Probabilistic DHP Adaptive Critic for Nonlinear Stochastic Control Systems

Randa Herzallah

Abstract

Following the recently developed algorithms for fully probabilistic control design for general dynamic stochastic systems [15], [18], this paper presents the solution to the probabilistic dual heuristic programming (DHP) adaptive critic method [15] and randomized control algorithm for stochastic nonlinear dynamical systems. The purpose of the randomized control input design is to make the joint probability density function of the closed loop system as close as possible to a predetermined ideal joint probability density function. This paper completes the previous work [15], [18] by formulating and solving the fully probabilistic control design problem on the more general case of nonlinear stochastic discrete time systems. A simulated example is used to demonstrate the use of the algorithm and encouraging results have been obtained.

Index Terms

nonlinear stochastic systems; fully probabilistic design; nonlinear randomized control input design; adaptive critic.

I. Introduction

The mean variance [2] and utility function in linear quadratic optimal control [1], [31], have been firstly introduced to characterize the performance of the closed loop stochastic control systems. In more recent work, stochastic adaptive control [22], stochastic linear quadratic martingale [25], [6], sliding mode control for stochastic systems [23], and predictive stochastic control [9], [5] have been proposed. In most of previous works the system under control has been assumed to be linear or of Gaussian probability density function (pdf). However, it has been shown [14], [30], [7] that in systems where the stochastic signals is non Gaussian or the system dynamics have strong nonlinearity, existing control methods do not generally yield optimization of the control system.

Consequently, in the past few years four groups of control algorithms for general stochastic systems with inherent models' and parameters' uncertainty have been developed: (1) the control of the shape of the output pdf [27], [29], (2) the minimum entropy control [28], (3) the Bayesian techniques for modelling and control [13], [26] and (4) the control of the closed loop pdf [18], [17]. The control objective in the first group is to find a control input which makes the shape of the measured output pdf follows a given desired distribution. The second group is a generalization of the minimum variance control for linear Gaussian systems. The entropy in this group of control algorithms is used to characterize the performance of the closed loop systems and the controller is designed such that the shape of the pdf of the closed loop system is made as narrow as possible. In the third group, a general class of stochastic estimation and control problems is formulated from the Bayesian decision-theoretic viewpoint which is shown to be a general framework to solve stochastic estimation and control problems. Motivated by the probabilistic description of the closed loop control system, the Kullback–Leibler divergence distance has been proposed in the fourth group of control as a performance measure rather than the mean variance. This method of control is known as fully probabilistic Design (FPD). For stochastic systems with measurable states x_t, the objective of the FPD is to determine the pdf of a randomized optimal control law, u_t described by,

$$c(u_t \mid x_{t-1}), \tag{1}$$

that minimizes the discrepancy between the actual joint pdf of the closed loop system, f and an ideal joint pdf, ${}^I f$ measured by the Kullback Leibler divergence distance,

$$\mathcal{D}\left(f || {}^I f\right) \equiv \int f(x_t, u_t) \ln \left(\frac{f(x_t, u_t)}{{}^I f(x_t, u_t)} \right) \mathrm{d}x_t \mathrm{d}u_t. \tag{2}$$

The FPD problem is further simplified by the assumption that all pdfs needed in the design paradigm are existent and known. The main advantage of the FPD is that it provides an explicit form of the randomized optimal controller. However, since the evaluation of the randomized optimal controller involves multivariate integration steps which need to be computed by backward recursion the problem renders to be nontrivial and computationally very intensive. To overcome the difficulties arising in the FPD, a probabilistic DHP adaptive critic method is proposed in [15]. The DHP adaptive critic method uses a critic network to circumvent the need for explicitly evaluating the optimal value function, therefore, dramatically reducing computational requirements. Further more, unlike the FPD method all pdfs in the probabilistic DHP adaptive critic approach are assumed to be unknown, therefore, estimated using recent development in neural networks. However, up to now, the previous methods of

R. Herzallah is with Al-Balqa' Applied University, Jordan.

FPD [18] and probabilistic DHP critic [15] are demonstrated only on linear stochastic Gaussian systems where the means of the associated density functions are restricted to be linear functions, hence the solution to the control problem is constrained to the linear Gaussian control theory. This is restrictive in many real world applications that are characterized by strong nonlinearity and uncertainty. In practical processes where the forward and inverse dynamics of the system have strong nonlinearities, the means of the associated density functions are in general going to be nonlinear functions. This motivates the work in the current paper, which is concerned with the development of the solution to the FPD problem [18] for nonlinear stochastic systems where the means of the various density functions are allowed to be general nonlinear functions.

