2011 American Control Conference
on O'Farrell Street, San Francisco, CA, USA
June 29 - July 01, 2011

Stochastic Optimal Control of Unknown Linear Networked Control System using *Q*-Learning Methodology

Hao Xu and S. Jagannathan

Abstract—**In this paper, the Bellman equation is utilized forward-in-time for the stochastic optimal control of Networked Control System (NCS) with unknown system dynamics in the presence of random delays and packet losses which are unknown. The proposed stochastic optimal control approach, referred normally as adaptive dynamic programming, uses an adaptive estimator (AE) and ideas from *Q*-learning to solve the infinite horizon optimal regulation control of NCS with unknown system dynamics. Update laws for tuning the unknown parameters of the adaptive estimator (AE) online to obtain the time-based *Q*-function are derived. Lyapunov theory is used to show that all signals are asymptotically stable (AS) and that the approximated control signals converge to optimal control inputs. Simulation results are included to show the effectiveness of the proposed scheme.**

Index Terms— **Networked Control System (NCS), *Q*-function, Adaptive Estimator (AE), Optimal Control.**

I. Introduction

Feedback control systems with control loops closed through a real-time network are called networked control systems (NCS) [1]. In NCS, a communication packet carries the reference input, plant output, and control input which are exchanged using a network among control system components such as sensors, controller, and actuators. The primary advantages of NCS are reduced system wiring, ease of system diagnosis and maintenance, and increased system agility. However, insertion of the communication network in the feedback control loop brings many challenging issues.

The first issue being the network-induced delay that occurs while exchanging data among devices connected to the shared medium. This delay, either constant or random, can degrade the performance of control system and even destabilize the system when the delay is not explicitly considered in the design process. The second issue is packet losses due to unreliable transmission which can cause a loss in control input resulting in instability. These issues have been identified in the literature and are being studied.

Recently, Nilsson [1] analyzed the stability of linear NCS with random delays. Walsh [2] considered stability performance of linear NCS with constant delays. Zhang [3] conducted the stability analysis of linear NCS with delays and packet losses and proposed a stability region.

However, optimality is generally preferred for linear NCS which is very difficult to attain. Lian [4] proposed the optimal controller design by using classical optimal control theory [7] for linear NCS with multiple constant delays into its NCS representation. Using the stochastic optimal control theory [7], Nilsson [1] proposed the optimal and suboptimal controller design for linear NCS with random delays. Although these optimal and suboptimal controller designs have resulted in satisfactory performance, they all require information about the system dynamics of linear NCS and information on delays and packet losses which are not commonly known beforehand. Even when the dynamics of the linear system is known, closing the loop over a communication network with random delays and packet losses can make the overall linear NCS dynamics uncertain.

On the other hand, approximate/adaptive dynamic programming (ADP) schemes proposed by Werbos [8], intend to solve optimal control problems forward-in-time. In ADP, one combines adaptive critics, a reinforcement learning technique, with dynamic programming. Recently Lewis [9] introduced the methods of reinforcement learning and ADP for feedback control to obtain the optimal controller for systems with unknown dynamics.

Tamimi [9] used the *Q*-learning method to solve the optimal strategies for discrete-time linear system quadratic zero-sum games in forward-in-time without knowing the system dynamics wherein the dynamics are defined as constant matrices. In [10], Dierks and Jagannathan used two neural networks (NN) to solve the Hamilton Jacobi Bellman (HJB) equation forward-in-time for optimal control of a class of general nonlinear affine discrete-time systems. While [9] is mainly addresses linear time-invariant systems, work in [10] targets optimal control for nonlinear systems. However, these papers [8-10] did not consider the effects of delays and packet losses which are normally found in a NCS. The delays and packet losses can cause instability [3] if they are not considered which in turn make the optimal controller design more involved and different than [9].

Thus, this paper introduces ADP techniques for the optimal control of linear NCS with uncertain system dynamics and in the presence of random networked-induced delays and packet losses which are unknown. In other words, a linear NCS with random delays and packet losses will be represented by a linear time-varying system with unknown system matrices. Consequently, the suboptimal approach in [9] is not directly applicable.

Therefore, first, a novel approach is undertaken to the optimal regulation of linear NCS with random delays and packet losses to solve the Bellman equation [7] online and forward-in-time. Using an initial stabilizing control, an adaptive estimator (AE) [11] is tuned online to learn the cost function since solving the stochastic Ricatti equation (SRE) requires the system matrixes. Then, using the idea of *Q*-

Hao Xu and S. Jagannathan are with the Dept. of Electrical and Computer Engineering, Missouri University of Science and Technology, USA. ({hx6h7, sarangap}@mst.,edu). Research in part by NSF grant ECCS# 0624644and Intelligent Systems Center

learning, the optimal controller which minimizes the cost function can be calculated based on the information provided by the adaptive estimator (AE). Thus the proposed *Q*-learning based scheme relaxes the need for system dynamics and information on random delay and packet losses. Next, the NCS background representation is presented..

