Distributed Adaptive Optimal Regulation of Uncertain Large-Scale Linear Networked Control Systems Using Q-Learning

Vignesh Narayanan*, S Jagannathan*
* Department of Electrical and Computer Engineering
Missouri University of Science and Technology
Rolla, Missouri- 65401
email: vnxv4, sarangap@mst.edu

***Abstract*—A novel Q-learning approach is presented for the design of an adaptive optimal regulator for linear large-scale interconnected system. The subsystems communicate among each other through a communication network while another communication network is inserted within the feedback loop of each subsystem. The network induced random delays and data dropouts of the network in the feedback are modelled along with the system dynamics. Stochastic Q-learning is used to adaptively learn the Q-function parameters with periodic and intermittent feedback. For efficient parameter learning with event-sampled feedback, a novel hybrid learning algorithm is proposed. Boundedness of the estimated parameters and asymptotic convergence of state vector in the mean square is achieved and it is demonstrated using Lyapunov stability analysis. Moreover, if the regression function of the QFE is persistently exciting (PE), the estimated parameters converge to their expected target values. The proposed analytical design is validated using a numerical example via simulation.**

I. Introduction

Large scale systems [1]–[3] such as electric power grids, transportation and automotive systems, in general consist of geographically distributed subsystems. From the control point of view, these systems require substantially higher computational and communication resources to ensure stable operation with a desired performance. The conventional design approach for a decentralized controller is to decouple the interconnected subsystems by assuming that subsystems have weak coupling [3] and use only the local state vector to obtain the control policy. In some cases, an additional learning mechanism is designed at each subsystem to adaptively learn the coupling terms to counter the effect of interactions [2].

If the interactions between the subsystems are neglected during control design, they may destabilize the overall system when uncertainties or disturbances are present [1], [2]. For large scale systems, the effect of interconnection is difficult to know beforehand. In [2], the importance of communication between the subsystems is discussed and it is reported that when the subsystems do not share their sensor measurements with each other and instead use reference models to obtain the state information of other subsystems, unsatisfactory transient performance can be observed.

The event-based sampling and controller execution is demonstrated to have advantageous over periodic time-driven sampling counterpart in terms of communication and computational costs [3]–[5]. The aperiodic event-based sampling time-instants are dynamically decided using an event-sampling condition such that stable operation of the system is preserved. Similar event-sampled framework for the design of controllers was extended to large scale interconnected systems in [3] with the objective to stabilize the subsystems assuming weak interconnections and known subsystem dynamics.

Optimal control [6] based on adaptive dynamic programming (ADP) [5], [7], [8] can be utilized to obtain forward-in-time solution for the optimal control problems when the system dynamics are uncertain. Among the ADP based Q-learning schemes, the authors in [5], [8] proposed a time-based algorithm to solve the ADP and Q-learning based optimal control, in an on-line manner without using iterations. Such a formulation for a large scale system with interconnected subsystems requires the inclusion of a communication network through which the information of state vector of each subsystem will be shared among them.

Networked control systems [8]–[10] (NCS) integrates the effects of communication with the control systems. The communication network in the feedback loop introduces random time delays and packet data dropouts, which might degrade the control performance. In [9], various issues related to NCS are discussed in detail. To the best knowledge of the authors, a time-based Q-learning scheme with intermittent feedback for a large-scale interconnected system with network induced losses is not reported.

In view of the above, in this paper, the sampled data modelling approach which uses augmented states and past input values introduced in [9] to include network induced delays, and extended in [8] to include the packet data drop outs is used to model the large-scale interconnected system. Stochastic Q-learning technique [8] is used to compute the control gains without a priori knowledge of the large scale system dynamics including the interconnection terms with periodic and intermittent feedback. Due to the network losses and aperiodic parameter updates, the learning process of the

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.92

Q-function slows down and to counter this effect a novel hybrid learning algorithm is introduced. Analytical results are presented for the proposed design and they are verified with a numerical example via simulation. In this paper, Euclidean norm for the case of vectors and Frobenius norm for matrices are used.

