STOCHASTIC CONTROL STRATEGIES AND ADAPTIVE CRITIC METHODS

Randa Herzallah
Faculty of Engineering Technology, Al-Balqa' Applied University, Jordan

David Lowe
NCRG, Aston University, U.K.

Keywords: Adaptive critic methods, functional uncertainty, stochastic control.

Abstract: Adaptive critic methods have common roots as generalizations of dynamic programming for neural reinforcement learning approaches. Since they approximate the dynamic programming solutions, they are potentially suitable for learning in noisy, nonlinear and nonstationary environments. In this study, a novel probabilistic dual heuristic programming (DHP) based adaptive critic controller is proposed. Distinct to current approaches, the proposed probabilistic (DHP) adaptive critic method takes uncertainties of forward model and inverse controller into consideration. Therefore, it is suitable for deterministic and stochastic control problems characterized by functional uncertainty. Theoretical development of the proposed method is validated by analytically evaluating the correct value of the cost function which satisfies the Bellman equation in a linear quadratic control problem. The target value of the critic network is then calculated and shown to be equal to the analytically derived correct value.

1 INTRODUCTION

In recent research of stochastic control systems, much attention has been paid to the problem of characterizing and incorporating functional uncertainty of dynamical control systems. This is because there is an increasing demand for high reliability of complex control systems which are accompanied by high level of inherent uncertainty in modeling and estimation and are characterized by intrinsic nonlinear dynamics involving unknown functionals and latent processes. Several methods have been developed, and examples include feedback linearization techniques (Botto et al., 2000; Hovakimyan et al., 2001), backstepping techniques (Sastry and Isidori, 1989; Zhang et al., 2000; Lewis et al., 2000), neural network based methods (Wang and Huang, 2005; Ge and Wang, 2004; Ge et al., 2001; Murray-Smith and Sbarbaro, 2002; Fabri and Kadirkamanathan, 1998), stochastic adaptive control methods (Karny, 1996; Wang and Zhang, 2001; Wang, 2002; Herzallah and Lowe, 2007; Herzallah and Lowe,), and adaptive critic based methods (Herzallah, 2007).

In the feedback linearization, backstepping and neural network based methods, only parameters or forward model uncertainty have been considered. The inverse controller has been assumed to be deterministic or dependent on the forward model. Stochastic adaptive control methods on the other hand have considered modeling the distribution of the inverse controller. However uncertainty in the stochastic adaptive control methods proposed in (Karny, 1996; Wang and Zhang, 2001; Wang, 2002), has been treated as a nuisance or perturbation therefore; did not affect the derivation of the optimal control law. In other words, uncertainty has been assumed to be input–independent and consequently they did not contribute to the derivation of the optimal control law. The stochastic adaptive control methods developed in (Herzallah and Lowe, 2007; Herzallah and Lowe,) on the other hand have considered input–dependent uncertainty and the methods are proven to significantly improve the performance of the controller.

Selected adaptive Critic (AC) methods, known as action–independent adaptive critic methods, have been shown to implement useful approximations of Dynamic Programming, a method for designing

Herzallah R. and Lowe D. (2008).
STOCHASTIC CONTROL STRATEGIES AND ADAPTIVE CRITIC METHODS.
In *Proceedings of the Fifth International Conference on Informatics in Control, Automation and Robotics - ICSO*, pages 281-288
DOI: 10.5220/0001481902810288

optimal control policies in the context of nonlinear plants (Werbos, 1992). However in their conventional form, the action–independent adaptive critic methods do not take into consideration model uncertainty. In most recent development to these methods, a novel dual heuristic programming (DHP) adaptive–critic–based cautious controller is proposed (Herzallah, 2007). The proposed controller avoids the preidentification training phase of the forward model and inverse controller by taking into consideration model uncertainty when calculating the control law. Only forward model uncertainty has been considered in (Herzallah, 2007). The inverse controller is assumed to be accurate and no knowledge of uncertainty needed to be characterized. However, similar to the forward model, the parameters of the inverse controller of the nonlinear dynamical systems are usually optimized using nonlinear optimization methods. This inevitably leads to uncertain model of the inverse controller. Consequently, uncertainty of the inverse controller should be estimated and considered in the derivation of the optimal control law.

As a result, the dual heuristic programming (DHP) adaptive–critic–based cautious control method (Herzallah, 2007) is still in need of further development. This forms the main purpose of this paper, where functional uncertainty of both the forward model and the inverse controller is characterized and used in deriving the optimal control law. Hence the novelty of this work stems from considering functional uncertainty in the inverse controller as well as the forward model. Furthermore, a new method for estimating functional uncertainty of the models will be introduced in this work. In contrast to the method proposed in (Herzallah, 2007) this method allows for considering multiplicative noise on both the state and the control law. Also it guarantees the positivity of the covariance matrix of the errors. This well lead to a novel theoretical development for the stochastic adaptive control. Moreover, the Riccati solution for a quadratic linear infinite horizon control problem will also be derived and compared to the solution of the developed probabilistic (DHP) adaptive critic method. The method developed in this paper, enhances the performance of the system by utilizing more fully the probabilistic information provided by the forward model and the inverse controller. No pre–identification will be needed for neither the forward model, the critic or the inverse controller. All networks in the new developed framework will be adapted at each instant of time.

