2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

Stochastic Optimal Control Design for Nonlinear Networked Control System via Neuro Dynamic Programming Using Input-Output Measurements

Hao Xu and S. Jagannathan

Abstract—**Neuro dynamic programming (NDP) techniques for optimal control of nonlinear network control system (NNCS) are not addressed in the literature. Therefore, in this paper, a novel NNCS representation incorporating the unknown system uncertainties and network imperfections is introduced first by using input and output measurements. Then, an online neural network (NN) identifier is introduced to estimate the control coefficient matrix. Subsequently, the critic NN and action NN are employed along with the NN identifier to determine the forward-in-time, time-based stochastic optimal control of NNCS without using value and policy iterations. Instead, value function and control inputs are updated at every sampling instant. Lyapunov theory is used to show that all the closed-loop signals and NN weights are uniformly ultimately bounded (UUB) while the approximated control input converges close to its target value with time.**

I. Introduction

Feedback control systems with control loops closed through a real-time communication 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 by using a communication network among control system components such as sensor, controller and actuators. The primary advantages of NCS are reducing system wiring, ease of system diagnosis and maintenance, and increasing system agility. Adding communication network in the feedback loop brings challenging issues such as random delays and packet losses which can cause the system unstable.

Therefore, Walsh [1] considered performance of linear NCS with constant delays and packet losses. Zhang [2] conducted a general stability analysis of NCS with delays and packet losses and proposed a stability region. These papers [1-2] focused on stability analysis of known linear NCS connected by a communication network with known network imperfections.

The work in [3] extended the controller design to a nonlinear NCS when the dynamics are considered known. However, optimal controller design is generally preferable for NCS and especially for NNCS, which is very difficult to attain. The uncertain dynamics and unknown network imperfections in the case of NNCS further complicates the optimal controller design. By using the stochastic optimal control theory, Nilsson [5] proposed the optimal and suboptimal controller design for linear NCS with random delays. Although these optimal [5] and suboptimal controller designs [5] have resulted in satisfactory performance, they are all based on known linear NCS dynamics and require information on delays which are not known beforehand.

Neuro dynamics programming (NDP) technique, on the other hand, proposed by Werbos [6] intends to solve optimal control problem forward-in-time. In NDP, one combines adaptive critics, a reinforcement learning technique, with dynamic programming. Although NDP is an effective technique to solve the optimal control of NNCS, traditional NDP techniques [6] requires partial knowledge of system dynamics which becomes a problem for NNCS due to network imperfections that are not known. In addition, NDP techniques-based on value and/or policy iterations [6-9] are not useful for real-time control since a significant number of iterations may be needed within any sampling interval. Also, in some cases [8-9], a model may be needed to iterate the value and policies. Finally, existing state-feedback based NDP schemes [6-9] are not applicable for NNCS since the unknown network imperfections such as delays and packet losses can cause instability [2] if they are not considered carefully in the controller design.

Thus, in this paper, a novel time-based NDP algorithm is derived for NNCS with uncertain dynamics and in the presence of unknown network imperfections such as random delays and packet losses by using input-output measurements. To learn the partial knowledge of NNCS, an online NN identifier is introduced first. Then by using an initial stabilizing control, a critic NN is tuned online to learn the value function of NNCS since solving the discrete-time Hamilton-Jacobi-Bellman (HJB) equation requires system dynamics. Subsequently, an action NN is utilized to minimize the value function based on the information provided by the critic NN. Therefore, the proposed novel input-output feedback-based NDP algorithm relaxes the need for system dynamics and information on random delays and packet losses. Value and policy iterations [8-9] are not used and the value function and control input are updated at each sampling instant making the proposed NDP scheme a time-based model-free optimal controller for uncertain NNCS with unknown network imperfections.

II. Nonlinear Networked Control System (NCS) Background

The feedback control loop in the NNCS considered in this paper is closed over a wireless network. Due to unreliability of wireless network, two types of networked-induced delay and one type of packet losses are included in NCS: (1) $\tau_{sc}(t)$: sensor-to-controller delay, (2) $\tau_{ca}(t)$: controller-to-actuator delay, and (3) $\gamma(t)$: indicator of packet losses at actuator.

Research supported in part by NSF ECCS #1128281 and Intelligent Systems Center. 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).

Next the following assumption is made similar as the literature in NCS [12]:

Assumption 1: a). Sensor is time-driven while the controller and actuator are event-driven [12]; b). Communication network is a wide area wireless network so that the two network-induced delays are considered independent and unknown whereas their probability distribution functions are considered known [12]; c). the sum of both the delay types is bounded [12] while the initial state of the nonlinear system is deterministic [12].

In this paper, a continuous-time nonlinear affine system of the form $\dot{x} = f(x) + g(x)u, y = Cx$ is considered, where x, y and u denotes system states, output and input while $f(\bullet)$ and $g(\bullet)$ are smooth nonlinear functions of the state. When the network-induced random delays and packet losses of the wireless network are considered, the control input $u(t)$ is delayed and can be lost at times due to packet losses. Therefore the original nonlinear affine system by incorporating the delay and packet loss effects can be expressed as

$$\dot{x}(t) = f(x(t)) + \gamma(t)g(x(t))u(t-\tau(t)),\ y(t) = Cx(t) \qquad (1)$$

where

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

with $\mathbf{I}^{n\times n}$ is identity matrix, $u(t-\tau(t))$ is delayed control input vector, $x(t) \in \Re^n, u(t) \in \Re^m, y(t) \in \Re^n, f(t) \in \Re^n, g(t) \in \Re^{n\times m}$ and $C \in \Re^{n\times n}$ which is invertible. From Assumption 1, sum of network-induced delays is bounded above, i.e. $\tau(t) = \tau_{sc}(t) + \tau_{ca}(t) < \bar{d}T_s$ where $\bar{d}$ represents the delay bound with T_s being the sampling interval.

For wireless network-based NNCS, information is communicated in the form of packets. As a result, the remote controller has to convert the control inputs into packets and transmit them to the actuator through the wireless network. Then actuator applies the control inputs in response to a received control command packet. Consequently, the controller for NNCS is normally referred to as event-driven and the control input $u(t)$ to the plant is considered piecewise constant [12] during each sampling interval, i.e. $u(t) = u_k,\ t \in [kT_s, (k+1)T_s)\ \forall k$ ..