II. System Description

Consider a linear time invariant continuous-time system having N interconnected subsystems, each of the form

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

where $\Re$ denotes the set of all real numbers. $x_i, \dot{x}_i \in \Re^{n_i \times 1}$ are the states and state derivatives of the i^{th} subsystem respectively. $u_i \in \Re^{m_i}, A_i \in \Re^{n_i \times n_i}, B_i \in \Re^{m_i \times 1}, A_{ij} \in \Re^{n_i \times n_j}$ denote control input, internal dynamics and control gain matrices of the i^{th} subsystem, A_{ij} represents the interconnection dynamics between the i^{th} and j^{th} subsystem. The overall system description can be expressed in a compact form as

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

where $X \in \Re^n, U \in \Re^m, B \in \Re^{n\times m}, A \in \Re^{n\times n}, \dot{X} = [\dot{x}_1, ., \dot{x}_N]^T$, $A = \begin{pmatrix} A_1 & ... & A_{1N} \\ : & ... & : \\ A_{N1} & ... & A_N \end{pmatrix}, B = diag[B_1, ., B_N], U = [u_1, ., u_N]^T$. The system dynamics are considered uncertain with the following assumption.

Assumption 1. *The system in (2) is considered controllable and the states are measurable with the control coefficient matrix satisfying $\|B\| \leq B_{\max}$, where $B_{\max} > 0$ being a known constant. Further, the order of subsystems is considered known.*

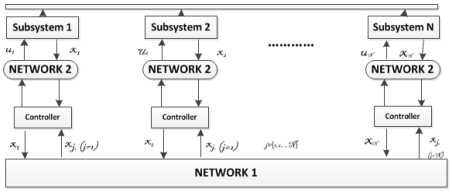

Fig. 1: Large-scale interconnected system

The Fig. 1 shows the block diagram of the large-scale system considered in this paper. The effects of the network induced delays and dropouts can be modelled along with the system dynamics, with the standard assumptions [8], [9]. With the network induced delays and data-dropout in Network 2, the original plant can be represented as,

$$\dot{X}(t) = AX(t) + \gamma_{ca}(t) BU(t - \tau(t)) \qquad (3)$$

where $\gamma_{ca}(t)$ indicates if the control input is lost in the transmission or not and $\tau(t)$ is the total loop delay.

Integrating the system (3) as in [8], we can define an augmented state vector consisting of states and past control inputs as $\bar{X}(k) = [X_k^T ... U_{k-d}^T] \in \Re^{n+lm}$.The new augmented system representation is given as

$$\bar{X}_{k+1} = A_{zk}\bar{X}_k + B_{zk}U_k \qquad (4)$$

with the matrices

$$A_{\bar{x}k} = \begin{bmatrix} A_d & \gamma_{ca,k-1}B_1^k & \cdots & \gamma_{ca,k-l}B_d^k \\ 0 & \cdots & 0 & 0 \\ \vdots & I_m & \vdots & \vdots \\ 0 & \cdots & I_m & 0 \end{bmatrix}, B_{\bar{x}k} = \begin{bmatrix} \gamma_{ca,k}B_0^k \\ I_m \\ \vdots \\ 0 \end{bmatrix}$$

Remark 1. If the delay is constant and there are no data dropouts, the system description will still be linear time-invariant. Since the system matrices in (4) are no longer linear time-invariant, the assumptions regarding the controllability and observability are now dependent upon the respective Grammian functions [6], therefore, we need the following assumption to proceed further.

Assumption 2. *[6] The system is uniformly completely observable and uniformly completely controllable.*

The Q-learning and adaptive optimal regulation for the NCS is presented next.

III. Problem Formulation

For the stochastic system dynamics (4), now represented with the matrices which are functions of stochastic variables, the cost function over the infinite time horizon, is defined as

$$J_k = \underset{\tau,\gamma}{E}\left[\sum_{t=k}^{\infty} \bar{X}_t^T P_{\bar{x}} \bar{X}_t + U_t^T R_{\bar{x}} U_t\right] \qquad (5)$$

where, $P_{\bar{x}} = diag(P, \frac{R}{d}, .., \frac{R}{d}), R_{\bar{x}} = \frac{R}{d}$.The penalty matrices are positive semi-definite and positive definite respectively with P, R being the penalty matrices of the original system states and control input defined in (2), $\underset{\tau,\gamma}{E}(\beta)$ denotes the expected value of the stochastic process β which is a function of stochastic variables γ, τ.