2 PRELIMINARIES

This preparatory section recalls basic elements of modeling conditional distributions of system outputs and inverse controller and the aim of fully probabilistic control.

2.1 Basic Elements

The behavior of a stochastic general class discrete time system with input $\mathbf{u}^{op}(k)$ and measurable state vector $\mathbf{x}(k)$ is described by a stochastic model of the following form

$$\mathbf{x}(k+1) = g[\mathbf{x}(k), \mathbf{u}^{op}(k)] + \tilde{\eta}(k+1) \qquad (1)$$

where $\tilde{\eta}(k+1)$ is random independent noise which has zero mean and covariance $\tilde{\mathbf{P}}$.

This can generally be expressed as:

$$\mathbf{x}(k+1) = f[\mathbf{x}(k), \mathbf{u}^{op}(k), \tilde{\eta}(k+1)]. \qquad (2)$$

The randomized controller to be designed is described by the following stochastic model

$$\mathbf{u}^{op}(k) = \quad c[\mathbf{x}(k)] + \tilde{\mathbf{e}}(k) \qquad (3)$$

where $\tilde{\mathbf{e}}(k)$ represents random independent noise of zero mean and $\tilde{\mathbf{Q}}$ covariance matrix. Notice that only state dependent controllers are considered. However, assuming state dependent controller can be shown to represent no real restrictions (Mine and Osaki, 1970) provided that the state can be measured. The stochastic model of the controller can be reexpressed in the following general form:

$$\mathbf{u}^{op}(k) = \quad h[\mathbf{x}(k), \tilde{\mathbf{e}}(k)]. \qquad (4)$$

All probability density functions in this paper are assumed to be unknown and need to be estimated. The estimation method of these probability density functions will be discussed in Section 2.3, but first we introduce the aim of designing a probabilistic control.

2.2 Problem Formulation

In dynamic programming, the randomized controller of the above stochastic control problem is obtained by minimizing the expected value of the Bellman equation

$$J[(\mathbf{x}(k)] = \left\langle \left\{ U(\mathbf{x}(k), \mathbf{u}^{op}(k)) + \gamma J[\mathbf{x}(k+1)] \right\} \right\rangle \quad (5)$$

where $< . >$ is the expected value, $J[\mathbf{x}(k)]$ is the cost to go from time k to the final time, $U(\mathbf{x}(k), \mathbf{u}^{op}(k))$ is the utility which is the cost from going from time k to time $k+1$, and $J[\mathbf{x}(k+1)]$ is assumed to be the minimum cost from going from time $k+1$ to the final time. The term γ is a discount factor $(0 \leq \gamma \leq 1)$ which allows the designer to weight the relative importance of present versus future utilities. The objective is then to choose the control sequence $\mathbf{u}(k)$, $k = 1, 2, \ldots$, so that the function J in (5) is minimized.

The critic network in the DHP scheme, estimates a variable called $\lambda[\mathbf{x}(k)]$ as the derivatives of $J(\mathbf{x}(k))$ with respect to the vector $\mathbf{x}(k)$.

$$\begin{aligned}\lambda[\mathbf{x}(k)] &= \frac{\partial U[\mathbf{x}(k),\mathbf{u}^{op}(k)]}{\partial \mathbf{x}(k)} \\ &+ \frac{\partial U[\mathbf{x}(k),\mathbf{u}^{op}(k)]}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{x}(k)} \\ &+ < \lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \mathbf{x}(k)} > \\ &+ < \lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{x}(k)} > \qquad (6)\end{aligned}$$

where γ has been given the value of 1. Since $\langle\lambda[\mathbf{x}(k+1)]\rangle, U[\mathbf{x}(k),\mathbf{u}^{op}(k)]$ and the system model derivatives are known, then $\lambda[\mathbf{x}(k)]$ can be calculated. The optimality equation is defined as

$$\begin{aligned}\frac{\partial J[\mathbf{x}(k)]}{\partial \mathbf{u}^{op}(k)} &= \frac{\partial U[\mathbf{x}(k),\mathbf{u}^{op}(k)]}{\partial \mathbf{u}^{op}(k)} \\ &+ < \lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \mathbf{u}^{op}(k)} > \\ &= 0 \qquad (7)\end{aligned}$$

The above two equations are usually used in dynamic programming to solve an infinite or finite horizon control policy.

If the nonlinear function $f[\mathbf{x}(k),\mathbf{u}^{op}(k),\tilde{\eta}(k+1)]$ was known or the system was noiseless, and given a deterministic function for the inverse controller, the optimal control law which achieves the above objective is shown to be derived using techniques of dynamic programming or DHP adaptive critic methods as an approximation methods to dynamic programming (Herzallah, 2007). Even if the function $f[\mathbf{x}(k),\mathbf{u}^{op}(k),\tilde{\eta}(k+1)]$ was unknown, researchers in the model based adaptive critic field would simply adapt a forecasting network which predicts the conditional mean of the state vector. This means that only deterministic models were considered in the conventional theory of the adaptive critic methods. Recently (Herzallah, 2007), it has been proved that the control law of the DHP adaptive critic methods which is derived based on the assumption of deterministic forward model is suboptimal. It has been shown in (Herzallah, 2007) that if the function of the controlled system was unknown then the problem should be formulated in an adaptive control scheme which is known to have functional uncertainty. Therefore, forward model uncertainty was quantified and used in their developed control algorithm. However, only forward model uncertainty was considered in (Herzallah, 2007). This forward model uncertainty was assumed to follow Gaussian distribution. The inverse controller on the other hand was assumed to be deterministic function.