II. BACKGROUND

The basic structure of NCS considered in this paper is shown as Figure 1 where the feedback control loop is closed over a wireless network. Since wireless network bandwidth is limited, two types of network-induced delays and one type of packet losses are included in this structure: (1) $\tau_{sc}(t)$: sensor-to-controller delay, (2) $\tau_{ca}(t)$:controller-to-actuator delay, and (3) $\gamma(t)$: indicator of packet received.

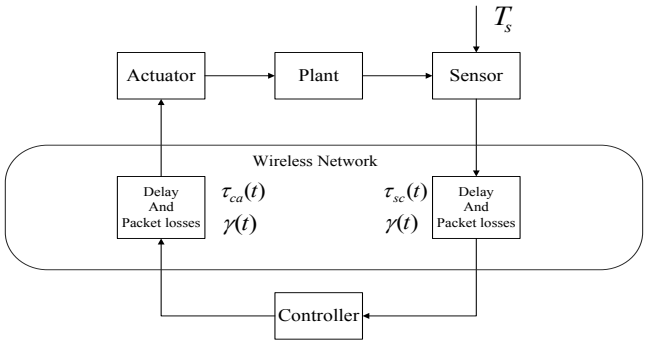

Fig. 1. Networked Control System (NCS).

The following mild assumption is made as [7]:

Assumption 1: a) Sensor is time-driven; controller and actuator are event-driven [4]; b) Communication network is a wide area wireless network so that two types of network-induced delays are independent and unknown whereas their probability distribution functions are considered known [7].; c) The sum of two delay types is bounded [7], while the initial state of system is deterministic [7].

Assuming that the controlled plant is a linear time-invariant system, when the networked-induced delays and packet losses are considered, the NCS can be expressed as

$$\dot{x}(t) = Ax(t) + \gamma(t)Bu(t-\tau(t)) \tag{1}$$

where

$$\gamma(t) = \begin{cases} \mathbf{I}^{n\times n} & \text{if the control input is received at time } t \\ \mathbf{0}^{n\times n} & \text{if the control input lost at time } t \end{cases},$$

$x(t) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m$ and the original system matrices are $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}$. From Assumption 1, we can assume the sum of network-induced delays $\tau(t) = \tau_{sc}(t) + \tau_{ca}(t) < \bar{d}T_s$, where $\bar{d}$ represents the delay bound while T_s is the sampling interval.

During a sampling interval $[kT_s,(k+1)T_s)\ \forall k$, the controller input $u(t)$ to the plant is a piecewise constant. According to Assumption 1, there are at most $\bar{d}$ various current and previous control input values that can be received at the actuator. If many control inputs are received at the same time, only the newer control input is allowed to act on the plant during any sampling interval $[kT_s,(k+1)T_s)$ $\forall k$, and other previous control inputs are deduced. Since control input is based on event driven and can be only received at random instant (Assumption 1), the changes in $x(t)$ occur at the random instants $kT_s + t_i^k, i = 0, 1,\ldots,\bar{d}-1$ and $t_i^k < t_{i-1}^k$ where $t_i^k = \tau_i^k - iT_s$ as illustrated in Figure 2 [7].

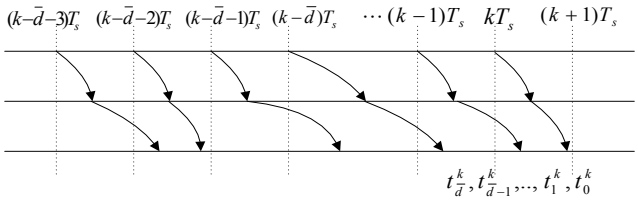

Fig. 2. Timing diagram of signals transmitting in NCS.

Since the controller is event driven, the term u_k can be used to express the controller when the sensor signal x_k is transmitted to the controller. Integration of (1) over a sampling interval $[kT_s,(k+1)T_s)\ \forall k$ yields

$$x_{k+1} = A_s + \sum_{i=0}^{\bar{d}-1} \gamma_{k-i} B_i^k u_{k-i} \tag{2}$$

where

$$x_k = x(kT_s), y_k = y(kT_s), A_s = e^{AT_s};$$

$$B_i^k = \int_{\tau_i^k - iT_s}^{\tau_{i-1}^k - (i-1)T_s} e^{A(T_s - s)} ds B \bullet \mathbf{1}(iT_s - \tau_{i-1}^k) \bullet \mathbf{1}(\tau_i^k - iT_s) \quad ; \quad \forall i = 1,2\ldots,\bar{d}-1$$

$$B_0^k = \int_{\tau_0^k - kT_s}^{T} e^{A(T_s - s)} ds B \bullet \mathbf{1}((k+1)T_s - \tau_0^k),$$

and

$$\gamma_{k-i} = \begin{cases} 0, \text{if } u_{k-i} \text{ is received at } [kT_s,(k+1)T_s) \\ 1, \text{if } u_{k-i} \text{ is lost at } [kT_s,(k+1)T_s) \end{cases}; \mathbf{1}(x) = \begin{cases} 1, & x \ge 0 \\ 0, & x < 0 \end{cases}$$

