SCIENCE CHINA
Information Sciences

• **Supplementary File** •

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.*

Appendix A Motivation of the work

In the last decades, estimating and negating uncertain dynamics as well as disturbances has been shown to be an effective approach in dealing with uncertain control systems. The fundamental issue of such approach is to design observers/estimators for uncertain dynamics or disturbances. As is well known, many popular observers have been developed for different kinds of systems, including uknown input observer (UIO) [1], disturbance observer based control (DOBC) [2, 3], extended state observer (ESO) [4], etc. ESO is designed to take the external disturbances and internal uncertainties as an extend state "total disturbance" and estimate it in real time together with system states. It has been proven that linear ESO (LESO) can deal with nonlinear unknown dynamics and piecewise disturbances [5]. In several industrial sectors, ESO and ESO based control had been applied widely, such as two-mass actuator systems [6], dc motor system [7], and spacecraft systems [8], as well as incipient fault diagnosis [9].

In the past years, substantial developments have been made for the continuous-time ESO for both stability and convergence, including both linear ESO [10, 11], nonlinear ESO (NLESO) [12] and time-varying ESO [13]. Aiming to get desired estimation performance, many modified ESOs with time-varying gains have been proposed. Liu et al. [14] combined LESO with an adaptive approach and applied it to electro-mechanical servo system. But the optimization for ESO's parameters under measurement noise is not considered. Wei et al. [13] designed LESO with decreasing gains to reduce the influence of measurement noise in system steady state. But the tuning law of gains is predetermined, which means it cannot adjust for disturbances outside the model. There are analogous attempts in [15] and [16]. However, most of these papers are based on systems with continuous input and output. On the contrary, in application, the system control input and output are mostly sampled-data. Therefore, it is necessary to explore how to design discrete-time ESO.

Compared with continuous ESO, there is still little relevant research about the optimality and stability of discrete-time ESO in the existing literatures. In [17], the stability of a third order discrete-time LESO, which has third order dynamics and was used to estimate the states and the "total disturbance" of a 2 order plant, was analyzed and the effect of ESO gains and sampling period on stability is revealed. Li et al. [18] proposed a design method of NLESO over discrete domain and obtained a sufficient condition to guarantee the absolutely stability of the estimation error system. Combined with the Kalman filter, [19] proposed extended state based Kalman filter (ESKF) to weaken the influence of noises and proved its stability. However, the upper bound on the variance of noise is needed to be the prior information. Thus, data-driven based tuning law for ESO without assuming model information of disturbances and noises seems to be an important open issue.

Reinforcement learning (RL) is an important segment of machine learning [20]. With the rise of reinforcement learning, many researchers try to apply it to the design of controllers and achieve great success [21–26]. In a general reinforcement learning problem, an agent in a certain state has several actions to choose and receive the next state and reward after interacting with the environment. The objective of reinforcement learning is to search a policy under which the agent can maximize the cumulative rewards. Thus, reinforcement learning approach seems to be an appealing tool for learning based tuning law for ESO. On the other hand, due to the unmeasurable property of the uncertain dynamics and disturbances, the principles of choosing states, reward and actions in reinforcement learning algorithm for ESO are not obvious. Additionally, since the output measurements are sample-data, it is straightforward to design a discrete ESO based on the discretized system model and tune the gains of ESO on-line through RL. As a result, the reinforcement learning ESO (RLESO) will be discrete and time-varying. However, the existing results, almost all of which are for continuous-time constant gain ESO, cannot be applied. Thus, RLESO design with guaranteed stability is an important and challenging issue to be studied.

Appendix B Explanations for Assumptions 1

Notice that $d_1(x(t),t)$ is used to model the dynamic or signal in physical world, then it is reasonable to assume $d_1(x(t),t)$ to be a m-times differentiable function and d_{m-1} can be approximated by a piece-wise linear function, which means $\dot{d}_m \triangleq \frac{\mathrm{d}^m}{\mathrm{d}t^m}(d_1)$ can be approximated by a constant c_i in every p sampling intervals. The $\sigma(t)$ can be seen as the approximation error. In the existing papers, it is popular to assume that $d^{(m)}$ is bounded [27]. The bounded assumptions of the variance and the integral of the approximation error can be regarded as the equivalent form of the bounded assumption of $d^{(m)}$ in the discrete case.

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

Appendix C Mathematical notations

Appendix C.1 Exact discrete model

Since the exact information of $x(kh)$ and $d(x(kh),kh)$ cannot be available in practice, it is necessary to estimate $x(kh)$ and $d(x(kh),kh)$ in real time. In the past years, many effective observer/estimator design methods have been proposed. In this paper, we will consider the discrete-time ESO design for the hybrid system (1), and adopt a reinforcement learning method for tuning the parameters of ESO. Define $z=[x,d]$ and combine (1) with (3) , we have

$$\dot{z}=A_1z+B_1u+B_2(c_i+\sigma), \tag{C1}$$

where

$$A_1=\begin{bmatrix}0&1&\cdots&\cdots&0\\0&0&1&\cdots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\cdots&1\\0&0&0&\cdots&0\end{bmatrix}_{(n+m)\times(n+m)},\quad B_1=\begin{bmatrix}0_{(n-1)\times1}\\1\\0_{m\times1}\end{bmatrix},\quad B_2=\begin{bmatrix}0_{(n+m-1)\times1}\\1\end{bmatrix}.$$

For $\forall t\in[t_0,t_0+h]$, the solution of (C1) is

$$z(t)=e^{A_1(t-t_0)}z(t_0)+\int_{t_0}^{t}e^{A_1(t-\tau)}B_1u(\tau)d\tau+\int_{t_0}^{t}e^{A_1(t-\tau)}B_2(c_i+\sigma)d\tau.$$