In the current paper, the forward model and the inverse controller are described by probability density functions as shown in Equations (2) and (4). These probability density functions are not limited to Gaussian density, they can be of any shape. As mentioned in Section 2.1, the probability density functions of the forward model and the inverse controller are assumed to be unknown and need to be estimated in this paper. The objective of the current paper is then to develop an appropriate method for estimating the non Gaussian distributions of both the forward model and the inverse controller and then use these probabilistic information in the derivation of the optimal control law. This yields a novel DHP adaptive critic control algorithm which we refer to as probabilistic DHP adaptive critic method. The developed theory will be illustrated on linear infinite horizon quadratic control problem. The Riccati solution for this linear problem will also be derived.

2.3 Stochastic Model Estimation

In the neurocontrol field researchers usually adapt forecasting networks to predict the conditional mean of the system output or state vector, $\hat{\mathbf{x}}(k+1)$. In most control applications this is probably enough. However; with the growing complexity of control systems and because of the inherent uncertainty in modeling and estimation, researchers recently considered modeling the conditional distribution of the stochastic systems rather than relying on the single estimate of the neural networks.

To estimate the conditional distribution of the system output, a neural network model is optimized such that its output approximates the conditional expectation of the system output. Once the output of the neural network model has been optimized the stochastic model of the system is simply shown to be given by (Herzallah and Lowe, 2007),

$$\mathbf{x}(k+1) = \hat{\mathbf{x}}(k+1) + \eta(k+1), \qquad (8)$$

where $\hat{\mathbf{x}}(k+1) = \hat{g}[\mathbf{x}(k),\mathbf{u}^{op}(k)]$, and $\eta(k+1)$ represents an input dependent random noise. The stochastic model in Equation (8) can in turn be reexpressed in the following general form:

$$\mathbf{x}(k+1) = \hat{f}[\mathbf{x}(k),\mathbf{u}^{op}(k),\eta(k+1)]. \qquad (9)$$

Usually the noise $\eta(k+1)$ is assumed to follow Gaussian distribution of zero mean and covariance matrix $\mathbf{P}$. In this work the assumption of Gaussian distribution is relaxed. In other words $\eta(k+1)$ is an input dependent random noise which could follow any non-Gaussian distribution of zero mean. This is a more realistic assumption, since a nonlinear mapping of random variable is non-Gaussian. This non-Gaussian distribution will be identified by evaluating the expectation and moments of the distribution. For example the second moment of the distribution is represented by its covariance matrix $\mathbf{P}$. This covariance matrix represents the covariance of the error in predicting $\mathbf{x}(k+1)$.

The method proposed in (Herzallah and Lowe, 2007) estimates the conditional distribution of the system output by using another neural network model to provide a prediction for the input dependent covariance matrix $\mathbf{P} =< \eta(k+1)\eta^T(k+1) >$. In the current paper we propose a different method for estimating the conditional distribution of the system output which could be non-Gaussian as well. This novel proposed method is based on estimating the distribution of the input dependent error $\eta(k+1)$ and not the input dependent covariance matrix $\mathbf{P}$. Since the covariance matrix $\mathbf{P}$ can be evaluated the distribution of the input dependent error is estimated by using a Gaussian Radial Basis Function neural network which has the important property of linear transformation.

$$\eta(\mathbf{x}(k), \mathbf{u}^{op}(k)) = \mathbf{w}\, \phi(\mathbf{x}(k), \mathbf{u}^{op}(k)). \tag{10}$$

where $\mathbf{w}_i$ is a random vector which has zero mean and a covariance matrix $\Sigma_i =< \phi^{\dagger^T} \eta_i^T \eta_i \phi^{\dagger} >$, and i is the output index. Here the RBFNN is taken to be a probabilistic rather than deterministic model. To adapt this probabilistic neural network model the following conditions are assumed to hold for the neural network:

Assumption 1. The state and control is always confined within the network approximation region defined by subset $\mathcal{Z}$ whose boundaries are known. This approximation region is a design parameter and could be made arbitrarily large.

Assumption 2. The basis function centers and width parameters ensuring that condition 1 is satisfied are known a priori.

The second assumption is justified by the universal approximation property of neural networks with well known developed methods of choosing appropriate basis function centers and width parameters a priori (Sanner and Slotine, 1992).

Using the neural network as a probabilistic model for the input dependent error allows us to consider multiplicative noise on both the state and control. Besides, it ensures the positivity of the error covariance matrix $\mathbf{P}$. Following the same procedure of the forward model, the stochastic model of the inverse controller is given by

$$\mathbf{u}^{op}(k) = \mathbf{u}(k) + \mathbf{e}(k). \tag{11}$$

The distribution of the error in predicting the control law is also estimated using the same method of predicting the distribution of the error of the forward model.