The stochastic cost function [6] can be represented as $J_k = \underset{\tau,\gamma}{E}[\bar{X}_k^T S_k \bar{X}_k]$ with S_k being the symmetric positive semi-definite solution of the stochastic Riccati equation. Since, the system now is no longer described by the linear time-invariant dynamics, the existence of the unique solution for the Riccati equation is not guaranteed by the controllability of $(A_{\bar{x}k}, B_{\bar{x}k})$ and observability of $(A_{\bar{x}k}, \sqrt{P_{\bar{x}}})$ and requires Assumption 2. Now, defining the optimal action dependent value function or the Q-function for the stochastic system described in (4) with the cost to go function of the form (5)

$$\begin{aligned} Q(\bar{X}, U) &= \underset{\tau,\gamma}{E}\left[r(\bar{X}_k, U_k) + J_{k+1} \,\middle|\, \bar{X}_k\right] \\ &= \underset{\tau,\gamma}{E}\left\{[\bar{X}_k^T \quad U_k^T] G_k [\bar{X}_k^T \quad U_k^T]^T\right\} \end{aligned} \qquad (6)$$

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. Using the Bellman equation (6) and the

definition of the stochastic cost function (5), and the system dynamics (4), we get

$$Q(\bar{X},U)=\begin{bmatrix}\bar{X}_k\\U_k\end{bmatrix}^T\begin{bmatrix}P_{\bar{x}}+\underset{\tau,\gamma}{E}(A_{\bar{x}k}^TS_{k+1}A_{\bar{x}k}) & \underset{\tau,\gamma}{E}(A_{\bar{x}k}^TS_{k+1}B_{\bar{x}k})\\ \underset{\tau,\gamma}{E}(B_{\bar{x}k}^TS_{k+1}A_{\bar{x}k}) & R_{\bar{x}}+\underset{\tau,\gamma}{E}(B_{\bar{x}k}^TS_{k+1}B_{\bar{x}k})\end{bmatrix}\begin{bmatrix}\bar{X}_k\\U_k\end{bmatrix} \tag{7}$$

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

From the matrix equation (7), the time varying control gain can be expressed, without using the system dynamics as

$$K_k=\underset{\tau,\gamma}{E}\{(G_k^{UU})^{-1}G_k^{U\bar{x}}\} \tag{8}$$

Traditionally, the sequence of control policies, $U(k)$, which minimizes the value function (5) can be obtained by solving the Riccati equation (RE) [6]. The control policy is given by

$$U^*(k)=-\underset{\tau,\gamma}{E}\{K_k{}^*\bar{X}(k)\} \tag{9}$$

where $K_k^*=(R_{\bar{x}}+B_{\bar{x}k}{}^TS_kB_{\bar{x}k})^{-1}B_{\bar{x}k}{}^TS_kA_{\bar{x}k}$. In the Q-learning based ADP algorithms, the initial control policy is assumed to be admissible to keep the cost function (5) finite and the control policy is obtained using (8) without the system dynamics. For the case of interconnected systems, the overall cost function (5) for the system given by (4), can be represented as the sum of the individual cost of all the subsystems as

$$J(k)=\sum\nolimits_{i=1}^{N}J_i(k) \tag{10}$$

where $J_i(k)=\underset{\tau,\gamma}{E}\{\frac{1}{2}\sum_{s=k}^{\infty}\bar{x}_i^T(s)P_{\bar{x},i}\bar{x}_i(s)+u_i^T(s)R_{\bar{x},i}u_i(s)\}$ is the quadratic cost function for i^{th} subsystem. The relation (10) will hold by choosing the penalty matrices $P_{\bar{x}}=diag\{P_{\bar{x},1}\cdots P_{\bar{x},N}\}$ and $R_{\bar{x}}=diag\{R_{\bar{x},1}\cdots R_{\bar{x},N}\}$.

The optimal control problem with the objective to minimize the quadratic cost function (10), for the large scale interconnected system, in a decentralized framework is non- trivial due to the interconnection dynamics. The optimal control policy for each subsystem which minimizes the cost function (10), is obtained by using the Riccati equation of the overall system given the system dynamics $A_{\bar{x}k}$ and $B_{\bar{x}k}$. It is given by

$$u_i^*(k)=-K_i^*\bar{x}_i(k)-\sum\nolimits_{j=1,j\neq i}^{N}K_{ij}^*\bar{x}_j(k) \tag{11}$$

where K_i^* are the diagonal elements, and K_{ij}^* are the off diagonal elements of the overall gain matrix K_k^* in (9). In the following lemma, it is shown that, with the control law (11) designed at each subsystem, the overall system is asymptotically stabilized in the mean square.

Lemma 1. Consider the i^{th} subsystem of the large scale interconnected system (4). The optimal control policy obtained from (11), which uses the system dynamics, renders the individual subsystems asymptotically stable in mean-square.