Let $t=(k+1)h$ and $t_0=kh$, then we have

$$z((k+1)h)=A_zz(kh)+B_xu(kh)+B_dc(kh)+l(kh), \tag{C2}$$

where

$$\begin{aligned}A_z&=\begin{bmatrix}1&h&\frac{h^2}{2!}&\cdots&\frac{h^{m+n-1}}{(m+n-1)!}\\0&1&h&\cdots&\frac{h^{m+n-2}}{(m+n-2)!}\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\cdots&h\\0&0&0&\cdots&1\end{bmatrix},\quad B_x=\begin{bmatrix}\frac{h^n}{n!}\\\vdots\\\frac{h^2}{2!}\\h\\0_{m\times1}\end{bmatrix},\quad B_d=\begin{bmatrix}\frac{h^{m+n}}{(m+n)!}\\\vdots\\\frac{h^2}{2!}\\h\end{bmatrix},\\l(kh)&=\begin{bmatrix}\int_{kh}^{(k+1)h}\frac{((k+1)h-\tau)^{m+n-1}}{(m+n-1)!}\sigma(\tau)d\tau\\\vdots\\\int_{kh}^{(k+1)h}((k+1)h-\tau)\sigma(\tau)d\tau\\\int_{kh}^{(k+1)h}\sigma(\tau)d\tau\end{bmatrix}.\end{aligned} \tag{C3}$$

Appendix C.2 Notation in equation (12)

$$A_k=\begin{bmatrix}1-h\beta_{k,1}&h&\cdots&\frac{h^{m+n}}{(m+n)!}\\\vdots&\vdots&\ddots&\vdots\\-h\beta_{k,m+n}&0&\cdots&h\\-h\beta_{k,m+n+1}&0&\cdots&1\end{bmatrix}, \tag{C4}$$

$$\xi_{k+1}=\bar{B}_kv(kh)+\delta_k,\quad \bar{B}_k=-h\begin{bmatrix}\beta_1(kh)\\\beta_2(kh)\\\vdots\\\beta_{m+n+1}(kh)\end{bmatrix},\quad \delta_k=\begin{bmatrix}l(kh)\\c((k+1)h)-c(kh)\end{bmatrix}. \tag{C5}$$

Appendix D More details for the design of RLESO

Taking the jth to $(j+2)$th tunings as an example, the working mechanism of RLESO is shown in Fig. C1.

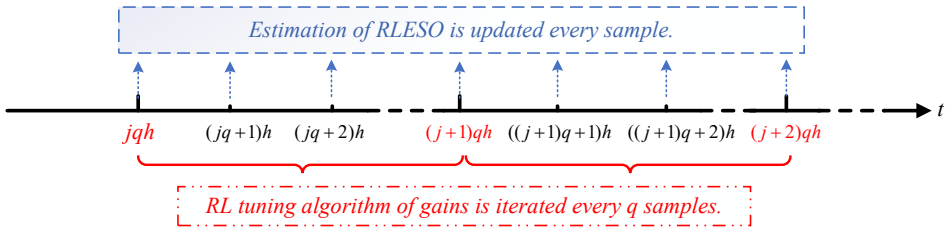

Figure D1 The working mechanism of RLESO from $t = jqh$ to $t = (j+2)qh$

Remark 1. The current ESO gains and the estimation errors are the two elements that reflect the current working condition of ESO. However, the estimation error of the total disturbance cannot be accessed because the total disturbance is unknown. In our experience, when ESO's estimation error of the total disturbance is small, its estimation error of x_1 is also small, and vice versa. This fact motivate us to replace the estimation error of the total disturbance with the estimation error of x_1, which is a key step in designing RLESO. Hence, the state contains two elements: i) The current observer gains, i.e., $s_{j,1}$. ii) The average of the cumulative estimation error for x_1 over the past q sampling periods, i.e., $s_{j,2}$.

Remark 2. For general reinforcement learning algorithms, the reward function r_j is the most important factor to be designed. As discussed in Remark 1, this paper uses the estimation error of x_1 to reflect the estimation effect of ESO. Naturally, $s_{j,2}$ is used to design the reward function. Let $\bar{s}$ be the maximum estimation error to be tolerated. On the one hand, $s_{j,2} > \bar{s}$ indicates that the system may be in a transient state, or a sudden disturbance has occurred. In this case, the penalty term $-r_p$ will give a quick response and the RLESO can adjust the observer gains. On the other hand, $s_{j,2} \leqslant \bar{s}$ indicates that the ESO's estimation is in steady state and its estimation error is small. Thus, a larger reward $-s_{j,2}/\bar{s} \in [-1, 0]$ is given.

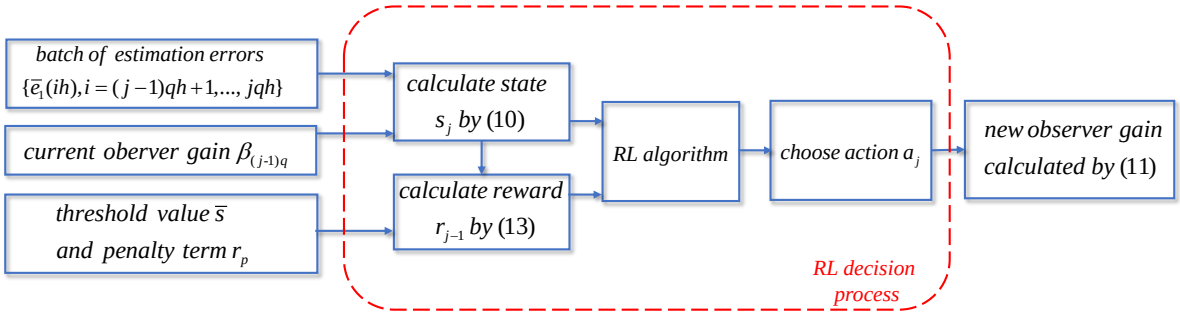

Figure D2 The frame of RLESO at $t = jqh$

The frame of RL is shown in Fig. D2. For the RL decision process in Figure 2, notice that we did not specify a particular RL algorithm. In fact, this is one of the characteristics of our work: we give a reasonable way to design states, actions, and rewards, and one can combine them with different RL algorithms depending on the situation, rather than sticking to one algorithm.