To reemphasize, the method proposed in this section for estimating the conditional distributions of the models: ensures the positivity of the covariance matrix of the errors, it uses the neural network as a probabilistic models, and allows considering multiplicative noise on both the state and the control. However, the method proposed in (Herzallah and Lowe, 2007) does not guarantee the positivity of the covariance matrix, and it uses the neural network a deterministic model.

The theory developed in this section will be used in the next section for developing the theory behind the probabilistic DHP adaptive critic method proposed in this paper.

3 PROPOSED PROBABILISTIC ADAPTIVE CRITIC METHOD

In this section we propose a probabilistic type DHP adaptive critic controller which takes uncertainty of the forward model and the inverse controller into consideration when calculating the control law. The proposed controller can be obtained directly by optimally solving the adaptive critic problem which considers stochastic models rather than deterministic models. In the proposed probabilistic DHP adaptive critic method the control law is derived such as to minimize the expected value of the cost–to–go $J[\mathbf{x}(k)]$ given in (5) using $\gamma = 1$, but with the uncertainty of the models' estimates being taken into consideration. This is accomplished by treating the forward model and the inverse controller as random variables.

Following the procedure presented in Section 2.3 the conditional distributions of the forward model and the inverse controller are estimated. Using this in equation (5), Bellman's equation could be reexpressed as:

$$\begin{aligned} J[\mathbf{x}(k)] &=< U(\mathbf{x}(k), \mathbf{u}^{op}(k)) > + < J[\mathbf{x}(k+1)] > \\ &=< U(\mathbf{x}(k), \mathbf{u}^{op}(k)) > \\ &\quad + < J[\hat{f}(\mathbf{x}(k), \mathbf{u}^{op}(k), \eta(k+1))] > \end{aligned} \tag{12}$$

Since the errors $\eta(k+1)$ and $\mathbf{e}(k)$ of the forward model and the inverse controller respectively are state dependent, the variable $\lambda[\mathbf{x}(k)]$ is shown to be given by the following theorem.

Theorem 1. The variable $\lambda[\mathbf{x}(k)]$ of the cost function of equation (12) subject to the stochastic models of equations (9) and (11), is given by

$$
\begin{aligned}
\lambda[\mathbf{x}(k)] =&< \frac{\partial U[\mathbf{x}(k),\mathbf{u}^{op}(k)]}{\partial \mathbf{x}(k)} \\
&+ \frac{\partial U[\mathbf{x}(k),\mathbf{u}^{op}(k)]}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{u}(k)} \frac{\partial \mathbf{u}(k)}{\partial \mathbf{x}(k)} \\
&+ \frac{\partial U[\mathbf{x}(k),\mathbf{u}^{op}(k)]}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{e}(k)} \frac{\partial \mathbf{e}(k)}{\partial \mathbf{x}(k)} > \\
&+ < \lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \mathbf{x}(k)} \\
&+ \lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{u}(k)} \frac{\partial \mathbf{u}(k)}{\partial \mathbf{x}(k)} \\
&+ \lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{e}(k)} \frac{\partial \mathbf{e}(k)}{\partial \mathbf{x}(k)} \\
&+ \lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \eta(k+1)} \frac{\partial \eta(k+1)}{\partial \mathbf{x}(k)} \\
&+ \lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \eta(k+1)} \frac{\partial \eta(k+1)}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{u}(k)} \frac{\partial \mathbf{u}(k)}{\partial \mathbf{x}(k)} \\
&+ \lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \eta(k+1)} \frac{\partial \eta(k+1)}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{e}(k)} \frac{\partial \mathbf{e}(k)}{\partial \mathbf{x}(k)} > \quad (13)
\end{aligned}
$$

Proof. To prove the above theorem we simply derive the cost function of equation (12) with respect to the state $\mathbf{x}(k)$ at time k.

The error in predicting the state vector $\eta(k+1)$ is dependent on the control signal as well, so the optimality equation can be seen to be given by the following theorem.

Theorem 2. The optimality equation of the cost function of equation (12) subject to the stochastic models of equations (9) and (11), is given by

$$
\begin{aligned}
\frac{\partial J[\mathbf{x}(k)]}{\partial \mathbf{u}(k)} =&< \frac{\partial U[\mathbf{x}(k),\mathbf{u}^{op}(k)]}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{u}(k)} \\
&+ \lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{u}(k)} \\
&+ \lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \eta(k+1)} \frac{\partial \eta(k+1)}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{u}(k)} >= 0
\end{aligned} \quad (14)
$$

Proof. To prove the above theorem we simply derive the cost function of equation (12) with respect to the optimal control $\mathbf{u}(k)$ at time k.

The training process for the probabilistic type DHP adaptive critic proposed in this section is exactly the same as that for the conventional DHP adaptive critic. It consists of training the action network which outputs the optimal control policy $\mathbf{u}[\mathbf{x}(k)]$ and the critic network which approximates the derivative of the cost function $\lambda[\mathbf{x}(k)]$. As a first step both networks' parameters are initially randomized. Next, the difference between the target value of the critic, $\lambda^*[\mathbf{x}(k)]$ calculated from Equation (13) and the critic network output $\lambda[\mathbf{x}(k)]$ is used to correct the critic network until it converges. The output from the converged critic is used in (14) solving for the target $\mathbf{u}^{op}(k)$ which is then used to correct the action network. These two steps continue until a predetermined level of convergence is reached.