Proof: Note that the optimal control input (11) is stabilizing [6]. Therefore, the closed-loop system matrix $(A_{\bar{x}k}-B_{\bar{x}k}K_k^*)$ is Schur. The Lyapunov equation is given by $(A_{\bar{x}k}-B_{\bar{x}k}K_k^*)^T\bar{P}(A_{\bar{x}k}-B_{\bar{x}k}K_k^*)-\bar{P}=-\bar{F}$, has a positive definite solution $\bar{F}$. The matrix $\bar{F}$ can be chosen diagonal. Consider the Lyapunov function candidate $L(k)=\underset{\tau,\gamma}{E}(\bar{X}^T(k)\bar{P}\bar{X}(k))$, with $\bar{P}$ being a positive definite matrix of appropriate dimension. The first difference, along the overall system dynamics and the optimal control input

$$\Delta L(k)=-\underset{\tau,\gamma}{E}(\bar{X}^T(k)\bar{F}\bar{X}(k)) \tag{12}$$

Since, $\bar{F}$ is a diagonal matrix, the first difference in terms of the subsystems can be expressed as

$$\Delta L(k)=-\sum\nolimits_{i=1}^{N}\underset{\tau,\gamma}{E}(x_i^T(k)\bar{F}_ix_i(k))\leq-\sum\nolimits_{i=1}^{N}\bar{q}_{\min}\underset{\tau,\gamma}{E}\|x_i(k)\|^2 \tag{13}$$

where $\bar{q}_{\min}$ is the minimum singular value of $\bar{F}_i$. The results of this lemma will be used in the stability analysis of the interconnected system and the requirement of the system dynamics will be relaxed. Thus, the problem is now reduced to the following objective, which is to design a controller for the overall system by minimizing (10), so that each subsystem control policy is of the form (11) with event-sampled feedback.

IV. Controller Design

To address the above mentioned issues, a Q-learning scheme can be designed for the interconnected system by estimating the values of the Q-function of the overall system at each subsystem. Since, subsystems broadcast their state vector, every subsystem can adaptively learn the optimal Q-function of the overall system. The estimation of the Q-function at every subsystem increases the computation. This additional computation can be viewed as a trade-off for relaxing the assumption on strength of interconnection terms and estimating optimal control.

With the following assumption, the Q-function estimator design will be presented for periodic and intermittent feedback.

Assumption 3. *[11] The unknown parameters are assumed to vary slowly.*

A. Periodic feedback

The Q-function (6) in parametric form is given by

$$Q^*(\bar{X}(k),U(k))=\underset{\tau,\gamma}{E}(z^T(k)G_kz(k))=\underset{\tau,\gamma}{E}(\Theta_k^T\xi(k)) \tag{14}$$

where, $z(k)=[\gamma_{sc}(k)\bar{X}^T(k)\quad U^T(k)]^T\in\Re^{\bar{l}}$ with $\bar{l}=m+n+ml$, $\xi(k)=z(k)\otimes z(k)$ is a quadratic polynomial or regression vector, $\otimes$ denotes Kronecker product and $\Theta_k\in\Omega_\Theta\subset\Re^{l_g}$ is the Q-function parameter vector formed by vectorization of the parameter matrix G_k , $\gamma_{sc}(k)$ is packet loss indicator from sensor to controller with appropriate dimension. The estimate of the Q-function is expressed as

$$\hat{Q}(\bar{X}(k),U(k))=\underset{\tau,\gamma}{E}(z^T(k)\hat{G}_kz(k))=\underset{\tau,\gamma}{E}(\hat{\Theta}^T(k)(\xi(k))) \tag{15}$$

where,$\hat{\Theta}(k) \in \Re^{l_g}$ is the estimate of Q-function parameter vector. By Bellman's principle of optimality, the optimal value function satisfies

$$0 = \underset{\tau,\gamma}{E}\left(r(\bar{X}(k), U(k))\right) + \underset{\tau,\gamma}{E}\left(\Theta_k^T \Delta\xi(k)\right) \tag{16}$$

where, $\Delta\xi(k) = \xi(k+1) - \xi(k)$. Since the estimated Q-function does not satisfy (16), the temporal difference (TD) or the Bellman error will be seen which is given by

$$e_B(k) = \underset{\tau,\gamma}{E}\left(r(\gamma_{sc}(k)\bar{X}(k), U(k)) + \hat{\Theta}^T(k)(\Delta\xi(k))\right) \tag{17}$$

The QFE parameter vector $\hat{\Theta}^i(k)$, at the i^{th} subsystem, is tuned by using the history of the Bellman error. Therefore, the auxiliary Bellman error at the sampling instants k is expressed as

$$\Xi_B{}^i(k) = \underset{\tau,\gamma}{E}\left(\Pi^i(k)\right) + \underset{\tau,\gamma}{E}\left(\hat{\Theta}^{i^T}(k) Z^i(k)\right) \tag{18}$$

where $\Pi^i(k) = [r(X^i(k), U^i(k)) \;\; r(X^i(k-1), U^i(k-1)) \;\; \cdots \;\; r(X^i(k-\nu-1), U^i(k-\nu-1))] \in \Re^{1\times\nu}$ and $Z^i(k) = [\Delta\xi^i(k) \;\; \Delta\xi^i(k-1) \;\; \cdots \;\; \Delta\xi^i(k-1-v)] \in \Re^{l_g\times\nu}$ with $0 < \nu < l$. The auxiliary Bellman error (18) uses the current estimated QFE parameter vector $\hat{\Theta}^i(k) \in \Re^{l_g}$ to evaluate the error.