Appendix E Supplement to Theorems 1-2

Before stating the main theorem, we introduce the L_2-norm:

Definition 1. For a vector or matrix X, its L_2-norm in the sense of mean square is defined as:

$$\|X\|_{L_2} \triangleq \{E[\|X\|_2^2]\}^{1/2}. \tag{E1}$$

The dynamics of estimation error is a time-varying random system whose stability is jointly determined by time-varying matrix, measurement noise, and discretization error. RLESO is time-varying and discrete, so its stability cannot be guaranteed simply by making the eigenvalues of A_k in the unit circle in every iteration. Next, the properties of the time-varying matrix will be analyzed.

Lemma 1. Define $-$, where $\rho(A_k)$ means the spectral radius of A_k. For any $\zeta > 0$ such that $\zeta + \bar{\rho} < 1$, there exists a positive number N such that $\forall n > N$, it holds

$$\|{A_k}^n\|_2 < (\bar{\rho} + \zeta)^n, \quad \forall k > 0. \tag{E2}$$

Proof of Lemma 1. By Gelfand's Formula,

$$\rho(A_k) = \lim_{n\to\infty} \|{A_k}^n\|_2^{1/n}.$$

As a result, for any $\zeta > 0$ s.t. $\zeta + \bar{\rho} \in (0, 1)$, there exists a positive number N such that $\forall n > N$,

$$\left|\rho(A_k) - \|{A_k}^n\|_2^{1/n}\right| < \zeta, \quad \forall k,$$

so

$$\|{A_k}^n\|_2 < (\rho(A_k) + \zeta)^n \leqslant (\bar{\rho} + \zeta)^n. \qquad \blacksquare$$

Appendix E.1 Proof of Theorem 1.

Step 1. First, the uniform exponential stability of $e_{k+1} = A_k e_k$ will be proved.

By the fact that $\|A_k\|_2 \leqslant \sqrt{tr(A_k^T A_k)}$, we have

$$\max_k \|A_k\|_2 \leqslant \widetilde{A} \triangleq \max_k \sqrt{tr(A_k^T A_k)}. \tag{E3}$$

Considering the Lyapunov equation

$$A_k^T P_{k+1} A_k + I = P_{k+1}, \tag{E4}$$

it can be verified that the following symmetric positive definite matrix P_{k+1} is the solution of (E4):

$$P_{k+1} = I + \sum_{i=1}^{\infty} (A_k^T)^i (A_k)^i.$$

Suppose P_{k+1}^1 and P_{k+1}^2 are all solutions of (E4), then

$$P_{k+1}^1 - P_{k+1}^2 = A_k^T \left(P_{k+1}^1 - P_{k+1}^2\right) A_k = \left(A_k^T \otimes A_k^T\right) vec(P_{k+1}^1 - P_{k+1}^2),$$

where $\otimes$ is the Kronecker product. By Assumption 3, any two eigenvalues of A_k multiplied by each other cannot be 1, which means that $A^T \otimes A^T$ is full rank. Therefore, $P_{k+1}^1 - P_{k+1}^2$ must be 0 and P_{k+1} is the unique solution of (E4).

The upper bound of $\|P_{k+1}\|_2$ is:

$$\begin{aligned}
&\|P_{k+1}\|_2 \\
\leqslant& 1 + \sum_{i=1}^{\infty} \|{A_k}^i\|_2^2 \\
\leqslant& 1 + \sum_{i=1}^{N} \|A_k\|_2^{2i} + \sum_{i=N+1}^{\infty} \|{A_k}^i\|_2^2 \\
\leqslant& 1 + \sum_{i=1}^{N} \widetilde{A}^{2i} + \sum_{i=N+1}^{\infty} (\bar{\rho} + \zeta)^{2i} \\
\leqslant& 1 + \frac{\widetilde{A}^2 - \widetilde{A}^{2N}}{1 - \widetilde{A}^2} + \frac{(\bar{\rho} + \zeta)^{2(N+1)}}{1 - (\bar{\rho} + \zeta)^2} = P.
\end{aligned} \tag{E5}$$

Obviously, $\{P_k\}$ is absolutely convergent and the norm of P_k is bounded:

$$1 \leqslant \|P_k\|_2 \leqslant P, \quad \forall k. \tag{E6}$$

Defining the Lyapunov function $V_k(e) = e^T P_k e$, it is easy to see that

$$\|e\|_2^2 \leqslant V_k(e) \leqslant P\|e\|_2^2.$$

Next, by subtracting two consecutive instances of (E4), we have

$$\begin{aligned}
&P_{k+1} - P_k \\
=& A_k^T P_{k+1} A_k + I - A_{k-1}^T P_k A_{k-1} - I \\
=& A_k^T P_{k+1} A_k - A_k^T P_k A_k + A_k^T P_k A_k - A_k^T P_k A_{k-1} + A_k^T P_k A_{k-1} - A_{k-1}^T P_k A_{k-1} \\
=& A_k^T \left(P_{k+1} - P_k\right) A_k + A_k^T P_k \left(A_k - A_{k-1}\right) + \left(A_k - A_{k-1}\right)^T P_k A_{k-1},
\end{aligned} \tag{E7}$$

and then

$$A_k^T (P_{k+1} - P_k) A_k - (P_{k+1} - P_k) = -A_k^T P_k (A_k - A_{k-1}) - (A_k - A_{k-1})^T P_k A_{k-1}. \tag{E8}$$

Define

$$L_k = A_k^T P_k (A_k - A_{k-1}) + (A_k - A_{k-1})^T P_k A_{k-1}.$$

Obviously,

$$\|L_k\|_2 \leqslant 2\|A_k - A_{k-1}\|_2 P \widetilde{A}. \tag{E9}$$

It can be verified that the solution of (E8) is

$$P_{k+1} - P_k = L_k + \sum_{i=1}^{\infty} (A_k^T)^i L_k (A_k)^i, \tag{E10}$$

and the uniqueness of the solution is verifiable. The verification method is the same as (E4).

