2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Finite Horizon Stochastic Optimal Control of Nonlinear Two-Player Zero-Sum Games under Communication Constraint

Hao Xu and S. Jagannathan

***Abstract*—In this paper, the finite horizon stochastic optimal control of nonlinear two-player zero-sum games, referred to as Nonlinear Networked Control Systems (NNCS) two-player zero-sum game, between control and disturbance input players in the presence of unknown system dynamics and a communication network with delays and packet losses is addressed by using neuro dynamic programming (NDP). The overall objective being to find the optimal control input while maximizing the disturbance attenuation. First, a novel online neural network (NN) identifier is introduced to estimate the unknown control and disturbance coefficient matrices which are needed in the generation of optimal control input. Then, the critic and two actor NNs have been introduced to learn the time-varying solution to the Hamilton-Jacobi-Isaacs (HJI) equation and determine the stochastic optimal control and disturbance policies in a forward-in-time manner. Eventually, with the proposed novel NN weight update laws, Lyapunov theory is utilized to demonstrate that all closed-loop signals and NN weights are uniformly ultimately bounded (UUB) during the finite horizon with ultimate bounds being a function of initial conditions and final time. Further, the approximated control input and disturbance signals tend close to the saddle-point equilibrium within finite-time. Simulation results are included.**

I. INTRODUCTION

NONLINEAR Networked Control Systems (NNCS) [1], which brings in a communication network to close the feedback loop between the nonlinear system and the controller, has been considered as the next-generation control system since many benefits such as high efficiency, flexibility, low installation cost, so on can be harvested. However, the unreliable communication network in the feedback-loop causes several challenging issues such as performance degradation and instability due to network imperfections such as random delays and packet losses while exchanging feedback among the plant and remote controller.

Recently, authors in [2] considered asymptotic behavior of NNCS with network-induced delays alone. A discrete-time framework has been proposed in [3] to analyze NNCS stability in presence of both delays and packet losses. All the schemes [2-3] need the knowledge of system dynamics and network imperfections for maintaining stability of NNCS. However, due to uncertain network imperfections, the NCS dynamics cannot be assumed to be known a priori. Moreover, besides stability, optimality is more preferred [4-5] for NCS. Thus, authors in [5] proposed an infinite horizon optimal control of NNCS under uncertain dynamics and unknown network imperfections. However, finite horizon optimal of NCS is not considered.

Finite horizon optimal problem is more difficult due to terminal constraints [8]. In addition, existing designs ignored the worst case disturbance issues referred to as NCS two-player zero-sum game [6] whereas there are several practical engineering applications including smart grid, aircraft, where the influence of disturbance has to be considered with the controller design. Other practical examples of two-player zero sum game include rock-paper-scissors, chess, go, or checkers to name a few. To optimize performance of zero-sum game, two-player min-max optimization H_∞ control problem [6] is introduced, where controller is a minimizing player and the disturbance is a maximizing player. The overall objective of the two-player zero-sum game is the force the players to attain the saddle-point equilibrium [11] by using adaptive decision strategy within an infinite horizon.

This paper for the first time considers the finite horizon optimal solution for NNCS two-player zero-sum games with uncertain dynamics by using neuro dynamic programming [7] (NDP) while incorporating terminal constraints. In [8], the disturbance effects are not considered. In [9], authors utilized synchronous policy iteration to attain optimal design of nonlinear two-player zero-sum games. However, iteration-based NDP methods [7-9] require significant number of iterations [10] within a fixed sampling interval which is not practical and hence time-based ADP approach is needed [10].

However, the existing NDP schemes [7-10] are unsuitable for finite horizon NNCS two-player zero-sum game since a) worst case disturbance has not been considered even though the authors in [6] had proven that there exists a saddle-point equilibrium in two-player zero-sum game [11], b) network imperfections stemming from unreliable communication network are ignored, and c) only infinite horizon optimal control [11] is addressed with partial system dynamics.

In contrast, a time-based NDP based approach is undertaken with the overall objective, being to find the optimal control input while maximizing the disturbance input, formulated as a two-player zero sum game in the presence of uncertain system dynamics and network imperfections such as delays and packet losses. First, to relax the need for partial system dynamics, a novel NN identifier is proposed to learn both control and disturbance coefficient matrices online. Then, a critic NN is introduced and tuned forward-in-time to learn stochastic value function of NNCS two-player zero-sum game within a finite time by using Hamilton-Jacob-Isaacs (HJI) equation [6], given the terminal constraint. Eventually, the proposed two actor NNs are utilized to estimate both the stochastic optimal control and disturbance inputs by minimizing and maximizing tuned stochastic value function respectively. It can be shown that the overall game reaches

Hao Xu and S. Jagannathan are with the Department of Electrical and Computer Engineering, Missouri University of Science and Technology, MO 65409 USA (e-mail: hx6h7@mst.edu, sarangap@mst.edu). Research is supported in part by NSF ECCS #1128281.

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

close to the saddle-point equilibrium [11].

II. BACKGROUND

A. NNCS Zero-sum Game

The block diagram of NNCS is same as [5] where the nonlinear feedback control-loop is closed by using a communication network. Since the communication network is shared [5], the NNCS in this paper considers the network-induced delays and packet losses including: (a) $\tau_{sc}(t)$: sensor-to-controller delay, (b) $\tau_{ca}(t)$: controller-to-actuator delay, and (c) $\gamma(t)$: indicator of network-induced packet losses.

Assumption 1: a) Due to the wide area communication network, the two types of network-induced delays are considered independent, ergodic and unknown whereas their probability distribution functions are considered known. The sensor-to-controller delay is assumed to be less than a sampling interval. b) The sum of two delays is considered to be bounded while initial state of system is assumed to be deterministic [4].

Incorporating the network-induced delays and packet losses, the original nonlinear two-player zero-sum game in affine form can be represented as

$$\dot{x}(t) = f(x(t)) + \gamma(t)g(x(t))u(t-\tau(t)) + \gamma(t)h(x(t))d(t-\tau(t)) \quad (1)$$

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

$\mathbf{I}^{n\times n}$ is $n\times n$ identity matrix, $x(t)\in\mathbb{R}^n$, $u(t)\in\mathbb{R}^m$, $d(t)\in\mathbb{R}^l$, $f(x)\in\mathbb{R}^{n\times n}$, $g(x)\in\mathbb{R}^{n\times m}$ and $h(x)\in\mathbb{R}^{n\times l}$ represent system state, control inputs, disturbance inputs, nonlinear internal dynamics, control coefficient and disturbance coefficient matrix respectively.