Because the proposed probabilistic DHP adaptive critic method takes model uncertainty into consideration, it is recommended to be implemented on–line. The forward model of the plant to be controlled the controller and the critic networks can all be adapted on–line.

4 LINEAR QUADRATIC MODEL

Stochastic linear quadratic models is one of the most widely used models in modern control engineering and finance. To understand and prove the validity of the proposed probabilistic DHP adaptive critic methods, the theory developed in the previous section is applied here to infinite horizon linear quadratic control problem. Before we evaluate the proposed methods themselves, the correct values of various functions in this problem will be calculated, so that we have something to check the proposed method against. Besides evaluating the correct values of various functions, we also derive the Riccati solution of this nonstandard stochastic control problem.

4.1 Dynamic Programming Solution for the Linear Quadratic Model

Suppose that the vector of observable, $\mathbf{x}(k)$ is the same as the state vector of the plant. Since we consider an infinite horizon problem, the objective is to minimize a measure of utility, $U(k)$, summed from the present time to the infinite future, which is defined by:

$$U(k) = \mathbf{x}^T\mathbf{O}\mathbf{x} + \mathbf{u}^{op^T}\mathbf{G}\mathbf{u}^{op}. \quad (15)$$

Suppose that the plant is described by the following stochastic model:

$$\mathbf{x}(k+1) = \mathbf{S}\mathbf{x}(k) + \mathbf{R}\mathbf{u}^{op}(k) + \eta(k+1), \quad (16)$$

where the error of the prediction, $\eta(k+1)$ is estimated as described in Section 2.3. This error is shown in Section 2.3 to be control signal and state dependent.

Since it should have the same structure and same inputs as the forward model of the plant, it is taken in this linear quadratic problem to be linear with two inputs, the state vector and the control signal:

$$\eta(k+1) = \mathbf{Dx}(k) + \mathbf{Eu}^{op}(k). \tag{17}$$

where **D** and **E** are matrices of random numbers which contain the parameters of the error model. Suppose that the action network is described by the following stochastic model:

$$\mathbf{u}^{op}(k) = \mathbf{u}(k) + \mathbf{e}(k), \tag{18}$$

where

$$\mathbf{u}(k) = \mathbf{Ax}(k) \tag{19}$$

and where **A** is the matrix of the controller parameters, and $\mathbf{e}(k)$ is the error in predicting the optimal control estimated as discussed in Section 2.3 and assumed to have the following form

$$\mathbf{e}(k) = \mathbf{Qx}(k) \tag{20}$$

where **Q** is a matrix of random numbers that describes the mapping from the state space to the error in predicting the optimal control law. Using the control expression of Equation (19) in Equation (18) and substituting back in Equation (16) yields:

$$\mathbf{x}(k+1) = \tilde{\mathbf{S}}\mathbf{x}(k) + \mathbf{Re}(k) + \eta(k+1), \tag{21}$$

where

$$\tilde{\mathbf{S}} = \mathbf{S} + \mathbf{RA}. \tag{22}$$

Similarly the expression of the error in predicting the state vector as defined in Equation (17) can be rewritten in the following form

$$\eta(k+1) = \tilde{\mathbf{D}}\mathbf{x}(k), \tag{23}$$

where we have used Equations (18), (19), and (20) and where

$$\tilde{\mathbf{D}} = \mathbf{D} + \mathbf{EA} + \mathbf{EQ}. \tag{24}$$

As a preliminary step to calculating the correct value of the cost function of Bellman's equation let us define **M** as the matrix that solves the following equation:

$$\begin{aligned}\mathbf{M} = \quad & \mathbf{O} + \mathbf{A}^T\mathbf{GA} + <\mathbf{Q}^T\mathbf{GQ}> + \tilde{\mathbf{S}}^T\mathbf{M}\tilde{\mathbf{S}} \\ + \quad & <\mathbf{Q}^T\mathbf{R}^T\mathbf{MRQ}> + <\tilde{\mathbf{D}}^T\mathbf{M}\tilde{\mathbf{D}}>. \end{aligned} \tag{25}$$

Following all the above definitions, the true value of the cost function J satisfying the Bellman equation is given in the following theorem:

