IET Control Theory & Applications

Special Issue: Data-driven Control, and Data-Based System Modelling, Monitoring, and Control

IET Journals
The Institution of Engineering and Technology

Distributed adaptive optimal regulation of uncertain large-scale interconnected systems using hybrid Q-learning approach

ISSN 1751-8644
Received on 24th September 2015
Revised on 19th February 2016
Accepted on 22nd February 2016
doi: 10.1049/iet-cta.2015.0943
www.ietdl.org

Vignesh Narayanan ✉, *Sarangapani Jagannathan*

Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65401, USA
✉ *E-mail: vnxv4@mst.edu*

Abstract: A novel hybrid Q-learning algorithm is introduced for the design of a linear adaptive optimal regulator for a large-scale interconnected system with event-sampled inputs and state vector. Here, the time-driven Q-learning along with proposed iterative parameter learning updates are utilised within the event-sampled instants to both improve efficiency of the optimal regulator and obtain a more generalised online Q-learning framework. The network-induced losses due to the presence of a communication network among the subsystems are considered along with the uncertain system dynamics. Stochastic model-free Q-learning and dynamic programming are utilised in the hybrid learning mode for the optimal regulator design. The asymptotic convergence of the system state vector and boundedness of the parameter vector is demonstrated using Lyapunov analysis. Further, when the regression vector of the Q-function estimator satisfies the persistency of excitation condition, the Q-function parameters converge to the expected target values. The analytical design is evaluated using numerical examples via simulation. The net result is the design of a data-driven event-sampled adaptive optimal regulator for an uncertain large-scale interconnected system.

1 Introduction

Optimal control [1] using adaptive dynamic programming (ADP) [2–7] has drawn more attention because of the forward-in-time solution to the optimal control problems for uncertain systems. The ADP-based control schemes use reinforcement learning to solve the Bellman or Hamilton–Jacobi–Bellman equation [4] through online parameterisation and obtain optimal control policy. Among the ADP-based Q-learning schemes, Bradtke *et al.* [2] proposed a policy iteration approach using the Bellman equation. Later, the Q-learning scheme was extended in [3] to zero-sum-game formulation by using model-free policy iteration.

Policy/value iteration-based techniques use a significant number of iterative parameter updates within a sampling interval to maintain system stability, and its online implementation is not practically viable [5]. Therefore, online implementation for such iterative techniques was presented in [4], where the parameters are updated after collecting sufficient data-points. In contrast, the effort from [5] followed by the authors [6, 7] introduced a time-based model-free ADP scheme where the past data of the cost-to-go errors are used for constructing the optimal value function.

On the other hand, control of large-scale interconnected systems [8] has been an active area of research. Large-scale systems are complex systems composed of geographically distributed subsystems connected through a communication network. The traditional centralised controller design for such systems is often impractical for computational reasons and lack of control integrity [9]. Therefore, various decentralised/distributed control schemes have been developed in the literature such that each subsystem has an independent controller [9–22]. The complexity in the control design arises due to the structural constraint in the form of interconnection/coupling matrix [8–20], which determines how the states/control of one subsystem influence the dynamics of the other subsystems.

Over the years, the controllers for large-scale systems have evolved to stabilise the subsystems in the presence of uncertain interconnection matrix with limited communication [8–10]. Adaptive controllers were proposed to learn the interconnection terms, with which suitable compensation was provided [9–12], but they were limited to handle weak interconnections. Later, reference models were utilised to provide information about the other subsystems. However, it is reported in [13] that if the subsystems do not communicate their state information with each other and use a reference model to obtain this information, unsatisfactory transient performance will occur. Further, utilising the communication network connecting the subsystems, several distributed control algorithms to solve optimisation problem for large-scale system using model-predictive control (MPC) have been proposed in [14–18] and the references therein.

Although MPC-based control algorithms are popular due to their inherent ability to handle input and state constraints efficiently, distributed MPC algorithms are not as efficient as their centralised counterpart due to the effect of coupling between the subsystems in the large-scale systems [14, 16, 18]. Also, MPC-based algorithms in general requires system model to predict the future output over a limited time horizon with which a desired cost-function is minimised iteratively [14–18]. In contrast, the Q-function-based control algorithm developed in this work neither requires an accurate model for the system nor utilises significant iterations to solve the optimisation problem. It should be noted that the above mentioned works [14, 16–19] use periodic feedback and utilise the system dynamics to generate control. However, it is not feasible to communicate the state information periodically due to the communication cost involved.

Recently, it was demonstrated that event-based sampling is advantageous over periodic sampling in terms of computational cost [6, 21–23]. The aperiodic event-based sampling instants are determined by using a trigger condition while maintaining stability of the system. Such an event-sampled approach for control design was extended to large-scale interconnected systems in [20–22] by assuming either weak interconnections [20, 21] or control gain satisfying a strong matching condition, in order to decouple the subsystems [22].

The presence of a communication network among the subsystems and in the feedback-loop introduces random time delays and data-dropouts [24–26], which degrades control performance. It was shown in [7] that a linear time-invariant system with a communication network within its feedback loop can be represented as a

stochastic time-varying linear system with uncertain dynamics. To the best knowledge of the authors, a time-based Q-learning scheme with intermittent feedback is not reported for such uncertain large-scale interconnected systems.

Therefore, in this paper, a novel hybrid model-free Q-learning scheme using event-sampled state and input vector is introduced for a large-scale interconnected system that is enclosed by a communication network. This algorithm enables a finite number of proposed Q-function parameter updates iteratively within the event-sampled instants to attain optimality faster without explicitly increasing the events when compared with the algorithm in [6].

In the proposed algorithm, the temporal-difference (TD)-based ADP schemes [6, 7] and the policy/value iterations-based ADP schemes [4, 27] become special cases. It also relaxes the assumptions on the estimated control input utilised in [6, 7]. This makes the learning algorithm more flexible than the existing model-free Q-learning-based ADP schemes for online control. Since the Q-function parameters at each subsystem are estimated online with event-sampled input, state information along with past history and the data obtained from other subsystems through the communication network, an overall system model is not required. This makes the control scheme data-driven [28]. It is important to note that the infinite horizon cost function associated can be evaluated only for an admissible control policy [4]. This requires the control policy obtained using the learning process to be admissible at every step.

The contributions of the paper include: (a) development of a novel hybrid Q-learning scheme using event-sampled states, input vector and their history; (b) the derivation of a time-driven and hybrid Q-learning scheme for an uncertain large-scale interconnected system enclosed by a communication network without any assumptions on coupling terms; (c) a decentralised event-sampling condition based on Lyapunov function without needing a mirror estimator at the sensor; and (d) demonstration of closed-loop stability for such system using Lyapunov analysis.

This paper uses, $\Re$ to denote the set of all real numbers, Euclidean norm for vectors and Frobenius norm for matrices. The following section introduces the system description followed by the derivation of time-driven Q-learning scheme for large-scale interconnected systems with periodic feedback.

2 Background

2.1 System description

Consider a linear time-invariant continuous-time system having N interconnected subsystems shown in Fig. 1 with subsystem dynamics described by

$$\dot{x}_i(t) = A_i x_i(t) + B_i u_i(t) + \sum_{\substack{j=1 \\ j\neq i}}^{N} A_{ij} x_j(t), \quad x_i(0) = x_{i0}, \tag{1}$$

where $x_i, \dot{x}_i \in \Re^{n_i \times 1}$ represent the state vector and state derivatives, respectively, $u_i \in \Re^{m_i}, A_i \in \Re^{n_i \times n_i}$ and $B_i \in \Re^{n_i \times m_i}$ denote control input, internal dynamics and control gain matrices of the ith subsystem, respectively, $A_{ij} \in \Re^{n_i \times n_j}$ represents the interconnection matrix between the ith and the jth subsystems, $i \in 1, 2, \ldots, N$. The overall system description can be expressed in a compact form as

$$\dot{X}(t) = AX(t) + BU(t), \quad X(0) = X_0, \tag{2}$$

where $X \in \Re^n, U \in \Re^m, B \in \Re^{n\times m}, A \in \Re^{n\times n}, \dot{X} = [\dot{x}_1^T, \ldots, \dot{x}_N^T]^{\mathrm{T}}$

$$A = \begin{pmatrix} A_1 & \cdots & A_{1N} \\ \vdots & \cdots & \vdots \\ A_{N1} & \cdots & A_N \end{pmatrix},$$

$$B = \mathrm{diag}[B_1, \ldots, B_N], \quad U = \left[u_1^T, \ldots, u_N^T\right]^{\mathrm{T}}.$$

The system dynamics A_i, B_i and the interconnection matrix A_{ij} are considered uncertain. In the large-scale interconnected system, the subsystems communicate with each other via Network 1, while each subsystem is also enclosed by Network 2. Effects of the network-induced losses can be modelled along with the system dynamics by utilising the standard and mild assumptions as listed in [7, 24].

Assumption 1: The system (2) is considered controllable and the states are measurable. Further, the order of subsystems is considered known.