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 several control inputs are received at the same time, only the latest control input will be applied to the plant during any sampling interval $[kT_s, (k+1)T_s)\ \ \forall k$ while the others are ignored. System states change at time instants $kT_s + t_i^k,\ \ i = 0,1,..,\bar{d}$ and $t_i^k < t_{i-1}^k$ where $t_i^k = \tau_i^k - iT_s$ as illustrated in [12].

Since the controller is event-driven, (the controller updates the command signal based on the receipt of a sensor measurement), the term u_k can be used to express the controller when the sensor signal x_k is transmitted to the controller. Thus, integration (1) over a sampling interval $[kT_s, (k+1)T_s)$ yields

$$x_{k+1} = Z_{\tau,\gamma}(x_k, u_{k-1}, \cdots u_{k-\bar{d}}) + P_{\tau,\gamma}(x_k, u_{k-1}, \cdots u_{k-\bar{d}})u_k, y_k = Cx_k \ (2)$$

where $x(kT_s) = x_k, y(kT_s) = y_k, \gamma((k-i)T_s) = \gamma_{k-i}$and $u((k-i)T_s) = u_{k-i}\ \ i = 0,1,2,\ldots,\bar{d}$ are pervious control inputs, and $P_{\tau,\gamma}(x_k, u_{k-1}, \cdots u_{k-\bar{d}}) = \gamma_k\left(\int_{\tau_0}^{(k+1)T_s} g(x(t))dt\right), Z_{\tau,\gamma}(x_k, u_{k-1}$ $\cdots u_{k-\bar{d}}) = x_k + \int_{kT_s}^{(k+1)T_s} f(x(t))dt + \gamma_{k-\bar{d}}\left(\int_{kT_s}^{\tau_d - \bar{d}T_s} g(x(t))dt\right)u_{k-\bar{d}-1} +$ $\cdots + \gamma_{k-1}\left(\int_{\tau_2-2T_s}^{\tau_1-T_s} g(x(t))dt\right)u_{k-1}$.

Using (2), define a new augment state variable $z_k = \left[x_k^T\ u_{k-1}^T \cdots u_{k-\bar{d}}^T\right]^T \in \Re^{n+\bar{d}m}$ and a modified state vector consisting of the current output and input vectors as $y_k^o = \left[y_k^T\ u_{k-1}^T \cdots u_{k-\bar{d}}^T\right]^T \in \Re^{n+\bar{d}m}$, where $u_{k-i}, i = 1,\ldots,\bar{d}$ are previous control inputs Now equation (2) can be represented

$$z_{k+1} = H(z_k) + L(z_k)u_k,\ y_k^o = C_o z_k \qquad (3)$$

where $H(z_k) = [Z_{\tau,\gamma}^T(x_k, u_{k-1} \cdots u_{k-\bar{d}})\ 0\ u_{k-1}^T \cdots u_{k-\bar{d}}^T]^T$, $L(z_k)$ $= [P_{\tau,\gamma}^T(x_k, u_{k-1}, \cdots u_{k-\bar{d}})\ I_m\ 0\ 0\cdots 0]^T$, $I_m \in \Re^{m\times m}$ is the identity matrix and $C_o = \begin{bmatrix} C & 0 & \cdots & 0 \\ 0 & I_m & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I_m \end{bmatrix}$.

It is important to note that $H(z_k) \in \Re^{n+\bar{d}m}$ and $L(z_k) \in \Re^{(n+\bar{d}m)\times m}$ are nonlinear matrix functions in terms of newly defined augmented state vector z_k . Hence, the NNCS dynamics (3) is still in nonlinear affine form in terms of the augmented state vector. The output matrix C_o is known and invertible since the output matrix C is considered known and invertible.

Next, the nonlinear NCS can be expressed in the input-output form as

$$y_{k+1}^o = C_o H(C_o^{-1} y_k^o) + C_o L(C_o^{-1} y_k^o)u_k = F(y_k^o) + G(y_k^o)u_k \quad (4)$$

where $F(y_k^o) = C_o H(C_o^{-1} y_k^o), G(y_k^o) = C_o L(C_o^{-1} y_k^o),\ \left\| G(y_k^o) \right\|_F \le G_M$, with $\|\bullet\|_F$ denoting the Frobenius norm [11]. Here due to random delays and packet losses, $F(y_k^o)$ and $G(y_k^o)$ are real-valued functions and $F(y_k^o), G(y_k^o)$ can be calculated based on equation (2) and (3) provided information on random delays and packet losses are available. In other words, the network imperfections can make the nonlinear dynamics uncertain requiring adaptive techniques.

We derive the optimal adaptive controller to minimize the stochastic value function [12] as

$$V_k = \underset{\tau,\gamma}{E}\left[\sum_{i=k}^{\infty}\left(x_i^T Q x_i + u_i^T R u_i\right)\right]\ \ k = 0,1,2... \qquad (5)$$

where Q and R are symmetric positive semi-definite and symmetric positive definite constant matrices respectively and $\underset{\tau,\gamma}{E}(\bullet)$ is the expectation operator (in this case the mean value) of $\sum_{i=k}^{\infty}\left(x_i^T Q x_i + u_i^T R u_i\right)$ based on random networked-induced delays and packet losses at different time interval.