Theorem 3. The true value of the cost function J, satisfying the Bellman equation (with $\gamma = 1$) subject to the system of equation (16) and uncertainty models of the forward model and the inverse controller defined in Equations (17) and (20) respectively and all other definitions previously mentioned is given by:

$$J(\mathbf{x}) = \mathbf{x}^T\mathbf{Mx}. \tag{26}$$

Proof. To prove the above theorem we simply substitute into Bellman's equation (12) and verify that it is satisfied. For the left hand side of the equation, we get:

$$J[\mathbf{x}(k)] = \mathbf{x}^T(k)\mathbf{Mx}(k). \tag{27}$$

For the right hand side, we get:

$$\begin{aligned} & < U(\mathbf{x}(k), \mathbf{u}^{op}(k)) + J[\mathbf{x}(k+1)] >= \\ & \quad < \mathbf{x}^T\mathbf{Ox} + (\mathbf{Ax}+\mathbf{e})^T\mathbf{G}(\mathbf{Ax}+\mathbf{e}) > \\ & \quad + < (\tilde{\mathbf{S}}\mathbf{x} + \mathbf{Re} + \eta)^T\mathbf{M}(\tilde{\mathbf{S}}\mathbf{x} + \mathbf{Re} + \eta) > \\ & = \mathbf{x}^T\mathbf{Ox} + \mathbf{x}^T\mathbf{A}^T\mathbf{GAx} + \mathbf{x}^T <\mathbf{Q}^T\mathbf{GQ}> \mathbf{x} + \mathbf{x}^T\tilde{\mathbf{S}}^T\mathbf{M}\tilde{\mathbf{S}}\mathbf{x} \\ & \quad + \mathbf{x}^T <\mathbf{Q}^T\mathbf{R}^T\mathbf{MRQ}> \mathbf{x} + \mathbf{x}^T <\tilde{\mathbf{D}}^T\mathbf{M}\tilde{\mathbf{D}}> \mathbf{x}, \end{aligned} \tag{28}$$

where we used Equations (20) and (23) and where we made use of the fact that η, and **e** are uncorrelated random variables of zero mean. Making use of Equation (25) in Equation (28), yields

$$J(\mathbf{x}) = \mathbf{x}^T\mathbf{Mx}. \tag{29}$$

Comparing Equations (26) and (29) we can see that Bellman's equation is satisfied. Howards has proven (Howard, 1960) that the optimal control law, based on the policy iteration method, can be derived by alternately calculating the cost function J for the current control law, modify the control law so as to minimize the cost function J, recalculate J for the new control law, and so on.

4.2 Proposed Probabilistic DHP Adaptive Critic in the Linear Quadratic Model

The objective of this section is to calculate the targets for the output of the critic network $\lambda^*[\mathbf{x}(k)]$ as they would be generated by the proposed probabilistic DHP adaptive critic, and then check them against the correct values. In other words we need to check that if the critic was initially correct it will stay correct after one step of adaptation.

From (29), the correct value of $J(\mathbf{x})$ is $\mathbf{x}^T\mathbf{Mx}$, and consequently the correct value of $\lambda(\mathbf{x})$ is simply the gradient of $J(\mathbf{x})$, i.e $\lambda(\mathbf{x}) = 2\mathbf{Mx}(k)$. Hence $\lambda(k+1)$ is given by,

$$\lambda(k+1) = 2\mathbf{Mx}(k+1).$$

Next we carry out the calculations implied by equation (13) but with the expectation of the derivatives being evaluated at the end.

To calculate the first term on the right hand side of (13), we simply calculate the gradient of $U(\mathbf{x}(k), \mathbf{u}^{op}(k))$ with respect to $\mathbf{x}(k)$:

$$\frac{\partial U[\mathbf{x}(k), \mathbf{u}^{op}(k)]}{\partial \mathbf{x}(k)} = 2\mathbf{O}\mathbf{x}(k).$$

For the second term the value of the partial derivatives of $\mathbf{u}(k)$, $\mathbf{u}^{op}(k)$ and $U(\mathbf{x}(k), \mathbf{u}^{op}(k))$ with respect to $\mathbf{x}(k)$, $\mathbf{u}(k)$and $\mathbf{u}^{op}(k)$ respectively need to be evaluated:

$$\frac{\partial U[\mathbf{x}(k), \mathbf{u}^{op}(k)]}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{u}(k)} \frac{\partial \mathbf{u}(k)}{\partial \mathbf{x}(k)} = 2\mathbf{G}(\mathbf{u}+\mathbf{e})\mathbf{A}.$$

The third term can be evaluated by calculating the partial derivatives of $\mathbf{e}(k)$, $\mathbf{u}^{op}(k)$ and $U(\mathbf{x}(k), \mathbf{u}^{op}(k))$ with respect to $\mathbf{x}(k)$, $\mathbf{e}(k)$and $\mathbf{u}^{op}(k)$ respectively:

$$\frac{\partial U[\mathbf{x}(k), \mathbf{u}^{op}(k)]}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{e}(k)} \frac{\partial \mathbf{e}(k)}{\partial \mathbf{x}(k)} = 2\mathbf{G}(\mathbf{u}+\mathbf{e})\mathbf{Q}.$$

The fourth term requires propagating $\lambda(k+1)$ through the model of equation (16) back to $\mathbf{x}(k)$, which yields

$$\lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \mathbf{x}(k)} = 2\mathbf{M}\mathbf{x}(k+1)\mathbf{S}.$$

The fifth term requires propagating $\lambda(k+1)$ through the model of the plant, $\mathbf{x}(k+1)$, back to $\mathbf{u}^{op}(k)$ and then through the action network, which yields

$$\lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{u}(k)} \frac{\partial \mathbf{u}(k)}{\partial \mathbf{x}(k)} \\ = 2\mathbf{M}\mathbf{x}(k+1)\mathbf{R}\mathbf{A}.$$