For this purpose we adopt the probabilistic DHP adaptive critic method proposed in [15]. Although systems under consideration are nonlinear stochastic systems, all pdfs are assumed to be Gaussians. This assumption can be shown to represent no real restriction provided that the conditional expectations of these pdfs are estimated using nonlinear models. However, it is worth mentioning that although the pdfs of the nonlinear stochastic systems are approximated by Gaussian density functions, it is in general difficult to integrate the nonlinear conditional expectations of these density functions to the probabilistic DHP adaptive critic method or even the FPD method. As such, radial basis function (RBF) neural network with Gaussian basis functions [4] is proposed in this paper to approximate the conditional expectations of the nonlinear stochastic models. An important property of such a neural network is that it forms a unifying link with density estimation. As will be clear from subsequent developments, the use of RBF networks with Gaussian basis functions to approximate the unknown nonlinear models facilitates the evaluation of the Gaussian integrations and integrates naturally to the framework of probabilistic DHP adaptive critic paradigm.

The main achievement of this paper is the solution of the nonlinear fully probabilistic optimal control problem for stochastic nonlinear systems. Although, the previous methods [15], [18] have discussed the general framework of nonlinear control systems, the problem rendered to be very hard to implement even under the Gaussian and linear models assumptions. Hence, the control solution to these methods are derived and demonstrated on linear systems only. By using the probabilistic DHP critic method proposed in [15] and RBF networks to estimate all required density functions, we develop and demonstrate the solution to the FPD control problem on stochastic nonlinear systems. The derived solution provides an efficient approach to the solution of the fully probabilistic control design for nonlinear stochastic systems.

To achieve the objective of this paper, it will be organized as follows. Section II formulates the problem and discusses the problem of estimating the pdf of the system dynamics. Section III presents the probabilistic DHP adaptive critic solution to nonlinear stochastic systems. The control algorithm of nonlinear control problems based on the probabilistic DHP adaptive critic method is discussed in Section IV. Section V contains a simulation example to show the effectiveness of the proposed probabilistic controller. The conclusion is provided in Section VI.

II. Problem Formulation and Preliminaries

A. Model Description and Control objective

The system considered in this paper is a nonlinear stochastic dynamical control system described by the following general stochastic equation,

$$x_t = \tilde{h}(x_{t-1}) + \tilde{g}(x_{t-1})u_t + \tilde{\epsilon}_t, \tag{3}$$

where $x_t \in \mathcal{R}^n$ is the measured state vector, $u_t \in \mathcal{R}^r$ is the control input vector, $\tilde{h}(x_{t-1}) : \mathcal{R}^n \longmapsto \mathcal{R}^n$ and $\tilde{g}(x_{t-1}) : \mathcal{R}^n \longmapsto \mathcal{R}^r$ are unknown nonlinear functions of the state, and $\tilde{\epsilon}_t \in \mathcal{R}^n$ is an additive noise vector. The control problem confronted here is to design a control strategy for the system in (3) to control the state of the system to a predefined desired state value. However, because of the noise input $\tilde{\epsilon}_t$ the previous state and present and future controls do not completely specify the present state, but instead determine only the probability distribution of these states, $s(x_t \mid u_t, x_{t-1})$. It is assumed that the noise input $\tilde{\epsilon}_t$ is unknown and hence the probability distribution of the states is unknown.