The stochastic value function (5) can be expressed in terms of the augmented state variable z_k as

$$V_k = \underset{\tau,\gamma}{E}\left[\sum_{i=k}^{\infty}\left(z_i^T Q_z z_i + u_i^T R_z u_i\right)\right] \qquad k = 0,1,2,\ldots \tag{6}$$

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

Using the input-output form of NNCS (4), the value function (6) can be represented as

$$V_k = \underset{\tau,\gamma}{E}\left[\sum_{i=k}^{\infty}\left(y_i^{oT} Q_y y_i^o + u_i^T R_y u_i\right)\right] \qquad k = 0,1,2,\ldots \tag{7}$$

where $Q_y = (C_o^{-1})^T Q_z C_o^{-1}$, $R_y = R_z$. Note the matrices Q_y and R_y are still symmetric positive semi-definite and symmetric positive definite respectively. Equation (7) can be also expressed as

$$V_k = \underset{\tau,\gamma}{E}(y_k^{oT} Q_y y_i^o + u_k^T R_y u_k) + V_{k+1} \tag{8}$$

Based on the observability condition [4], when $y^o = 0$, $V = 0$, the stochastic value function V_k serves as a Lyapunov function [13]. According to Bellman principle of optimality [4], the optimal stochastic value function V_k^* satisfies the discrete-time Hamilton-Jacob-Bellman (HJB) equation in the infinite horizon optimization case as $V_k^* = \min_{u_k}(\underset{\tau,\gamma}{E}\left(y_k^{oT} Q_y y_k^o + u_k^T R_y u_k\right) + V_{k+1}^*)$, the discrete-time HJB [13] can be represented by using the system inputs and outputs as

$$V_k^* = \underset{\tau,\gamma}{E}\left(y_k^{oT} Q_y y_k^o + \frac{1}{4}\frac{\partial V_{k+1}^{*T}}{\partial y_{k+1}^o} G(y_k^o) R^{-1} G^T(y_k^o) \frac{\partial V_{k+1}^*}{\partial y_{k+1}^o}\right) + V_{k+1}^* \tag{9}$$

where V_k^* is the stochastic value function corresponding to the optimal control input $u^*(y_k^o)$.

In the general nonlinear case, discrete-time HJB equation develops backward-in-time and cannot be solved exactly. This paper introduces a novel scheme in contrast to [7], which is based on time-based NDP scheme, and that can solve the optimal control for unknown NNCS when information on random delays and packet losses are not accurately known.

III. STOCHASTIC OPTIMAL REGULATION

In this section, for solving the drawbacks of existing NDP-based schemes and to utilize the HJB equation forward-in-time, first a novel online identifier is introduced to relax the requirement on the partial knowledge of NNCS dynamics. Subsequently, a novel optimal control of NNCS is proposed by using critic and action NNs.

A. Online NN-Based Identifier for $G(y_k^o)$.

In this part, a novel online NN-based identification is proposed to identify $G(y_k^o)$. According to [9], NNCS (4) can be expressed by using following approximation representation on a compact set Ω as

$$y_k^o = F(y_{k-1}^o) + G(y_{k-1}^o)u_{k-1} = W_C^T \psi_C(y_{k-1}^o) U_{k-1} + \bar{\varepsilon}_{Ck-1} \tag{10}$$

where $W_C = [W_F^T \quad W_G^T]^T$, $\psi_C(y_{k-1}^o) = [\theta_F^T(y_{k-1}^o) \quad \theta_G^T(y_{k-1}^o)]^T$, $U_{k-1} = [I, u_{k-1}]^T$, $\varepsilon_{Ck-1} = [\varepsilon_{Fk-1} \quad \varepsilon_{Gk-1}]$, and $\bar{\varepsilon}_{Ck-1} = \varepsilon_{Ck-1} U_{k-1}$, with $\left\|\psi_C(y_{k-1}^o)\right\| \le \psi_M$ and $\left\|\psi_C(y_{k-1}^o) U_{k-1}\right\| \le \Psi_M$ are the bounds while the estimation error satisfies $\left\|\bar{\varepsilon}_{Ck-1}\right\| < \varepsilon_{CM}$, $\forall k$. Since the NN activation functions $\theta_F(\bullet), \theta_G(\bullet)$, and $\psi_C(\bullet)$ are known, NNCS dynamics $G(y_k^o)$ can be identified, when NN-based identifier weights W_C are learned. Hence, in this section, a suitable update law will be proposed to tune the NN weights. Here, in Theorem 1, the inputs are assumed to the bounded for purpose of the identifier proof whereas it is relaxed during the controller design and proof in Theorem 3.

The output y_k^o can be estimated at time k by using a NN-based identifier as

$$\hat{y}_k^o = \hat{W}_{Ck}^T \psi_C(y_{k-1}^o) U_{k-1} \tag{11}$$

Using (10) and (11), the identification error is defined as

$$e_{yk} = y_k^o - \hat{y}_k^o = y_k^o - \hat{W}_{Ck}^T \psi_C(y_{k-1}^o) U_{k-1} \tag{12}$$

The identification error dynamics (12) are expressed as

$$e_{yk+1} = y_{k+1}^o - \hat{y}_{k+1}^o = y_{k+1}^o - \hat{W}_{Ck+1}^T \psi_C(y_k^o) U_k \tag{13}$$

Based on [10], an auxiliary identification error vector can be

$$\Sigma_{yk} = Y_k^o - \hat{W}_{Ck}^T \Delta\psi_{Ck-1} \bar{U}_{k-1} \tag{14}$$

where $Y_k^o = [y_k^o \;\; y_{k-1}^o \cdots y_{k-l}^o]$, $\Delta\psi_{Ck-1} = [\psi_C(y_{k-1}^o) \; \psi_C(y_{k-2}^o) \cdots \psi_C(y_{k-1-l}^o)]$ and $\bar{U}_{k-1} = \left[U_{k-1} \; U_{k-2} \cdots U_{k-1-l}\right]$, $0 < l < k-1$. Note equation (14) represents l previous identification errors which are recalculated by using most recent NN-based weights $\hat{W}_{Ck}$.

Similar to (14), the auxiliary identification error dynamics are revealed to be

$$\Sigma_{yk+1} = Y_{k+1}^o - \hat{W}_{Ck+1}^T \Delta\psi_{Ck} \bar{U}_k \tag{15}$$

It is desired to tune the NN identifier weights $\hat{W}_{Ck}$ such that the identification error e_{yk} converges to zero asymptotically, i.e. $k \to \infty, e_{yk} \to 0$. Hence, the update law for NN weights can be defined as

$$\hat{W}_{Ck+1} = \bar{U}_k \Delta\psi_{Ck} (\Delta\psi_{Ck}^T \bar{U}_k^T \bar{U}_k \Delta\psi_{Ck})^{-1} (Y_{k+1}^o - \alpha_C \Sigma_{yk})^T \tag{16}$$

where α_C is the tuning parameter of the NN-based identifier satisfying $0 < \alpha_C < 1$. Substituting (16) into (15)

$$\Sigma_{yk+1} = \alpha_C \Sigma_{yk} \tag{17}$$

Remark 1: We can define $\beta_k = \psi_C(y_k^o)U_k$, and β_k has to persistently exiting [11] long enough for the online NN-based identifier to learn the NNCS dynamics $G(y_k^o)$.