Similar to [5], integrating (1) over a sampling interval $[kT_s,(k+1)T_s)$ with network-induced delays and packet losses and introducing a new augment state as $z_k = [x_k^T \; u_{k-1}^T \cdots u_{k-\bar{b}}^T \quad d_{k-1}^T \cdots d_{k-\bar{b}}^T]^T \in \mathbb{R}^{n+\bar{b}(m+l)}$, the NNCS two-player zero-sum game can be represented in compact form as

$$z_{k+1} = F(z_k) + G(z_k)u_k + H(z_k)d_k \quad (2)$$

with $I_m \in \mathbb{R}^{m\times m}$, $I_l \in \mathbb{R}^{l\times l}$ are the identity matrices,

$F(z_k) = \left[Z_{\tau,\gamma}^T(z_k) \;\; 0 \;\; u_{k-1}^T \;\; \cdots \;\; u_{k-\bar{b}}^T \;\; d_{k-1}^T \;\; \cdots \;\; d_{k-\bar{b}}^T\right]^T$,

$G(z_k) = \left[P_{\tau,\gamma}^T(z_k) \;\; I_m \;\; 0 \;\; \cdots \;\; 0 \;\; 0 \;\; \cdots \;\; 0\right]^T$, and

$H(z_k) = \left[K_{\tau,\gamma}^T(z_k) \;\; 0 \;\; 0 \;\; \cdots \;\; I_l \;\; 0 \;\; \cdots \;\; 0\right]^T$. Moreover, $\bar{b}\,T_s$ is upper bound on network-induced delay, T_s is sampling interval, $x_k = x(kT_s)$, $u_{k-i} = u((k-i)T_s)$, $d_{k-i} = d((k-i)T_s)$ $\forall i = 0,1,\ldots\bar{b}$ are discretized current NNCS system state, previous control inputs and disturbance signals. Additionally, $X_{\tau,\gamma}(\bullet)$, $P_{\tau,\gamma}(\bullet)$, $K_{\tau,\gamma}(\bullet)$ for NNCS two-play zero-sum game has been derived as

$$X_{\tau,\gamma}(z_k) = x_k + \int_{kT_s}^{(k+1)T_s} f(x(t))dt + \gamma_{k-\bar{d}}\left(\int_{kT_s}^{\tau_d-\bar{b}T_s} g(x(t))dt\right)u_{k-\bar{b}} + \cdots +$$
$$\gamma_{k-1}\left(\int_{\tau_2-2T_s}^{\tau_1-T_s} g(x(t))dt\right)u_{k-1} + \gamma_{k-d}\left(\int_{kT_s}^{\tau_d-\bar{b}T_s} h(x(t))dt\right)d_{k-\bar{b}} + \cdots + \cdots$$
$$+ \gamma_{k-1}\left(\int_{\tau_2-2T_s}^{\tau_1-T_s} h(x(t))dt\right)d_{k-1}, \; P_{\tau,\gamma}(z_k) = \gamma_k\left(\int_{\tau_0}^{(k+1)T_s} g(x(t))dt\right) \text{and}$$
$$K_{\tau,\gamma}(z_k) = \gamma_k\left(\int_{\tau_0}^{(k+1)T_s} h(x(t))dt\right).$$

Further, note that $F(\bullet)$, $G(\bullet)$ and $H(\bullet)$ in (2) represent NNCS two-player zero-sum game internal dynamics, control and disturbance coefficient matrices with $\|G(z_k)\|_F \le G_M$ and $\|H(z_k)\|_F \le H_M$ [10] where $\|\bullet\|_F$ denotes the Frobenius norm [12] and G_M, H_M are positive constants [10]. Moreover, since network-induced delays and packet losses have been considered in the NNCS two-player zero-sum game representation, equation (2) becomes uncertain and stochastic thus needing adaptive control methods.

B. Traditional Stochastic Optimal Strategies

Consider the nonlinear two-player zero-sum game as

$$x_{DT,k+1} = f_{DT}(x_{DT,k}) + g_{DT}(x_{DT,k})u_{DT,k} + h_{DT}(x_{DT,k})d_{DT,k} \quad (3)$$

where $x_{DT,k}$, $u_{DT,k}$, $d_{DT,k}$ represent the system state, control input and disturbance signals and $f_{DT}(x_{DT,k})$, $g_{DT}(x_{DT,k})$, $h_{DT}(x_{DT,k})$ denote the internal dynamics, control and disturbance coefficient matrices respectively in discrete-time. According to traditional optimal control [6][11] of two-player zero-sum game, the finite horizon optimal strategies can be derived to minimize the value function which is expressed as

$$\begin{cases} V_k(x_{DT,k},k) = E[\phi_N(x_{DT,N}) + \sum_{l=k}^{N-1} r(x_{DT,l},u_{DT,l},d_{DT,l})], k = 0,..,N-1 \\ V_N(x_{DT,N},N) = E[\phi_N(x_{DT,N})] \end{cases} \quad (4)$$

with cost-to-go is denoted as $r(x_{DT,k},u_{DT,k},D_{DT,k}) = Q_{DT}(x_{DT,k})$ $+u_{DT,k}^T R_{DT} u_{DT,k} - \gamma^2 d_{DT,k}^T S_{DT} d_{DT,k}$, $\forall k = 0,...,N-1$, NT_s is the final time instant, γ is a positive constant (i.e. $\gamma > 0$), $Q_{DT}(x) \ge 0$, $\varphi_N(x) \ge 0$ and R_{DT}, S_{DT} are symmetric positive definite matrices. In contrast to infinite horizon design [6], $\phi_N(x)$ is the terminal constraint which needs to be satisfied in the finite horizon two-player zero-sum game optimal design. According to dynamic programming technique [7], equation (4) can also be represented for $k = 0,...,N-1$ as

$$V_k(x_{DT,k},k) = E\left\lfloor r(x_{DT,k},u_{DT,k},d_{DT,k}) + V_{k+1}(x_{DT,k+1},k+1)\right\rfloor \quad (5)$$