With the network-induced delays and data-dropout, the original plant can be represented as

$$\dot{X}(t) = AX(t) + \gamma_{ca}(t) BU(t - \tau(t)), \quad X(0) = X_0, \tag{3}$$

where $\gamma_{ca}(t)$ is the data-dropout indicator, which becomes $I^{n\times n}$ when the control input is received at the actuator and $0^{n\times n}$ when the control policy is lost at time t. This only includes the data loss in Network 2 and $\tau(t)$ is the total delay. Now, integrating the system dynamics with network parameters over the sampling interval [7, 24], we obtain

$$\begin{aligned} X_{k+1} = {} & A_d X_k + \gamma_{ca,k} B_0^k U_k + \gamma_{ca,k-1} B_1^k U_{k-1} \\ & + \cdots + \gamma_{ca,k-\bar{d}} B_{\bar{d}}^k U_{k-\bar{d}}, \quad X(0) = X_0, \end{aligned} \tag{4}$$

where $X_k = X(kT_s), A_d = e^{AT_s}$, $\bar{d}$ is the delay bound and U_k is the control input. B_0^k, B_i^k, $\forall i = \{1, 2, \ldots \bar{d}\}$ are all defined as in [7]. From the discretised system representation, we can define an augmented state vector consisting of state and past control inputs

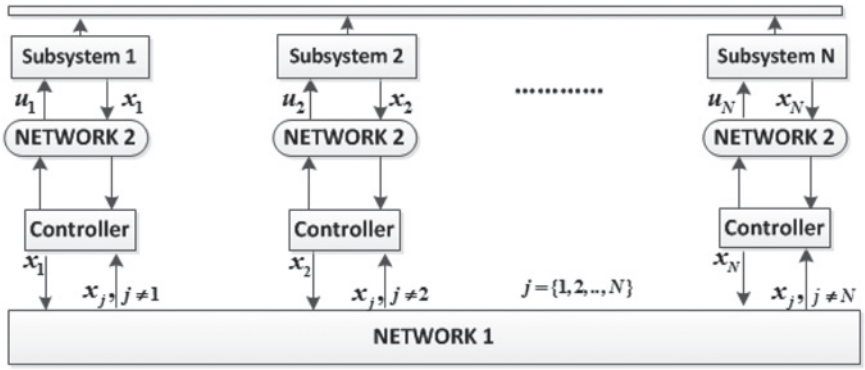

Fig. 1 *Large-scale interconnected system*

as $\bar{X}(k) = [X_k^T \; U_{k-1}^T \; \dots \; U_{k-\bar{d}}^T]^T \in \Re^{n+\bar{d}m}$. The new augmented system representation is given by

$$\bar{X}_{k+1} = A_{\bar{x}k}\bar{X}_k + B_{\bar{x}k}U_k, \quad \bar{X}(0) = \bar{X}_0, \tag{5}$$

with the system matrices given by

$$A_{\bar{x}k} = \begin{bmatrix} A_d & \gamma_{ca,k-1}B_1^k & \cdots & \gamma_{ca,k-\bar{d}}B_{\bar{d}}^k \\ 0 & \cdots & 0 & 0 \\ \vdots & I_m & \vdots & \vdots \\ 0 & \cdots & I_m & 0 \end{bmatrix} \quad \text{and}$$

$$B_{\bar{x}k} = \begin{bmatrix} \gamma_{ca,k}B_0^k \\ I_m \\ \vdots \\ 0 \end{bmatrix}.$$

Remark 1: Note that the system dynamics are now stochastic due to the network-induced delays and data-dropouts. The assumptions regarding the controllability, observability and the existence of unique solution for the stochastic Riccati equation (SRE) are now dependent on the Grammian functions [1].

Hence, the following assumption is needed to proceed further.

Assumption 2: The system is both uniformly completely observable and controllable [1].

The time-driven Q-learning and adaptive optimal regulation of such stochastic linear time-varying interconnected system is presented as follows.

2.2 Periodically sampled time-driven Q-learning

For the system dynamics (5), the infinite horizon cost function is defined as

$$J_k = \underset{\tau,\gamma}{E}\left[\frac{1}{2}\sum_{t=k}^{\infty}\bar{X}_k^T P_{\bar{x}}\bar{X}_k + U_k^T R_{\bar{x}}U_k\right] \tag{6}$$

where $P_{\bar{x}} = \text{diag}(P, R/\bar{d}, \dots, R/\bar{d}), R_{\bar{x}} = R/\bar{d}$. The penalty matrices P and R are positive semi-definite and positive definite, respectively. $E_{\tau,\gamma}(.)$ denotes the expected value of the stochastic process (.).

The cost function (6) can also be represented as $J_k = E_{\tau,\gamma}[\bar{X}_k^T S_k \bar{X}_k]$ with S_k being the symmetric positive semi-definite solution of the SRE [1]. The next step is to define the action-dependent Q-function for the stochastic system (5) with the cost-to-go function (6) as

$$\begin{aligned} Q(\bar{X}_k, U_k) &= \underset{\tau,\gamma}{E}[r(\bar{X}_k, U_k) + J_{k+1} \,|\bar{X}_k] \\ &= \underset{\tau,\gamma}{E}\{[\bar{X}_k^T \quad U_k^T]G_k[\bar{X}_k^T \quad U_k^T]^{\mathrm{T}}\} \end{aligned} \tag{7}$$

where $r(\bar{X}_k, U_k) = \bar{X}_k^T P_{\bar{x}}\bar{X}_k + U_k^T R_{\bar{x}}U_k$ and G_k is a time-varying matrix. From the Bellman equation, we obtain

$$\begin{aligned} \begin{bmatrix}\bar{X}_k \\ U_k\end{bmatrix}^{\mathrm{T}} \underset{\tau,\gamma}{E}(G_k)\begin{bmatrix}\bar{X}_k \\ U_k\end{bmatrix} &= \begin{bmatrix}\bar{X}_k \\ U_k\end{bmatrix}^{\mathrm{T}} \\ \times \begin{bmatrix} P_{\bar{x}} + \underset{\tau,\gamma}{E}(A_{\bar{x}k}^T S_{k+1}A_{\bar{x}k}) & \underset{\tau,\gamma}{E}(A_{\bar{x}k}^T S_{k+1}B_{\bar{x}k}) \\ \underset{\tau,\gamma}{E}(B_{\bar{x}k}^T S_{k+1}A_{\bar{x}k}) & R_{\bar{x}} + \underset{\tau,\gamma}{E}(B_{\bar{x}k}^T S_{k+1}B_{\bar{x}k}) \end{bmatrix} &\begin{bmatrix}\bar{X}_k \\ U_k\end{bmatrix} \end{aligned} \tag{8}$$

where

$$\underset{\tau,\gamma}{E}(G_k) = \begin{bmatrix} \underset{\tau,\gamma}{E}(G_k^{\bar{x}\bar{x}}) & \underset{\tau,\gamma}{E}(G_k^{\bar{x}U}) \\ \underset{\tau,\gamma}{E}(G_k^{U\bar{x}}) & \underset{\tau,\gamma}{E}(G_k^{UU}) \end{bmatrix}$$

From the matrix equation (8), the time-varying control gain can be expressed as

$$K_k = \underset{\tau,\gamma}{E}\{[R_{\bar{x}} + B_{\bar{x}k}^T S_{k+1}B_{\bar{x}k}]^{-1}B_{\bar{x}k}^T S_{k+1}A_{\bar{x}k}\} = \underset{\tau,\gamma}{E}\{(G_k^{UU})^{-1}G_k^{U\bar{x}}\} \tag{9}$$

The Q-function (7) in parametric form is given by

$$Q(\bar{X}_k, U_k) = \underset{\tau,\gamma}{E}(z_k^T G_k z_k) = \underset{\tau,\gamma}{E}(\Theta_k^T \xi_k) \tag{10}$$

where $z_k = [(\gamma_{sc,k}\bar{X}_k)^T \quad U_k^T]^{\mathrm{T}} \in \Re^{\bar{l}}$ with $\bar{l} = m + n + m\bar{d}$, $\xi_k = z_k^T \otimes z_k$ is the regression vector. $\otimes$ denotes Kronecker product and $\Theta_k \in \Omega_\Theta \subset \Re^{l_g}$ is formed by vectorisation of the parameter matrix G_k. $\gamma_{sc,k}$ is a packet loss indicator, defined similar to $\gamma_{ca,k}$. The estimate of the optimal Q-function is expressed as

$$\hat{Q}(\bar{X}_k, U_k) = \underset{\tau,\gamma}{E}(z_k^T \hat{G}_k z_k) = \underset{\tau,\gamma}{E}(\hat{\Theta}_k^T \xi_k) \tag{11}$$

where $\hat{\Theta}_k \in \Re^{l_g}$ is the estimate of expected target parameter Θ_k. By Bellman's principle of optimality, the optimal value function satisfies

$$\begin{aligned} 0 &= \underset{\tau,\gamma}{E}(J_{k+1}^* \,|\bar{X}_k) - \underset{\tau,\gamma}{E}(J_k^*) + \underset{\tau,\gamma}{E}(r(\bar{X}_k, U_k)) \\ &= \underset{\tau,\gamma}{E}(r(\bar{X}_k, U_k)) + \underset{\tau,\gamma}{E}(\Theta_k^T \Delta\xi_k), \end{aligned} \tag{12}$$

where $\Delta\xi_k = \xi_{k+1} - \xi_k$, and $E_{\tau,\gamma}(J_{k+1}^* \,|\bar{X}_k)$ is the expected cost-to-go at $k+1$st instant, given the state information of the kth instant. Since the estimated Q-function does not satisfy (12), the TD error/Bellman error will be observed as

$$e_{\mathrm{B}}(k) = \underset{\tau,\gamma}{E}(r(\bar{X}_k, U_k) + \hat{\Theta}_k^T \Delta\xi_k) \tag{13}$$

Remark 2: In the iterative learning schemes [2, 4], the parameters of the Q-function estimator (QFE) is updated by minimising the error in (17) until the error converges to a small value for every time step k. On the contrary, time-driven ADP schemes [6, 7] calculate the Bellman error at each step and update once at the sampling instant and the stability of the closed-loop system is established under certain mild assumptions on the estimated control policy.

The overall cost function (6) for the large-scale system (5), can be represented as the sum of the individual cost of all the subsystems as, $J_k = \sum_{i=1}^N J_{i,k}$, where $J_{i,k} = E_{\tau,\gamma}\{(1/2)\sum_{s=k}^{\infty}\bar{x}_{i,k}^T P_{\bar{x},i}\bar{x}_{i,k} + u_{i,k}^T R_{\bar{x},i}u_{i,k}\}$ is the quadratic cost function for the ith subsystem with $\bar{x}_i$ representing the augmented states of the ith subsystem, $P_{\bar{x}} = \text{diag}\{P_{\bar{x},1}, \dots, P_{\bar{x},N}\}$ and $R_{\bar{x}} = \text{diag}\{R_{\bar{x},1}, \dots, R_{\bar{x},N}\}$.