Next, the update law [12] for the QFE parameter vector $\hat{\Theta}^i(k)$, is given by

$$\hat{\Theta}^i(k) = \hat{\Theta}^i(k-1) + \frac{W^i(k-2) Z^i(k-1) \Xi_B^{i^T}(k-1)}{1 + Z^{i^T}(k-1) W^i(k-2) Z^i(k-1)} \tag{19}$$

where

$$W^i(k) = W^i(k-1) - \frac{W^i(k-1) Z^i(k-1) Z^{i^T}(k-1) W^i(k-1)}{1 + Z^{i^T}(k-1) W^i(k-1) Z^i(k-1)} \tag{20}$$

with $W^i(0) = \beta I$, $\beta > 0$, a large positive value and I is the identity matrix of appropriate dimension.

Remark 2. The estimator should wait for all the subsystems to broadcast their feedback information. If γ_{sc} is unity, the Q-function estimator is updated and control law is sent to the actuator, as soon as it is computed. The broadcast scheme requires a suitable scheduling mechanism and acknowledgment signals. The computational delay can be added with controller-to-actuator delay.

Remark 3. The number of history values ν is not fixed and a value $\nu \leq l_g$ is found suitable during simulation studies. For the computation of $Z(k-1)$, the past values are required to be stored at the value function estimator.

Next, the case were subsystems broadcast their feedback information only at event-sampled instants is presented.

B. Event-based Feedback

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

$$e_{ET}(k) = \bar{X}(k) - \bar{X}^e(k), k_l \leq k < k_{l+1}, l = 1, 2, \cdots \tag{21}$$

Our objective is to design an optimal controller by minimizing (5) with the event sampled state vector. The event sampled optimal control input sequence, when used with a Q-function, can be re-written as

$$U^*(k) = -\underset{\tau,\gamma}{E}\left((G^{uu})^{-1} G^{ux}(\bar{X}(k)) + e_{ET}(k)\right) \tag{22}$$

for $k_l \leq k < k_{l+1}$, $l = 1, 2, \cdots$. This optimal control input (22) is governed by the error $e_{ET}(k)$. Further, since the estimation of K^* or G must use event sampled state vector, $\bar{X}^e(k)$, the Q-function estimate can be expressed as

$$\hat{Q}(\bar{X}^e(k), U(k)) = \underset{\tau,\gamma}{E}\left(z^{eT}(k)\hat{G}_k z^e(k)\right) = \underset{\tau,\gamma}{E}\left(\hat{\Theta}^T(k)\xi^e(k)\right) \tag{23}$$

for $k_l \leq k < k_{l+1}$, where $z^e(k) = [\gamma_{sc}(k)\bar{X}^{eT}(k) \;\; U^{eT}(k)]^T \in \Re^l$ and $\xi^e(k) = z^e(k) \otimes z^e(k)$ being the event sampled regression vector. The Bellman error with event sampled state can be represented as

$$e_B(k) = \underset{\tau,\gamma}{E}\left[r(\gamma_{sc}(k)\bar{X}^e(k), U(k)) + \hat{\Theta}^T(k)\Delta\xi^e(k)\right], \tag{24}$$

where, $k_l \leq k < k_{l+1}$, $r(\bar{X}^e(k), U(k)) = \bar{X}^{eT}(k) P_{\bar{x}} \bar{X}^e(k) + U^T(k) R_{\bar{x}} U(k)$ and $\Delta\xi^e(k) = \xi^e(k+1) - \xi^e(k)$. The Bellman error (24) in terms of the periodic system state is rewritten as

$$\begin{aligned} e_B(k) = \underset{\tau,\gamma}{E}\{&r(\bar{X}(k), U(k)) + \hat{\Theta}^T(k)\Delta\xi(k) \\ &+ \Xi_s\left(\bar{X}(k), e_{ET}(k), \hat{\Theta}(k)\right)\} \end{aligned} \tag{25}$$

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

Remark 4. By comparing (25) with (17), the Bellman error in (25) includes an additional term, which is driven by the event sampling error $e_{ET}(k)$. Hence, the accuracy of the estimation of QFE parameters depends upon the frequency of the event sampling instants.