Similar to [5], when $x = 0$, $V_k(x,k) = 0$, the value function $V_k(x,k)$ serves as the Lyapunov function [12]. Based on Bellman principle of optimality [7], the optimal value function also satisfies discrete-time Hamilton-Jacob-Isaacs (HJI) equation described by

$$\begin{cases} V^*(x_{DT,k},k) = \min_{u_{DT,k}} \max_{d_{DT,k}} \left\lfloor E\left[Q_{DT}(x_{DT,k}) + u_{DT,k}^T R_{DT} u_{DT,k} - d_{DT,k}^T S_{DT} d_{DT,k} + V^*(x_{DT,k+1})\right]\right\rfloor \\ V^*(x_{DT,N},N) = E\left[\phi_N(x_{DT,N})\right] \end{cases} \quad (6)$$

For finite horizon nonlinear two-player zero-sum game, controller u_k is viewed as the player minimizing the cost function while the disturbance d_k maximizes the cost. According to [6][11], this two-player zero-sum game optimal design has a unique solution while Nash condition holds as

$$\min_u \max_d V(x,0) = \max_d \min_u V(x,0) \tag{7}$$

According to [6][11], the optimal control and disturbance policies, $u^*_{DT,k}, d^*_{DT,k}$ can be derived by differentiating (6) as

$$u^*(x_{DT,k}) = -\frac{1}{2} E\left[R_{DT}^{-1} g_{DT}^T(x_{DT,k}) \frac{\partial V^*(x_{DT,k+1}, k+1)}{\partial x_{DT,k+1}} \right],\ k = 0,...,N-1 \tag{8}$$

$$d^*(x_{DT,k}) = \frac{1}{2\gamma^2} E\left[S_{DT}^{-1} h_{DT}^T(x_{DT,k}) \frac{\partial V^*(x_{DT,k+1}, k+1)}{\partial x_{DT,k+1}} \right],\ k = 0,...,N-1 \tag{9}$$

where γ is a positive parameter. Substituting (8),(9) into (6), the discrete-time HJI equation (6) is given by

$$V^*(x_{DT,k}, k) = E\Bigg[Q_{DT}(x_{DT,k}) + \frac{1}{4} \frac{\partial V^{*T}(x_{DT,k+1})}{\partial x_{DT,k+1}} g_{DT}(x_{DT,k}) R_{DT}^{-1} g_{DT}^T(x_{DT,k})$$
$$\times \frac{\partial V^{*T}(x_{DT,k+1})}{\partial x_{DT,k+1}} - \frac{1}{4\gamma^4} \frac{\partial V^{*T}(x_{DT,k+1})}{\partial x_{DT,k+1}} h_{DT}(x_{DT,k}) \tag{10}$$
$$\times S_{DT}^{-1} h_{DT}^T(x_{DT,k}) \frac{\partial V^{*T}(x_{DT,k+1})}{\partial x_{DT,k+1}} + V^*(x_{DT,k+1}, k+1) \Bigg], k = 0,..,N-1$$

III. Finite Horizon Stochastic Optimal Design

The control and disturbance coefficient matrices are needed for selecting an optimal strategy for the nonlinear two-player zero-sum game [6] whereas these are unknown. To circumvent this issue, a novel NN-based identifier is proposed.

A. Online NN-identifier Design

According to [5], the NNCS two-player zero-sum game internal dynamics, control and disturbance coefficient matrices can be represented on a compact set Ω as

$$\begin{aligned} F(z_k) &= \underset{\tau,\gamma}{E}[W_F^T \upsilon_F(z_k) + \varepsilon_{F,k}] \quad \forall k = 0,1,\ldots,N \\ G(z_k) &= \underset{\tau,\gamma}{E}[W_G^T \upsilon_G(z_k) + \varepsilon_{G,k}] \quad \forall k = 0,1,\ldots,N \\ H(z_k) &= \underset{\tau,\gamma}{E}[W_H^T \upsilon_H(z_k) + \varepsilon_{H,k}] \quad \forall k = 0,1,\ldots,N \end{aligned} \tag{11}$$

with $W_F \in \mathbb{R}^{p_f \times n}, W_G \in \mathbb{R}^{p_g \times n}, W_H \in \mathbb{R}^{p_h \times n}$ denote the target NN weights, $\upsilon_F(\bullet) \in \mathbb{R}^{p_f}, \upsilon_G(\bullet) \in \mathbb{R}^{p_g}, \upsilon_H(\bullet) \in \mathbb{R}^{p_h}$ are activation functions and $\varepsilon_{F,k} \in \mathbb{R}^{p_f}\ \varepsilon_{G,k} \in \mathbb{R}^{p_g}, \varepsilon_{H,k} \in \mathbb{R}^{p_h}$ are reconstruction errors respectively.

Substituting (11) into NNCS two-player zero-sum game dynamics (2), we get

$$z_k = \underset{\tau,\gamma}{E}[W_I^T \upsilon_I(z_{k-1}) \beta_{k-1} + \varepsilon_{I,k-1}],\ \forall k = 0,1,...,N \tag{12}$$

with $W_I = [W_F^T\ W_G^T\ W_H^T]^T$ and $\upsilon_I(z_{k-1}) = diag[\upsilon_F(z_{k-1})\ \upsilon_G(z_{k-1})\ \upsilon_H(z_{k-1})]$ are NN identifier target weight and activation function respectively, $\beta_{k-1} = [I_n^T\ u_{k-1}^T\ d_{k-1}^T]^T \in \mathbb{R}^{n+m+l}$ which includes $I_n = [1\ 1 \cdots 1]^T \in \mathbb{R}^{n \times 1}$, historical control and disturbance inputs u_{k-1}, d_{k-1} is defined as augment input at time $(k-1)T_s$ and $\varepsilon_{I,k-1} = \varepsilon_{F,k-1} + \varepsilon_{G,k-1} u_{k-1} + \varepsilon_{H,k-1} d_{k-1}$ denotes NN identifier reconstruction error. Moreover, $\underset{\tau,\gamma}{E}(\bullet)$ is expectation operator.