The optimal control sequence to minimise the quadratic cost function (6) in a decentralised framework is not straightforward because of the interconnection dynamics. The optimal control policy for each subsystem, which minimises the cost function (6), is obtained by using the SRE of the overall system given the system dynamics $A_{\bar{x}k}$ and $B_{\bar{x}k}$, as

$$u_{i,k}^* = \underset{\tau,\gamma}{E}\left\{-K_{i,k}^*\bar{x}_{i,k} - \sum_{j=1, j\neq i}^{N} K_{ij,k}^*\bar{x}_{j,k}\right\} \tag{14}$$

where $K_{i,k}^*$ are the diagonal elements and $K_{ij,k}^*$ are the off-diagonal elements, of K_k^* in (9). In the following lemma, it is shown that,

with the control law (14) designed at each subsystem, the overall system is asymptotically stabilised in the mean square.

Lemma 1: Consider the ith subsystem of the large-scale interconnected system (5). Assuming that the system matrices $A_{\bar{x}k}$ and $B_{\bar{x}k}$ are known along with Assumption 2. The optimal control policy obtained from (14) renders the individual subsystems asymptotically stable in the mean square.

Proof: Note that the optimal control input is stabilising [1]. Therefore, the closed-loop system matrix $(A_{\bar{x}k} - B_{\bar{x}k}K_k^*)$ is Schur. The Lyapunov equation $(A_{\bar{x}k} - B_{\bar{x}k}K_k^*)^T\bar{P}(A_{\bar{x}k} - B_{\bar{x}k}K_k^*) - \bar{P} = -\bar{F}$, has a positive definite solution $\bar{F}$. Consider the Lyapunov function candidate $L_k = E_{\tau,\gamma}(\bar{X}_k^T\bar{P}\bar{X}_k)$, with $\bar{P}$ being positive definite. The first difference, using the overall system dynamics with optimal control input is $\Delta L_k = -E_{\tau,\gamma}(\bar{X}_k^T\bar{F}\bar{X}_k)$. Since, $\bar{F}$ can be chosen as a diagonal matrix, the first difference in terms of the subsystems can be expressed as

$$\Delta L_k = -\sum_{i=1}^{N} \underset{\tau,\gamma}{E}(\bar{x}_{i,k}^T\bar{F}_i\bar{x}_{i,k}) \leq -\sum_{i=1}^{N}\bar{q}_{\min}\underset{\tau,\gamma}{E}\left\|\bar{x}_{i,k}\right\|^2 \tag{15}$$

where $\bar{q}_{\min}$ is the minimum singular value of $\bar{F}$. This implies the subsystems are asymptotically stable in the mean square. The results of this lemma will be used in the stability analysis of the interconnected system where the need for the accurate knowledge of $A_{\bar{x}k}, B_{\bar{x}k}$ will be relaxed. The controller design using a novel hybrid Q-learning-based ADP approach for such a large-scale interconnected system in the presence of network-induced losses and with intermittent feedback will be discussed as follows. □

3 Distributed event-based hybrid Q-learning scheme

In this section, a novel hybrid learning scheme, which utilises time-driven Q-learning-based ADP approach, for the control of large-scale interconnected system to improve the convergence time with event-sampled state and input vector will be introduced. In the proposed algorithm, the idle-time between two events is utilised to perform limited parameter updates iteratively in order to minimise the Bellman error. With the finite number of iterations between any two events varying, the control policy need not necessarily converge to an admissible policy and the stability of the closed-loop system cannot be established either using the traditional iterative ADP schemes [4] or the time-driven Q-learning schemes [6, 7].

An additional challenge is to estimate the Q-function parameters in (11) for the system defined in (5) with intermittent feedback and in the presence of network-induced losses. Since subsystems broadcast their states via the communication network, each local subsystem can estimate the Q-function of the overall system so that a predefined reference model is not needed. Subsequently, the optimal control gains and the decoupling gains for each subsystem can be computed without using the complete knowledge of the system dynamics and interconnection matrix.

Although, the estimation of the Q-function at each subsystem increases the computation, this additional computation can be considered as trade-off for relaxing the assumption on the strength of interconnection terms and estimating optimal control policy. With the following assumption, the QFE design will be presented for intermittent feedback.

Assumption 3: The target parameters are assumed to be slowly varying [29].

3.1 Time-driven Q-learning with intermittent feedback

In the case of an event-sampled system, the system state vector $\bar{X}_k$ is sent to the controller at event-sampled instants. To denote the event-sampling instants, we define a sub-sequence $\{k_l\}_{l\in\mathbb{N}}, \forall k \in \{0, \mathbb{N}\}$ with $k_0 = 0$ being the initial sampling instant and $\mathbb{N}$ is the set of natural numbers. The system state vector $\bar{X}_{k_l}$ sent to the controller is held by zero order hold (ZOH) until the next sampling instant, and it is expressed as $\bar{X}_k^e = \bar{X}_{k_l}, k_l \leq k < k_{l+1}$. The corresponding error referred to as an event-sampling error can be expressed as

$$e_{\mathrm{ET}}(k) = \bar{X}_k - \bar{X}_k^e, \quad k_l \leq k < k_{l+1}, \quad l = 1, 2, \ldots \tag{16}$$

Since the estimation of G_k must use $\bar{X}_k^e$, the QFE can be expressed as

$$\hat{Q}(\bar{X}_k^e, U_k) = \underset{\tau,\gamma}{E}(z_k^{e,T}\hat{G}_k z_k^e) = \underset{\tau,\gamma}{E}(\hat{\Theta}_k^T\xi_k^e), \quad k_l \leq k < k_{l+1} \tag{17}$$

where $z_k^e = [(\gamma_{sc,k}\bar{X}_k^e)^T \quad U_k^T]^{\mathrm{T}} \in \Re^{\bar{l}}$ and $\xi_k^e = z_k^{e,T} \otimes z_k^e$ being the event-sampled regression vector and $\hat{\Theta}$ is the result of vectorisation of the matrix $\hat{G}_k$. The Bellman error calculated with event-sampled state is

$$e_{\mathrm{B}}(k) = \underset{\tau,\gamma}{E}\left[r(\bar{X}_k^e, U_k) + \hat{\Theta}_k^T\Delta\xi_k^e\right], \quad k_l \leq k < k_{l+1} \tag{18}$$

where $r(\bar{X}_k^e, U_k) = \bar{X}_k^{e,T}P_{\bar{x}}\bar{X}_k^e + U_k^T R_{\bar{x}}U_k$, $\Delta\xi_k^e = \xi_{k+1}^e - \xi_k^e$. The Bellman error (18) can be rewritten as

$$e_{\mathrm{B}}(k) = \underset{\tau,\gamma}{E}\left\{r(\bar{X}_k, U_k) + \hat{\Theta}_k^T\Delta\xi_k + \Xi_s\left(\bar{X}_k, e_{\mathrm{ET}}(k), \hat{\Theta}_k\right)\right\} \tag{19}$$

where $\Xi_s\left(\bar{X}_k, e_{\mathrm{ET}}(k), \hat{\Theta}_k\right) = r(\bar{X}_k - e_{\mathrm{ET}}(k), U_k) - r(\bar{X}_k, U_k) + \hat{\Theta}_k^T(\Delta\xi_k^e - \Delta\xi_k)$.

Remark 3: By comparing (19) with (13), the Bellman error in (19) has an additional error term which is $\Xi_s\left(\bar{X}_k, e_{\mathrm{ET}}(k), \hat{\Theta}_k\right)$. This additional error consists of errors in cost-to-go, and the regression vector, which are driven by $e_{\mathrm{ET}}(k)$. Hence, the estimation of QFE parameters depends upon the frequency of the event-sampling instants.

The QFE estimated parameter vector, $\hat{\Theta}_k^i$, is tuned only at the event-sampling instants. The superscript i denotes the overall system parameters at the ith subsystem and the estimated control policy can be computed as

$$U_k^i = -\hat{K}_k^i\bar{X}_k^{i,e} = -(\hat{G}_k^{i,uu})^{-1}(\hat{G}_k^{i,ux})\bar{X}_k^{i,e} \tag{20}$$

By using (20), the event-based estimated control input for the ith subsystem is given by

$$u_{i,k} = -\hat{K}_{i,k}\bar{x}_{i,k}^e - \sum_{j=1, j\neq i}^{N}\hat{K}_{ij,k}\bar{x}_{j,k}^e, \quad k_l \leq k < k_{l+1}$$
$$\forall i \in \{1, 2, \ldots, N\} \tag{21}$$

Remark 4: It should be noted that the optimal controllers designed at each subsystem takes into account the structural constraint which are present in the form of the interconnection matrix. However, the consideration of input, state and time constraints [1] as a part of the optimal control problem is reserved for future work.

With the following assumption, the parameter update rule for the QFE will be presented.

Assumption 4: The target parameter vector Θ_k is assumed to be bounded by positive constant, such that $\|\Theta_k\| \leq \Theta_M$. The regression function $Z^i(\bar{X}_k)$ is locally Lipschitz for all $\bar{X}_k \in \Omega_x$.

3.2 Parameter update at event-sampling instants

The QFE parameter vector $\hat{\Theta}_k^i$, is tuned by using the past data of the Bellman error (19) that is available at the event-sampling instants. Therefore, the auxiliary Bellman error at the event-sampling instants is expressed as $\Xi_B{}^{i,e}(k) = \Pi_k^{i,e} + \hat{\Theta}_k^{i,T} Z_k^{i,e}$, for $k = k_l$, where $\Pi_k^{i,e} = [r(\bar{X}_{k_l^i}^i, U_{k_l^i}^i) r(\bar{X}_{k_{l-1}^i}^i, U_{k_{l-1}^i}^i, \ldots, r(\bar{X}_{k_{l-\nu-1}^i}^i, U_{k_{l-\nu-1}^i}^i)] \in \Re^{1\times\nu}$ and $Z_k^{i,e} = [\Delta\xi_{k_l^i}^i, \Delta\xi_{k_{l-1}^i}^i, \ldots, \Delta\xi_{k_{l-\nu-1}^i}^i] \in \Re^{l_g\times\nu}$.