Next, NN-based identifier weight estimation error is defined as $\widetilde{W}_{Ck} = W_C - \hat{W}_{Ck}$, and recalling (13), the identification error dynamics can be rewritten as

$$e_{yk+1} = y_{k+1}^o - \hat{y}_{k+1}^o = \widetilde{W}_{Ck+1}^T \psi_C(y_k^o)U_k + \bar{\varepsilon}_{Ck} \tag{18}$$

Using $e_{yk+1} = \alpha_C e_{yk}$ from (17), we have

$$\widetilde{W}_{Ck+1}^T \psi_C(y_k^o)U_k = \alpha_C(\widetilde{W}_{Ck}^T \psi_C(y_{k-1}^o)U_{k-1}) + \alpha_C \bar{\varepsilon}_{Ck-1} - \bar{\varepsilon}_{Ck} \tag{19}$$

Eventually, the boundedness of the identification error dynamics e_{yk} given by (12) and NN weights estimation error dynamics $\widetilde{W}_{Ck}$ given by (19) will be demonstrated.

Theorem 1 *(Boundedness of the identifier)*. Let the proposed NN-based identifier be defined as (11) and NN weights update law be given by (16). Under the assumption that β_k defined in Remark 1 satisfies the PE condition, there exists a positive constant α_C satisfying $0 < \alpha_C < \min\{1, \ \Psi_{\min} / \sqrt{2}\Psi_M\}$ and computable positive constants B_{WC}, B_{ey}, such that the identification error (12) and NN weights estimation errors $\widetilde{W}_{Ck}$ (19) are all uniformly ultimately bounded (*UUB*) [13] with ultimate bounds given by $\|e_{yk}\| \le B_{ey}$ and $\|\widetilde{W}_{Ck}\| \le B_{WC}$.

Proof: Combined with the overall stability.

B. NN Approximation of the Value Function and Control Policy

In [13], by using universal approximation property of NN, the stochastic value function (7) and control policy can be represented with critic NN and action NN as

$$V(y_k^o) = W_V^T \vartheta(y_k^o) + \varepsilon_{Vk}, u^*(y_k^o) = W_u^T \phi(y_k^o) + \varepsilon_{uk} \tag{20}$$

where W_V and W_u represent the constant target NN weights, $\varepsilon_{Vk}, \varepsilon_{uk}$ are the approximation errors for critic NN and action NN respectively, and $\vartheta(\bullet)$ and $\phi(\bullet)$ are the vector activation functions for two NNs, respectively. The upper bounds for the two target NNs weights are defined as $\|W_V\| \le W_{VM}$ and $\|W_u\| \le W_{uM}$ where W_{VM}, W_{uM} are positive constants [7], and the approximation errors are also considered bounded as $\|\varepsilon_{Vk}\| \le \varepsilon_{VM}$ and $\|\varepsilon_{uk}\| \le \varepsilon_{uM}$ where $\varepsilon_{VM}, \varepsilon_{uM}$ are also positive constants [7] respectively. Additionally, the gradient of approximation error is assumed to be bounded as $\|\partial\varepsilon_{Vk} / \partial y_{k+1}^o\| \le \varepsilon_{VM}^{'}$ with $\varepsilon_{VM}^{'}$ being a positive constant [10].

The critic NN and action NN approximation of (20) can be expressed as [13]

$$\hat{V}(y_k^o) = \hat{W}_{Vk}^T \vartheta(y_k^o), \hat{u}(y_k^o) = \hat{W}_{uk}^T \phi(y_k^o) \tag{21}$$

where $\hat{W}_{Vk}$ and $\hat{W}_{uk}$ are the approximation of the target weights W_V and W_u, respectively. In this work, the activation functions $\vartheta(\bullet), \phi(\bullet)$ are selected to be a basis function set and linearly independent [10]. Since it is required that $V(y_k^o = 0) = 0$ and $u(y_k^o = 0) = 0$, the basis functions $\vartheta(\bullet), \phi(\bullet)$ are chosen such that $\vartheta(y_k^o = 0) = 0, \phi(y_k^o = 0) = 0$.

Substituting (21) into equation (8), it can be rewritten as

$$W_V^T \Delta\vartheta(y_{k+1}^o) + r(y_k^o, u_k) = \Delta\varepsilon_{Vk} \tag{22}$$

where $r(y_k^o, u_k) = \underset{\tau,\gamma}{E}(y_k^{oT} Q_y y_k^o + u_k^T R_y u_k)$, $\Delta\vartheta(y_k^o) = \vartheta(y_{k+1}^o) - \vartheta(y_k^o)$ and $\Delta\varepsilon_{vk} = \varepsilon_{vk} - \varepsilon_{vk+1}$. However, when implementing the estimated value function (21), equation (22) does not hold. Therefore, using delayed values for convenience, the residual error or cost-to-go error with (24) can be expressed

$$e_{Vk} = r(y_k^o, u_k) + \hat{W}_{Vk}^T \Delta\vartheta(y_k^o) \tag{23}$$

Based on gradient descent algorithm, the update law of critic NN weights is given by

$$\hat{W}_{Vk+1} = \hat{W}_{Vk} - \alpha_v \frac{\Delta\vartheta(y_k^o)}{\Delta\vartheta^T(y_k^o)\Delta\vartheta(y_k^o) + 1} e_{Vk}^T \tag{24}$$

Remark 2: It is important to note that the stochastic value function (8) and critic NN approximation (21) all become zero only when $y_k^o = 0$. Therefore, once the system outputs have converged to zero, the value function approximation is no longer updated. This can be also viewed as a PE requirement for the inputs to the critic NN where the system outputs must be persistently exiting long enough for the approximation so that critic NN learns the optimal stochastic value function. In this paper, PE condition is met by introducing noise.