Using (2), a new augment state variable $z_k = [x_k^T \ u_{k-1}^T \ \cdots \ u_{k-\bar{d}-1}^T]^T$ is defined such that (2) can be expressed as

$$z_{k+1} = A_{zk} z_k + B_{zk} u_k \tag{3}$$

where the system matrices become time-varying

$$A_{zk} = \begin{bmatrix} A_s & \gamma_{k-1}B_1^k & \cdots & \gamma_{k-i}B_i^k & \cdots & \gamma_{k-\bar{d}-1}B_{\bar{d}-1}^k \\ 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & I_m & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & I_m & \cdots & 0 \\ 0 & 0 & \cdots & 0 & I_m & 0 \end{bmatrix} \quad B_{zk} = \begin{bmatrix} \gamma_k B_0^k \\ I_m \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

In this paper, we derive the optimal controller to minimize the cost function

$$J_k = \underset{\tau,\gamma}{E}\left[\sum_{i=k}^{\infty}\left(x_i^T S x_i + u_i^T R u_i\right)\right] \quad k = 0,1,2,\ldots \tag{4}$$

where S and R are symmetric positive semi-definite and symmetric positive definite matrices respectively and $\underset{\tau,\gamma}{E}(\bullet)$ is the expected operator (in this case the mean value) of $\sum_{i=0}^{\infty}\left(x_i^T S x_i + u_i^T R u_i\right)$ based on delays and packet losses. After

redefining the augment state variable z_k , original cost function, (4) can be expressed as

$$J_k = \underset{\tau,\gamma}{E}\left[\sum_{i=k}^{\infty}\left(z_i^T S_z z_i + u_i^T R_z u_i\right)\right] \quad k = 0,1,2,\ldots \tag{5}$$

where $S_z = \begin{bmatrix} S & 0 & \cdots & 0 \\ 0 & R/\bar{d} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R/\bar{d} \end{bmatrix}$, and $R_z = \frac{1}{\bar{d}} R$.

Note the matrices S_z and R_z are still symmetric semi-positive definite and symmetric positive definite respectively.

III. Stochastic Optimal Control

In this section, the idea of *Q*-learning [9] and concept of adaptive estimator is used to develop a stochastic optimal control design for NCS with unknown dynamics in the presence of random delays and packet losses. First, *Q*-function is setup for NCS with random delays and packet losses. Second, model-free online tuning of the parameters based on adaptive estimator and *Q*-learning algorithm will be proposed. Eventually the convergence proof is given.

A. Q-function Setup for NCS

Consider the NCS in the presence of random delays and packet losses described by equation (3) as $z_{k+1} = A_{zk} z_k + B_{zk} u_k$. Assume that the NCS system has $z = 0$ a unique equilibrium point on a set Ω while the states are considered measurable. According to these conditions, the stochastic optimal control input which minimizes the stochastic cost function J_k for NCS system (3) can be derived as $u_k^* = -K_k z_k$ with K_k being the optimal gain. According to the optimal control theory, the cost function can also be represented as

$$J_k^* = \underset{\tau,\gamma}{E}\left(z_k^T P_k z_k\right) \tag{6}$$

where P_k is a symmetric positive definite matrix and the solution to the SRE [7]. The optimal action dependent value function $Q(\bullet)$ of NCS is defined in terms of conditional expected value as

$$\begin{aligned} Q\left(z_k, u_k^*\right) &= \underset{\tau,\gamma}{E}\{z_k^T S_z z_k + u_k^T R_z u_k + J_{k+1}^*\} = \underset{\tau,\gamma}{E}\{r\left(z_k, u_k\right) + J_{k+1}^*\} \\ &= \underset{\tau,\gamma}{E}\{[z_k^T \quad u_k^T] H_k [z_k^T \quad u_k^T]^T\} \end{aligned} \tag{7}$$

Since stochastic optimal control, u_k , is dependent on state z_k which is known at time k , *Q*-function can be expressed as $Q\left(z_k, u_k^*\right) = [z_k^T \quad u^*\left(z_k\right)^T] \underset{\tau,\gamma}{E}(H_k) [z_k^T \quad u^*\left(z_k\right)^T]^T$. Then using Bellman equation and cost function definition, we can formulate the following equation by applying *Q*-function as

$$\begin{aligned} \begin{bmatrix} z_k \\ u_k^* \end{bmatrix}^T \underset{\tau,\gamma}{E}\left(H_k\right) \begin{bmatrix} z_k \\ u_k^* \end{bmatrix} &= z_k^T S_z z_k + u_k^{*T} R_z u_k^* + \underset{\tau,\gamma}{E}\left(z_{k+1}^T P_{k+1} z_{k+1}\right) \\ &= \begin{bmatrix} z_k \\ u_k^* \end{bmatrix}^T \begin{bmatrix} S_z + \underset{\tau,\gamma}{E}\left(A_{zk}^T P_{k+1} A_{zk}\right) & \underset{\tau,\gamma}{E}\left(A_{zk}^T P_{k+1} B_{zk}\right) \\ \underset{\tau,\gamma}{E}\left(B_{zk}^T P_{k+1} A_{zk}\right) & R_z + \underset{\tau,\gamma}{E}\left(B_{zk}^T P_{k+1} B_{zk}\right) \end{bmatrix} \begin{bmatrix} z_k \\ u_k^* \end{bmatrix} \end{aligned} \tag{8}$$