Then

$$\begin{aligned}
&\|P_{k+1}-P_k\|_2\\
\leqslant&\|L_k\|_2\left[1+\sum_{i=1}^{\infty}\|(A_k^T)^i\|_2\|A_k{}^i\|_2\right]\\
\leqslant&2\|A_k-A_{k-1}\|_2P\widetilde{A}\cdot P \qquad (by\quad (E5))\\
=&2P^2\widetilde{A}\|A_k-A_{k-1}\|_2\\
=&2P^2\widetilde{A}\sqrt{\sum_{i=1}^{m+n+1}(\beta_{k,i}-\beta_{k-1,i})^2}\\
\leqslant&2P^2\widetilde{A}\sqrt{\sum_{i=1}^{m+n+1}\max\{\underline{a}_i^2,\bar{a}_i^2\}}\\
<&1. \qquad (by\quad (13))
\end{aligned} \tag{E11}$$

As a result, there is a positive number $\mu\in(0,1)$ satisfying

$$\|P_{k+1}-P_k\|_2\leqslant\mu<1. \tag{E12}$$

Then

$$\begin{aligned}
&V_{k+1}(e_{k+1})-V_k(e_k)\\
=&e_k^T(-I+P_{k+1}-P_k)e_k\\
=&-e_k^Te_k+e_k^T(P_{k+1}-P_k)e_k\\
\leqslant&(-1+\mu)e_k^Te_k,
\end{aligned} \tag{E13}$$

which means that $e_{k+1}=A_ke_k$ is uniformly exponentially stable.

Step 2. Next step is to prove ξ_{k+1} is L_2 bounded: On one hand, according to Assumption 2, equation (8) and equation (E1), it can be calculated directly

$$\begin{aligned}
&\left\|\bar{B}_kv(kh)\right\|_{L_2}\\
=&\left\{E\left[h^2\sum_{i=1}^{m+n+1}\beta_i^2(kh)v^2(kh)\right]\right\}^{1/2}\\
=&\left\{h^2\sum_{i=1}^{m+n+1}\beta_i^2(kh)E\left[v^2(kh)\right]\right\}^{1/2}\\
\leqslant&\left\{h^2R_k^v\sum_{i=1}^{m+n+1}\bar{\beta}_i^2\right\}^{1/2}\\
\leqslant&\left\{h^2R^v\sum_{i=1}^{m+n+1}\bar{\beta}_i^2\right\}^{1/2}.
\end{aligned} \tag{E14}$$

On the other hand, according to Assumption 1 and (C3), it holds

$$\begin{aligned}
&\|\delta_k\|_{L_2}\\
=&\left\{E\left[\sum_{i=1}^{m+n+1}\left(\int_{kh}^{(k+1)h}\frac{((k+1)h-\tau)^i}{i!}\sigma(\tau)d\tau\right)^2+\left(c((k+1)h)-c(kh)\right)^2\right]\right\}^{1/2}\\
\leqslant&\left\{E\left[\sum_{i=1}^{m+n+1}\int_{kh}^{(k+1)h}\left(\frac{((k+1)h-\tau)^i}{i!}\right)^2\sigma(\tau)^2d\tau+\left(c((k+1)h)-c(kh)\right)^2\right]\right\}^{1/2}\\
\leqslant&\left\{E\left[\sum_{i=1}^{m+n+1}\left(\frac{h^i}{i!}\right)^2\int_{kh}^{(k+1)h}\sigma(\tau)^2d\tau+\left(c((k+1)h)-c(kh)\right)^2\right]\right\}^{1/2}\\
=&\left\{\sum_{i=1}^{m+n+1}\left(\frac{h^i}{i!}\right)^2E\left[\int_{kh}^{(k+1)h}\sigma(\tau)^2d\tau\right]+E\left[c^2((k+1)h)\right]+E\left[c^2(kh)\right]-2E\left[c((k+1)h)c(kh)\right]\right\}^{1/2}\\
\leqslant&\left\{\sum_{i=1}^{m+n+1}\left(\frac{h^i}{i!}\right)^2R_k^\sigma+R_{k+1}^c+R_k^c\right\}^{1/2}\\
\leqslant&\left\{\sum_{i=1}^{m+n+1}\left(\frac{h^i}{i!}\right)^2R^\sigma+2R^c\right\}^{1/2}.
\end{aligned} \tag{E15}$$

Define

$$H_{1,k}=h\left\{R_k^v\sum_{i=1}^{m+n+1}\bar{\beta}_i^2\right\}^{1/2}+\left\{\sum_{i=1}^{m+n+1}\left(\frac{h^i}{i!}\right)^2R_k^\sigma+R_{k+1}^c+R_k^c\right\}^{1/2} \tag{E16}$$

and

$$H_1=h\left\{R^v\sum_{i=1}^{m+n+1}\bar{\beta}_i^2\right\}^{1/2}+\left\{\sum_{i=1}^{m+n+1}(\frac{h^j}{i!})^2R^\sigma+2R^c\right\}^{1/2}, \tag{E17}$$

then

$$\|\xi_{k+1}\|_{L_2} \leqslant H_{1,k} \leqslant H_1.$$

Step 3. The boundedness of estimation error.
According to (E6) and (E13),

$$\begin{aligned} & V_{k+1}(e_{k+1}) \\ =& e_{k+1}^T P_{k+1} e_{k+1} \\ \leqslant& e_k^T P_k e_k - (1-\mu) e_k^T e_k \\ =& e_k^T P_k e_k - \frac{(1-\mu)}{P} P e_k^T e_k \ . \\ \leqslant& e_k^T P_k e_k - \frac{(1-\mu)}{P} e_k^T P_k e_k \\ =& \left(1 - \frac{1-\mu}{P}\right) V_k(e_k). \end{aligned} \tag{E18}$$

Define $\eta^2 = 1 - \frac{(1-\mu)}{P}$, then for $\forall i \leqslant k$ we have

$$\begin{aligned} & V_{k+1}(e_{k+1}) \leqslant \eta^2 V_k(e_k) \\ \Rightarrow & V_{k+1}(e_{k+1}) \leqslant \eta^{2(k+1-i)} V_i(e_i) \\ \Rightarrow & \|e_{k+1}\|_2^2 \leqslant V_{k+1}(e_{k+1}) \leqslant \eta^{2(k+1-i)} V_i(e_i) \\ \Rightarrow & \|e_{k+1}\|_2 \leqslant \eta^{k+1-i} \sqrt{P} \|e_i\|_2. \end{aligned} \tag{E19}$$