The sixth term can be calculated by propagating $\lambda(k+1)$ through the model of the plant, $\mathbf{x}(k+1)$, back to $\mathbf{u}^{op}(k)$ and then through the error network of the controller, which yields

$$\lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{e}(k)} \frac{\partial \mathbf{e}(k)}{\partial \mathbf{x}(k)} \\ = 2\mathbf{M}\mathbf{x}(k+1)\mathbf{R}\mathbf{Q}.$$

The seventh term can also be calculated by propagating $\lambda(k+1)$ through the model of the plant, $\mathbf{x}(k+1)$, back to the error network of the forward model, which yields

$$\lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \eta(k+1)} \frac{\partial \eta(k+1)}{\partial \mathbf{x}(k)} = 2\mathbf{M}\mathbf{x}(k+1)\mathbf{D}.$$

The eighth term is calculated by propagating $\lambda(k+1)$ through the model of the plant, $\mathbf{x}(k+1)$, back to the error network of the forward model and then $\mathbf{u}^{op}(k)$ and then through the action network. This yields

$$\lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \eta(k+1)} \frac{\partial \eta(k+1)}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{u}(k)} \frac{\partial \mathbf{u}(k)}{\partial \mathbf{x}(k)} \\ = 2\mathbf{M}\mathbf{x}(k+1)\mathbf{E}\mathbf{A}.$$

Finally the last term is calculated as follows

$$\lambda[\mathbf{x}(k+1)] \frac{\partial \mathbf{x}(k+1)}{\partial \eta(k+1)} \frac{\partial \eta(k+1)}{\partial \mathbf{u}^{op}(k)} \frac{\partial \mathbf{u}^{op}(k)}{\partial \mathbf{e}(k)} \frac{\partial \mathbf{e}(k)}{\partial \mathbf{x}(k)} \\ = 2\mathbf{M}\mathbf{x}(k+1)\mathbf{E}\mathbf{Q}.$$

Adding all terms together and taking the expectation, yields

$$\begin{aligned} \lambda^*(k) =& < 2\mathbf{O}\mathbf{x}(k) + 2\mathbf{A}^T\mathbf{G}\{\mathbf{A}\mathbf{x}(k) + \mathbf{Q}\mathbf{x}(k)\} \\ & + 2\mathbf{Q}^T\mathbf{G}\{\mathbf{A}\mathbf{x}(k) + \mathbf{Q}\mathbf{x}(k)\} + 2\mathbf{S}^T\mathbf{M}\{\tilde{\mathbf{S}}\mathbf{x}(k) + \mathbf{R}\mathbf{Q}\mathbf{x}(k) \\ & + \tilde{\mathbf{D}}\mathbf{x}(k)\} + 2\mathbf{A}^T\mathbf{R}^T\mathbf{M}\{\tilde{\mathbf{S}}\mathbf{x}(k) + \mathbf{R}\mathbf{Q}\mathbf{x}(k) + \tilde{\mathbf{D}}\mathbf{x}(k)\} \\ & + 2\mathbf{Q}^T\mathbf{R}^T\mathbf{M}\{\tilde{\mathbf{S}}\mathbf{x}(k) + \mathbf{R}\mathbf{Q}\mathbf{x}(k) + \tilde{\mathbf{D}}\mathbf{x}(k)\} + 2\mathbf{D}^T\mathbf{M}\{\tilde{\mathbf{S}}\mathbf{x}(k) \\ & + \mathbf{R}\mathbf{Q}\mathbf{x}(k) + \tilde{\mathbf{D}}\mathbf{x}(k)\} + 2\mathbf{A}^T\mathbf{E}^T\mathbf{M}\{\tilde{\mathbf{S}}\mathbf{x}(k) + \mathbf{R}\mathbf{Q}\mathbf{x}(k) \\ & + \tilde{\mathbf{D}}\mathbf{x}(k)\} + 2\mathbf{Q}^T\mathbf{E}^T\mathbf{M}\{\tilde{\mathbf{S}}\mathbf{x}(k) + \mathbf{R}\mathbf{Q}\mathbf{x}(k) + \tilde{\mathbf{D}}\mathbf{x}(k)\} >, \end{aligned} \tag{30}$$

where we used equations (21), (19), (20) and (23). Evaluating the expectation of Equation (30) yields,

$$\begin{aligned} \lambda^*(k) =& 2\mathbf{O}\mathbf{x}(k) + 2\mathbf{A}^T\mathbf{G}\mathbf{A}\mathbf{x}(k) \\ & + 2 < \mathbf{Q}^T\mathbf{G}\mathbf{Q} > \mathbf{x}(k) + 2\tilde{\mathbf{S}}^T\mathbf{M}\tilde{\mathbf{S}}\mathbf{x}(k) \\ & + 2 < \mathbf{Q}^T\mathbf{R}^T\mathbf{M}\mathbf{R}\mathbf{Q} > \mathbf{x}(k) + 2 < \tilde{\mathbf{D}}^T\mathbf{M}\tilde{\mathbf{D}} > \mathbf{x}(k), \end{aligned} \tag{31}$$

where we made use of the fact that the expected value of the random variables $\mathbf{Q}$, $\mathbf{E}$, and $\mathbf{D}$ is zero, that η and $\mathbf{e}$ are uncorrelated and finally that $\mathbf{Q}$ and $\mathbf{E}$ are uncorrelated random variables. Using equation (25) in (31) yields,

$$\lambda^*(k) = 2\mathbf{M}\mathbf{x}(k). \tag{32}$$

From (32) it can be clearly seen that the target vector of the proposed probabilistic critic network is equal to the correct value. This validates the theoretical development of the probabilistic DHP adaptive critic method proposed in this paper.

