_{(j-1)q}, \\ s_{j,2} = \dfrac{1}{q}\displaystyle\sum_{i=0}^{q-1} |y((j-1)qh + ih) - \hat{x}_1((j-1)qh + ih)|. \end{cases} \tag{7}$$

Action. The action $a_j = [a_{j,1}, \ldots, a_{j,m+n+1}]$ is chosen in $\Lambda_1 \times \cdots \times \Lambda_{m+n+1}$ and $\{\Lambda_i\}_{i=1}^{n+m+1}$ is a real number set with finite elements, in which the lower and upper bounds are represented as $\underline{a}_i$ and $\overline{a}_i$, respectively.

After the action is identified, the gains will be updated at $k = jq, j = 1, 2, \ldots, i = 1, 2, \ldots, m+n+1$,

$$\beta_{jq,i} = \begin{cases} \underline{\beta}_i, & \beta_{(j-1)q,i} + a_{j,i} < \underline{\beta}_i, \\ \beta_{(j-1)q,i} + a_{j,i}, & \underline{\beta}_i \leqslant \beta_{(j-1)q,i} + a_{j,i} \leqslant \overline{\beta}_i, \\ \overline{\beta}_i, & \beta_{(j-1)q,i} + a_{j,i} > \overline{\beta}_i, \end{cases} \tag{8}$$

where $\underline{\beta}_i$ and $\overline{\beta}_i$ are the predetermined bounds of $\beta_{k,i}$. In rest of the cases, β_k remains the same, namely,

$$\beta_k = \beta_{k-1}, \quad \mathrm{mod}(k, q) \neq 0. \tag{9}$$

Reward. The reward function is set as

$$r_j = \begin{cases} -s_{j,2}/\bar{s}, & \text{if} \quad s_{j,2} \leqslant \bar{s}, \\ -r_p, & \text{if} \quad s_{j,2} > \bar{s}, \end{cases} \tag{10}$$

where $\overline{s}$ is a pre-set threshold value of the estimation error and $r_p > 1$ is a penalty term.

Stability and convergence analysis for RLESO. Provided $e_{x_i} = x_i - \hat{x}_i, i = 1, \ldots, n$, $e_{d_i} = d_i - \hat{d}_i, i = 1, \ldots, m$, and $e_c = c - \hat{c}$, the corresponding dynamic equation of estimation error can be written as

$$e_{k+1} = A_k e_k + \xi_{k+1}, \tag{11}$$

where $e_k = [e_{x_1}, \ldots, e_{x_n}, e_{d_1}, \ldots, e_{d_m}, e_c]^{\mathrm{T}}(kh)$ and

$$A_k = \begin{bmatrix} 1 - h\beta_{k,1} & h & \cdots & \dfrac{h^{m+n}}{(m+n)!} \\ \vdots & \vdots & \ddots & \vdots \\ -h\beta_{k,m+n} & 0 & \cdots & h \\ -h\beta_{k,m+n+1} & 0 & \cdots & 1 \end{bmatrix}. \tag{12}$$

The definitions of the mathematical notation are listed in Appendix C.

Assumption 3. h, $\{\Lambda_i\}_{i=1}^{n+m}$, $\{\underline{\beta}_i\}_{i=1}^{m+m}$, and $\{\overline{\beta}_i\}_{i=1}^{m+m}$ are set to make the eigenvalues of A_k are all in the unit circle $\forall k$.

Theorem 1. Consider the sampled-data system (1) and the RLESO with Assumptions 1 and 2. Let the action set satisfy

$$2P^2\widetilde{A}\sqrt{\sum_{i=1}^{m+n+1} \max\{\underline{a}_i{}^2, \overline{a}_i{}^2\}} < 1, \tag{13}$$

where

$$\widetilde{A} = \max_k \sqrt{\mathrm{tr}(A_k^{\mathrm{T}} A_k)}, P = 1 + \frac{\widetilde{A}^2 - \widetilde{A}^{2N}}{1 - \widetilde{A}^2} + \frac{(\bar{\rho} + \zeta)^{2(N+1)}}{1 - (\bar{\rho} + \zeta)^2}. \tag{14}$$

Then

$$\sup_k \|e_k\|_{L_2} < \infty.$$

Theorem 2. Consider the sampled-data control system (1) and the RLESO (6) with Assumptions 1 and 2. Following the same parameter settings in Theorem 1, if the uncertain dynamics and measurement noise satisfy

$$\lim_{k\to\infty} \left(R_k^c{}^2 + R_k^v{}^2 + R_k^\sigma{}^2\right) = 0, \tag{15}$$

then

$$\lim_{k\to\infty} \|e_k\|_{L_2} = 0.$$

The proofs of Theorems 1 and 2 are provided in Appendix E.

Conclusion. This study proposes the RLESO to optimize the performance of the ESO. The framework of RLESO has been established, and mean square boundedness, as well as convergence of the estimation error, has been examined.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62122083, 61973299, 62103408) and Youth Innovation Promotion Association CAS.

Supporting information Appendixes A–F. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Han J. The "extended state observer" of a class of uncertain systems. Control and Decision, 1995, 1: 85–88
2 Wei W, Xue W, Li D. On disturbance rejection in magnetic levitation. Control Eng Pract, 2019, 82: 24–35
3 Liu C, Luo G, Duan X, et al. Adaptive LADRC-based disturbance rejection method for electromechanical Servo system. IEEE Trans Ind Applicat, 2020, 56: 876–889
4 Bai W Y, Xue W C, Huang Y, et al. On extended state based Kalman filter design for a class of nonlinear time-varying uncertain systems. Sci China Inf Sci, 2018, 61: 042201
5 Sutton R, Barto A. Reinforcement Learning: An Introduction. Cambridge: MIT Press, 2018
6 Shi J, Chen X, Yau S S T. High order linear extended state observer and error analysis of active disturbance rejection control. Asian J Math, 2019, 23: 631–650
7 Shao X, Wang H. Performance analysis on linear extended state observer and its extension case with higher extended order (in Chinese). Control and Decision, 2015, 30: 815–822