Therefore, $\underset{\tau,\gamma}{E}\left(H_k\right)$ (in contrast with a constant H in [9]) can be written as

$$\begin{aligned} \overline{H}_k = \underset{\tau,\gamma}{E}\left(H_k\right) &= \begin{bmatrix} \overline{H}_k^{zz} & \overline{H}_k^{zu} \\ \overline{H}_k^{uz} & \overline{H}_k^{uu} \end{bmatrix} \\ &= \begin{bmatrix} S_z + \underset{\tau,\gamma}{E}\left(A_{zk}^T P_{k+1} A_{zk}\right) & \underset{\tau,\gamma}{E}\left(A_{zk}^T P_{k+1} B_{zk}\right) \\ \underset{\tau,\gamma}{E}\left(B_{zk}^T P_{k+1} A_{zk}\right) & R_z + \underset{\tau,\gamma}{E}\left(B_{zk}^T P_{k+1} B_{zk}\right) \end{bmatrix} \end{aligned} \tag{9}$$

The optimal action dependent cost function $Q\left(z_k, u_k^*\right)$ is equal to the minimum of cost function J_k while the policy u_k is optimal. Therefore, we have

$$\begin{aligned} J_k^* &= \min_{u_k} J_k = \min_{u_k} Q\left(z_k, u_k\right) \\ &= \min_{u_k} \underset{\tau,\gamma}{E}([z_k^T \quad u_k^T] H_k [z_k^T \quad u_k^T]^T) = Q(z_k, u^*(z_k)) \end{aligned} \tag{10}$$

Then using (9) and stochastic control theory [7], the gain of the optimal control can be expressed in terms of $\overline{H}_k$ as

$$K_k = \left[R_z + \underset{\tau,\gamma}{E}\left(B_{zk}^T P_{k+1} B_{zk}\right)\right]^{-1} \underset{\tau,\gamma}{E}\left(B_{zk}^T P_{k+1} B_{zk}\right) = \left(\overline{H}_k^{uu}\right)^{-1} \overline{H}_k^{uz} \tag{11}$$

Note that if P_k is known, then one still need the time varying system matrices to compute the controller gains. On the other hand, if time varying matrix $\overline{H}_k$ can be learned online without the knowledge of linear time varying system dynamics, the NCS system matrices are not needed to compute the optimal controller gains.

B. Model-free Online Tuning based on Adaptive Estimator and Q-Learning

The proposed online tuning approach entails one adaptive estimator which is used to learn the *Q*-function. Since *Q*-function include $\overline{H}_k$ matrix, this matrix can be solved online and the control signal can be obtained using (11). We make the following assumption since the NCS is linear, the delays of NCS are upper bounded and packet losses satisfy the Bernoulli distribution, and both of them change slowly [5].

Assumption 2: The *Q*-function, $Q\left(z_k, u_k\right)$, can be expressed as the linear in the unknown parameters (LIP).

By using the stochastic adaptive control theory [7] and the definition of *Q*-function (7), the *Q*-function can be represented in vector form similar to the adaptive estimator representation on a set as

$$Q\left(z_k, u_k\right) = w_k^T \overline{H}_k w_k = \overline{h}_k^T w_k \tag{12}$$

where $\overline{h}_k = vec\left(H_k\right), w_k = \left[z_k^T \quad u\left(z_k^T\right)\right]^T, w_k \in \mathbb{R}^{n+\bar{d}m=l}$ and $\overline{w}_k = \left(w_{k1}^2, \ldots, w_{k1} w_{kl}, w_{k2}^2, \ldots, w_{l-1} w_l, w_l^2\right)$ is the Kronecker product quadratic polynomial basis vector.

Note: $vec(\bullet)$ function is constructed by stacking the columns of matrix into one column vector with off-diagonal elements which can be combined as $H_{mn} + H_{nm}$. The time-varying matrix $\overline{H}_k$ can be considered as slowly varying [5]. Then *Q*-function can be expressed as target unknown parameter vector and the regression function $\overline{w}_k$.

Now, the Q-function $Q(z_k,u_k)$ estimation will be considered.

C. Q-function Estimation for Optimal Regulator Design

According to the definition of Q-function [9] and relationship between Q-function and cost function [9], we can use matrix $\overline{H}_k$ in (9) to express the cost function as

$$J_k = w_k^T \overline{H}_k w_k = \overline{h}_k^T w_k \tag{13}$$

Then the Q-function $Q(z_k,u_k)$ can be approximated by an adaptive estimator in term of estimated parameters $\hat{\overline{h}}_k$ as

$$\hat{Q}(z_k,u_k) = \hat{\overline{h}}_k^T w_k \tag{14}$$

where $\hat{\overline{h}}_k$ is the estimated value of the target parameter vector $\overline{h}_k$ with basis function satisfying $\|\overline{w}_k\| = 0$ for $\|z_k\| = 0$.