5 CONCLUSIONS

The nonstandard formulation of the stochastic control design presented in this paper leads to a different form of optimal controller that depends on the solution of stochastic functional equations. It provides the complete solution for designing a stochastic controller for complex control systems accompanied by high levels of inherent uncertainty in modeling and estimation. All probability density functions needed in the proposed methods are assumed to be unknown. To estimate these probability density functions we propose using probabilistic neural network models to estimate errors in predicting conditional expectations of the functions. This proposed method always guarantees the positivity of the covariance of the errors and allows for considering multiplicative noise on both the state and control of the system.

The proposed probabilistic DHP critic method is suitable for deterministic and stochastic control problems characterized by functional uncertainty. Unlike current established control methods, it takes uncertainty of the forward model and inverse controller into consideration when deriving the optimal control law.

Theoretical development in this paper is demonstrated through linear quadratic control problem. There, the correct value of the cost function which satisfies the Bellman equation is evaluated and shown to be equal to its corresponding value produced by the proposed probabilistic critic network.

REFERENCES

Botto, M. A., Wams, B., van den Boom, and da Costa, J. M. G. S. (2000). Robust stability of feedback linearised systems modelled with neural networks: Dealing with uncertainty. *Engineering Applications of Artificial Intelligence*, 13(6):659–670.

Fabri, S. and Kadirkamanathan, V. (1998). Dual adaptive control of nonlinear stochastic systems using neural networks. *Automatica*, 34(2):245–253.

Ge, S. S., Hang, C. C., Lee, T. H., and Zhang, T. (2001). *Stable Adaptive Neural Network Control*. Kluwer, Norwell, MA.

Ge, S. S. and Wang, C. (2004). Adaptive neural control of uncertain mimo nonlinear systems. *IEEE Transactions on Neural Networks*, 15(3):674–692.

Herzallah, R. (2007). Adaptive critic methods for stochastic systems with input-dependent noise. *Automatica*. Accepted to appear.

Herzallah, R. and Lowe, D. A Bayesian perspective on stochastic neuro control. *IEEE Transactions on Neural Networks*. re-submited 2006.

Herzallah, R. and Lowe, D. (2007). Distribution modeling of nonlinear inverse controllers under a Bayesian framework. *IEEE Transactions on Neural Networks*, 18:107–114.

Hovakimyan, N., Nardi, F., and Calise, A. J. (2001). A novel observer based adaptive output feedback approach for control of uncertain systems. In *Proceedings of the American Control Conference*, volume 3, pages 2444–2449, Arlington, VA, USA.

Howard, R. A. (1960). *Dynamic Programming and Markov Processes*. The Massachusetts Institute of Technology and John Wiley and Sons, Inc., New York. London.

Karny, M. (1996). Towards fully probabilistic control design. *Automatica*, 32(12):1719–1722.

Lewis, F. L., Yesildirek, A., and Liu, K. (2000). Robust backstepping control of induction motors using neural netwoks. *IEEE Transactions on Neural Networks*, 11:1178–1187.

Mine, H. and Osaki, S., editors (1970). *Markovian Decision Processes*. Elsevier, New York, N.Y.

Murray-Smith, R. and Sbarbaro, D. (2002). Nonlinear adaptive control using non-parametric gaussian process prior models. In *15th IFAC Triennial World Congress*, Barcelona.

Sanner, R. M. and Slotine, J. J. E. (1992). Gaussian networks for direct adaptive control. *IEEE Transactions on Neural Networks*, 3(6).

Sastry, S. S. and Isidori, A. (1989). Adaptive control of linearizable systems. *IEEE Transactions on Automatic Control*, 34(11):1123–1131.

Wang, D. and Huang, J. (2005). Neural network-based adaptive dynamic surface control for a class of uncertain nonlinear systems in strict-feedback form. *IEEE Transactions on Neural Networks*, 16(1):195–202.

Wang, H. (2002). Minimum entropy control of non-gaussian dynamic stochastic systems. *IEEE Transactions on Automatic Control*, 47(2):398–403.

Wang, H. and Zhang, J. (2001). Bounded stochastic distribution control for pseudo armax stochastic systems. *IEEE Transactions on Automatic Control*, 46(3):486–490.

Werbos, P. J. (1992). Approximate dynamic programming for real-time control and neural modeling. In White, D. A. and Sofge, D. A., editors, *Handbook of Intillegent Control*, chapter 13, pages 493–526. Multiscience Press, Inc, New York, N.Y.

Zhang, Y., Peng, P. Y., and Jiang, Z. P. (2000). Stable neural controller design for unknown nonlinear systems using backstepping. *IEEE Transactions on Neural Networks*, 11:1347–1359.