Let $\Phi(k,i)$ be the transition matrix of $e_{k+1} = A_k e_k$, according to the definition of induced norm, we have

$$\|\Phi(k,i)\|_2 = \sup_{\|e_i\|=1} \|\Phi(k,i) e_i\|_2 \leqslant \sqrt{P} \eta^{k-i}. \tag{E20}$$

Define $H_2 = \frac{\sqrt{P}}{1-\eta}$, it is apparent that

$$\sum_{i=1}^{k} \|\Phi(k,i)\|_{L_2} \leqslant H_2, \quad \forall k. \tag{E21}$$

And then

$$\|e_{k+1}\|_{L_2} \leqslant \sqrt{P} \eta^{k+1} \|e_0\|_2 + \sum_{i=0}^{k} \|\Phi(kh, ih) \xi_{i+1}\|_{L_2} \leqslant \sqrt{P} \|e_0\|_2 + H_1 H_2, \quad \forall k. \tag{E22}$$

At this point, the proof is completed. ■

Appendix E.2 Proof of Theorem 2.

The proof is quite like Theorem 1. By condition (15), it holds

$$\lim_{k \to \infty} H_{1,k} = 0.$$

Thus,

$$\lim_{k \to \infty} \|e_{k+1}\|_{L_2} \leqslant \lim_{k \to \infty} \sqrt{P} \eta^{k+1} \|e_0\|_2 + H_{1,k} H_2 = 0.$$

Since $\|e_{k+1}\|_{L_2} \geqslant 0$,

$$\lim_{k \to \infty} \|e_{k+1}\|_{L_2} = 0.$$

Appendix E.3 The significance of Theorems 1-2

Remark 3. Theorem 1 gives a quantitative condition to guarantee the boundedness of the RLESO's estimation error. Actually, (13) clearly provides the design principles for the action set of RLESO.

Remark 4. Theorem 2 states that if the noise and the m-th derivative of total disturbance converge to zero, then the estimation error of RLESO will converge to 0. Actually, such conditions are satisfied when the system dynamics is in steady-state and there is no measurement noise.

Appendix F Application to velocity control system of UGV

In this section, the effect of the RLESO will be tested by simulation. In the simulation, the following longitudinal velocity control system of the unmanned ground vehicle (UGV) will be studied:

$$\begin{cases} \dot{v}(t) = \frac{b}{m} T(t) - \frac{F_1}{m}, \\ F_1(t) = k_f(t) m(t), \\ y(kh) = v(kh) + n(kh), \end{cases} \qquad t \geqslant 0, \quad k > 0, \tag{F1}$$

where $m = m_0 + m_p$ and the significances of the meanings are shown in Table F1:

For simplicity, the friction resistance is assumed to be affected only by road friction coefficient k_f and vehicle total load m. According to [28], friction coefficients under different road surfaces are shown in Table 2:

Table F1 Significance of symbol

Symbol	Physical Significance	Value
v	Longitudinal Velocity	25∼50 (km/h)
b	Control gain	5000
m_0	Mass of the Vehicle	1270 (kg)
m_p	Load of the Vehicle	140 ∼ 350 (kg)
T	Pedal Angle	Control input
F_1	Frictional Resistance	Depends on k_f and m
n	Measurement Noise	$\sim N(0, 0.01)$
k_f	Friction Coefficient	Depends on the road surface

Table F2 Friction coefficient under different road surfaces

Road Surface	Cement Concrete	Cement Concrete	Bituminous Concrete	Bituminous Concrete
Old/New	Old	New	Old	New
k_f	0.53	0.82	0.63	0.95

Next, the simulation will be carried out under three different vehicle driving conditions:

Table F3 Different driving conditions

	Types of Road Surface	Load Variation	Velocity (km/h)	Load (kg)	Noise
Case 1	1	No	50	1270	$N(0, 0.01)$
Case 2	4	No	50	1270	0
Case 3	4	Yes	0 ∼ 50	1480∼1620	$N(0, 0.01)$

Taking $F_2 = (\frac{b}{m} - \frac{b}{m_0})T - \frac{F_1}{m}$ as the total disturbance, the LESO is designed as follows:

$$\begin{bmatrix} \hat{z}_1((k+1)h) \\ \hat{z}_2((k+1)h) \\ \hat{z}_3((k+1)h) \end{bmatrix} = \begin{bmatrix} 1-h\beta_{k,1} & h & \frac{h^2}{2} \\ 1-h\beta_{k,2} & 1 & h \\ 1-h\beta_{k,3} & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{z}_1(kh) \\ \hat{z}_2(kh) \\ \hat{z}_3(kh) \end{bmatrix} + \begin{bmatrix} h \\ 0 \\ 0 \end{bmatrix} \frac{b}{m_0} T(kh) + h \begin{bmatrix} \beta_{k,1} \\ \beta_{k,2} \\ \beta_{k,3} \end{bmatrix} y(kh) \tag{F2}$$

where $\hat{z} = [\hat{z}_1, \hat{z}_2, \hat{z}_3]^T$ is the estimation value of $[v, F_2, \dot{F}_2]^T$.

For the convenience of adjusting parameters, β_k adjusted by the "bandwidth method":

$$\begin{cases} \beta_{k,1} = 3\omega_o(kh), \\ \beta_{k,2} = 3{\omega_o}^2(kh) - \frac{1}{2}h{\omega_o}^3(kh), \quad \omega_o(kh) \in [0.5, 10], \quad \forall k > 0, \\ \beta_{k,3} = {\omega_o}^3(kh), \end{cases} \tag{F3}$$

and the eigenvalues of

$$\begin{bmatrix} 1-h\beta_{k,1} & h & \frac{h^2}{2} \\ 1-h\beta_{k,2} & 1 & h \\ 1-h\beta_{k,3} & 0 & 1 \end{bmatrix}$$

are all $1 - h\omega_o(kh)$ and the bandwidth ω_o is the only gain to be adjusted.