It is observed that Bellman Equation can be rewritten as $J_{k+1} - J_k + r(z_k,u_k) = 0$ .This relationship, however, is not guaranteed to hold when we apply the estimated matrix $\hat{\overline{H}}_k$. Hence, using delayed values for convenience; the residual error associated with (14) can be expressed as $\hat{J}_k - \hat{J}_{k-1} + r(z_{k-1},u_{k-1}) = e_{hk}$, i.e.

$$e_{hk} = r(z_{k-1},u_{k-1}) + \hat{\overline{h}}_k^T \Delta W_{k-1} \tag{15}$$

where $\Delta W_{k-1} = \overline{w}_k - \overline{w}_{k-1}$.

The dynamics of (15) are then rewritten as

$$e_{hk+1} = r(z_k,u_k) + \hat{\overline{h}}_{k+1}^T \Delta W_k \tag{16}$$

Next, we define an auxiliary residual error vector as

$$\Xi_{hk} = \Gamma_{k-1} + \hat{\overline{h}}_k^T \Delta \mathbf{W}_{k-1} \in \Re^{1\times(1+j)} \tag{17}$$

where $\Gamma_{k-1} = [r(z_{k-1},u_{k-1})\ r(z_{k-2},u_{k-2}) \cdots r(z_{k-1-j},u_{k-1-j})]$, and $\Delta \mathbf{W}_{k-1} = [\Delta W_{k-1} \quad \Delta W_{k-2} \quad \cdots \quad \Delta W_{k-1-j}], 0 < j < k-1 \in \mathbb{N}$ with $\mathbb{N}$being the set of natural real numbers. It is important to note that (17) indicates a time history of the previous $j+1$ residual errors (15) recalculate by using the most recent $\hat{\overline{h}}_k$.

The dynamics of the auxiliary vector (17) are generated similar to (16) and revealed to be

$$\Xi_{hk+1} = \Gamma_k + \hat{\overline{h}}_{k+1}^T \Delta \mathbf{W}_k \tag{18}$$

Now define the update law of the time varying matrix $\overline{H}_k$ as

$$\hat{\overline{h}}_{k+1} = \Delta \mathbf{W}_k \left(\Delta \mathbf{W}_k \Delta \mathbf{W}_k^T\right)^{-1} \left(\alpha_h \Xi_{hk}^T - \Gamma_k^T\right) \tag{19}$$

where $0 < \alpha_h < 1$. Substituting (19) into (18) results

$$\Xi_{hk+1} = \alpha_h \Xi_{hk} \tag{20}$$

Remark 1: It is observed that the cost function J_k and adaptive estimation (13) will become zero only when $z_k = 0$. Hence, when the system states have converged to zero, the Q-function $Q(z_k,u_k)$ approximation is no longer updated. It can be seen as a persistency of excitation (PE) requirement for the inputs to the Q-function $Q(z_k,u_k)$ adaptive estimator wherein the system states must be persistently exiting long enough for the adaptive estimator to learn the optimal cost function. Here PE condition is met by introducing noise.

Therefore, we define the parameter estimation error to be $\tilde{\overline{h}}_k = \overline{h}_k - \hat{\overline{h}}_k$ and rewrite Bellman Equation using the target adaptive estimator representation (12) revealing $\overline{h}_{k+1}^T \overline{w}_{k+1} = r(z_k,u_k) + \overline{h}_{k+1}^T \overline{w}_{k+1}$, which can be expressed as

$$r(z_k,u_k) = \overline{h}_{k+1}^T \overline{w}_k - \overline{h}_{k+1}^T \overline{w}_{k+1} = -\overline{h}_{k+1}^T \Delta W_k \tag{21}$$

Substituting $r(z_k,u_k)$ into (16) and utilizing (15) with $e_{hk+1} = \alpha_h e_{hk}$ from (20) yields

$$\tilde{\overline{h}}_{k+1}^T \Delta W_k = -\alpha_h r(z_{k-1},u_{k-1}) - \alpha_h \hat{\overline{h}}_k^T \Delta W_{k-1} \tag{22}$$

Using the similar method as $r(z_k,u_k)$, we now form $r(z_{k-1},u_{k-1}) = -\overline{h}_k^T \Delta W_{k-1}$, and substitute this expression into (22), we have

$$\tilde{\overline{h}}_{k+1}^T \Delta W_k = \alpha_h \tilde{\overline{h}}_k^T \Delta W_{k-1} \tag{23}$$

Next, the convergence of the cost function estimation error dynamics with adaptive estimation error, dynamics $\tilde{\overline{h}}_k$ given by (23) is demonstrated given an initial admissible control [11] policy. The linear NCS time varying system dynamics are known to be asymptotically stable if an initial admissible control policy can be applied provided the system matrices are known. However, introducing the estimated Q-function results in estimation errors for the cost function J_k, whose stability need to be studied. Subsequently, the results of Theorem 1 will be used for proving the overall closed-loop system stability in Theorem 2 by using an initial admission control policy.

Theorem 1: *(Asymptotic stability of the Cost AE Errors)*. Let $u_0(z_k)$ be an initial admissible control policy for the linear NCS (3), and let the adaptive estimator (AE) parameter update law be given by (19). Then, there exists a positive constant α_h satisfying $0 < \alpha_h < 1$ such that the adaptive parameter estimator errors converge to zero asymptotically.

Next, we show that the estimated control input based on this estimated matrix will indeed converge to the optimal control input.