Since the NN activation function and augmented input from previous time instants are bounded, which will be proven via Lyapunov given a bounded initial condition, the term $\left\| \underset{\tau,\gamma}{E}[\upsilon_I(z_{k-1})\beta_{k-1}] \right\|$ will be bounded i.e. $\left\| \underset{\tau,\gamma}{E}[\upsilon_I(z_{k-1})\beta_{k-1}] \right\| \le \zeta_M$ where ζ_M is a positive constant [5][10]. In addition, the NN identifier reconstruction error is considered to be bounded such that $\left\| \underset{\tau,\gamma}{E}[\varepsilon_{I,k-1}] \right\| \le \varepsilon_{I,M}$ [10], with $\varepsilon_{I,M}$ is a positive constant. Using (11) and given NN activation functions $\upsilon_F(\bullet),\ \upsilon_G(\bullet), \upsilon_H(\bullet)$ and $\upsilon_I(\bullet)$, $G(z), H(z)$ can be identified while the NN identifier weight, W_I, is being tuned.

The NNCS two-player zero-sum game system state z_k can be approximated by using the NN identifier as

$$\hat{z}_k = \underset{\tau,\gamma}{E}[\hat{W}_{I,k}^T \upsilon_I(z_{k-1}) \beta_{k-1}],\ \forall k = 0,1,\ldots,N \tag{13}$$

with $\hat{W}_{I,k}$ is the estimated weight matrix of the NN identifier at time kT_s, $\underset{\tau,\gamma}{E}[\overline{\upsilon}_I(z_k)\beta_k]$ is activation function of NN identifier.

Next, the identification error can be expressed as

$$\underset{\tau,\gamma}{E}(e_{I,k}) = \underset{\tau,\gamma}{E}(z_k - \hat{z}_k) = \underset{\tau,\gamma}{E}(z_k) - \underset{\tau,\gamma}{E}[\hat{W}_{I,k}^T \upsilon_I(z_{k-1}) \beta_{k-1}] \tag{14}$$

Moreover, similar to [10] and using the history of NNCS two-player zero-sum game, the auxiliary identification error vector can be represented as

$$\underset{\tau,\gamma}{E}(\Xi_{I,k}) = \underset{\tau,\gamma}{E}(Z_k - \hat{Z}_k) = \underset{\tau,\gamma}{E}(Z_k) - \underset{\tau,\gamma}{E}[\hat{W}_{I,k}^T \overline{\upsilon}_I(Z_{k-1}) \overline{\beta}_{k-1}] \tag{15}$$

with $Z_k = [z_k\ \ z_{k-1} \cdots z_{k+1-i}], \overline{\upsilon}_I(Z_{k-1}) = [\upsilon_I(z_{k-1})\ \upsilon_I(z_{k-2}) \cdots \upsilon_I(z_{k-i})], \overline{\beta}_{k-1} = diag[\beta_{k-1}^T \cdots \beta_{k-i}^T]^T$ and $\overline{\varepsilon}_{I,k-1} = [\varepsilon_{I,k-1} \cdots \varepsilon_{I,k-i}]$ with $0 < i < k-1$. The i previous identification errors (14) are recomputed by using most recently NN identifier weights.

Next, auxiliary identification error dynamics can be derived

$$\underset{\tau,\gamma}{E}(\Xi_{I,k+1}) = \underset{\tau,\gamma}{E}(Z_{k+1}) - \underset{\tau,\gamma}{E}[\hat{W}_{I,k+1}^T \overline{\upsilon}_I(Z_k) \overline{\beta}_k], \forall k = 0,1,..,N-1 \tag{16}$$

For tuning NN identifier target weights within the finite horizon, the update law for $\underset{\tau,\gamma}{E}(\hat{W}_{I,k})$ can be expressed as

$$\underset{\tau,\gamma}{E}(\hat{W}_{I,k+1}) = \underset{\tau,\gamma}{E}[\overline{U}_k \overline{\upsilon}_I(Z_k)(\overline{\upsilon}_I^T(Z_k) \overline{U}_k^T \overline{U}_k \overline{\upsilon}_I(Z_k))^{-1}(Z_k - \alpha_I \Xi_{I,k})^T] \tag{17}$$

where the tuning parameter α_I satisfies $0 < \alpha_I < 1$. Utilizing the update law (17) into the auxiliary error dynamics (15), the auxiliary error dynamics $\underset{\tau,\gamma}{E}(\Xi_{I,k+1})$ can be represented as

$$\underset{\tau,\gamma}{E}(\Xi_{I,k+1}) = \alpha_I \underset{\tau,\gamma}{E}(\Xi_{I,k}), \quad \forall k = 0,1,\ldots,N-1 \tag{18}$$

In order to learn the NNCS control and disturbance coefficient matrices $G(z), H(z)$ by using proposed online NN identifier, $\underset{\tau,\gamma}{E}[\overline{\upsilon}_I(Z_k)\overline{\beta}_k]$ has to be persistently existing (PE) [5][10] long enough. In other words, there exists a positive constant $\zeta_{\min}$ such that $0 < \zeta_{\min} \le \left\| \underset{\tau,\gamma}{E}[\overline{\upsilon}_I(Z_k)\overline{\beta}_k] \right\|$ holds for $k = 0,1,\ldots,N$.

Next, identification error dynamics (15) can be expressed as

$$\underset{\tau,\gamma}{E}(e_{I,k+1}) = \underset{\tau,\gamma}{E}[\tilde{W}_{I,k+1}^T \upsilon_I(z_k) \beta_k + \varepsilon_{I,k}], \forall k = 0,1,.., N-1 \tag{19}$$

where $\widetilde{W}_{I,k} = W_I - \hat{W}_{I,k}$ is NN identifier weight estimation error at time kT_s.

B. Stochastic Value Function Setup and Critic NN Design

According to the value function defined in [5][10], the stochastic value function for NNCS two-player zero-sum game can be expressed in terms of augment state z_k as

$$\begin{cases} V(z_k,k) = \underset{\tau,\gamma}{E}\left[\phi_N(z_N) + \sum_{l=k}^{N-1}(Q_z(z_l) + u_l^T R_z u_l - \gamma^2 d_l^T S_z d_l)\right], k = 0,..,N-1 \\ V(z_N,N) = \underset{\tau,\gamma}{E}\left[\phi_N(z_N)\right] \end{cases} \quad (20)$$

with $Q_z(z_k) \geq 0$ and R_z, S_z are positive definite matrices, γ is a positive constant (i.e. $\gamma > 0$). In contrast to stochastic value function under infinite horizon, a terminal constraint (i.e. $V_N(z_N,N) = \underset{\tau,\gamma}{E}\left[\phi_N(z_N)\right]$) is incorporated while deriving the finite horizon stochastic optimal design.