Adopting the following control law

$$T(t) = \frac{m_0}{b}[-\hat{z}_2(kh) - k_v(\hat{z}_1(kh) - v^*(kh))], \quad t \in [kh, (k+1)h). \tag{F4}$$

The Q-learning algorithm combined with ϵ-greedy is selected to make a decision. The factors are designed as follows and the control block diagram is shown in Fig. F1. Details of the Q-learning based tuning algorithm are given in Algorithm 1.

Parameters: $\underline{\omega}_o = 0.5$, $\overline{\omega}_o = 10$, q=100, greedy rate $\epsilon = 0.9$, discount factor $\gamma = 0.95$ and learning rate $\alpha_n = \frac{1}{n^2}$ [29];

State: $s_j = [s_{j,1}, \widetilde{s}_{j,2}]$, where

$$\begin{cases} s_{j,1} = \omega_o((j-1)qh), \\ s_{j,2} = \frac{1}{q}\sum\limits_{i=0}^{q-1} \left|\overline{e}((j-1)qh + ih)\right|, \quad \overline{e} = y - \hat{x}_1, \end{cases}$$

and

$$\widetilde{s}_{j,2} = \begin{cases} 1, & s_{j,2} \in [0, 0.002), \\ 2, & s_{j,2} \in [0.002, 0.004), \\ \quad \vdots & \\ 9, & s_{j,2} \in [0.016, 0.018), \\ 10, & \text{else}; \end{cases}$$

Action : $\Lambda = \{-0.5, 0, 0.5\}$;

Reward:

$$r_{j-1} = \begin{cases} -s_{j,2}/0.2, & s_{j,2} \leqslant 0.2, \\ -10, & s_{j,2} > 0.2. \end{cases}$$

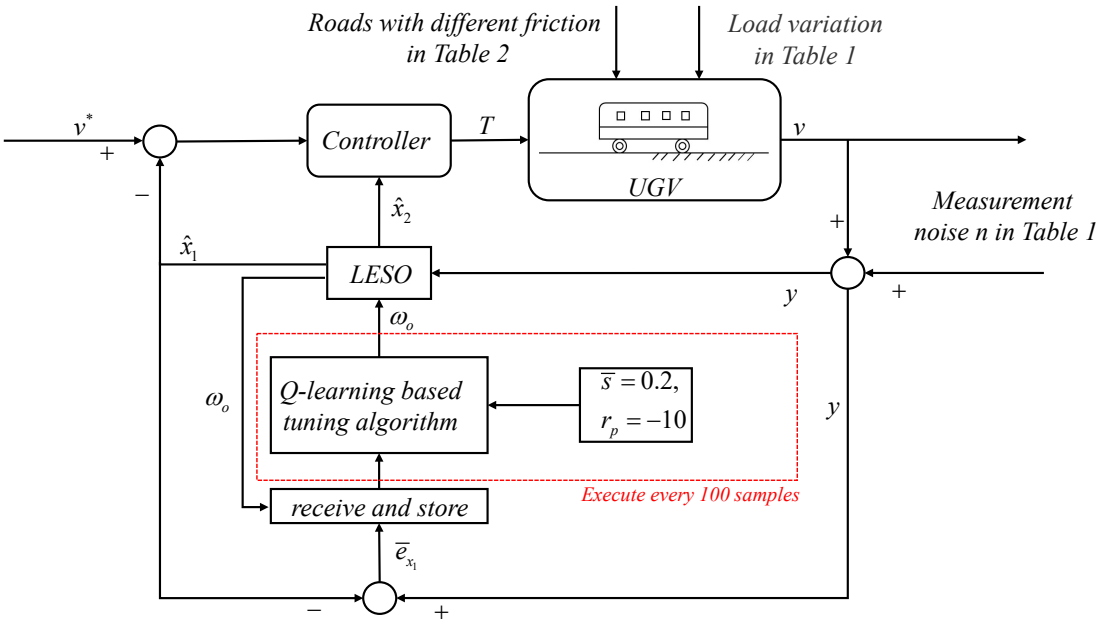

Figure F1 UGV longitudinal velocity control system based on RLESO

Algorithm F1 Tuning algorithm based on Q-learning with ϵ-greedy method

Initialize $j=0$, $Q(s,a)=0$ for all s and a. Then select the discount factor γ and learning rate sequence $\{\alpha_j\}$.
if the experiment has not been finished **then**
repeat for every q samples:

1. $j=j+1$;
2. Get state s_j;
3. Choose action a_j which satisfies

$$a_j = \underset{a'\in\Lambda}{argmax}\, Q(s_j, a') \tag{F5}$$

with probability ϵ, and choose action randomly with probability $1-\epsilon$;

4. Tune the bandwidth:

$$\omega_{o,jq} = \begin{cases} \overline{\omega}_o, & if \quad \omega_{o,(j-1)q} + a_j \geqslant \overline{\omega}_o \\ \omega_{o,(j-1)q} + a_j, & if \quad \omega_{o,(j-1)q} + a_j \in (\underline{\omega}_o, \overline{\omega}_o) \\ \underline{\omega}_o, & if \quad \omega_{o,(j-1)q} + a_j \leqslant \underline{\omega}_o \end{cases};$$

5. Observe the next state s_j and reward r_{j-1};
6. Update $Q(s_j, a_j)$ with

$$Q(s_j,a_j) = Q(s_j,a_j) + \alpha_j[r_{j-1} + \gamma \cdot \max_{a'\in A} Q(s_{j+1},a') - Q(s_j,a_j)]; \tag{F6}$$

end if

In Case 1, the vehicle is driving at a steady speed on the concrete road with no load and the total disturbance F_2 would be a constant and the measurement noise $n \sim N(0.01)$. For a constant disturbance, the ideal bandwidth is small to reduce the impact of measurement noise. Moreover, simulations with $\omega_o = 0.5$, 5 and 10 are set as comparisons. Combined with Fig. F2 and Fig. F3, it is easy to see the bandwidth is gradually declining and the steady estimation error of the RLESO is smaller and smaller.