As a final step in the critic NN design, define the weight estimation error as $\widetilde{W}_{Vk} = W_V - \hat{W}_{Vk}$. Since $r^T(y_k^o, u_k) = -\Delta\vartheta^T(y_{k+1}^o) \ W_V + \Delta\varepsilon_{Vk}^T$ in equation (24), the dynamics of the critic NN weights estimation error can be rewritten as

$$\widetilde{W}_{Vk+1} = \widetilde{W}_{Vk} - \alpha_V \frac{\Delta\vartheta(y_k^o)(\Delta\vartheta^T(y_k^o)\widetilde{W}_{Vk} + \Delta\varepsilon_{Vk})}{\Delta\vartheta^T(y_k^o)\Delta\vartheta(y_k^o) + 1} \tag{25}$$

Now we need to find the control policy via action NN which minimizes the approximated value function (21). First, the action NN estimation errors are defined to be the difference between the optimal controls applied to NNCS (4) and the control signal that minimizes the estimated value function (21) with identified NNCS dynamics $\hat{G}(y_k^o)$, which can be expressed as

$$e_{uk} = \hat{W}_{uk}^T \phi(y_k^o) + \frac{1}{2} R_y^{-1} \hat{G}^T(y_k^o) \frac{\partial\vartheta^T(y_{k+1}^o)}{\partial y_{k+1}^o} \hat{W}_{Vk} \tag{26}$$

Next, the update law for action NN weights is defined as

$$\hat{W}_{uk+1} = \hat{W}_{uk} - \alpha_u \frac{\phi(y_k^o)}{\phi^T(y_k^o)\phi(y_k^o) + 1} e_{uk}^T \tag{27}$$

where $0 < \alpha_u < 1$ is a positive constant. By selecting the control policy u_k to minimize the desired value function (20), the following equation results

$$W_u^T \phi(y_k^o) + \varepsilon_{uk} + \frac{1}{2} R_y^{-1} \hat{G}^T(y_k^o)(\frac{\partial\vartheta^T(y_{k+1}^o)}{\partial y_{k+1}^o} W_V + \frac{\partial\varepsilon_{Vk}^T}{\partial y_{k+1}^o}) = 0 \tag{28}$$

Substituting (28) into (26), the action NN estimation error dynamics can be rewritten as

$$e_{uk} = -\widetilde{W}_{uk}^{T}\phi(y_k^o) - \frac{1}{2}R_y^{-1}\hat{G}^T(y_k^o)\frac{\partial \vartheta^T(y_{k+1}^o)}{\partial y_{k+1}^o}\widetilde{W}_{Vk} + \frac{1}{2}R_y^{-1}\widetilde{G}^T(y_k^o)\frac{\partial \varepsilon_{Vk}^T}{\partial y_{k+1}^o} - \varepsilon_{ek} \tag{29}$$

where $\widetilde{G}(y_k^o) = G(y_k^o) - \hat{G}(y_k^o)$, $\varepsilon_{ek} = \varepsilon_{uk} + \frac{1}{2}R_y^{-1}G^T(y_k^o)\frac{\partial \varepsilon_{Vk}^T}{\partial y_{k+1}^o}$

satisfying $\|\varepsilon_{ek}\| \le \varepsilon_{eM}$ with ε_{eM} being a positive constant, and $\left\|\frac{\partial \varepsilon_{Vk}^T}{\partial y_{k+1}^o}\right\| \le \varepsilon_{VM}^{'}$.

The action NN weight estimation error dynamics can be represented as

$$\widetilde{W}_{uk+1} = \widetilde{W}_{uk} + \alpha_u \frac{\phi(y_k^o)}{\phi^T(y_k^o)\phi(y_k^o)+1}e_{uk}^T \tag{30}$$

C. Closed-loop Stability

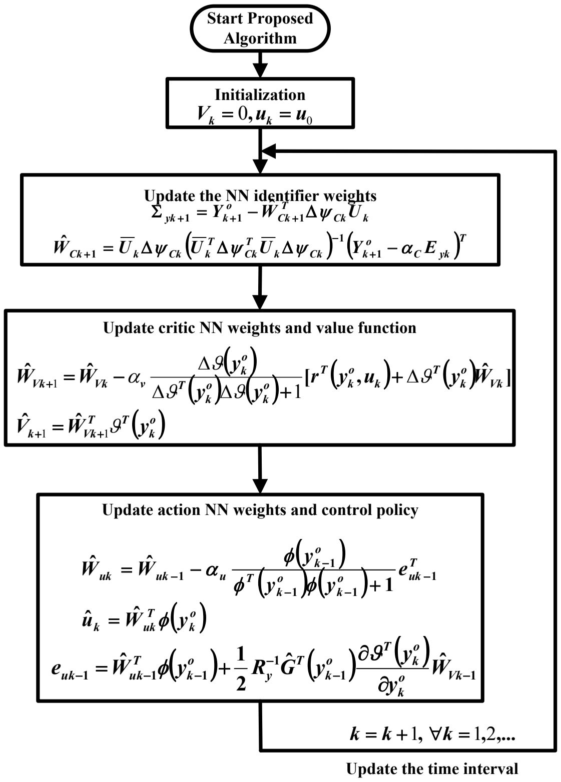
Start Proposed Algorithm
Initialization
$V_k = 0, u_k = u_0$
Update the NN identifier weights
$\Sigma_{yk+1} = Y_{k+1}^o - \hat{W}_{Ck+1}^T \Delta\psi_{Ck} \bar{U}_k$
$\hat{W}_{Ck+1} = \bar{U}_k \Delta\psi_{Ck}\left(\bar{U}_k^T \Delta\psi_{Ck}^T \bar{U}_k \Delta\psi_{Ck}\right)^{-1}\left(Y_{k+1}^o - \alpha_C E_{yk}\right)^T$
Update critic NN weights and value function
$\hat{W}_{Vk+1} = \hat{W}_{Vk} - \alpha_v \frac{\Delta\vartheta(y_k^o)}{\Delta\vartheta^T(y_k^o)\Delta\vartheta(y_k^o)+1}[r^T(y_k^o, u_k) + \Delta\vartheta^T(y_k^o)\hat{W}_{Vk}]$
$\hat{V}_{k+1} = \hat{W}_{Vk+1}^T \vartheta^T(y_k^o)$
Update action NN weights and control policy
$\hat{W}_{uk} = \hat{W}_{uk-1} - \alpha_u \frac{\phi(y_{k-1}^o)}{\phi^T(y_{k-1}^o)\phi(y_{k-1}^o)+1} e_{uk-1}^T$
$\hat{u}_k = \hat{W}_{uk}^T \phi(y_k^o)$
$e_{uk-1} = \hat{W}_{uk-1}^T \phi(y_{k-1}^o) + \frac{1}{2} R_y^{-1} \hat{G}^T(y_{k-1}^o) \frac{\partial\vartheta^T(y_k^o)}{\partial y_k^o} \hat{W}_{Vk-1}$
$k = k+1, \forall k = 1,2,...$
Update the time interval