According to [10], the stochastic value function (20) can be represented by using a critic NN as

$$V(z_k,k) = \underset{\tau,\gamma}{E}(W_V^T \varphi(z_k, N-k) + \varepsilon_{V,k}),\ \forall k = 0,1,\ldots,N \quad (21)$$

with $W_V \in \mathbb{R}^r, \varepsilon_{V,k} \in \mathbb{R}$ represent the critic NN target weight matrix and reconstruction error respectively, and $\varphi(z_k, N-k) \in \mathbb{R}^r$ denotes the time-dependent critic NN activation function. It is important to note that the activation function explicitly dependent upon time and this makes the finite horizon problem different and difficult over the infinite horizon case [10]. Moreover, the critic NN target weight and reconstruction errors are bounded in the mean as $\left\|\underset{\tau,\gamma}{E}(W_V)\right\| \leq W_{VM}, \left\|\underset{\tau,\gamma}{E}(\varepsilon_{V,k})\right\| \leq \varepsilon_{VM}$ with W_{VM}, ε_{VM} being positive constants [10]. In addition, the reconstruction error gradient is assumed to be bounded in the mean as $\left\|\underset{\tau,\gamma}{E}(\partial\varepsilon_{V,k}/\partial z_k)\right\| \leq \varepsilon_{VM}^{'}$ with $\varepsilon_{VM}^{'}$ being a positive constant [10].

Next, the critic NN approximation of stochastic value function (21) can be expressed as

$$\hat{V}(z_k,k) = \underset{\tau,\gamma}{E}(\hat{W}_{V,k}^T \varphi(z_k, N-k)),\ \forall k = 0,1,...,N \quad (22)$$

where $\underset{\tau,\gamma}{E}(\hat{W}_{V,k})$ is the estimated critic NN weight matrix and time dependent activation function $\varphi(z_k, N-k)$ has been selected from a basis function set whose elements in the set are linearly independent [10]. Also, since the activation function is continuous and smooth, two time independent functions $\varphi_{\min}(z_k)$, $\varphi_{\max}(z_k)$ can be found such that $\|\varphi_{\min}(z_k)\| \leq \|\varphi(z_k, N-k)\| \leq \|\varphi_{\max}(z_k)\|, k = 0,\ldots,N$. Moreover, the stochastic value function (15) can be bounded as $\left\|\underset{\tau,\gamma}{E}(W_V)\right\|\|\phi_{\min}(z_k)\| - \varepsilon_{VM}\Big|^2 \leq \|V(z_k,k)\|^2 \leq \left|\left\|\underset{\tau,\gamma}{E}(W_V)\right\|\|\phi_{\min}(z_k)\| + \varepsilon_{VM}\right|^2$. In addition, recalling (21) and (22), value function estimation error, $\widetilde{V}(z_k,k) = \widetilde{W}_{V,k}^T \varphi(z_k, N-k) + \varepsilon_{V,k}$, can also be bounded as $\left|\left\|\underset{\tau,\gamma}{E}(\widetilde{W}_{V,k})\right\|\|\phi_{\min}(z_k)\| - \varepsilon_{VM}\right|^2 \leq \|V(z_k,k)\|^2 \leq \left|\left\|\underset{\tau,\gamma}{E}(\widetilde{W}_{V,k})\right\|\|\phi_{\min}(z_k)\| + \varepsilon_{VM}\right|^2$

Recalling HJI equation [6][9], the following can be derived by substituting (20) into the HJI equation as

$$\underset{\tau,\gamma}{E}(\varepsilon_{V,k} - \varepsilon_{V,k+1}) = \underset{\tau,\gamma}{E}[W_V^T(\varphi(z_{k+1}, N-k-1) - \varphi(z_k, N-k))] + \underset{\tau,\gamma}{E}(z_k^T Q_z z_k + u_k^T R_z u_k - d_k^T S_z d_k),\quad \forall k = 0,1,..., N-1 \quad (23)$$

In the other words,

$$\underset{\tau,\gamma}{E}(W_V^T \Delta\varphi(z_k, N-k)) + r(z_k,u_k,d_k) = \Delta\varepsilon_{V,k}, k = 0,.., N-1 \quad (24)$$

where $\Delta\varphi(z_k, N-k) = \varphi(z_{k+1}, N-k-1) - \varphi(z_k, N-k)$, $r(z_k,u_k,d_k) = \underset{\tau,\gamma}{E}(z_k^T Q_z z_k + u_k^T R_z u_k - d_k^T S_z d_k)$ and $\Delta\varepsilon_{V,k} = \underset{\tau,\gamma}{E}(\varepsilon_{V,k} - \varepsilon_{V,k+1})$ with $\|\Delta\varepsilon_{V,k}\| = \Delta\varepsilon_{VM}, \forall k = 0,..., N-1$ [10]. However, when an estimated value of critic NN, $\hat{V}(z_k,k)$ from (22), is utilized instead of ideal critic NN output, $V(z_k,k)$, equation (23) does hold. Then using ideas similar to [10], and incorporating delay values for convenience, the temporal difference (TD) error dynamics associated with (24) are introduced as

$$\underset{\tau,\gamma}{E}(e_{HJI,k}) = \underset{\tau,\gamma}{E}(\hat{W}_{V,k}^T \Delta\psi(z_k, N-k)) + r(z_k,u_k,d_k),\ k = 0,...,N-1 \quad (25)$$

with $\underset{\tau,\gamma}{E}(e_{HJI,k})$ denotes the HJI equation TD error for the finite horizon case (i.e. $t \in [0, NT_s]$). Moreover, since $r(z_k,u_k,d_k) = \Delta\varepsilon_{V,k} - \underset{\tau,\gamma}{E}(W_V^T \Delta\varphi(z_k, N-k)), \forall k = 0,1,..., N$ (24), the HJI equation TD error dynamics can be derived as

$$\underset{\tau,\gamma}{E}(e_{HJI,k}) = -\underset{\tau,\gamma}{E}(\widetilde{W}_{V,k}^T \Delta\phi(z_k, N-k)) + \Delta\varepsilon_{V,k}, k = 0,..., N-1 \quad (26)$$

where $\underset{\tau,\gamma}{E}(\widetilde{W}_{V,k}) = \underset{\tau,\gamma}{E}(W_V) - \underset{\tau,\gamma}{E}(\hat{W}_{V,k})$ represents the critic NN weight estimation error.