Since only probability distribution of the states can be determined, the above objective of this control problem should be re-defined in terms of the probabilistic control theory. Therefore, to achieve this control objective we consider designing a probabilistic controller $c(u_t \mid x_{t-1})$ that shapes the joint pdf of the closed loop system, $f(x_t, u_t)$ and makes it as close as possible to a predefined desired pdf, ${}^If(x_t, u_t)$. This design method was originally presented in [18] where the probabilistic controller is obtained such that it minimizes the KullbackLeibler divergence distance defined in (2). The minimum cost function resulting from minimization of (2) with respect to admissible control sequence u_t, $t \in \{1, \dots, H\}$, with H being the control horizon, is then shown to be given by the following recurrence equation [15],

$$-\ln(\gamma(x_{t-1})) = \min_{c(u_t|x_{t-1})} \int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1}) \times \Bigg[\underbrace{\ln\left(\frac{s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})}{{}^Is(x_t|u_t, x_{t-1})\,{}^Ic(u_t|x_{t-1})}\right)}_{\equiv\text{partial cost} \Longrightarrow U(x_t,u_t)} - \underbrace{\ln(\gamma(x_t))}_{\text{optimal cost-to-go}} \Bigg] \mathrm{d}(x_t, u_t), \tag{4}$$

where $-\ln(\gamma(x_{t-1}))$ is the expected minimum cost–to–go function and

$$f(x_t, u_t) = s(x_t|u_t, x_{t-1})c(u_t|x_{t-1}), \tag{5}$$

is the decomposition of the actual joint pdf by the chain rule [24], which represents the most complete probabilistic description of the closed loop system. Here the pdf $s(x_t|u_t, x_{t-1})$ describes the dynamics of the observed state vector x_t. Similarly

$${}^{I}f(x_t, u_t) = {}^{I}s(x_t|u_t, x_{t-1})\,{}^{I}c(u_t|x_{t-1}), \tag{6}$$

is the decomposition of the ideal joint pdf of the closed loop system and ${}^{I}s(x_t|u_t, x_{t-1})$ and ${}^{I}c(u_t|x_{t-1})$ represent the pdfs of the desired dynamics of the observed state vector and ideal controller respectively. The solution of the FPD is given in the following proposition.

Proposition 1: The pdf of optimal controller minimizing the cost–to–go function (4) is given by

$$\begin{aligned} c^*(u_t|x_{t-1}) &= \frac{{}^{I}c(u_t|x_{t-1})\exp[-\beta(u_t, x_{t-1})]}{\gamma(x_{t-1})}, \\ \gamma(x_{t-1}) &= \int {}^{I}c(u_t|x_{t-1})\exp[-\beta(u_t, x_{t-1})]\mathrm{d}u_t, \\ \beta(u_t, x_{t-1}) &= \int s(x_t|u_t, x_{t-1})\left[\ln\frac{s(x_t|u_t, x_{t-1})}{{}^{I}s(x_t|u_t, x_{t-1})} - \ln(\gamma(x_t))\right]\mathrm{d}x_t. \end{aligned} \tag{7}$$

***Proof*:** This proposition can be proven by adapting the proof of Proposition 2 in [17].

It should be noted that although a closed form can be found for the pdf of the optimal controller, the multivariate integrations in (7) are only tractable for the linear Gaussian case, where the mean of the Gaussian distribution is linear in the state and control values. Besides even for the linear Gaussian case, the solution to the optimal control history need to be computed from (7) by backward recursion. This backward dynamic programming approach is computationally very expensive and grows exponentially with the dimensionality of the state vector. To avoid these difficulties of the FPD, a probabilistic DHP adaptive critic method is proposed in [15] to approximate the optimal cost-to-go function and the probabilistic controller. Unknown pdfs were also estimated using recent development from neural network models. However, numerical experiments and previous analytical studies have demonstrated the usefulness of this control approach to obtain the control efforts only on linear stochastic Gaussian systems. The solution to the probabilistic DHP adaptive critic methods for the nonlinear stochastic systems (3) on the other hand was not discussed. The objective of this paper is to discuss the various steps to obtain this solution and demonstrate the theoretical development on a nonlinear stochastic simulation example of the form given in (3).

We first start by discussing the estimation problem of unknown probabilistic models of the stochastic system defined in (3) and reviewing the probabilistic DHP adaptive critic methods that will be needed for further development in the article.