Remark 5: A larger time history may lead to faster convergence, but it results in higher computation. The number of history values ν is not fixed and a value $\nu < l$ is found suitable during simulation studies.

Next, select the update law [29] for the QFE parameter vector $\hat{\Theta}_k^i$ tuned only at the event-sampling instants, as

$$\hat{\Theta}_k^i = \hat{\Theta}_{k-1}^i + \frac{W_{k-2}^i Z_{k-1}^{i,e} \Xi_B^{i,e^T}(k-1)}{1 + Z_{k-1}^{i,e^T} W_{k-2}^i Z_{k-1}^{i,e}}, \quad k = k_l \tag{22}$$

where

$$W_k^i = W_{k-1}^i - \frac{W_{k-1}^i Z_{k-1}^{i,e} Z_{k-1}^{i,e^T} W_{k-1}^i}{1 + Z_{k-1}^{i,e^T} W_{k-1}^i Z_{k-1}^{i,e}}, \quad k = k_l \tag{23}$$

with $W_0^i = \beta I$, $\beta > 0$, a large positive value. The aperiodic execution of (22), saves computation, when compared to the traditional adaptive Q-learning techniques. The superscript i indicating the overall system parameters at the ith subsystem, will be dropped from hereon. In the time-driven Q-learning scheme [6], the parameters of the QFE are not updated during the inter-event period. On the contrary, in the hybrid learning algorithm, the parameters are updated during the inter-event period and the update rules are presented next.

3.3 Iterative parameter update

The recursive least square (RLS) algorithm was used in [2, 4] to perform iterative updates within any two periodic sampling instants, using policy iteration. The update equation iteratively searches for a control policy that minimises the Bellman error. Analytical results are provided in [2, 4] to show that each iterative update resulted in a control policy that is better than or as good as the existing control policy, in minimising the Bellman error.

Since a significant number of iterative updates are not viable for online control, the time-driven Q-learning [6, 7] was proposed which uses gradient descent-based update equations at the sampling instants to minimise the Bellman error. It was shown in [6, 7] that as the sampling instants increases, the parameter estimation error converges to zero. To improve the estimation error convergence rate, the RLS update (22) and (23) is used at the sampling instants in this work and convergence result similar to the time-driven Q-learning holds for the proposed algorithm without the iterative parameter updates presented next.

To utilise the time between two event-sampling instants, parameters are updated iteratively to minimise the error that was calculated during the previous event, which is expressed as $\Xi_B{}^{j,e}(k) = \Pi_k^{j,e} + \hat{\Theta}_k^{j,T} Z_k^{j,e}, k = k_l$, where j is the iteration index. The Q-function parameters are updated using the equations

$$\hat{\Theta}(k_l^j) = \hat{\Theta}(k_l^{j-1}) + \frac{W(k_{l-2}^{j-2}) Z(k_{l-1}^{j-1}) \Xi_B^T(k_{l-1}^{j-1})}{1 + Z^T(k_{l-1}^{j-1}) W(k_{l-2}^{j-1}) Z(k_{l-1}^{j-1})} \tag{24}$$

$$W(k_l^j) = W(k_{l-1}^{j-1}) - \frac{W(k_{l-1}^{j-1}) Z(k_{l-1}^{j-1}) Z^T(k_{l-1}^{j-1}) W(k_{l-1}^{j-1})}{1 + Z^T(k_{l-1}^{j-1}) W(k_{l-1}^{j-1}) Z(k_{l-1}^{j-1})} \tag{25}$$

Whenever there is an event, the Q-function parameter vector which is updated iteratively using (24) and (25) is passed on to the QFE to calculate the new Bellman error. The estimated control gain matrix can be obtained from the estimated parameter vector $\hat{\Theta}_k$ in (22) at each event-sampled instants. In terms of the estimated parameters, the control gains are given by (20), where

$$\hat{K}_k = (\hat{G}_k^{uu})^{-1} \hat{G}_k^{ux} = \begin{bmatrix} \hat{K}_1 & \cdots & \hat{K}_{1N} \\ \vdots & \ddots & \vdots \\ \hat{K}_{N1} & \cdots & \hat{K}_N \end{bmatrix} \tag{26}$$

is the estimated control gain. It is important to note that this control gain is obtained directly from the Q-function parameters which are constructed with the past data and the current feedback information, without using the system dynamics.

In the proposed algorithm, the update equations (24) and (25) together with (18) search for an improved control policy during every inter-event period. Utilising the Bellman error equation (18) to evaluate the existing control policy, the Q-function is iteratively updated between two event-sampling instants. However, in contrast to the algorithms in [2, 4], the iteration index j in (24) and (25) depends on the event-sampling mechanism, resulting in finite, varying number of iterative updates between any two events.

Remark 6: The control policy for the individual subsystem is given by (21). Since it is possible that $\hat{G}_k^{uu}$ might be rank-deficient during the learning phase, the following conditions are checked before the control law is updated. If $\hat{G}_{k-1}^{uu}$ is singular or if $\hat{G}_{k-1}^{uu} - R_{\bar{x}}$ is not positive definite, then, $\hat{G}_{k-1}^{uu}$ is replaced by $R_{\bar{x}}$ in the control policy. The conditions can be checked easily by calculating the eigenvalues of $\hat{G}_{k-1}^{uu}$.

Remark 7: The QFE parameter tuning law (22) and (23) requires the state vectors X_{k_l} to $X_{k_{l-\nu-1}}$ for the computation of regression vector at $k = k_l$. Therefore, the past values are required to be stored at the value function estimator.

With the update rules presented in this section and the control gains selected from (26), the assumption in [6, 7] that the inverse of $\hat{G}_k^{uu}$ exists when the updates utilise the time history of the regression function and Bellman error is also relaxed. The analytical results for the proposed learning algorithm is presented next.

3.4 Stability analysis

Defining the QFE parameter estimation error $E_{\tau,\gamma}(\tilde{\Theta}_k) = E_{\tau,\gamma}(\Theta_k - \hat{\Theta}_k)$, the error dynamics using (22), (24) can be represented as

$$\underset{\tau,\gamma}{E}(\tilde{\Theta}_{k_{l+1}}^0) = \underset{\tau,\gamma}{E}\left(\tilde{\Theta}_k^j + \frac{W_k^j Z_k^{j,e} \Xi_{\mathrm{B}}^{j,e^T}(k)}{1 + Z_k^{ej,T} W_k^j Z_k^{j,e}}\right), \quad k = k_l^0 \tag{27}$$

$$\underset{\tau,\gamma}{E}(\tilde{\Theta}_{k_l}^{j+1}) = \underset{\tau,\gamma}{E}\left(\tilde{\Theta}_k^j + \frac{W_k^j Z_k^{j,e} \Xi_{\mathrm{B}}^{j,e^T}(k)}{1 + Z_k^{ej,T} W_k^j Z_k^{j,e}}\right), \quad k_l^0 < k < k_{l+1}^0 \tag{28}$$

Remark 8: When there is no data-loss, the QFE is updated and the control policy is updated as soon as it is computed. This requires the broadcast scheme to generate an acknowledgment signal whenever the packets are successfully received at the subsystems [22]. A suitable scheduling protocol has to ensure that the data lost in the network is kept minimal.

Next an event-sampling condition has to be selected for the proposed scheme to work. Consider a quadratic function $f^i(k) = \bar{x}_i(k)^T \Gamma_i \bar{x}_i(k)$, with $\Gamma_i > 0$, for the ith subsystem. The event-sampling condition should satisfy

$$f^i(k) \le \lambda f^i(k_l + 1), \quad \forall k \in [k_l + 1, k_{l+1}), \tag{29}$$

for stability, when $\lambda < 1$, as shown in the following section.

IET Control Theory Appl., 2016, Vol. 10, Iss. 12, pp. 1448–1457

Remark 9: The event-sampling condition presented here depends only on the local subsystem state information. The Lyapunov function-based event-sampling condition is also presented in [23] for a single system. The hybrid learning algorithm presented in this paper is independent of the event-sampling condition.

The following result will be used to prove the stability of the closed-loop system during the learning period.

Lemma 2: Consider the system in (5) and the QFE (17). Define $\tilde{U}(k_{l-1}) = U(k_{l-1}) - \hat{U}(k_{l-1})$ and $\tilde{G}^{ux}_{k_{l-1}} = G^{ux}_{k_{l-1}} - \hat{G}^{ux}_{k_{l-1}}$. If the control policy is updated such that, whenever $\hat{G}^{uu}_{k_{l-1}} - R_{\bar{x}}$ is not positive definite or $\hat{G}^{uu}_{k_{l-1}}$ is singular, $\hat{G}^{uu}_{k_{l-1}}$ is replaced by $R_{\bar{x}}$ in the control policy, then

$$\underset{\tau,\gamma}{E}\left(\tilde{U}(k_{l-1})\right) \leq \underset{\tau,\gamma}{E}\left\{2\left\|R_{\bar{x}}^{-1}\right\|\left\|G^{ux}_{k_{l-1}}\right\|\left\|\bar{X}_{k_{l-1}}\right\| + \left\|R_{\bar{x}}^{-1}\right\|\left\|\tilde{G}^{ux}_{k_{l-1}}\right\|\left\|\bar{X}_{k_{l-1}}\right\|\right\} \tag{30}$$

Proof: See the Appendix. □

Definition 1 [29]: A regression vector $\varphi(x_k)$ is said to be persistently exciting if there exists positive constants δ, $\underline{\alpha}$, $\bar{\alpha}$ and $k_d \geq 1$ such that $\underline{\alpha} I \leq \sum_{k=k_d}^{k+\delta} \varphi(x_k)\varphi^T(x_k) \leq \bar{\alpha} I$, where I is the identity matrix of appropriate dimension.