Fig.1. Flowchart of the proposed optimal control for unknown NNCS.

In this section before introducing the theorem on system stability, we present the flowchart in Fig. 1 of the proposed time-based NDP for NNCS with uncertain system dynamics and unknown network imperfections.

For the closed-loop stability and convergence proof, the initial system outputs are considered to reside in a compact set $\Omega \in \Re^n$ due to the initial admissible control input $u_0(y_k^o)$. Also, in compact set Ω, the critic NN basis function and its gradient as well as the activation function of the action NN are considered bounded with $\|\vartheta(y_k^o)\| \le \vartheta_M, \|\partial\vartheta(y_k^o)/\partial y_k^o\| \le \vartheta_M^{'}$, and $\|\phi(y_k^o)\| \le \phi_M$, respectively. Further, sufficient conditions for the three NNs tuning parameters, α_C, α_V and α_u, are derived to guarantee that all future output states never leave the compact set.

Theorem 2: *(Convergence of the Optimal Control Signal).* Let $u_0(y_k^o)$ be any initial stabilizing control policy for the NNCS described in (4) when $0 < k^* < 1/2$. Let the NN weight tuning for the identifier, critic and the action NN be provided by (16), (24) and (27), respectively. Then, there exists positive constant $\alpha_C, \alpha_u, \alpha_V$ satisfying $0 < \alpha_C < \min\{1, \frac{\Psi_{\min}}{2\sqrt{2}\Psi_M}\}$, $\frac{1}{4} < \alpha_V < \frac{3+\sqrt{3}}{12}$ and $\frac{1}{6} < \alpha_u < \frac{1}{3}$, and positive constants b_y, b_V, b_{WC}, b_{ey} and b_u such that the system output vector y_k^o, NN identification error e_{yk}, weight estimation errors $\widetilde{W}_{Ck}$, critic NN and action NN weight estimation errors $\widetilde{W}_{Vk}$ and $\widetilde{W}_{uk}$, respectively, are all *UUB* for all $k \ge k_0 + T$ with ultimate bounds given by $\|y_k^o\| \le b_y, \|e_{yk}\| \le b_{ey}, \|\widetilde{W}_{Ck}\| \le b_{WC}, \|\widetilde{W}_{Vk}\| \le b_V$ and $\|\widetilde{W}_{uk}\| \le b_u$.

Proof: Consider the Lyapunov function candidate

$$L = L_{DN} + L_{uN} + L_{VN} + L_{CN} + L_{AN} + L_{BN} \tag{31}$$

where $L_{DN} = (y_k^o)^T y_k^o$, $L_{uN}, L_{VN}, L_{CN}, L_{AN}$ and L_{BN} are defined

$$L_{uN} = tr\{\widetilde{W}_{uk}^T \Omega \widetilde{W}_{uk}\}, L_{VN} = tr\{\widetilde{W}_{Vk}^T \Lambda \widetilde{W}_{Vk}\} \tag{32}$$

$$L_{CN} = tr\{e_{yk}^T e_{yk}\} + tr\{\widetilde{W}_{Ck}^T \mathrm{O} \widetilde{W}_{Ck}\}$$

$$L_{AN} = (tr\{\widetilde{W}_{uk}^T \Gamma \widetilde{W}_{uk}\})^2, L_{BN} = (tr\{\widetilde{W}_{Ck}^T \Theta \widetilde{W}_{Ck}\})^2$$

with $\Omega = \frac{24G_M^2\phi_M^2(\phi_M^2+1)}{\phi_{\min}^2}\mathbf{I}$, $\Lambda = \frac{288\phi_M^2\Xi^2(\Delta\vartheta_M^2+1)}{\phi_{\min}^2\Delta\vartheta_{\min}^2} \times \left(G_M^2 + 12\varepsilon_{CM}^2 \times \frac{2\psi_M^2}{\Psi_{\min}^2}\right)\mathbf{I}$, $\mathrm{O} = 2[\Psi_M^2 + \frac{9(\varepsilon_{VM}^{'}\psi_M)^2\phi_M^2\Xi^2}{2\phi_{\min}^2} + 6(\Xi\psi_M)^2\phi_M^2\frac{\phi_M^2\Delta\bar{\varepsilon}_{CM}^2}{\phi_{\min}^2\Psi_{\min}^2}]\mathbf{I}$,

$\Gamma = (85\sqrt{\frac{(\Xi\psi_M\phi_M)^2(\Delta\vartheta_M^2+1)^2}{\phi_{\min}^2\Delta\vartheta_{\min}^4}})\mathbf{I}$ and $\Theta = (\sqrt{\frac{24(\Xi\psi_M\phi_M)^2}{\phi_{\min}^2}})\mathbf{I}$ are positive definite matrices, $\mathbf{I}$ is identity matrix, Ξ is defined as $\lambda_{\max}(R^{-1})\vartheta_M^{'}G_M$, and $\lambda_{\max}(R^{-1})$ is the maximum singular value of R. The first difference of (31) is given by $\Delta L = \Delta L_{DN} + \Delta L_{uN} + \Delta L_{VN} + \Delta L_{CN} + \Delta L_{AN} + \Delta L_{BN}$.