The Q-function at i^{th} subsystem can be expressed as in (23). The QFE estimated parameter vector $\hat{\Theta}^i(k)$ is tuned only at the event sampling instants. The superscript i denotes the overall system parameters at the i^{th} subsystem and the overall estimated control input can be computed at each subsystem as

$$U^i(k) = -\hat{K}^i(k)\bar{X}^{ie}(k) = -\{(\hat{G}^{i,uu}(k))\}^{-1}\hat{G}^{i,ux}(k)\bar{X}^{ie}(k) \tag{26}$$

By using (26), the event-based estimated control input for the i^{th} subsystem can be written as

$$u_i(k) = -\hat{K}_i \bar{x}_i^e(k) - \sum\nolimits_{j=1, j\neq i}^{N} \hat{K}_{ij} \bar{x}_j^e(k), \tag{27}$$

for $k_l^i \leq k < k_{l+1}^i$, $i \in \{1, 2, ..N\}$ and $\forall l \in \mathbb{N}$.

With the event sampled feedback, the QFE is updated only at event-sampled instants. This increases the parameter convergence time, as the updates are aperiodic. In order to facilitate the learning process, an event-sampling condition which uses a mirror estimator and designed such that it explicitly increases the sampling instants was proposed in [5]. In the following section, a novel hybrid learning algorithm is proposed to improve the learning process in the event-sampled framework without explicitly increasing the event-sampling instants without using a mirror estimator.

C. Hybrid-learning algorithm

In the traditional event-driven Q-learning scheme, the Q-function is updated at every event-sampling instant. These events are generally spaced out and dynamically decided. Since the events are spaced out, there are sampling instants where the Q-function is idle.

In the proposed hybrid-learning algorithm, the time-driven Q-learning based ADP [5], [8] is used along with the proposed iterative parameter updates within the inter-event period. The Q-function is updated at the event-sampling instant and the control law is updated at the actuator. Between the event-sampling instants, the Q-function is not idle, but iterative parameter updates are performed.

Whenever there is a new event, the Q-function which is updated iteratively is passed on to the QFE to calculate the new target and the Bellman error. By this approach, the convergence of the parameters should be faster than the traditional Q-learning algorithms.

Next, the error dynamics is defined and the Lyapunov analysis is used to analyze the stability of the closed loop system.

The Bellman error $\underset{\tau,\gamma}{E}({e_B}^i(k))$, in terms of $\underset{\tau,\gamma}{E}(\tilde{\Theta}^i(k))$ can be computed as $\underset{\tau,\gamma}{E}({e_B}^i(k)) = -\underset{\tau,\gamma}{E}(\tilde{\Theta}^{i,T}(k)\Delta\xi^i(k))$ and the auxiliary Bellman error is given by $\underset{\tau,\gamma}{E}(\Xi^i_{_B}(k)) = -\underset{\tau,\gamma}{E}(\tilde{\Theta}^{i,T}(k)Z^i(k))$. Defining the QFE parameter estimation error $\underset{\tau,\gamma}{E}(\tilde{\Theta}^i(k)) = \underset{\tau,\gamma}{E}(\Theta) - \hat{\Theta}^i(k)$, the error dynamics using (19), can be represented as

$$\underset{\tau,\gamma}{E}(\tilde{\Theta}^i(k+1)) = \underset{\tau,\gamma}{E}\left[\tilde{\Theta}^i(k) + \frac{W^i(k)Z^i(k)\Xi_B^{i^T}(k)}{1+Z^{i^T}(k)W^i(k)Z^i(k)}\right] \tag{28}$$

Remark 5. The QFE parameter estimation error $\tilde{\Theta}^i_{_k}$ will converge to zero if the augmented matrix $Z^i(k)$ satisfies the PE condition. A PE like condition for the regression vector $\xi(k)$ can be satisfied by adding an exploration noise to the control input $U(k)$ during the estimation process [11].

Before going to the main results, the request based event-sampling algorithm proposed is presented next.

For estimating the overall Q-function locally, we will use the following request based broadcast scheme and event-sampling algorithm. Consider an event-sampling instant ${k_l}^i$ at the i^{th} subsystem. The i^{th} subsystem generates a request signal and it is broadcast along with the sensor measurements to the other subsystems. On receiving the request signal, all the other subsystems broadcast their respective sensor measurements to all the subsystems. This can be considered as a forced event at the other subsystems. For the event sampling algorithm, consider a quadratic function $f^i(k) = \bar{x}_i(k)^T\Gamma_i\bar{x}_i(k)$, for the i^{th} subsystem. The event sampling condition satisfies the following inequality at every subsystem

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

Remark 6. It is important to mention that the subsystems broadcast their states whenever there is a local event or when there is a request from the other subsystem. The request signal is considered to be broadcast without any delay in Network 1.