B. pdf of the system dynamics

To estimate the probabilistic model of the nonlinear stochastic system (3) we adopt the method proposed in [15], where neural network models are used to provide a prediction for the conditional expectation of the system state values and calculating the global average variance of its residual error. For such a system, there exists a neural network model [15] such that the inequality,

$$\mid x_t - N_f(u_t, x_{t-1}) \mid \leq \delta, \tag{8}$$

holds, where $\delta > 0$ is a known small number and $N_f(u_t, x_{t-1}) = \hat{x}_t$ is a neural network approximation of the state x_t. Assuming a RBF neural network model of the form $N_f(u_t, x_{t-1}) = h(x_{t-1}) + g(x_{t-1})u_t$ in which $h(x_{t-1})$ and $g(x_{t-1})$ are estimates of the nonlinear functions $\tilde{h}(x_{t-1})$ and $\tilde{g}(x_{t-1})$ respectively, the stochastic system (3) can be re-expressed as

$$x_t = h(x_{t-1}) + g(x_{t-1})u_t + e(x_{t-1}, u_t). \tag{9}$$

Here, $e(x_{t-1}, u_t)$ represents the approximation error satisfying $\mid e(x_{t-1}, u_t) \mid \leq \delta$. This means that the resulting conditional distribution of the system dynamics $s(x_t \mid u_t, x_{t-1})$ is Gaussian distribution function with conditional expectation of the distribution being given by the neural network approximation and a global average covariance given by [15],

$$\Sigma = E\left((x_t - \hat{x}_t)(x_t - \hat{x}_t)^T\right), \tag{10}$$

with $E(.)$ denoting the expected value.

C. Review of the Probabilistic DHP Adaptive Critic Method

The probabilistic DHP adaptive critic method uses, two neural networks: an adaptive critic network that approximates the derivative of the optimal cost-to-go function (4) with respect to the state, $\lambda^*[x_{t-1}] = \partial[-\ln(\gamma(x_{t-1}))]/\partial x_{t-1}$ and an action network that produces optimal randomized control inputs, u_t^*. In the probabilistic DHP critic method, the optimal control law is computed by deriving (4) with respect to the control input [15],

$$\frac{\partial[-\ln(\gamma(x_{t-1}))]}{\partial u_t}\bigg|_{u_t=u_t^*} = \int s(x_t|u_t,x_{t-1})c(u_t|x_{t-1}) \times \left[\frac{\partial U(x_t,u_t)}{\partial x_t}\frac{\partial x_t}{\partial u_t} + \frac{\partial U(x_{t-1},u_t)}{\partial u_t} \right.$$
$$\left. +\lambda[x_t]\frac{\partial x_t}{\partial u_t}\right] \mathrm{d}(x_t,u_t) = 0. \tag{11}$$

So the action network is optimized such that the error between optimal control input u_t^*, obtained from (11) and estimated control input u_t from the neural network is minimized. Once this network is optimized information about the error between optimal control u_t^* and estimated control u_t will become available. This allows estimation of the conditional distribution of the randomized controller $c(u_t \mid x_{t-1})$ which is assumed to be Gaussian with mean computed from the output of the controller network and a global covariance matrix computed from the residual error between the output of the controller network and the optimal control signal, $E\left((u_t^* - u_t)(u_t^* - u_t)^T\right)$.

Given estimation of control law from the controller network and the derivative of the output of the critic network $\lambda[x_t]$, the critic network is then optimized by computing its desired value as follows [15],

$$\lambda^*[x_{t-1}] = \int s(x_t|u_t,x_{t-1})c(u_t|x_{t-1}) \left[\frac{\partial U(x_t,u_t)}{\partial x_t}\frac{\partial x_t}{\partial x_{t-1}} + \frac{\partial U(x_t,u_t)}{\partial x_t}\frac{\partial x_t}{\partial u_t}\frac{\partial u_t}{\partial x_{t-1}} + \frac{\partial U(x_t,u_t)}{\partial u_t}\frac{\partial u_t}{\partial x_{t-1}} \right.$$
$$\left. + \quad \lambda[x_t]\frac{\partial x_t}{\partial x_{t-1}} + \lambda[x_t]\frac{\partial x_t}{\partial u_t}\frac{\partial u_t}{\partial x_{t-1}}\right] \mathrm{d}(x_t,u_t). \tag{12}$$