D. Estimation of the Optimal Feedback Control Signal

There are two ways to estimate the optimal control signal for regulating the NCS. One of them is based on time varying matrix $\overline{H}_k$, the other one is based on standard optimal theory by minimizing the cost function. The difference being that the second method requires the system dynamics and it solves the optimal controller backward. However, it is shown that ultimately both are equivalent which can be used in the proofs.

Method I: As mentioned before, time varying matrix $\overline{H}_k$ can be estimated by an adaptive estimator (AE). According to Q-learning and equation (11), the estimated optimal control input can be expressed by the adaptive estimation $\overline{H}_k$ as

$$\hat{u}_{1k} = -\hat{K}_k z_k = -(\hat{\bar{H}}_k^{uu})^{-1}\hat{\bar{H}}_k^{uz} z_k \quad (24)$$

Method II: Alternatively, the estimated optimal control signal which minimizes the estimated cost function (13) with the adaptive estimation (AE) $\hat{\bar{H}}_k$ is written as

$$\hat{u}_{2k} = -\frac{1}{2} R_z^{-1} B_{zk}^T \frac{\partial \hat{J}_{k+1}}{\partial z_{k+1}} \quad (25)$$

where $\hat{J}_{k+1} = \underset{\tau,\gamma}{E}\left(w_{k+1}^T \hat{H}_{k+1} w_{k+1}\right) = \underset{\tau,\gamma}{E}\left(x_{k+1}^T \hat{P}_{k+1} x_{k+1}\right)$ .Next, it will be shown that the optimal control input obtained by method I and II are equivalent.

***Lemma 1*:** The optimal control estimation calculated with the adaptive estimation of $Q(z_k, u_k)$ is equal to the optimal control calculated by minimizing the cost function J_k , i.e. $\hat{u}_{1k} = \hat{u}_{2k}$.

Since the equality proven in this lemma is in both ways and noting the drawback of second method, we use the first method to solve the optimal controller design for the NCS. However, we will use the Lemma 1 to complete the convergence proof since they are equivalent. Next, the stability of the cost estimation, control estimation, and adaptive estimation error dynamics are considered.

E. Closed-loop System Stability

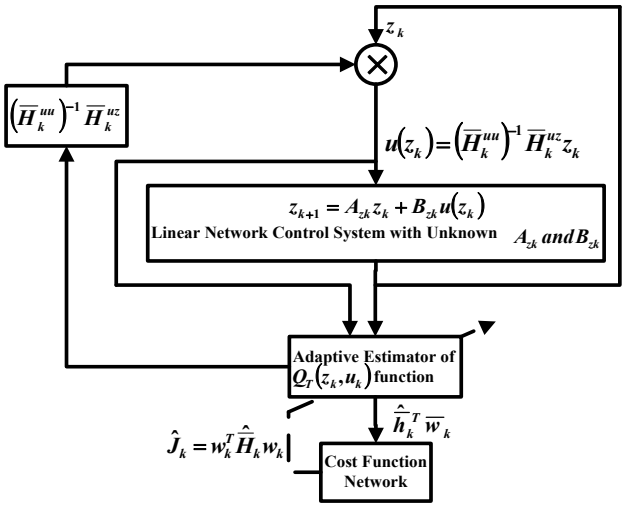

Fig. 3. Stochastic optimal regulator block diagram

In this section, it will be shown that time-varying matrix $\bar{H}_k$ and related cost function estimation error dynamics are asymptotically stable (AS). Further, the estimated control input of NCS (25) will approach the optimal control signal asymptotically. Before introducing the theorem on system stability, we present the block diagram for the stochastic optimal regulator of linear NCS with unknown system dynamics which is shown in Figure 3.

Next, the initial system states are considered to reside in the set which in turn is stabilized by using the initial stabilizing control input u_{0k} . Further sufficient condition for the adaptive estimator tuning gain α_h is derived to ensure the all future states will converge to zero. Then it can be shown that the actual control input approaches the optimal control asymptotically.

Theorem 2 *(Convergence of the Optimal Control Signal):* Let u_{0k} be any initial admissible control policy for the NCS (3) with random delays and packet losses with $0 < k^* < 1/2$. Let the parameters be tuned and estimation control policy be provided by (19) and (25) respectively. Then, there exist positive constants α_h given by Theorem 1 such that the system states z_k and cost function parameter estimator errors $\tilde{\bar{h}}_k$ are all asymptotic stable. In other words, as $k \to \infty, z_k \to 0, \tilde{\bar{h}}_k \to 0, \hat{J}_k \to 0$ and $\hat{u}_{1k} \to u_k^*, \hat{u}_{2k} \to u_k^*$.

***Proof*:** Consider the following positive definite Lyapunov function candidate

$$V = V_D(z_k) + V_J(\tilde{\bar{h}}_k) \quad (26)$$

where $V_D(z_k)$ is defined as $V_D(z_k) = z_k^T z_k$ and $V_J(z_k)$ is defined as $V_J(\tilde{\bar{h}}_k) = (\tilde{\bar{h}}_k^T \bar{w}_k - \tilde{\bar{h}}_k^T \bar{w}_{k-1})^2 = (\tilde{\bar{h}}_k^T \Delta W_{k-1})^2$.