The stability of the closed-loop system with the proposed distributed control scheme using event-sampled feedback information is presented next.

Theorem 1. *Consider the closed-loop system (4), QFE and QFE parameter estimation error dynamics (28) along with the control policy (26). Let all the Assumptions 1 through 3 hold, $U(0) \in \Omega_u$ be an initial admissible control policy. Suppose the QFE parameter vector, $\hat{\Theta}(k)$ are updated at event-sampling instants by using (19) and (20), at every subsystem and the event-sampling condition satisfies (29). Then, there exists a constant $\Gamma_{\min} > 0$ such that the closed-loop system state vector $\underset{\tau,\gamma}{E}(\bar{X}(k))$ for all $x_0 \in \Omega_x$ converge to zero asymptotically and the QFE parameter estimation error $\underset{\tau,\gamma}{E}(\tilde{\Theta}(k))$ for all $\hat{\Theta}(0) \in \Omega_\Theta$ remain bounded, provided the inequalities $\Gamma_{\min} > \mu$ and $\bar{\pi} > C_1$ are satisfied. Further, with the PE condition, the estimated Q-function $\hat{Q}(\bar{X}(k), U(k)) \rightarrow \underset{\tau,\gamma}{E}(Q^*(\bar{X}(k), U(k)))$ and estimated control input $U(k) \rightarrow \underset{\tau,\gamma}{E}(U^*(k))$ as event-sampling instants $l \rightarrow \infty$.*

Outline of the proof: During the event-sampled instant, due to the updated control policy (26), the Lyapunov function decreases. Due to the event-sampling condition (29), the Lyapunov function is bounded during the inter-sampling period. When the new event occurs, the controller is updated and the Lyapunov function decreases further leaving a new threshold which is lesser than the previous threshold. This process continues until the Lyapunov function converges to zero.Therefore, it can be concluded that as the event-sampled instants ${k_l}^i \rightarrow \infty$, $\underset{\tau,\gamma}{E}\{L(\bar{x}, \tilde{\Theta})\} \rightarrow 0$.

Remark 7. The inequalities $\Gamma_{\min} > \mu$ and $\bar{\pi} > C_1$ required for the stable operation of the closed-loop system will be satisfied by choosing W and Γ_i appropriately.

Remark 8. For the time-driven Q-learning algorithm, the QFE is not updated in the inter-event period. Therefore the parameter error is constant and hence the Lyapunov function is negative semidefinite. For the hybrid learning algorithm, due to the iterative parameter update, the parameter error decreases in the inter-event period, giving a stronger condition, which helps in faster parameter convergence.

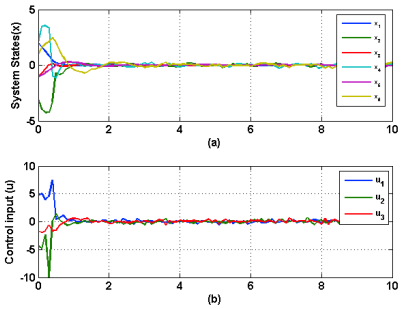

Fig. 2: Time history plot of states and control inputs

The proposed control scheme is tested with a numerical example in the next section.

V. Simulation Analysis

A system of 3 interconnected inverted pendulums, coupled by spring is considered for the verification of the analytical design described in this paper. The dynamics is $\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ \frac{g}{l} - \frac{a_i k}{ml^2} & 0 \end{bmatrix} x_i(t) + \begin{bmatrix} 0 \\ \frac{1}{ml^2} \end{bmatrix} u_i(t) + \sum_{j \in N_i} \begin{bmatrix} 0 & 0 \\ \frac{h_{ij} k}{ml^2} & 0 \end{bmatrix} x_j(t)$ where $l = 2$,$g = 10$,$m = 1$,$k = 5$ and $h_{ij} = 1$, $\forall j \in \{1, 2, .., N\}$. The system is open loop unstable. The system is discretized with a sampling time of 0.1 sec. The cost function was chosen with $P_i = I_{2\times 2}$ and $R_i = 1$ for $\forall i = 1, 2, 3$. The initial states for the system was selected as $x_1 = [2 \quad -3]^T$,$x_2 = [-1 \quad 2]^T$ and $x_3 = [-1 \quad 1]^T$.The initial value of W in the update law is chosen to be 10^6 . For the PE condition Gaussian white noise with zero mean and 0.2 standard deviation, was added to the control inputs. Fig. 2 presents the system response with periodic feedback when there are no network induced losses.