The training process for the adaptive critic network is a two stage process. The training of the action network, which outputs the optimal control policy $u[x_t]$ and the training of the critic network, which approximates the derivative of the cost function $\lambda[x_{t-1}]$. As a first step in the training process, the critic and the action networks need to be designed and the initial weights of these networks should be randomized. Since the derivative of the partial cost function can be calculated, this in combination with the critic outputs and the system model derivatives, allows the use of (12) to calculate the target value of the critic, $\lambda^*[x_{t-1}]$. The difference between $\lambda^*[x_{t-1}]$ and the output of the critic, $\lambda[x_{t-1}]$ is used to correct the critic network, until it converges. The output from the converged critic is used in (11) solving for the target u_t^*, which is then used to correct the action network. This alternating process of training the action and the critic networks is repeated until an acceptable performance is reached.

III. Solution to the Problem

A. Basic Elements

In this section we derive the probabilistic DHP adaptive critic solution to the nonlinear stochastic system defined in (3). For presentations clarity and simplicity the solution to this problem will be developed for a regulation problem where the objective is to reach a zero state with a spread determined by a specified covariance matrix. Generalization to a state value that is different than zero is straight forward.

As discussed in Section II, the conditional distribution of the nonlinear system (3) is estimated as a Gaussian distribution described by,

$$x_t = h(x_{t-1}) + g(x_{t-1})u_t + e(x_{t-1},u_t)$$
$$s(x_t \mid u_t, x_{t-1}) \rightsquigarrow \mathcal{N}_{x_t}(h(x_{t-1}) + g(x_{t-1})u_t, \Sigma). \tag{13}$$

For the considered regulation problem, the system is initially in state x_{t-1} and the aim is to return the system state to the origin. Hence, the distribution of the ideal state of the system is taken to be,

$${}^I s(x_t|u_t,x_{t-1}) = \mathcal{N}_{x_t}(0,\Sigma), \tag{14}$$

where here the desired mean value of the state is taken to be zero and where Σ specifies the covariance of the innovation of the state values.

The stochastic model of the randomized controller to be designed is estimated as discussed in Section II-C by the well known RBF neural network,

$$\begin{aligned} u_t^k &= \sum_{j=0}^{M} w_{kj}\psi_j(x_{t-1}) + \omega_t^k \\ u_t &= W\psi(x_{t-1}) + \omega_t \\ c(u_t|x_{t-1}) &\rightsquigarrow \mathcal{N}_{u_t}(W\psi(x_{t-1}),\Gamma), \end{aligned} \tag{15}$$

where $W = [w_{kj}]$ is the matrix of the weight parameters, M is the number of basis functions of the controller network, ω_t is the residual error of the control input vector, Γ is the covariance of the residual error of control, and the RBF activation functions, $\psi_j(x_{t-1})$ are Gaussian basis functions [4],

$$\psi_j(x_{t-1}) = \exp\bigg(- (x_{t-1} - \mu_j)^T \rho_j^{-1}(x_{t-1} - \mu_j)\bigg). \tag{16}$$

The distribution of the ideal controller is also assumed to be

$$^{I}c(u_t|x_{t-1}) = \mathcal{N}_{u_t}(0, \Gamma). \tag{17}$$

The desired value of the critic network given in (12) is also taken to be the target of an RBF neural network as follows [4],

$$\begin{aligned} \lambda^m[x_{t-1}] &= \sum_{l=0}^{L} \chi_{ml}\phi_l(x_{t-1}), \\ \lambda[x_{t-1}] &= \chi\phi(x_{t-1}), \end{aligned} \tag{18}$$

where χ is the matrix of weight parameters of the critic network, L is the number of basis functions, and the basis functions, $\phi_l(x_{t-1})$ are taken to be Gaussian basis functions [4],

$$\phi_l(x_{t-1}) = \exp\bigg(- (x_{t-1} - z_l)^T \gamma_l^{-1}(x_{t-1} - z_l)\bigg). \tag{19}$$

Having defined those elements, the solution to the probabilistic DHP adaptive critic can now be obtained by calculating the desired value of the critic network and the optimal control inputs. We start in the next section by calculating the desired value of the critic. The optimal control input will be calculated in Section III-C.