Considering first difference $\Delta L_{DN} = (y_{k+1}^o)^T y_{k+1}^o - (y_k^o)^T y_k^o$, using the NNCS dynamics (4), and applying the Cauchy-Schwartz inequality reveals that the first difference becomes

$$\Delta L_{DN} \le \left\| \begin{matrix} F(y_k^o) + G(y_k^o)u^*(y_k^o) - G(y_k^o)u^*(y_k^o) \\ + G(y_k^o)\hat{u}(y_k^o) - G(y_k^o)\hat{u}(y_k^o) + \hat{G}(y_k^o)\hat{u}(y_k^o) \end{matrix} \right\|^2 - (y_k^o)^T y_k^o \tag{33}$$

$$\le -(1-2k^*)\|y_k^o\|^2 + 4\Psi_M^2\|\widetilde{W}_{Ck}^T\|^2 + 8G_M^2\phi_M^2\|\widetilde{W}_{uk}\|^2 + 8G_M^2\varepsilon_{uM}^2$$

Next, first different L_u can be expressed as

$$\Delta L_{uN} \le -\frac{(3\alpha_u - 6\alpha_u^2)\phi_{\min}^2\|\Omega\|}{\phi_M^2+1}\|\widetilde{W}_{uk}\|^2 + \frac{(\alpha_u^2+\alpha_u)\Xi^2\|\Omega\|}{2(\phi^T(y_k^o)\phi(y_k^o)+1)}\|\widetilde{W}_{Vk}\|^2$$

$$+\|\Omega\|\Delta\varepsilon_{eM}^2+\frac{(2\alpha_u^2+\alpha_u)(\Xi\psi_M)^2\|\Omega\|}{2(\phi_M^2+1)G_M^2}\left\|\widetilde{W}_{Ck}\right\|^4 \tag{34}$$

$$+\frac{(2\alpha_u^2+\alpha_u)(\Xi\psi_M)^2\|\Omega\|}{4(\phi_M^2+1)G_M^2}\left\|\widetilde{W}_{Vk}\right\|^4+\frac{(2\alpha_u^2+\alpha_u)(\Xi\varepsilon_{VM}^{'}\psi_M)^2\|\Omega\|}{4(\phi_M^2+1)G_M^2}\left\|\widetilde{W}_{Ck}\right\|^2$$

where $0<\phi_{\min}<\left\|\phi(y_k^o)\right\|$ is ensured by the PE condition described in *Remarks1 and 2*, $\left\|\Delta\varepsilon_{euk}\right\|^2=\dfrac{\alpha_u\left\|\varepsilon_{ek}\right\|^2}{(\phi^T(y_k^o)\phi(y_k^o)+1)}$ $\le\Delta\varepsilon_{eM}^2$, which is a bounded positive constant.

Next, first difference L_{AN}, L_{BN} can be expressed as

$$\Delta L_{AN}\le-\frac{(\alpha_V-2a_V^2-\frac{1}{12})\Delta\vartheta_{\min}^2\|\Gamma\|^2\left\|\widetilde{W}_{Vk}\right\|^4}{(\Delta\vartheta_M^2+1)}$$

$$+\|\Gamma\|^2\frac{4\Delta\varepsilon_{VM}^2}{3(\Delta\vartheta_M^2+1)}\left\|\widetilde{W}_{Vk}\right\|^2+\frac{4\|\Gamma\|^2\Delta\varepsilon_{VM}^4}{9(\Delta\vartheta_M^2+1)^2} \tag{35}$$

$$\Delta L_{BN}\le-\|\Theta\|^2(1-4\alpha_C^4\frac{\Psi_M^4}{\Psi_{\min}^4})\left\|\widetilde{W}_{Ck}\right\|^4+\frac{8\alpha_C^2\Psi_M^2\|\Theta\|^2}{\Psi_{\min}^2}\left\|\widetilde{W}_{Ck}\right\|^2\Delta\bar{\varepsilon}_{CM}^2$$

$$+\frac{4\|\Theta\|^2}{\Psi_{\min}^4}\Delta\bar{\varepsilon}_{CM}^4 \tag{36}$$

Next, using (33), (34), (35) and (36) to form ΔL as

$$\Delta L\le-(1-2k^*)\left\|y_k^o\right\|^2-288\rho(\alpha_V-2a_V^2-\frac{1}{12})\left\|\widetilde{W}_{Vk}\right\|^2-(1-\alpha_C^2)\left\|e_{yk}\right\|^2$$

$$-24G_M^2\phi_M^2(3\alpha_u-6\alpha_u^2-\frac{1}{3})\left\|\widetilde{W}_{uk}\right\|^2-4\eta(1-4\alpha_C^2\frac{\Psi_M^2}{\Psi_{\min}^2})\left\|\widetilde{W}_{Ck}\right\|^2$$

$$-\frac{216(\alpha_V-2a_V^2-\frac{1}{9})(\Xi\psi_M\phi_M)^2}{\phi_{\min}^2}\left\|\widetilde{W}_{Vk}\right\|^4 \tag{37}$$

$$-(1-8\alpha_C^4\frac{\Psi_M^4}{\Psi_{\min}^4})\frac{6(\Xi\psi_M\phi_M)^2}{\phi_{\min}^2}\left\|\widetilde{W}_{Ck}\right\|^4+\varepsilon_{TM}$$

where, $\eta=[\Psi_M^2+\dfrac{9(\varepsilon_{VM}^{'}\psi_M)^2\phi_M^2\Xi^2}{2\phi_{\min}^2}+\dfrac{6(\Xi\psi_M)^2\phi_M^2\Delta\bar{\varepsilon}_{CM}^2}{\phi_{\min}^2\Psi_{\min}^2}]$ and