Lemma 3: Consider both the QFEs in (17) with an initial admissible control policy $U_0 \in \Re^m$. Let Assumptions 1–4 hold, and the QFE parameter vector $\hat{\Theta}(0)$ be initialised in a compact set Ω_Θ. When the QFE is updated at the event-sampling instants using (22) and (23) and during the inter-sampling period using (24) and (25), the QFE parameter estimation error $E_{\tau,\gamma}(\tilde{\Theta}^j_{k_l})$ is bounded. Under the assumption that the regression vector $\xi^j_{k_l}$ satisfies the persistency of excitation (PE) condition, the QFE parameter estimation error $\tilde{\Theta}^j_{k_l}$ for all $\hat{\Theta}(0) \in \Omega_\Theta$ converges to zero asymptotically in the mean square, with event-sampled instants $k_l \to \infty$.

Proof: See the Appendix. □

Remark 10: Covariance resetting technique [29] is used to reset W whenever $W \leq W_{\min}$. This condition will also be used in the Lyapunov analysis to ensure stability of the closed-loop system. With the covariance resetting, the parameter convergence proof in Lemma 3 will still be valid [29].

Next, the Lyapunov analysis is used to derive the conditions for the stability of the closed-loop system, with the controller designed in this section.

Theorem 1: Consider the closed-loop system (5), parameter estimation error dynamics (27) along with the control input (20). Let Assumptions 1–4 hold, and let $U(0) \in \Omega_u$ be an initial admissible control policy. Suppose the last held state vector, $\bar{X}^{e,j}_{k_l}$, and the QFE parameter vector, $\hat{\Theta}^j_{k_l}$ are updated by using, (22) and (23) at the event-sampled instants, and (24) and (25) during the inter-sampling period. Then, there exists a constant $\gamma_{\min} > 0$ such that the closed-loop system state vector $\bar{X}^j_{k_l}$ for all $\bar{X}(0) \in \Omega_x$ converges to zero asymptotically in the mean square and the QFE parameter estimation error $\tilde{\Theta}^j_{k_l}$ for all $\hat{\Theta}(0) \in \Omega_\Theta$ remains bounded. Further, under the assumption that the regression vector $\xi^j_{k_l}$ satisfies the PE condition, the QFE parameter estimation error $\tilde{\Theta}^j_{k_l}$ for all $\hat{\Theta}(0) \in \Omega_\Theta$ converges to zero asymptotically in the mean square, with event-sampled instants $k_l \to \infty$, provided the inequality $\gamma_{\min} > \mu + \rho_1$ is satisfied. Further, the estimated Q-function $\hat{Q}(\bar{X}(k), U(k)) \to E_{\tau,\gamma}\{Q^*(\bar{X}(k), U(k))\}$ and estimated control input $U(k) \to E_{\tau,\gamma}\{U^*(k)\}$. μ, ρ_1 are positive constants, defined in the proof.

Proof: See the Appendix. □

The evolution of the Lyapunov function is depicted in Fig. 2*a*. During the event-sampled instant, due to the updated control policy (21), the Lyapunov function decreases. Due to the event-sampling condition (29) and the iterative learning within the event-sampling instants, the Lyapunov function decreases during the inter-sampling period.

Since the iterative learning does not take place in the time-driven Q-learning [6], the first difference of the parameter estimation error is zero for the inter-event period. This makes the Lyapunov function negative semi-definite during this period. The evolution of the Lyapunov function is depicted in Fig. 2*b* for the time-driven Q-learning.

Remark 11: The design constants $R_{\bar{x}}, W_{\min}, W_0$ are selected based on the inequalities that are analytically derived in Theorem 1 using the bounds on $A_{\bar{x}k}, B_{\bar{x}k}, S_k$. Then, the constants Γ and $\bar{\Pi}$ can be found to ensure closed-loop system stability.

Remark 12: The requirement of PE condition is necessary so that the regression vector is non-zero until the parameter error goes to zero. By satisfying the PE condition in the regression vector, the expected value of the parameter estimation error $\tilde{\Theta}_k$ will converge to zero. This PE signal is viewed as the exploration signal in the reinforcement learning literature [4].

Remark 13: An initial identification process can be used to obtain the nominal values of $A_{\bar{x}k}, B_{\bar{x}k}$ which can be used to initialise the Q-function parameters.

Remark 14: The algorithm proposed in this section can be used as a time-driven Q-learning scheme by not performing the iterative learning between the event-sampling instants, in stochastic framework. Also, if the iteration index, $j \to \infty$, for each k_l, the algorithm becomes the traditional policy iteration-based ADP scheme.

The event-sampling and broadcast algorithm for the subsystems followed by the proposed hybrid learning algorithm is summarised next.

3.5 Proposed algorithm

For estimating the overall Q-function locally, we will use the following request-based event-sampling algorithm. Consider an event occurring at the ith subsystem at the sampling instant k_l. This subsystem generates a request signal and broadcasts it with its state information to the other subsystems. Upon receiving the broadcast request, the other subsystems broadcast their respective state information to all the subsystems. This can be considered as a forced event at the other subsystems.

Remark 15: The events at all the subsystem occur asynchronously based on the local event-sampling condition, whereas the QFE and control policy remain synchronised at each subsystem due to the forced event. The request signal is considered to be broadcasted without any delay in Network 1 in Fig. 1.

The algorithm for the hybrid learning scheme is summarised in Fig. 3.

The proposed control scheme is tested via simulation and the results are presented next.

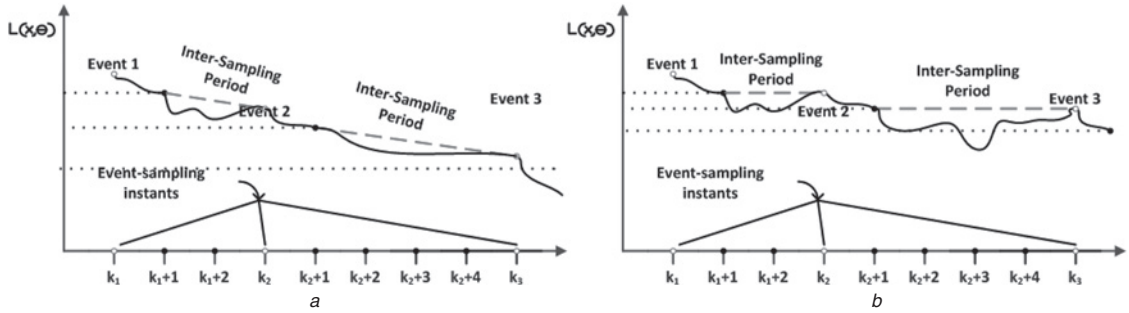

Fig. 2 *Evolution of the Lyapunov function*

a With hybrid learning scheme
b Without hybrid learning scheme

Algorithm 1

Initialise $\hat{\Theta}_0^j, W_0^j, U_0$
for Event-sampling instants: $l = 0 \rightarrow \infty$ **do**
  if Event = Yes **then**
    Calculate Bellman Error $e_B(k_l^j)$
    Update $\hat{\Theta}_{k_l}^j, W_{k_l}^j$
    Update the control input at the actuator U_{k_l}
    Pass the parameters $\hat{\Theta}_{k_l}^0, W_{k_l}^0, e_B(k_l^0)$ for iterations
  else
    for Iterative Index: $j = 0 \rightarrow \infty$ **do**
      Update $\hat{\Theta}_{k_l}^j, W_{k_l}^j$ with $e_B(k_l^j)$
      Calculate $e_B(k_l^{j+1})$
      if $e_B(k_l^{j+1})$-$e_B(k_l^j) < \epsilon$ or Event = Yes **then**
        Pass the Parameters $\hat{\Theta}_{k_l}^j, W_{k_l}^j$ to QFE
        Goto 4:
      end if
      $j = j + 1$
    end for
  end if
  if $e_B(k_{l+1}^0)$-$e_B(k_l^0) < \epsilon$ **then**
    Stop PE Condition
  end if
  $l = l + 1$
end for

Fig. 3 *Hybrid Q-learning for intermittent feedback*

4 Simulation results and discussion

A system of N interconnected inverted pendulums, coupled by a spring is considered for the verification of the analytical design. The dynamics are

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ \frac{g}{l} - \frac{a_i k}{ml^2} & 0 \end{bmatrix} x_i(t) + \begin{bmatrix} 0 \\ \frac{1}{ml^2} \end{bmatrix} u_i(t) + \sum_{j \in N_i} \begin{bmatrix} 0 & 0 \\ \frac{h_{ij} k}{ml^2} & 0 \end{bmatrix} x_j(t)$$

where $l = 2$, $g = 10$, $m = 1$, $k = 5$ and $h_{ij} = 1$ for $\forall j \in \{1, 2, \ldots, N\}$. The system is open loop unstable.

4.1 Ideal network

The system is discretised with a sampling time of 0.1 s. With $P_i = I_{2\times 2}$ and $R_i = 1$, $\forall i = 1, 2, 3$, the initial states for the system was selected as $x_1 = [2\ -3]^{\mathrm{T}}$, $x_2 = [-1\ 2]^{\mathrm{T}}$ and $x_3 = [-1\ 1]^{\mathrm{T}}$ and $W(0) = 500, \lambda = 0.6, W_{\min} = 250$. For the PE condition, Gaussian white noise with zero mean and 0.2 standard deviation was added to the control inputs. The initial parameters of the QFE are obtained by solving the SRE of the nominal model of the system. Under the ideal case, without network-induced losses, the comparison between the time-driven Q-learning against the proposed hybrid learning scheme shown in Fig. 4*a*, verifies that the convergence rate is faster in the hybrid learning scheme with event-sampled feedback. This is due to the iterative parameter update within the inter-event period.