B. Desired Value of the Critic Network

For the conditional distribution of nonlinear system (13), conditional distribution of nonlinear controller (15), and nonlinear critic model (18), the desired target value of the critic network can be calculated by carrying out the calculations implied by (12). Starting by the first term on the right hand side of (12) we get,

$$\begin{aligned} &\int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})\frac{\partial U(x_t, u_t)}{\partial x_t}\frac{\partial x_t}{\partial x_{t-1}}\mathrm{d}(x_t, u_t) = \int \exp[-(u_t - \hat{u}_t)^T\Gamma^{-1}(u_t - \hat{u}_t)] \times \\ &\bigg\{\int \exp[-(x_t - \hat{x}_t)^T\Sigma^{-1}(x_t - \hat{x}_t)]2\hat{x}_t^T\Sigma^{-1}[h'(x_{t-1}) + g'(x_{t-1})u_t]\mathrm{d}x_t\bigg\}\mathrm{d}u_t \\ &= \int \exp[-(u_t - \hat{u}_t)^T\Gamma^{-1}(u_t - \hat{u}_t)]2\hat{x}_t^T\Sigma^{-1}[h'(x_{t-1}) + g'(x_{t-1})u_t]\mathrm{d}u_t \\ &= 2[h(x_{t-1}) + g(x_{t-1})\hat{u}_t]^T\Sigma^{-1}[h'(x_{t-1}) + g'(x_{t-1})\hat{u}_t], \end{aligned} \tag{20}$$

where we have introduced the definitions $\hat{x}_t = h(x_{t-1}) + g(x_{t-1})u_t$ and $\hat{u}_t = W\psi(x_{t-1})$ and where $h'(x_{t-1}) = \frac{\partial h(x_{t-1})}{\partial x_{t-1}}$ and $g'(x_{t-1}) = \frac{\partial g(x_{t-1})}{\partial x_{t-1}}$. The definition of the partial cost $U(x_t, u_t)$ is given in (4). For the considered regularization problem, it evaluates to $U(x_t, u_t) = 2\hat{x}_t^T\Sigma^{-1}x_t - \hat{x}_t^T\Sigma^{-1}\hat{x}_t + 2\hat{u}_t^T\Gamma^{-1}u_t - \hat{u}_t^T\Gamma^{-1}\hat{u}_t$.

For the second term the partial derivatives of $U(x_t, u_t)$, x_t, and u_t with respect to x_t, u_t, and x_{t-1} respectively need to be calculated,

$$\begin{aligned} &\int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})\frac{\partial U(x_t, u_t)}{\partial x_t}\frac{\partial x_t}{\partial u_t}\frac{\partial u_t}{\partial x_{t-1}}\mathrm{d}(x_t, u_t) = \int \exp[-(u_t - \hat{u}_t)^T\Gamma^{-1}(u_t - \hat{u}_t)] \times \\ &\bigg\{\int \exp[-(x_t - \hat{x}_t)^T\Sigma^{-1}(x_t - \hat{x}_t)]2\hat{x}_t^T\Sigma^{-1}g(x_{t-1})W\psi'(x_{t-1})\mathrm{d}x_t\bigg\}\mathrm{d}u_t \\ &= \int \exp[-(u_t - \hat{u}_t)^T\Gamma^{-1}(u_t - \hat{u}_t)]2\hat{x}_t^T\Sigma^{-1}g(x_{t-1})W\psi'(x_{t-1})\mathrm{d}u_t \\ &= 2[h(x_{t-1}) + g(x_{t-1})\hat{u}_t]^T\Sigma^{-1}g(x_{t-1})W\psi'(x_{t-1}), \end{aligned} \tag{21}$$

where $\psi'(x_{t-1}) = \frac{\partial\psi(x_{t-1})}{\partial x_{t-1}}$.