The first difference of (26) can be expressed as $\Delta V = \Delta V_D(z_k) + \Delta V_J(\tilde{\bar{h}}_k)$, and $\Delta V_J(\tilde{\bar{h}}_k) = (\tilde{\bar{h}}_{k+1}^T \Delta W_k)^2 - (\tilde{\bar{h}}_k^T \Delta W_{k-1})^2$ with the adaptive estimator, we have

$$\Delta V_J(\tilde{\bar{h}}_k) = -(1-\alpha_h^2)(\tilde{\bar{h}}_k^T \Delta W_{k-1})^2 \le -(1-\alpha_h^2)\Delta W_{\min}^2 \left\|\tilde{\bar{h}}_k\right\|^2 \quad (27)$$

Next, considering the first part $\Delta V_D(z_k) = z_{k+1}^T z_{k+1} - z_k^T z_k$ and applying the NCS and Cauchy-Schwartz inequality reveals

$$\begin{aligned}\Delta V_D(z_k) &\le \left\|A_{zk} z_k + B_{zk}\hat{u}_{2k} - B_{zk}\tilde{u}_k\right\|^2 - z_k^T z_k \\ &\le 2\left\|A_{zk} z_k + B_{zk}\hat{u}_{2k}\right\|^2 + 2\left\|B_{zk}\tilde{u}_k\right\|^2 - z_k^T z_k \end{aligned} \quad (28)$$

Applying the bounds on closed-loop dynamics with optimal control, and recalling $\hat{u}_{1k} = \hat{u}_{2k}$ from Lemma 1 (i.e. $\tilde{u}_k = 0$).

Therefore, $\Delta V_D(z_k)$ is expressed in terms as the adaptive estimator (AE) error dynamics of the matrix $\bar{H}_k$ and the relationship between $Q(z_k, u_k), \tilde{\bar{h}}_k$ and $\tilde{J}_k$, (28) revealing

$$\Delta V_D(z_k) \le -(1-2k^*)\left\|z_k\right\|^2 + 2\left\|B_{zk}\tilde{u}_k\right\|^2 \le -(1-2k^*)\left\|z_k\right\|^2 \quad (29)$$

At final step, combining the equation (27) and (29), we have

$$\Delta V \le -(1-2k^*)\left\|z_k\right\|^2 - (1-\alpha_h^2)\Delta W_{\min}^2 \left\|\tilde{\bar{h}}_k\right\|^2 \quad (30)$$

Since $0 < k^* < 1/2$ and $0 < \alpha_h < 1$, ΔV is negative definite (See Remark 1) and V is positive definite. Therefore, system states z_k and $\tilde{\bar{h}}_k$ are all asymptotically stable. In other words, as $k \to \infty, z_k \to 0, \tilde{\bar{h}}_k \to 0$ $\hat{J}_k \to J_k^*$ and $\hat{u}_{1k} \to u_k^*, \hat{u}_{2k} \to u_k^*$.

IV. Simulation Results

In this section, stochastic optimal control of NCS is evaluated and compared with a pole placement controller. The networked control batch reactor system [2] with random delays and packet losses is given

$$\dot{x}=\begin{bmatrix} 1.38 & -0.2077 & 6.715 & -5.676 \\ -0.5814 & -4.29 & 0 & 0.675 \\ 1.067 & 4.273 & -6.654 & 5.893 \\ 0.048 & 4.273 & 1.343 & -2.104 \end{bmatrix} x+\begin{bmatrix} 0 & 0 \\ 5.679 & 0 \\ 1.136 & -3.146 \\ 1.136 & 0 \end{bmatrix} u \quad (31)$$

The parameters of this NCS are given as: the sampling time: $T_s = 100ms$; the bound of delays is two, i.e. $\bar{d} = 2$; the random delay: $E(\tau_1) = 80ms, E(\tau_2) = 150ms$; packet Losses follow Bernoulli distribution with $p = 0.3$.

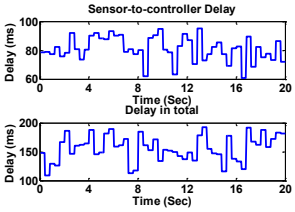

Fig. 4. The distribution of delays.

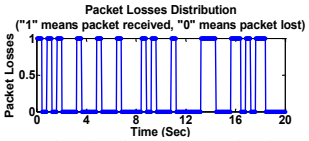

Fig. 5. The distribution of packet losses.

The distributions of two random delays between sensor and actuator are shown in Figure 4. On the other hand, the packet losses are shown in Figure 5. Based on NCS parameters, the NCS model (31) can be discretized as (3).