Table 1 Comparison of parameter error convergence time

Mean-delay, ms	%Data-drop out	Convergence time, s	
		Time-driven Q-learning	Hybrid learning algorithm
0	0	10.7	13.6
30	10	61.0	36.9
	25	246	190.0
80	10	632.0	317.0
	25	486.5	269.0
100	10	239.0	198.0
	25	637.3	239.8

4.2 Monte Carlo analysis

The simulation is carried out with random delays ($\bar{d} = 2$) introduced by the network. The delay is characterised by normal distribution with 80 ms mean and 10 ms standard deviation and a Monte Carlo analysis is carried out for 500 iterations. In the case where the random delays are considered, the state and control trajectories are stable during the learning period as seen in Fig. 4*b*. The comparison between the time-driven Q-learning and the proposed hybrid learning schemes as seen in Fig. 5*b* shows that parameter error convergence in the hybrid scheme is much faster, which shows that the hybrid learning algorithm is more robust than the time-driven Q-learning in the presence of delays. This is partly due to augmented state vector and iterative parameter learning within the event-sampled instants.

Random packet-losses characterised with Bernoulli distribution is introduced keeping the probability of data lost as 10%. All design parameters are kept the same. Table 1 lists the convergence time for the parameter estimation error for the existing time-driven Q-learning algorithm and the proposed hybrid learning algorithm. The error threshold was defined as 10^{-2} and the design parameters were unchanged. In the ideal case, when there are no network losses, the difference in the convergence time for the two algorithms is small. As the network losses are increased, the parameter error converges to the threshold much faster with the proposed hybrid learning algorithm. It is clear that with the hybrid learning scheme the estimation error converges much quicker than the time-driven Q-learning scheme per the information given in Table 1.

The total number of events, the state and control policies during the learning period are shown in Figs. 6*a* and *b*, respectively. With the hybrid learning algorithm, the stability of the system is not affected during the learning period. As the events are spaced out, the more number of iterative parameter updates takes place within the inter-event period. Simulation figures for all the cases are not included due to space consideration.

IET Control Theory Appl., 2016, Vol. 10, Iss. 12, pp. 1448–1457

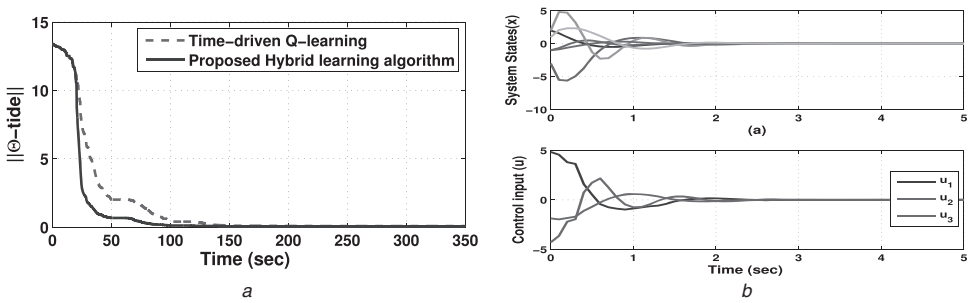

Fig. 4 *Comparison between the time-driven Q-learning against the proposed hybrid learning scheme*

a Estimation error comparison for ideal network
b Controller performance with delays

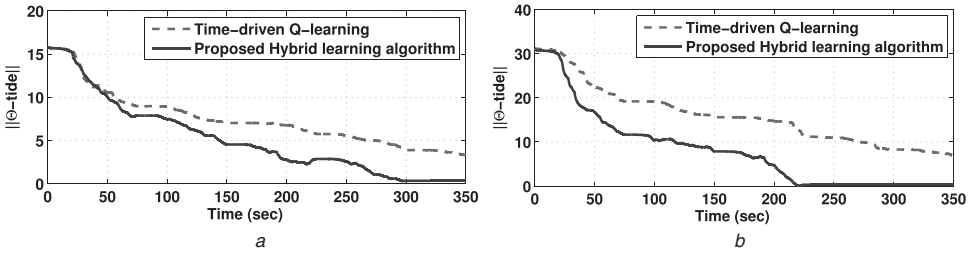

Fig. 5 *Estimation error comparison*

a With 10% packet loss
b Without packet loss

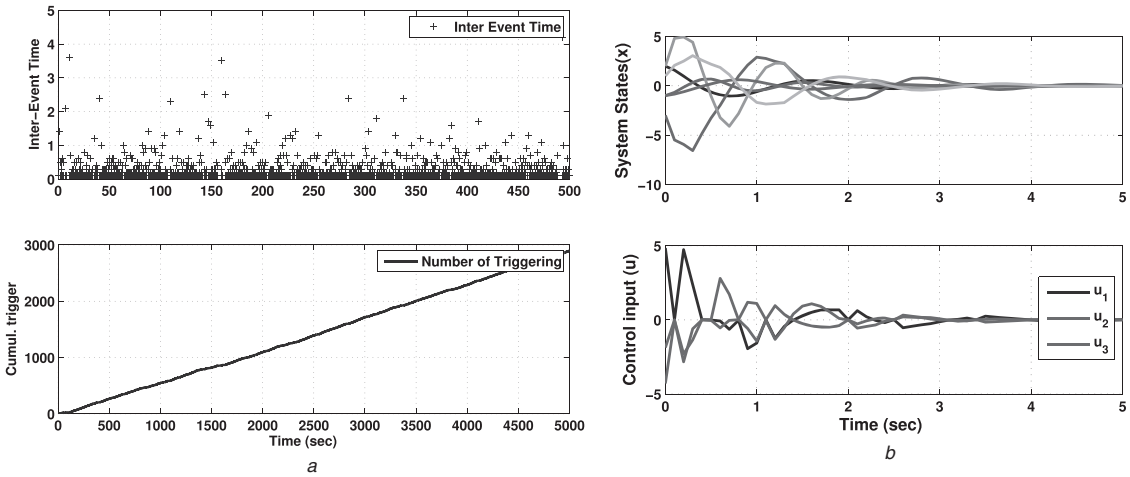

Fig. 6 *Total number of events, the state and control policies during the learning period*

a Inter-event time and cumulative trigger instants
b Controller performance with packet-loss

5 Conclusions

The proposed hybrid Q-learning-based scheme for a large-scale interconnected system appears to guarantee a desired performance. The stability conditions for the closed-loop system during the learning period is derived using the Lyapunov stability analysis. Q-function parameters for the entire system are estimated at each subsystem with the event-sampled inputs, states and past state vectors. This control scheme does not impose any assumptions on the interconnection strengths. The mirror estimator is not used in the event-sampling mechanism and reference models for each subsystems are not needed. With the help of the simulation study, the proposed analytical design is verified. From the simulation results, the proposed algorithm appears to provide advantages over the existing model-free Q-learning scheme for online control.

The proposed hybrid approach utilises past input and state information for each subsystems and state information from other systems via communication network and therefore the net result is the design of a data-driven optimal regulator for a class of large-scale interconnected systems.

6 Acknowledgments

This research supported in part by NSF ECCS #1128281 and #1406533 and Intelligent Systems Center, at the Missouri University of Science and Technology, Rolla.

7 References

1 Lewis, F.L., Syrmos, V.L.: ‘Optimal control’ (John Wiley & Sons, 1995)

2 Bradtke, S.J., Ydstie, B.E., Barto, A.G.: 'Adaptive linear quadratic control using policy iteration'. Proc. American Control Conf., July 1994, pp. 3475–3479
3 Al-Tamimi, A., Lewis, F.L., Abu-Khalaf, M.: 'Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control', *Automatica*, 2007, **43**, (3), pp. 473–481
4 Lewis, F.L., Vrabie, D., Vamvoudakis, K.G.: 'Reinforcement learning and feedback control: using natural decision methods to design optimal adaptive controllers', *IEEE Control Syst.*, 2012, **32**, (6), pp. 76–105
5 Dierks, T., Jagannathan, S.: 'Online optimal control of affine nonlinear discrete-time systems with unknown internal dynamics by using time-based policy update', *IEEE Trans. Neural Netw. Learn. Syst.*, 2012, **23**, (7), pp. 1118–1129
6 Sahoo, A., Jagannathan, S.: 'Event-triggered optimal regulation of uncertain linear discrete-time systems by using Q-learning scheme', Proc. IEEE 53rd Annual Conf. on Decision and Control, Los Angeles, CA, December 2014, pp. 1233–1238
7 Xu, H., Jagannathan, S., Lewis, F.L.: 'Stochastic optimal control of unknown linear networked control system in the presence of random delays and packet losses', *Automatica*, 2012, **48**, (6), pp. 1017–1030
8 Jamshidi, M.: 'Large-scale systems: modeling, control, and fuzzy logic' (Prentice-Hall, Inc., 1996)
9 Ioannou, P.: 'Decentralized adaptive control of interconnected systems', *IEEE Trans. Autom. Control*, 1986, **31**, (4), pp. 291–298
10 Šiljak, D.D., Zečević, A.I.: 'Control of large-scale systems: beyond decentralized feedback', *Ann. Rev. Control*, 2005, **29**, (2), pp. 169–179
11 Mehraeen, S., Jagannathan, S.: 'Decentralized optimal control of a class of interconnected nonlinear discrete-time systems by using online Hamilton–Jacobi–Bellman formulation', *IEEE Trans. Neural Netw.*, 2011, **22**, (11), pp. 1757–1769
12 Liu, D., Wang, D., Li, H.: 'Decentralized stabilization for a class of continuous-time nonlinear interconnected systems using online learning optimal control approach', *IEEE Trans. Neural Netw. Learn. Syst.*, 2014, **25**, (2), pp. 418–428
13 Narendra, K.S., Mukhopadhyay, S.: 'To communicate or not to communicate: A decision-theoretic approach to decentralized adaptive control'. Proc. American Control Conf., Baltimore, MD, July 2010, pp. 6369–6376
14 Venkat, A.N., Hiskens, I.A., Rawlings, J.B., *et al.*: 'Distributed MPC Strategies with application to power system automatic generation control', *IEEE Trans. Control Syst. Technol.*, 2008, **16**, (6), pp. 1192–1206
15 Camponogara, E., De, L., Marcelo, L.: 'Distributed optimization for MPC of linear networks with uncertain dynamics', *IEEE Trans. Autom. Control*, 2012, **57**, (3), pp. 804–809
16 Song, Y., Fang, X.: 'Distributed model predictive control for polytopic uncertain systems with randomly occurring actuator saturation and packet loss', *IET Control Theory Appl.*, 2014, **8**, (5), pp. 297–310
17 Zhou, X., Li, C., Huang, T., *et al.*: 'Fast gradient-based distributed optimisation approach for model predictive control and application in four-tank benchmark', *IET Control Theory Appl.*, 2015, **9**, (10), pp. 1579–1586
18 Zheng, Y., Li, S., Qiu, H.: 'Networked coordination-based distributed model predictive control for large-scale system', *IEEE Trans. Control Syst. Technol.*, 2013, **21**, (3), pp. 991–998
19 Wang, X., Hong, Y., Huang, J., *et al.*: 'A distributed control approach to a robust output regulation problem for multi-agent linear systems', *IEEE Trans. Autom. Control*, 2010, **55**, (12), pp. 2891–2895
20 Chen, M.Z.Q., Liangyin, Z., Su, H., *et al.*: 'A distributed control approach to a robust output regulation problem for multi-agent linear systems', *IET Control Theory Appl.*, 2015, **9**, (5), pp. 755–765
21 Wang, X., Lemmon, M.D.: 'Event-triggering in distributed networked control systems', *IEEE Trans. Autom. Control*, 2011, **56**, (3), pp. 586–601
22 Guinaldo, M., Lehmann, D., Sanchez, J., *et al.*: 'Distributed event-triggered control with network delays and packet losses', Proc. IEEE 51st Annual Conf. on Decision and Control, Maui, HI, July 2012, pp. 1–6
23 Meng, X., Chen, T.: 'Event-driven communication for sampled-data control systems'. Proc. American Control Conf., Washington, DC, June 2013, pp. 3002–3007
24 Halevi, Y., Ray, A.: 'Integrated communication and control systems: Part I – analysis', *J. Dyn. Syst. Meas. Control*, 1988, **110**, (4), pp. 367–373
25 Zhang, W., Branicky, M.S., Phillips, S.M.: 'Stability of networked control systems', *IEEE Control Syst.*, 2001, **21**, (1), pp. 84–99
26 Shousong, H., Qixin, Z.: 'Stochastic optimal control and analysis of stability of networked control systems with long delay', *Automatica*, 2003, **39**, (11), pp. 1877–1884
27 Xiangnan, Z., Zhen, Ni, Haibo, He., *et al.*: 'Event-triggered reinforcement learning approach for unknown nonlinear continuous-time system'. Proc. Int. Joint Conf. on Neural Networks, Beijing, July 2014, pp. 3677–3684
28 Hou, Z., Wang, Z.: 'From model-based control to data-driven control: survey, classification and perspective', *Inf. Sci.*, 2013, **235**, (1), pp. 3–35
29 Goodwin, G.C., Sin, K.S.: 'Adaptive filtering prediction and control' (Courier Corporation, 2014)
30 Horn, R.A., Johnson, C.R.: 'Matrix analysis' (Cambridge University Press, 2012)