Next, in order to incorporate the terminal constraint, the estimation error $\underset{\tau,\gamma}{E}(e_{FC,k})$ can be defined as

$$\underset{\tau,\gamma}{E}(e_{FC,k}) = \underset{\tau,\gamma}{E}\left[\phi_N(z_N)\right] - \underset{\tau,\gamma}{E}(\hat{W}_{V,k}^T \varphi(\hat{z}_{N,k}, 0)),\ \forall k = 0,1,\ldots,N \quad (27)$$

where $\hat{z}_{N,k}$ is the estimated final NNCS two-player zero-sum game system state at time kT_s by using NN identifier (i.e. $\hat{F}(\bullet), \hat{G}(\bullet), \hat{H}(\bullet)$).

Considering the HJI TD error and terminal constraint estimation error jointly and using the gradient descent technique, update law for critic NN weight can be derived by

$$\underset{\tau,\gamma}{E}(\hat{W}_{V,k+1}) = \underset{\tau,\gamma}{E}(\hat{W}_{V,k}) + \alpha_V \underset{\tau,\gamma}{E}\left(\frac{\varphi(\hat{z}_{N,k},0) e_{FC,k}^T}{\varphi^T(\hat{z}_{N,k},0)\varphi(\hat{z}_{N,k},0) + 1}\right) - \alpha_V \underset{\tau,\gamma}{E}\left(\frac{\Delta\varphi(z_k, N-k) e_{HJI,k}^T}{\Delta\varphi^T(z_k, N-k)\Delta\varphi(z_k, N-k) + 1}\right), k = 0,.., N-1 \quad (28)$$

C. Actor NN Estimation of Control and Disturbance Inputs

Similar to [10], the ideal finite horizon NNCS optimal control and disturbance inputs can be expressed by using actor NNs as

$$\begin{aligned} u^*(z_k) &= \underset{\tau,\gamma}{E}[W_u^T \vartheta(z_k,k) + \varepsilon_{u,k}],\ \forall k = 0,1,..., N \\ d^*(z_k) &= \underset{\tau,\gamma}{E}[W_d^T \psi(z_k,k) + \varepsilon_{d,k}],\ \forall k = 0,1,..., N \end{aligned} \quad (29)$$

where $W_u \in \mathbb{R}^s, \varepsilon_{u,k} \in \mathbb{R}$ represent two actor NN target weight matrix and reconstruction errors for control input

respectively, $W_d \in \mathbb{R}^b, \varepsilon_{d,k} \in \mathbb{R}$ denote actor NN target weight matrix and reconstruction error, and $\vartheta(z_k,k) \in \mathbb{R}^s$, $\psi(z_k,k) \in \mathbb{R}^b$ represent smooth time-varying activation function for control and disturbance actor NNs respectively. Moreover, time independent functions $\vartheta_{\min}(z_k)$, $\vartheta_{\max}(z_k)$ $\psi_{\min}(z_k)$ and $\psi_{\max}(z_k)$ can be found such that $\|\vartheta_{\min}(z_k)\| \le \|\vartheta(z_k,k)\| \le \|\vartheta_{\max}(z_k)\|$ and $\|\psi_{\min}(z_k)\| \le \|\psi(z_k,k)\| \le \|\psi_{\max}(z_k)\|$ $k=0,...,N$. Additionally, actor NN weight matrices, activation functions and reconstruction errors are considered.

Next, similar to [5][10], estimated actor NNs for control and disturbance inputs can be represented as

$$\hat{u}(z_k) = \underset{\tau,\gamma}{E}[\hat{W}_{u,k}^T \vartheta(z_k,k)], \hat{d}(z_k) = \underset{\tau,\gamma}{E}[\hat{W}_{d,k}^T \psi(z_k,k)], \quad \forall k = 0,1,...,N \tag{30}$$

with $\underset{\tau,\gamma}{E}(\hat{W}_{u,k})$, $\underset{\tau,\gamma}{E}(\hat{W}_{d,k})$ denote the estimated weights for control and disturbance actor NNs respectively. Further, the actor NN estimation errors will be considered as the difference between the actual control and disturbance (30) inputs applied to the NNCS where the control policy is considered to be obtained by minimizing the tuned stochastic value function whereas the disturbance policy can be obtained by maximizing the tuned stochastic value function (22), i.e. $\min_u \max_d V(z)$, with identified control and disturbance coefficient matrices (i.e. $\hat{G}(z_k)$, $\hat{H}(z_k)$) within the finite horizon. The estimation errors are expressed as

$$\underset{\tau,\gamma}{E}(e_{u,k}) = \underset{\tau,\gamma}{E}\left[\hat{W}_{u,k}^T \vartheta(z_k,k) + \frac{1}{2} R_z^{-1} \hat{G}^T(z_k) \frac{\partial \phi^T(z_{k+1}, N-k-1)}{\partial z_{k+1}} \hat{W}_{V,k}\right]$$

$$\underset{\tau,\gamma}{E}(e_{d,k}) = \underset{\tau,\gamma}{E}\left[\hat{W}_{d,k}^T \psi(z_k,k) - \frac{1}{2\gamma^2} S_z^{-1} \hat{H}^T(z_k) \frac{\partial \phi^T(z_{k+1}, N-k-1)}{\partial z_{k+1}} \hat{W}_{V,k}\right] \tag{31}$$

Using gradient descent scheme, the update law for the two estimated actor NNs weight matrices can be represented as

$$\begin{aligned} \underset{\tau,\gamma}{E}(\hat{W}_{u,k+1}) &= \underset{\tau,\gamma}{E}(\hat{W}_{u,k}) - \alpha_u \underset{\tau,\gamma}{E}\left[\frac{\vartheta(z_k,k)}{\vartheta^T(z_k,k)\vartheta(z_k,k)+1} e_{u,k}^T\right], k = 0,.., N-1 \\ \underset{\tau,\gamma}{E}(\hat{W}_{d,k+1}) &= \underset{\tau,\gamma}{E}(\hat{W}_{d,k}) - \alpha_d \underset{\tau,\gamma}{E}\left[\frac{\psi(z_k,k)}{\psi^T(z_k,k)\psi(z_k,k)+1} e_{d,k}^T\right], k = 0,.., N-1 \end{aligned} \tag{32}$$

where the two actor NNs tuning parameters α_u, α_d satisfies $0 < \alpha_u < 1$ and $0 < \alpha_d < 1$.