$\rho=\dfrac{\phi_M^2\Xi^2}{\phi_{\min}^2}(G_M^2+12\dfrac{\psi_M^2\Delta\varepsilon_{VM}^2}{\Delta\vartheta_{\min}^2})$ are positive constant and ε_{TM} is

$$\varepsilon_{TM}=8G_M^2\varepsilon_{uM}^2+16\frac{\Psi_M^2}{\Psi_{\min}^2}\Delta\bar{\varepsilon}_{CM}^2+\frac{24(\phi_M^2+1)G_M^2\phi_M^2}{\phi_{\min}^2}\Delta\varepsilon_{eM}^2$$

$$+\frac{(23G_M\phi_M\Xi)^2}{\phi_{\min}^2\Delta\vartheta_{\min}^2}\Delta\varepsilon_{VM}^2+\frac{72(\varepsilon_{VM}^{'}\psi_M)^2\phi_M^2\Xi^2}{\phi_{\min}^2}\Delta\bar{\varepsilon}_{CM}^2+\frac{48(\Xi\psi_M\phi_M)^2}{\phi_{\min}^2\Psi_{\min}^4}\Delta\bar{\varepsilon}_{CM}^4$$

$$+\frac{96(\Xi\psi_M\phi_M)^2\Delta\varepsilon_{VM}^4}{\phi_{\min}^2\Delta\vartheta_{\min}^2(\Delta\vartheta_M^2+1)}+\frac{(68\Xi\psi_M\phi_M)^2\Delta\varepsilon_{VM}^4}{\phi_{\min}^2\Delta\vartheta_{\min}^4}+\frac{96(\Xi\psi_M)^2\phi_M^2\Delta\bar{\varepsilon}_{CM}^4}{\phi_{\min}^2\Psi_{\min}^4}$$

Therefore, ΔL is less than zero when the following inequalities hold

$$\left\|\widetilde{e}_{yk}\right\|>\sqrt{\frac{\varepsilon_{TM}}{(1-\alpha_C^2)}}\equiv b_{ey}\text{ or }\left\|\widetilde{W}_{Ck}\right\|>\max\{\sqrt{\frac{\varepsilon_{TM}\Psi_{\min}^2}{4(\Psi_{\min}^2-4\alpha_C^2\Psi_M^2)\eta}} \tag{38}$$

$$,\sqrt[4]{\frac{\varepsilon_{TM}\Psi_{\min}^4\phi_{\min}^2}{6(\Psi_{\min}^4-8\alpha_C^4\Psi_M^4)(\Xi\psi_M\phi_M)^2}}\}\equiv b_{WC}\quad\text{or}\quad\left\|\widetilde{W}_{Vk}\right\|>\max\{$$

$$\sqrt{\frac{\varepsilon_{TM}}{288\rho(\alpha_V-2a_V^2-\frac{1}{12})}},\sqrt[4]{\frac{\phi_{\min}^2}{216(\alpha_V-2a_V^2-\frac{1}{9})(\Xi\psi_M\phi_M)^2}}\}\equiv b_V\text{ or}$$

$$\left\|\widetilde{W}_{uk}\right\|>\sqrt{\frac{\varepsilon_{TM}}{8G_M^2\phi_M^2(9\alpha_u-18\alpha_u^2-1)}}\equiv b_u\text{ or}\left\|y_k^o\right\|>\sqrt{\frac{\varepsilon_{TM}}{(1-2k^*)}}\equiv b_y$$

provided the tuning gains are selected according to (16), (24) and (27). Using the standard Lyapunov extension [10], the system outputs, NN identifier and weight estimation errors, critic NN and action NN estimation errors are *UUB*.

Simulation results are not included due to space constraints.

IV. Conclusions

In this work, an online time-based approximate dynamic programming technique for NNCS was proposed using three NNs to solve the stochastic optimal regulation of NNCS with unknown dynamics in presence of random delays and packet losses. The NN identifier relaxed the requirement of input gain matrix for NNCS. The history of past cost-to-go values relaxed the need for value and policy iterations while rendering a truly forward-in-time scheme that can be implemented in practical NNCS.

The initial admissible control policy ensured that NNCS is stable while NN identifier learns the input gain matrix, the critic NN approximates the stochastic value function $V(y_k^o)$, and the action NN generates the approximate stochastic optimal control. All NN weights were tuned online using proposed update laws and Lyapunov theory demonstrates the asymptotic convergence of the approximated control input to its optimal value over time.

References

[1] G. C. Walsh, O. Beldiman, and L. Bushnell. "Asymptotic behavior of networked control systems". in *Proc. IEEE Int. Conf. Contr. App.*, pp. 1448–1453, 1999.

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

[3] G. P. Liu, Y. Xia, D. Rees, and W. S. Hu, "Design and stability criteria of networked predictive control systems with random network delay in the feedback channel," *IEEE Trans. Syst., Man Cybern.*, vol. 37, pp. 173–184, 2007.

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

[5] 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.

[6] P. J. Werbos, "A menu of designs for reinforcement learning over time". *J. Neural Networks Contr.* pp. 67–95. MA: MIT Press, 1991.

[7] C. Zheng and S. Jagannathan, "Generalized hamilton–jacobi–bellman formulation-based neural network control of affine nonlinear discrete-time systems," *IEEE Trans. Neural Networks*, vol. 19, no. 1, pp. 90–106, 2008.

[8] A. Al-Tamimi and F. L. Lewis, "Discrete-time nonlinear HJB solution using approximate dynamics programming: convergence proof*," IEEE Trans. Syst., Man, Cybern. B*, vol. 38, pp. 943-949, 2008.

[9] H. Zhang, Y. Luo, and D. Liu, "Neural network based near optimal control for a class discrete-time affine nonlinear system with control constraints," IEEE Trans. Neur. Netwo., vol. 20, pp 1490-1503, 2009.

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

[11] R. K. Lim and M. Q. Phan, "Identification of a multistep-ahead observer and its application to predictive control," *J. Guid. Control Dyn.*, vol. 20, pp. 1200–1206, 1997.

[12] L. W. Liou, and A. Ray, "A stochastic regulator for integrated communication and control systems: part I—formulation of control law". *ASME J. Dynamic Syst., Measure. Contr.*, vol. 4, pp. 604–611, 1991.

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