8 Appendices

8.1 Proof of Lemma 2

The control input U_{k_l} always satisfies $\|U_{k_l}\| < \|R_{\bar{x}}^{-1}\|\|\hat{G}_{k_l}^{ux}\|\|X_{k_l}\|$. From Remark 6, two different possible control laws can emerge: one possibility is $\hat{G}_{k_{l-1}}^{uu}$ is non-singular and $\hat{G}_{k_{l-1}}^{uu} > R_{\bar{x}}$. Therefore, $\|U_{k_{l-1}}\| = \|(\hat{G}_{k_{l-1}}^{uu})^{-1}\hat{G}_{k_{l-1}}^{ux}\bar{X}_{k_{l-1}}\| \leq \|(\hat{G}_{k_{l-1}}^{uu})^{-1}\|\|\hat{G}_{k_{l-1}}^{ux}\|\|\bar{X}_{k_{l-1}}\|$. Since, Frobenius norm is used [30], we get $\|U_{k_{l-1}}\| \leq \|R_{\bar{x}}^{-1}\|\|\hat{G}_{k_{l-1}}^{ux}\|\|\bar{X}_{k_{l-1}}\|$. For the other possible condition

$$\|U_{k_{l-1}}\| = \|R_{\bar{x}}^{-1}\hat{G}_{k_{l-1}}^{ux}X_{k_{l-1}}\| \leq \|R_{\bar{x}}^{-1}\|\|\hat{G}_{k_{l-1}}^{ux}\|\|\bar{X}_{k_{l-1}}\| \tag{31}$$

Therefore, the error in the control law can be written as $E_{\tau,\gamma}(\tilde{U}_{k_{l-1}}) = E_{\tau,\gamma}(U_{k_{l-1}}^{*} - \hat{U}_{k_{l-1}})$. Taking the norm operator, using the definitions from (20) to obtain

$$\|\underset{\tau,\gamma}{E}(\tilde{U}_{k_{l-1}})\| \leq \underset{\tau,\gamma}{E}\left\{\left(\|(G_{k_{l-1}}^{uu})^{-1}G_{k_{l-1}}^{ux}\| + \|(\hat{G}_{k_{l-1}}^{uu})^{-1}\hat{G}_{k_{l-1}}^{ux}\|\right)\|\bar{X}_{k_{l-1}}\|\right\} \tag{32}$$

Since $G_{k_{l-1}}^{uu} > R_{\bar{x}}$, we get $\|E_{\tau,\gamma}(\tilde{U}_{k_{l-1}})\| \leq E_{\tau,\gamma}\|R_{\bar{x}}^{-1}\|(\|G_{k_{l-1}}^{ux}\| + \|\hat{G}_{k_{l-1}}^{ux}\|)\|\bar{X}_{k_{l-1}}\|$. By using the triangle inequality after representing the estimates in terms of the parameter error, we obtain

$$\|\underset{\tau,\gamma}{E}(\tilde{U}_{k_{l-1}})\| \leq \underset{\tau,\gamma}{E}\left\{2\|R_{\bar{x}}^{-1}\|\|G_{k_{l-1}}^{ux}\|\|\bar{X}_{k_{l-1}}\| + \|R_{\bar{x}}^{-1}\|\|\tilde{G}_{k_{l-1}}^{ux}\|\|\bar{X}_{k_{l-1}}\|\right\} \tag{33}$$

Thus the required inequality is obtained.

8.2 Proof of Lemma 3 [29]

• During the event-sampling instants: let the Lyapunov candidate function be

$$L_{i,\tilde{\Theta}}(k_l^j) = \underset{\tau,\gamma}{E}\tilde{\Theta}^T(k_l^j)W^{-1}(k_{l-1}^j)\tilde{\Theta}(k_l^j) \tag{34}$$

where j is the iteration index. From (23), using matrix inversion lemma, we have

$$W(k_l^j) = \frac{W(k_{l-1}^j)}{1 + Z(k_l^j)W(k_{l-1}^j)Z^T(k_l^j)} \tag{35}$$

Substituting in the error dynamics (27), we get $\tilde{\Theta}(k_l^j) = W(k_{l-1}^j)W^{-1}(k_{l-2}^j)\tilde{\Theta}(k_{l-1}^j)$. Using the definition of $\tilde{\Theta}$ and using (35), for all the value function estimators, the first difference becomes

$$\sum_{i=1}^{N}\Delta L_{i,\tilde{\Theta}}(k_l^j) \leq -N\underset{\tau,\gamma}{E}\frac{\tilde{\Theta}^T(k_{l-1}^j)Z(k_{l-1}^j)Z^T(k_{l-1}^j)\tilde{\Theta}(k_{l-1}^j)}{1 + Z^T(k_{l-1}^j)W(k_{l-1}^j)Z(k_{l-1}^j)} \tag{36}$$

• During the inter-sampling instants: the parameters are updated iteratively using (24) and (25). Let the Lyapunov candidate function be (34). Using similar arguments as in the previous case, the first difference is

$$\Delta L_{i,\tilde{\Theta}} = -\underset{\tau,\gamma}{E}\frac{\tilde{\Theta}^T(k_{l-1}^{j-1})Z(k_{l-1}^{j-1})Z^T(k_{l-1}^{j-1})\tilde{\Theta}(k_{l-1}^{j-1})}{1 + Z^T(k_{l-1}^{j-1})W(k_{l-2}^{j-2})Z(k_{l-1}^{j-1})} \tag{37}$$

Since the regression vector can become zero, we can only conclude that the Lyapunov function (34) is negative semi-definite. However, if the regression vector satisfies PE condition (Definition 1)

$$0 < \frac{Z(k_{l-1}^{j-1})Z^T(k_{l-1}^{j-1})}{1 + Z^T(k_{l-1}^{j-1})W(k_{l-2}^{j-2})Z(k_{l-1}^{j-1})} \leq 1$$

in (36) and (37), this results in

$$\sum_{i=1}^{N}\Delta L_{i,\tilde{\Theta}}(k_l^j) \leq -N\kappa_{\min}\underset{\tau,\gamma}{E}\|\tilde{\Theta}^T(k_{l-1}^j)\|^2 \tag{38}$$

with $0 < \kappa_{\min} \leq 1$. Thus, with the regression vector satisfying PE condition, the parameter estimation error is strictly decreasing both

during the event-sampling instants and the inter-event period. This implies that as $k_l^j \to \infty$, the QFE parameter estimation error converges to zero asymptotically in the mean square. This completes the proof.