D. Closed-loop Stability

The initial NNCS two-player zero-sum system state is considered to reside in a compact set Ω [5][10] due to the initial admissible policy $u_0(z_k)$, $d_0(z_k)$. In addition, two actor NNs activation functions, the critic NN activation function and its gradient are all considered to be bounded in a compact set Ω satisfying. $\left\|\underset{\tau,\gamma}{E}(\varphi(z_k,k))\right\| \le \varphi_M$, $\left\|\underset{\tau,\gamma}{E}\left[\frac{\partial \varphi(z_k,k)}{\partial z_k}\right]\right\| \le \varphi_M'$, $\left\|\underset{\tau,\gamma}{E}(\vartheta(z_k,k))\right\| \le \vartheta_M$ and $\left\|\underset{\tau,\gamma}{E}(\psi(z_k,k))\right\| \le \psi_M$ [5][10]) since the activation functions of NNs are selected to a bounded smooth continuous functions. Moreover, the PE condition will be held by adding exploration noise [5][10] and NN tuning parameters α_I, α_V and α_u, α_d will be chosen to ensure that all future system state remain in the compact set. In order to proceed, the following lemma is needed before demonstrating the main theorem.

Lemma 3 [5][10]: Let the optimal control and disturbance policies be utilized on the NNCS two-player zero-sum game (2) such that (2) is asymptotically stable in the mean [5]. Then, the closed-loop NNCS two-player zero-sum game dynamics, $\underset{\tau,\gamma}{E}[F(z_k) + G(z_k)u^*(z_k) + H(z_k)d^*(z_k)]$ can be expressed as

$$\left\|\underset{\tau,\gamma}{E}[F(z_k) + G(z_k)u^*(z_k) + H(z_k)d^*(z_k)]\right\|^2 \le l_o \left\|\underset{\tau,\gamma}{E}(z_k)\right\|^2 \tag{33}$$

with $u^*(z_k), d^*(z_k)$ are optimal control and disturbance signal policies respectively where $0 < l_o < 1/2$ is a positive constant.

Theorem 1 *(Convergence of optimal control and disturbance inputs)* Given $u_0(z_k)$, $d_0(z_k)$ be any initial admissible control and disturbance policy for the NNCS two-player zero-sum game (2) such that (36) holds with $0 < l_o < 1/2$. Consider the NN weight update laws for identifier, critic and two actor NNs as (17), (30) and (35) respectively, then there exists positive tuning parameters $\alpha_I, \alpha_V, \alpha_u, \alpha_d$ satisfying $0 < \alpha_I < \min\left\{\frac{1}{2\zeta_M}, \frac{\zeta_{\min}}{\sqrt{2}\zeta_M}\right\}$, $0 < \alpha_V < \frac{2-\chi}{\chi+5}$ with $0 < \chi = \frac{\psi_{\min}^2 + \Delta\psi_{\min}^2 + 2}{(\psi_{\min}^2+1)(\Delta\psi_{\min}^2+1)} < 2$ and $0 < \alpha_u < \frac{1}{3}$, $0 < \alpha_d < \frac{1}{3}$ such that the system state $\underset{\tau,\gamma}{E}(z_k)$, identification error $\underset{\tau,\gamma}{E}(e_{I,k})$, NN identifier weight estimation error $\underset{\tau,\gamma}{E}(\tilde{W}_{I,k})$, critic and two actor NNs weight estimation errors $\underset{\tau,\gamma}{E}(\tilde{W}_{V,k})$, $\underset{\tau,\gamma}{E}(\tilde{W}_{u,k})$, $\underset{\tau,\gamma}{E}(\tilde{W}_{d,k})$ are all *UUB* in the mean within $t \in [0, NT_s]$. Further, the ultimate bounds are a function of the final time, NT_s, bounded initial system state $B_{z,0}$, identification error $B_{I,0}$ and initial weight estimation error for NN identifier, critic and two actor NNs $B_{WI,0}$, $B_{WV,0}$, $B_{Wu,0}$, $B_{Wd,0}$ respectively. Moreover, $\left\|u_k^* - \hat{u}_k\right\| \le B_u$ and $\left\|d_k^* - \hat{d}_k\right\| \le B_d$ with B_u and B_d are small bounds.

Proof: Omitted due to space limitation.

IV. Simulation Results

Consider continuous-time version of the original nonlinear two-player zero-sum game in affine form from [9] given by

$$\dot{x} = f(x) + g(x)u + h(x)d \tag{34}$$

where $f(x) = \begin{bmatrix} -x_1 + x_2 \\ -0.5x_1 - 0.5x_2(1-(\cos(2x_1)+2)^2) \end{bmatrix}$,

$g(x) = \begin{bmatrix} 0 \\ \cos(2x_1)+2 \end{bmatrix}$ and $h(x) = \begin{bmatrix} 0 \\ \sin(4x_1)+2 \end{bmatrix}$.

The parameter of NNCS two-player zero-sum game are selected as [5]: 1) Sampling time: $T_s = 100ms$; 2) The upper bound of network-induced delay is given as two, i.e. $\bar{b} = 2$; 3) The network-induced delays are $E(\tau_{sc}) = 80ms$ and $E(\tau) = 150ms$; 4) Network-induced packet losses follow Bernoulli distribution with $\bar{\gamma} = 0.3$ and 5) Final time is set as $t_f = 20s$.