Vehicle in Case 2 drives on different roads without measurement noise. Furthermore, it is assumed that the friction changes linearly during the transition of the road surfaces. As a result, F_2 is time-varying red and the bandwidth should be maintained at a high level to get a good transient performance. And the simulation results in Fig. F4 and Fig. F5 meet expectations: The bandwidth starts at the minimum value and gradually increases to a high level. And the estimation performance of RLESO is improved.

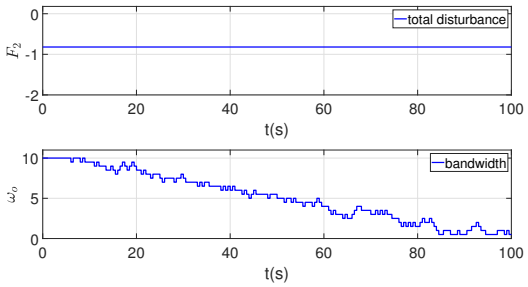

Figure F2 The curve of F_2 and ω_o (Case 1)

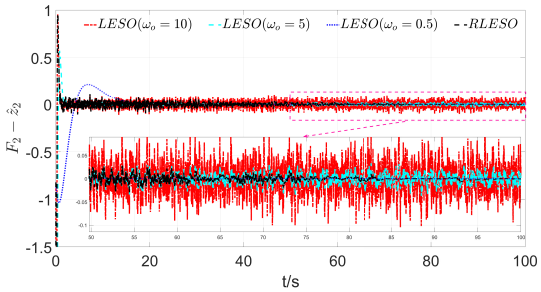

Figure F3 The estimation error of the total disturbance (Case 1)

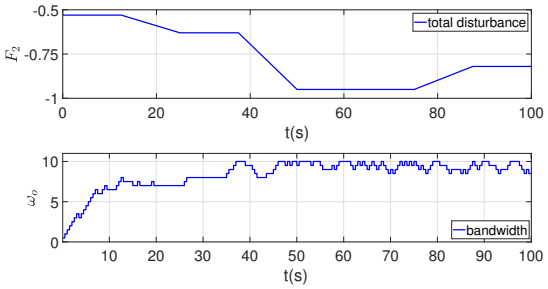

Figure F4 The curves of F_2 and ω_o (Case 2)

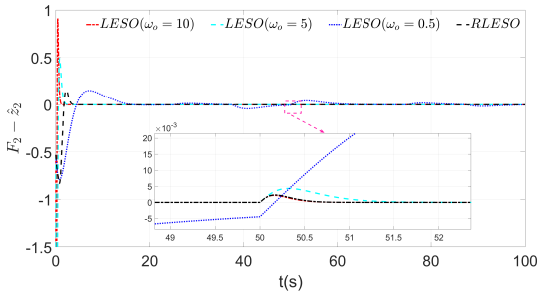

Figure F5 The estimation error of the total disturbance (Case 2)

Finally, Case 3 considers that the vehicle is driving on different roads, with loads getting on and off and velocity changing. To simulate passengers getting on and off and the vehicle stopping and starting, the expected velocity and mass of the vehicle are shown in Fig. F6. Under this condition, the RLESO is expected to adjust the bandwidth to get a good performance both in transient state and steady state. Fig. F8 shows a satisfying result: When the disturbance changes, the RLESO can quickly track the disturbance so the transient performance is great. When the disturbance becomes constant, the influence of noise can be reduced and a smooth estimation curve can be obtained.

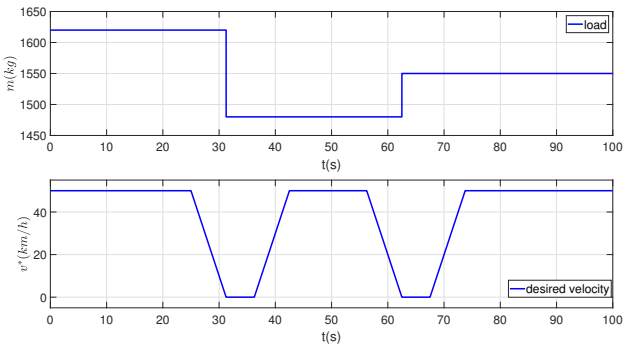

Figure F6 The curves of load and velocity (Case 3)

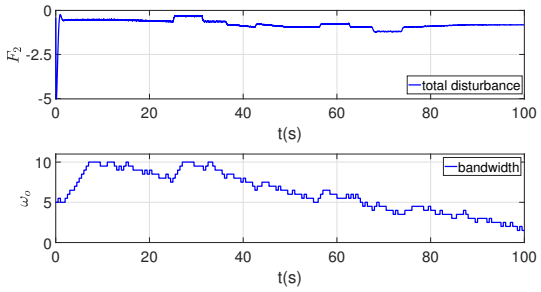

Figure F7 The curves of the F_2 and ω_o (Case 3)

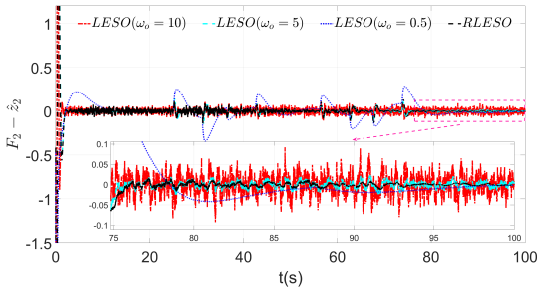

Figure F8 The estimation errors of the total disturbance (Case 3)

In addition to the above studies, considering that the RLESO is based on sampled-data, the influence of the sampling period h on it will be studied. With the same settings to Case 1, Fig. F9 shows the estimation effect of RLESO under different sampling period. In the same time, the shorter the sampling period, the more times RLESO can learn and tune, so as to obtain better estimation effect.