The third term requires calculation of the partial derivatives of $U(x_t, u_t)$ and u_t with respect to u_t and x_{t-1} respectively,

$$\begin{aligned}
&\int s(x_t|u_t, x_{t-1}) c(u_t|x_{t-1}) \frac{\partial U(x_t, u_t)}{\partial u_t} \frac{\partial u_t}{\partial x_{t-1}} \mathrm{d}(x_t, u_t) = \int \exp[-(u_t - \hat{u}_t)^T \Gamma^{-1} (u_t - \hat{u}_t)] \times \\
&\left\{ \int \exp[-(x_t - \hat{x}_t)^T \Sigma^{-1} (x_t - \hat{x}_t)] 2\hat{u}_t^T \Gamma^{-1} W \psi'(x_{t-1}) \mathrm{d}x_t \right\} \mathrm{d}u_t \\
&= \int \exp[-(u_t - \hat{u}_t)^T \Gamma^{-1} (u_t - \hat{u}_t)] 2\hat{u}_t^T \Gamma^{-1} W \psi'(x_{t-1}) \mathrm{d}u_t \\
&= 2\hat{u}_t^T \Gamma^{-1} W \psi'(x_{t-1}).
\end{aligned} \tag{22}$$

The propagation of $\lambda[x_t]$ through the stochastic model of (13) back to x_t yields the fourth term

$$\begin{aligned}
&\int s(x_t|u_t, x_{t-1}) c(u_t|x_{t-1}) \lambda[x_t] \frac{\partial x_t}{\partial x_{t-1}} \mathrm{d}(x_t, u_t) = \int \exp[-(u_t - \hat{u}_t)^T \Gamma^{-1} (u_t - \hat{u}_t)] \times \\
&\alpha(u_t, x_{t-1}) \mathrm{d}u_t,
\end{aligned} \tag{23}$$

where we used

$$\alpha(u_t, x_{t-1}) = \int \exp[-(x_t - \hat{x}_t)^T \Sigma^{-1} (x_t - \hat{x}_t)] (\chi\phi(x_t))^T [h'(x_{t-1}) + g'(x_{t-1}) u_t] \mathrm{d}x_t, \tag{24}$$

and where we used (18) with x_t as input to obtain $\lambda[x_t] = \chi\phi(x_t)$. Using (19) but again with x_t as an input in (24) yields,

$$\begin{aligned}
\alpha(u_t, x_{t-1}) = \int &\begin{bmatrix} \exp[-(x_t - \hat{x}_t)^T \Sigma^{-1} (x_t - \hat{x}_t)] \exp[-(x_t - z_1)^T \gamma_1^{-1} (x_t - z_1)] \\ \exp[-(x_t - \hat{x}_t)^T \Sigma^{-1} (x_t - \hat{x}_t)] \exp[-(x_t - z_2)^T \gamma_2^{-1} (x_t - z_2)] \\ \vdots \\ \exp[-(x_t - \hat{x}_t)^T \Sigma^{-1} (x_t - \hat{x}_t)] \exp[-(x_t - z_L)^T \gamma_L^{-1} (x_t - z_L)] \end{bmatrix}^T \chi^T \\
&\times [h'(x_{t-1}) + g'(x_{t-1}) u_t] \mathrm{d}x_t
\end{aligned} \tag{25}$$

$$= \begin{bmatrix} \exp[-(\hat{x}_t - z_1)^T (\Sigma + \gamma_1)^{-1} (\hat{x}_t - z_1)] \\ \exp[-(\hat{x}_t - z_2)^T (\Sigma + \gamma_2)^{-1} (\hat{x}_t - z_2)] \\ \vdots \\ \exp[-(\hat{x}_t - z_L)^T (\Sigma + \gamma_L)^{-1} (\hat{x}_t - z_L)] \end{bmatrix}^T \chi^T [h'(x_{t-1}) + g'(x_{t-1}) u_t], \tag{26}$$

where the sum of two quadratics as a result of the multiplication of the two exponentials of the first term in brackets in (25) is rewritten in the following form $(x_t - \hat{x}_t)^T \Sigma^{-1} (x_t - \hat{x}_t) + (x_t - z_l)^T \gamma_l^{-1} (x_t - z_l) = (x_t - \mathbf{E}_l)^T (\Sigma^{-1} + \gamma_l^{-1})(x_t - \mathbf{E}_l) + (\hat{x}_t - z_l)^T (\Sigma + \gamma_l)^{-1} (\hat{x}_t - z_l)$ and the integration with respect to x_t is evaluated. Here $\mathbf{E}_l = (\Sigma^{-1} + \gamma_l^{-1})^{-1} (\Sigma^{-1} \hat{x}_t + \gamma_l^{-1} z_l)$. The substitution of (26) in (23) yields the value of the fourth term