For incorporating network parameters and proposed design into NNCS two-player zero-sum game, the augment state is defined as $z_k = [x_k^T \; u_{k-1}^T \; u_{k-2}^T \; d_{k-1}^T \; d_{k-2}^T]^T \in \mathbb{R}^{6\times 1}$ and the initial state is chosen as $x_0 = [6 \; -3.5]^T$ while the initial admissible control and disturbance policies are given as $u_o(z_k) = [-2 \; -2.5 \; -1 \; -1 \; 0 \; 1] z_k$ and $d_o(z_k) = [-1 \; -2 \; -1 \; -1 \; 1 \; 0] z_k$ respectively. Similar to [5], activation function for NN identifier is taken as $\tanh\{(z_{k,1})^2, z_{k,1}z_{k,2}, \cdots, (z_{k,1})^5(z_{k,2}), ..., (z_{k,6})^6\}$, the critic NN state dependent part activation function is chosen as sigmoid of sixth order polynomial, that is, $\text{sigmoid}\{(z_{k,1})^2, z_{k,1}z_{k,2}, \ldots (z_{k,1})^5(z_{k,2}), ..., (z_{k,6})^6\}$ and time-dependent part of critic NN activation function is given as saturation polynomial time function, that is, $sat\{(N-k)^{10}, (N-k)^9, ..., 1; \cdots; 1, (N-k)^{10}, \ldots, N-k\}$, and activation function of two actor NNs are all selected as the gradient of critic NN activation function.

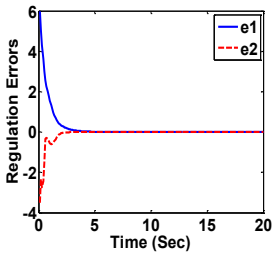

Fig. 1. NNCS state regulation errors.

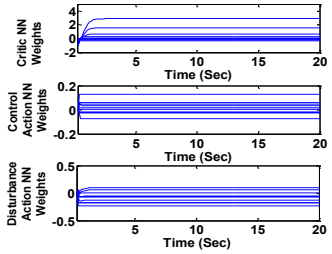

Fig. 2. The NN weights.

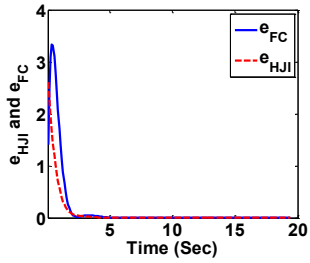

Fig. 3. HJI equation and terminal errors

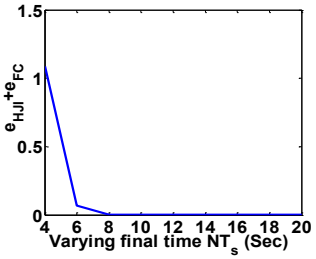

Fig. 4. Increasing final time

For the NN identifier, 39 hidden neurons are used. Hidden layer NN weights are initialized at random within (0,1] and input layer NN weights are all ones. Moreover, critical NN and two actor NNs use 55 hidden layer neurons. Further, hidden layer NN weights for critic NN are initialized to zero whereas the initial weights for two actor NNs weights are selected to reflect initial admissible policy. The input layer NN weights for critic NN and actor NNs are set as all ones.

As show in Figures 1, the proposed stochastic optimal control and disturbance inputs can force the NNCS two-player zero-sum game state regulation errors tend to zero closely within finite horizon. In other words, the proposed finite horizon strategies can maintain the NNCS two-player zero-sum game UUB in the mean even in the presence of uncertain NNCS dynamics and network imperfections. Moreover, in Figure 2, the approximated weights for critic and two actor NNs are demonstrated. It is important to note that the estimated weights of critic and two actor NNs converge to constant values and maintain UUB in the mean within the finite horizon, which is consistent as Theorem 1. Next, as shown in Figure 3, during the interval $t \in [0,20s]$, both HJI equation and terminal errors converge close to zero, which indicates the game solution satisfies optimality and the terminal constraint. Moreover, when the final time instant NT_s increases, the upper bound of sum of HJI equation and terminal constraint errors will decrease as shown in Figure 4.

V. Conclusions

In this paper, a novel time-based finite horizon NDP scheme was proposed for NNCS two-player zero-sum game by using NN identifier, critic and two actor NNs to solve the NNCS two-player zero-sum game in the presence of uncertain system dynamics and network imperfections. By using historical inputs and NN identifier, the requirement on both system internal dynamics, control and disturbance coefficient matrices was relaxed. Further, critic NN approximated the HJI equation solution online while satisfying the terminal constraint. An initial admissible control ensured that the system is stable when NNs were trained. Using Lyapunov theory, the closed-loop signals were shown to be *UUB* in the mean. When the final time NT_s increases, all the ultimate bounds will converge to zero as time goes to infinity.

References

[1] J. Hespanha, P. Naghshtabrizi, and Y. Xu, "A survey of recent results in networked control systems," *Proc. of IEEE*, vol.95, 2007, pp. 138–162.

[2] G. C. Walsh, O. Beldiman, and L.G. Bushnell, "Asymptotic behavior of nonlinear networked control systems," *IEEE Trans. Automat. Contr.*, vol. 46, 2001, pp. 1093–1097.

[3] N. V. D. Wouw, D. Nesic, and W. P. H. Heemels, "Stability analysis for nonlinear networked control systems: a discrete-time approach," in *Proc. IEEE Contr. Decision. Conf.*, 2010, pp. 7557–7563.

[4] H. Shousong and Z. Qixin, "Stochastic optimal control and analysis of stability of networked control systems with long delay," *Automatica*, vol. 39, 2003, pp. 1877–1884.

[5] H. Xu, and S. Jagannathan, "Stochastic optimal controller design for uncertain nonlinear networked control system via neuro dynamic programming," *IEEE Trans. on Neur. Net. Learn Syst.*, vol. 24, 2013, pp. 471-484.

[6] T. Basar, and G. J. Olsder, *Dynamic non-cooperative game theory*, 2nd Edition, Academic, New York, 1995.

[7] D. P. Bertsekas and J. Tsitsiklis, *Neuro-dynamics programming*, Athena Scientic, 1996.

[8] F. Y. Wang, N. Jin, D. Liu, and Q. L. Wei, "Adaptive dynamic programming for finite horizon optimal control of discrete-time nonlinear systems with ε-error bound," *IEEE Trans. On Neur. Net.*, vol. 22, 2011, pp. 24-36.

[9] K. Vamvoudakis, and F. L. Lewis, "Online solution of nonlinear two-player zero-sum games using synchronous policy iteration," in *Proc. Contr. Decision Conf.*, 2010, pp. 3040-3047.

[10] T. Dierks, and S. Jagannathan, "Online optimal control of affine nonlinear discrete-time systems with unknown internal dynamics by using time-based policy update", *IEEE Trans. Neur. Netw. Learn. Syst.*, vol. 23, 2012, pp. 1118-1129.

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

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