The simulations were carried out with the random loop delays with an upper bound of 200 ms. The delay was characterized by normal distribution with 0.8 as expected values and packet losses were characterized with Bernoulli distribution. Monte-Carlo analysis was carried out for 500 iterations. In the simulations, the PE condition was removed as soon as the Bellman error is reduced to 10^{-3}. Simulation figures for all the cases are not included due to space consideration.

The parameter estimation error comparison between time-driven ADP and proposed hybrid algorithm is given in Fig.3. The hybrid learning algorithm facilitates faster convergence compared to the traditional time-driven ADP algorithm. The decentralized event-sampling algorithm in Figure 4 shows that there is a reduction in the communication and computational cost.

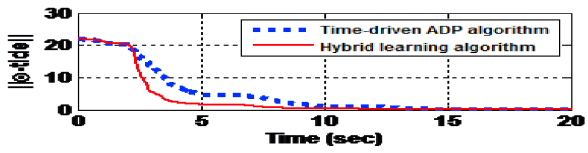

Fig. 3: Parameter estimation error for $case1$

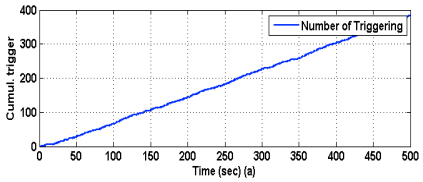

Fig. 4: Total number of feedback instants

VI. Conclusion

The Q-learning based scheme is extended to a large scale system by modeling the system as a centralized system and estimated the Q-function parameters for the entire system at each subsystem. No assumptions on the interconnection strengths is required for this control scheme. The mirror estimator is not used in the event sampling mechanism. The network induced losses and the aperiodic weight tuning increases the learning period and therefore, a novel hybrid Q-learning algorithm is proposed. The proposed algorithm gives a stronger stability results in the inter event period due to the iterative parameter updates. This improves the convergence time of the QFE at each subsystem.

Acknowledgment

This research supported in part by NSF ECCS 1128281, 1406533 and Intelligent Systems Center, MST, Rolla.

References

[1] M. Jamshidi, "Large-scale systems: modeling, control, and fuzzy logic," 1996.

[2] K. S. Narendra and S. Mukhopadhyay, "To communicate or not to communicate: A decision-theoretic approach to decentralized adaptive control," in *American Control Conference (ACC), 2010*. IEEE, 2010, pp. 6369–6376.

[3] X. Wang and M. D. Lemmon, "Event-triggering in distributed networked control systems," *Automatic Control, IEEE Transactions on*, vol. 56, no. 3, pp. 586–601, 2011.

[4] X. Meng and T. Chen, "Event-driven communication for sampled-data control systems," in *American Control Conference (ACC), 2013*. IEEE, 2013, pp. 3002–3007.

[5] A. Sahoo and S. Jagannathan, "Event-triggered optimal regulation of uncertain linear discrete-time systems by using q-learning scheme," in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*. IEEE, 2014, pp. 1233–1238.

[6] F. L. Lewis and V. L. Syrmos, *Optimal control*. John Wiley & Sons, 1995.

[7] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," in *American Control Conference, 1994*, vol. 3. IEEE, 1994, pp. 3475–3479.

[8] H. Xu, S. Jagannathan, and F. L. Lewis, "Stochastic optimal control of unknown linear networked control system in the presence of random delays and packet losses," *Automatica*, vol. 48, no. 6, pp. 1017–1030, 2012.

[9] Y. Halevi and A. Ray, "Integrated communication and control systems: Part ianalysis," *Journal of Dynamic Systems, Measurement, and Control*, vol. 110, no. 4, pp. 367–373, 1988.

[10] W. Zhang, M. S. Branicky, and S. M. Phillips, "Stability of networked control systems," *Control Systems, IEEE*, vol. 21, no. 1, pp. 84–99, 2001.

[11] L. Guo, "Estimating time-varying parameters by the kalman filter based algorithm: stability and convergence," *Automatic Control, IEEE Transactions on*, vol. 35, no. 2, pp. 141–147, 1990.

[12] G. C. Goodwin and K. Sin, "Adaptive control of nonminimum phase systems," *Automatic Control, IEEE Transactions on*, vol. 26, no. 2, pp. 478–483, 1981.