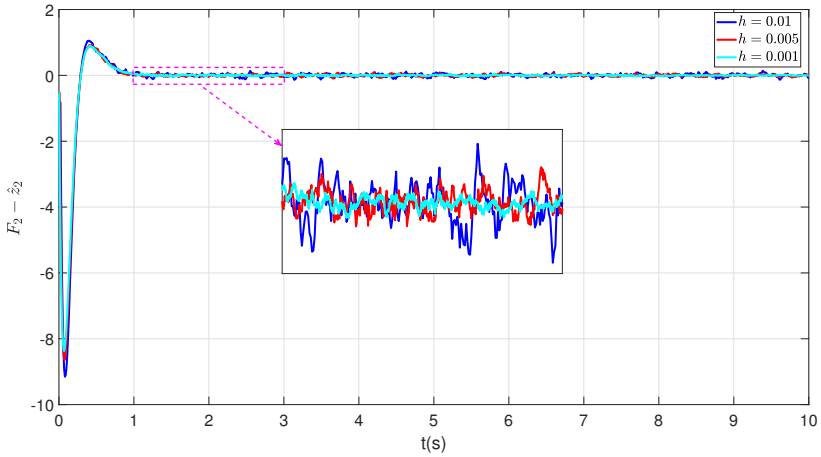

Figure F9 The curves of estimation error of RLESO under different sampling periods

References

1 C. Johnson. Optimal control of the linear regulator with constant disturbances. *IEEE Transactions on Automatic Control*, 13(4):416–421, 1968.

2 N. H. Jo, H. Shim, and Y. I. Son. Disturbance observer for non-minimum phase linear systems. *International Journal of Control Automation Systems*, 8(5):994–1002, 2010.

3 W. Chen. Disturbance observer based control for nonlinear systems. *IEEE/ASME Transactions on Mechatronics*, 9(4):706–710, 2004.

4 J. Han. The "extended state observer" of a class of uncertain systems. *Control and Decision*, 1995.

5 W. Xue and Y. Huang. Performance analysis of 2-DOF tracking control for a class of nonlinear uncertain systems with discontinuous disturbances. *International Journal of Robust and Nonlinear Control*, 28(4):1456–1473, 2018.

6 M. Ruderman and M. Iwasaki. Sensorless control of motor velocity in two-mass actuator systems with load sensing using extended state observer. In *2014 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, pages 360–365, 2014.

7 J. Linares-Flores, J. L. Barahona-Avalos, H. Sira-Ramirez, and M. A. Contreras-Ordaz. Robust passivity-based control of a buck–boost-converter/DC-motor system: An active disturbance rejection approach. *IEEE Transactions on Industry Applications*, 48(6):2362–2371, 2013.

8 Y. Xia, Z. Zhu, M. Fu, and S. Wang. Attitude tracking of rigid spacecraft with bounded disturbances. *IEEE Transactions on Industrial Electronics*, 58(2):647–659, 2011.

9 D. Huang, X. Hua, B. Mi, Y. Liu, and Z. Zhang. Incipient fault diagnosis on active disturbance rejection control. *Science China Information Sciences*, 65(9):1–2, 2022.

10 W. Xue and Y. Huang. On performance analysis of ADRC for a class of MIMO lower-triangular nonlinear uncertain systems. *ISA Transactions*, 53(4):955–962, 2014.

11 Q. Zheng. *On active disturbance rejection control: stability analysis and applications in disturbance decoupling control*. PhD thesis, Cleveland State University, 2009.

12 B. Guo and Z. Zhao. On the convergence of an extended state observer for nonlinear systems with uncertainty. *Systems & Control Letters*, 60(6):420–430, 2011.

13 W. Wei, W. Xue, and D. Li. On disturbance rejection in magnetic levitation. *Control Engineering Practice*, 82:24–35, 2019.

14 C. Liu, G. Luo, X. Duan, Z. Chen, and C. Qiu. Adaptive LADRC-based disturbance rejection method for electromechanical servo system. *IEEE Transactions on Industry Applications*, 56(1):876–889, 2020.

15 Z. Pu, R. Yuan, J. Yi, and X. Tan. A class of adaptive extended state observers for nonlinear disturbed systems. *IEEE Transactions on Industrial Electronics*, 62(9):5858–5869, 2015.

16 W. Xue, W. Bai, S. Yang, K. Song, Y. Huang, and H. Xie. ADRC with adaptive extended state observer and its application to air–fuel ratio control in gasoline engines. *IEEE Transactions on Industrial Electronics*, 62(9):5847–5857, 2015.

17 L. Shao, X. Liao, Y. Xia, and J. Han. Stability analysis and synthesis of third order discrete extended state observer. *Information and Control*, 02:135–139, 2008.

18 J. Li, Y. Xia, X. Qi, and H. Wan. On convergence of the discrete-time nonlinear extended state observer. *Journal of the Franklin Institute*, 355(1):501–519, 2017.

19 W. Bai, W. Xue, Y. Huang, and H. Fang. On extended state based Kalman filter design for a class of nonlinear time-varying uncertain systems. *Science China Information Sciences*, 61(4):1–16, 2018.

20 R.S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

21 M. M. Noel and B. J. Pandian. Control of a nonlinear liquid level system using a new artificial neural network based reinforcement learning approach. *Applied Soft Computing*, 23:444–451.

22 L. Wang, Y. Liu, and X. Zhai. Design of reinforce learning control algorithm and verified in inverted pendulum. In *Chinese Control Conference (CCC)*, 2015.

23 Z. An and D. Zhou. Deep reinforcement learning for quantum gate control. *EPL (Europhysics Letters)*, 126(6):60002, 2019.

24 C. Hu, H. Wang, and H. Shi. Robotic arm reinforcement learning control method based on autonomous visual perception. *Xibei Gongye Daxue Xuebao/Journal of Northwestern Polytechnical University*, 39(5):1057–1063, 2021.

25 F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis. Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers. *IEEE Control Systems Magazine*, 32(6):76–105, 2012.

26 B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis. Optimal and autonomous control using reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 29(6):2042–2062, 2017.

27 J. Shi, X. Chen, and S. T. Yau. High order linear extended state observer and error analysis of active disturbance rejection control. *Asian Journal of Mathematics*, 23(4):631–650, 2019.

28 P. Hu and X. Pan. Field test of pavement friction coefficient under different conditions. *Highway*, 2(2):5, 2011.

29 J. N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. In *Conference on Decision and Control*, pages 395–400, 1993.