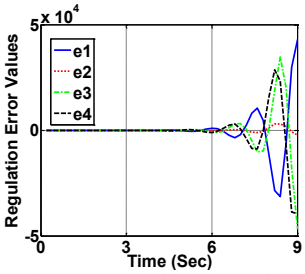

Fig. 6. State tracking error of NCS

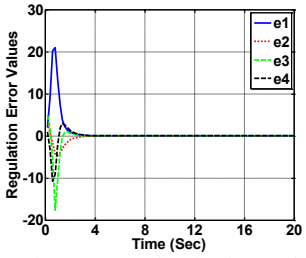

Fig. 7. Proposed optimal control

First, the effective of random delays and packet losses for NCS is studied. The traditional control input $u_k = -\begin{bmatrix} 0.88 & 0.77 & -0.11 & 1.07 \\ -1.65 & -0.08 & -2.93 & 2.61 \end{bmatrix} x_k$, designed by pole placement method maintains batch reactor system stable without any delays and packet losses, while it renders an unstable system in the presence of random delays and packet losses as shown in Figure 6. Secondly, when designing the control for NCS, packet losses and delays are normally unknown. The proposed Q-learning-based adaptive optimal controller is implemented to the NCS with unknown system dynamics in presence of random delays and packet losses. The augment state z_k is generated as $z_k = [x_k \; u_{k-1} \; u_{k-2}]^T \in \mathbb{R}^{8\times 1}$ and $w = [z \; u] \in \mathbb{R}^{10\times 1}$ the initial stabilizing policy for the algorithm was selected as $u_0(z_k) = -\begin{bmatrix} 0.88 & 0.77 & -0.11 & 1.07 & 0.25 & 0.01 & 0.14 & 0.02 \\ -1.65 & -0.08 & -2.93 & 2.61 & -0.02 & 0.68 & -0.03 & 0.51 \end{bmatrix} z_k$ while the regression function for the Q-function was generated as following $\{w_1^2, w_1w_2, w_1w_3, \ldots, w_2^2, \ldots, w_9^2, \ldots, w_{10}^2\}$ which is general defined as (12).

The design parameter for the Q-function $Q(z_k, u_k)$ was selected as $\alpha_h = 10^{-6}$ while the initial parameters for the adaptive estimator were set to zero at the beginning of the simulation. The initial parameters of the action control network were chosen to reflect the initial stabilizing control. The simulation was run for 200 times steps, and for the first 120 times steps, exploration noise with mean zero and variance 0.06 was added to the system in order to ensure the persistency of excitation (PE) condition holds (Remark 1).

In Figure 7, the proposed Q-learning based stochastic optimal controller makes the NCS state tracking errors converge to zero even when the NCS dynamics are unknown. According to the above results the proposed Q-learning based adaptive optimal control algorithm will have nearly the same performance to the NCS with unknown dynamics as that of an optimal controller for NCS when the system dynamics, delays and packet losses are known.

V. Conclusions

In this work, a direct dynamic programming scheme is proposed which combines the adaptive estimator (AE) and the concept of Q-learning to solve the Bellman equation in real time for the stochastic optimal regulation of NCS with random delays and packet losses. The availability of past state values ensured that NCS system dynamics were not needed when an adaptive estimator (AE) generates an estimated Q-function and a novel stochastic optimal control law based on the estimation of $Q(z_k, u_k)$. An initial admissible control policy ensures that the system is stable while the adaptive estimator learns the Q-function $Q(z_k, u_k)$ and the matrix $\bar{h}_k$, cost function and optimal control signal. All adaptive estimator (AE) parameters were tuned online using proposed update law and Lyapunov theory demonstrated the asymptotic stability of the overall closed-loop system.

References

[1] J. Nilsson, B. Bernhardsson, and B. Wittenmark, "Stochastic analysis and control of real-time systems with random time delays". *Automatica*, vol. 1, pp. 57–64, 1998.

[2] G. C. Walsh, H. Ye, and L. Bushnell. "Stability analysis of networked control systems". in *Proc. Amer. Contr. Conf.*, pp. 2876–2880, 1999.

[3] W. Zhang, M. S. Branicky, and S. Phillips, "Stability of networked control systems," *IEEE Contr. Syst. Mag.*, vol. 21, pp. 84–99, 2001.

[4] F. Lian, J. Moyne, and D. Tilbury. "Optimal controller design and evaluation for a class of networked control systems with distributed constant delays". in *Proc. Amer. Contr. Conf.*, pp.3009-3014, 2002.

[5] L. W. Liou, and A. Ray, "A stochastic regulator for integrated communication and control systems". *ASME J. Dynamic Syst., Measure. Contr.*, vol. 4, pp. 604–611, 1991.

[6] F.L. Lewis, and V.L. Syrmos, *Optimal Control*, 2nd ed., Wiley, New York, 1995.

[7] K. J. Åstrom, *Introduction to Stochastic Control Theory*. Academic Press, New York. 1970.

[8] P. J. Werbos, "A menu of designs for reinforcement learning over time". *J. Neural Networks Contr.* vol. 3, pp. 835–846, 1983.

[9] A. A. Tamimi, F. L. Lewis, and M. A. Khalaf. "Model-free Q-learning designs for linear discrete-time zero-sum games with application to *H*-infinity control". *Automatica*, vol. 3, pp. 473–481, 2007.

[10] T. Dierks, and S. Jagannathan, "Optimal control of affine nonlinear discrete-time systems with unknown internal dynamics", in *Proc. Conf. Decisi. Contr.*, pp. 6750-6755, 2009.

[11] S. Jagannathan, *Neural network control of nonlinear discrete-time systems*, CRC Press, 2006.