8.3 Proof of Theorem 1

Case 1: The periodic feedback case will be analysed first. Let the Lyapunov function be

$$L(\bar{X}, \tilde{\Theta}) = \underset{\tau,\gamma}{E} \bar{X}_{k-1}^T \Gamma \bar{X}_{k-1} + \underset{\tau,\gamma}{E} \bar{\Pi} \sum_{i=1}^{N} L_{i,\tilde{\Theta}} \tag{39}$$

$\bar{\Pi} = \eta(\|W_0\|\rho_2/N)$ with $\eta > 1$. Consider the first term, the first difference is written as $\Delta L_{\bar{x}} = E_{\tau,\gamma}\{\bar{X}_k^T \Gamma \bar{X}_k - \bar{X}_{k-1}^T \Gamma \bar{X}_{k-1}\}$. Substituting the system dynamics with the estimated control input, with the definition $K_k{}^* = \hat{K}_k - \tilde{K}_k$, we obtain

$$\Delta L_{\bar{x}} = \underset{\tau,\gamma}{E} \left\{ \bar{X}_{k-1}^T (A_{\bar{x}k,c}^T - (B_{\bar{x}k}\tilde{K}_{k-1})^T) \Gamma (A_{\bar{x}k,c} - B_{\bar{x}k}\tilde{K}_{k-1}) \bar{X}_{k-1} - \bar{X}_{k-1}^T \Gamma \bar{X}_{k-1} \right\} \tag{40}$$

where $A_{\bar{x}k,c} = A_{\bar{x}k} - B_{\bar{x}k}K_k^*$ is Schur with the optimal control policy U_k^* and there exists a positive definite solution $\bar{\Gamma}$ for the Lyapunov equation. The first difference is given by

$$\begin{aligned} \Delta L_{\bar{x}} \leq & -\gamma_{\min} \underset{\tau,\gamma}{E} \left\| \bar{X}_{k-1} \right\|^2 \\ & + 2 \underset{\tau,\gamma}{E} \| \bar{X}_{k-1}^T (B_{\bar{x}k}\tilde{K}_{k-1})^T \Gamma \| \left\| A_{\bar{x}k,c} \bar{X}_{k-1} \right\| \\ & + \underset{\tau,\gamma}{E} \| \Gamma B_{\bar{x}k} \tilde{K}_{k-1} \bar{X}_{k-1} \|^2 \end{aligned} \tag{41}$$

where $\gamma_{\min}$ is the minimum singular value of $\bar{\Gamma}$ and ϵ_2 is a positive constant. Applying Young's inequality, we obtain

$$\begin{aligned} \Delta L_{\bar{x}} \leq & -\gamma_{\min} \underset{\tau,\gamma}{E} \left\| \bar{X}_{k-1} \right\|^2 + \underset{\tau,\gamma}{E} \left\| \epsilon_2 A_c \bar{X}_{k-1} \right\|^2 \\ & + \underset{\tau,\gamma}{E} \left\| \left(\Gamma + \frac{\Gamma^2}{\epsilon_2} \right) B_{\max} \tilde{U}_{k-1} \right\|^2, \end{aligned}$$

Recalling Lemma 2

$$\begin{aligned} \underset{\tau,\gamma}{E} \| \tilde{U}(k-1) \|^2 \leq \underset{\tau,\gamma}{E} \Big\{ & 2 \| R_{\bar{x}}^{-1} \| \left\| G_{k-1}^{ux} \right\| \left\| \bar{X}_{k-1} \right\| \\ & + \| R_{\bar{x}}^{-1} \| \| \tilde{G}_{k-1}^{ux} \| \left\| \bar{X}_{k-1} \right\| \Big\}^2 \end{aligned} \tag{42}$$

Using Assumption 4, and with G_M as the bound on G_k^{ux}, we obtain

$$\begin{aligned} \leq \underset{\tau,\gamma}{E} \Big\{ & 4 G_M \| R_{\bar{x}}^{-1} \|^2 \left\| \bar{X}_{k-1} \right\|^2 + \| R_{\bar{x}}^{-1} \|^2 \| \tilde{G}_{k-1}^{ux} \|^2 \left\| \bar{X}_{k-1} \right\|^2 \\ & + 2\varepsilon \left\| \bar{X}_{k-1} \right\|^2 + \frac{2 G_M^2 \| R_{\bar{x}}^{-1} \|^4}{\varepsilon} \| \tilde{G}_{k-1}^{ux} \|^2 \left\| \bar{X}_{k-1} \right\|^2 \Big\} \end{aligned} \tag{43}$$

where ϵ is a positive constant. On simplification, it yields that (see (44))

$$\underset{\tau,\gamma}{E} \| \tilde{U}(k-1) \|^2 \leq \underset{\tau,\gamma}{E} \left\{ (4 G_M \| R_{\bar{x}}^{-1} \|^2 + 2\varepsilon) \left\| \bar{X}_{k-1} \right\|^2 + \left(\| R_{\bar{x}}^{-1} \|^2 + \frac{2 G_M^2 \| R_{\bar{x}}^{-1} \|^4}{\varepsilon} \right) \| \tilde{G}_{k-1}^{ux} \|^2 \left\| \bar{X}_{k-1} \right\|^2 \right\} \tag{44}$$

Using the fact that $E_{\tau,\gamma} \| \tilde{G}_{k-1}^{ux} \| < \underset{\tau,\gamma}{E} \| \tilde{G}_{k-1} \|$, we obtain (see (45))

$$\underset{\tau,\gamma}{E} \| \tilde{U}(k-1) \|^2 \leq \underset{\tau,\gamma}{E} \left\{ (4 G_M \| R_{\bar{x}}^{-1} \|^2 + 2\varepsilon) \left\| \bar{X}_{k-1} \right\|^2 + \left(\| R_{\bar{x}}^{-1} \|^2 + \frac{2 G_M^2 \| R_{\bar{x}}^{-1} \|^4}{\varepsilon} \right) \| \tilde{\Theta}_{k-1} \|^2 \left\| \bar{X}_{k-1} \right\|^2 \right\} \tag{45}$$

Using (45), the first difference of the Lyapunov function becomes

$$\begin{aligned} \Delta L_x \leq & -(\gamma_{\min} - \mu - \rho_1) \underset{\tau,\gamma}{E} \left\| \bar{X}_{k-1} \right\|^2 \\ & + \rho_2 \| \Gamma \| \underset{\tau,\gamma}{E} \| \tilde{\Theta}_{k-1} \|^2 \left\| \bar{X}_{k-1} \right\|^2, \end{aligned} \tag{46}$$

where

$$\rho_1 = \left\| \left(\Gamma + \frac{\Gamma^2}{\epsilon_2} \right) B_{\max} \right\|^2 (4 G_M \| R_{\bar{x}}^{-1} \|^2 + 2\varepsilon),$$

$$\rho_2 = \| \Gamma \| \left\| \left(1 + \frac{\Gamma}{\epsilon_2} \right) B_{\max} \right\|^2 \left(\| R_{\bar{x}}^{-1} \|^2 + \frac{2 G_M^2 \| R_{\bar{x}}^{-1} \|^4}{\varepsilon} \right),$$

$\mu = \|\epsilon_2 A_c\|^2$. Recalling Lemma 3, when $0 < \|\Gamma\| \leq W_{\min}$ (from Remark 10), substitute (35) in place of $\|\Gamma\|$ in (46). Using the fact that $\|W_0\| > \|W(k_l^j)\|, \forall k, l, j$, and since the history values are used, $\left\| Z_{k-1} \right\|^2 \geq \left\| \bar{X}_{k-1} \right\|^2$, then the first difference becomes

$$\begin{aligned} \Delta L \leq & -(\gamma_{\min} - \mu - \rho_1) \underset{\tau,\gamma}{E} \left\| \bar{X}(k-1) \right\|^2 \\ & - (\bar{\Pi} N - \|W_0\| \rho_2) \kappa_{\min} \underset{\tau,\gamma}{E} \| \tilde{\Theta}(k-1) \|^2, \end{aligned} \tag{47}$$

with $0 \leq \alpha \leq 1$. Substituting the value of $\bar{\Pi}$, the second term is always negative. Therefore, $L(k+1) < L(k), \forall k \in \mathbb{N}$.

Case 2: To extend the stability results for the event-based control scheme, it is required to prove that between any two aperiodic sampling instants, the Lyapunov function is non-increasing. Let the Lyapunov function be given by (39). Taking the first difference to get

$$\begin{aligned} \Delta L_k = & \underset{\tau,\gamma}{E} \left\{ \bar{X}_k^T \Gamma \bar{X}_k - \bar{X}_{k-1}^T \Gamma \bar{X}_{k-1} + \bar{\Pi} \sum_{i=1}^{N} \Delta L_{i,\tilde{\Theta}} \right\} \\ & k_l \leq k < k_{l+1}, \quad \forall l \in \mathbb{N} \end{aligned} \tag{48}$$

When the events occurring at k_l and $k_{l+1} = k_l + 1$, the Lyapunov function is decreasing due to (47). When the event-sampling does not occur consecutively at $k_l, k_l + 1$, the interval $[k_l, k_{l+1}) = [k_l, k_l + 1) \cup [k_l + 1, k_{l+1})$. During $[k_l, k_l + 1)$, the Lyapunov function is decreasing because of the control policy updated at k_l. In the interval $[k_l + 1, k_{l+1})$ due to the event-sampling algorithm the inequality in (29) is satisfied. Therefore, $\Delta L(\bar{X}, \tilde{\Theta}) = E_{\tau,\gamma}\{\bar{X}_k^T \Gamma \bar{X}_k - \lambda \bar{X}_{k_l+1}^T \Gamma \bar{X}_{k_l+1}\} + \Delta L_{i,\tilde{\Theta}}$. Using the results from Lemma 3 and for $\bar{\lambda} < 1$, we obtain

$$\Delta L_k = -(1 - \bar{\lambda}) \underset{\tau,\gamma}{E} \{\bar{X}_{k_l}^T \Gamma \bar{X}_{k_l}\} - N \kappa_{\min} \underset{\tau,\gamma}{E} \| \tilde{\Theta}^T (k_{l-1}^j) \|^2 \tag{49}$$

Therefore, $\Delta L(\bar{x}, \tilde{\Theta}) < 0$ during the inter-sampling period. From Lemma 1, $\sum_{i=1}^{N} \Delta L(\bar{x}_i, \tilde{\Theta}^i) < 0$. Combining Cases 1 and 2, the Lyapunov equation satisfies the following inequality

$$L(k_{l+1}) < L(k_l + 1) < L(k_l), \quad \forall \{k_l\}_{l \in \mathbb{N}} \tag{50}$$

This completes the proof.