$$\begin{aligned}
&\int s(x_t|u_t, x_{t-1}) c(u_t|x_{t-1}) \lambda[x_t] \frac{\partial x_t}{\partial x_{t-1}} \mathrm{d}(x_t, u_t) = \int \exp[-(u_t - \hat{u}_t)^T \Gamma^{-1} (u_t - \hat{u}_t)] \times \\
&\begin{bmatrix} \exp[-(\hat{x}_t - z_1)^T (\Sigma + \gamma_1)^{-1} (\hat{x}_t - z_1)] \\ \exp[-(\hat{x}_t - z_2)^T (\Sigma + \gamma_2)^{-1} (\hat{x}_t - z_2)] \\ \vdots \\ \exp[-(\hat{x}_t - z_L)^T (\Sigma + \gamma_L)^{-1} (\hat{x}_t - z_L)] \end{bmatrix}^T \chi^T [h'(x_{t-1}) + g'(x_{t-1}) u_t] \mathrm{d}u_t \\
&= \int \begin{bmatrix} \exp[-(\hat{x}_t - z_1)^T (\Sigma + \gamma_1)^{-1} (\hat{x}_t - z_1) - (u_t - \hat{u}_t)^T \Gamma^{-1} (u_t - \hat{u}_t)] \\ \exp[-(\hat{x}_t - z_2)^T (\Sigma + \gamma_2)^{-1} (\hat{x}_t - z_2) - (u_t - \hat{u}_t)^T \Gamma^{-1} (u_t - \hat{u}_t)] \\ \vdots \\ \exp[-(\hat{x}_t - z_L)^T (\Sigma + \gamma_L)^{-1} (\hat{x}_t - z_L) - (u_t - \hat{u}_t)^T \Gamma^{-1} (u_t - \hat{u}_t)] \end{bmatrix}^T \chi^T \\
&\times [h'(x_{t-1}) + g'(x_{t-1}) u_t] \mathrm{d}u_t.
\end{aligned} \tag{27}$$

To simplify the integration in (27), the square of the argument of the exponential is completed,

$$\begin{aligned}
&(\hat{x}_t - z_l)^T (\Sigma + \gamma_l)^{-1} (\hat{x}_t - z_l) - (u_t - \hat{u}_t)^T \Gamma^{-1} (u_t - \hat{u}_t) = \\
&(u_t - H_l)^T \Omega_l (u_t - H_l) + (z_l - h(x_{t-1}))^T (\Sigma + \gamma_l)^{-1} (z_l - h(x_{t-1})) \\
&+ \hat{u}_t^T \Gamma^{-1} \hat{u}_t - H_l^T \Omega_l H_l,
\end{aligned} \tag{28}$$

where $\Omega_l = g^T(x_{t-1})(\Sigma + \gamma_l)^{-1} g(x_{t-1}) + \Gamma^{-1}$, and $H_l = \Omega_l^{-1} [g^T(x_{t-1})(\Sigma + \gamma_l)^{-1}(z_l - h(x_{t-1})) + \Gamma^{-1} \hat{u}_t]$. Denoting the exponential of the last constant three terms (the terms that are independent of u_t) on the right hand side of (28) by

$\digamma_l = \exp[-(z_l - h(x_{t-1}))^T(\Sigma + \gamma_l)^{-1}(z_l - h(x_{t-1})) - \hat{u}_t^T\Gamma^{-1}\hat{u}_t + H_l^T\Omega_l H_l]$ and substituting (28) in (27) yields,

$$\int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})\lambda[x_t]\frac{\partial x_t}{\partial x_{t-1}}\mathrm{d}(x_t, u_t) =$$
$$\int \begin{bmatrix} \exp[-(u_t - H_1)^T\Omega_1(u_t - H_1)] \\ \exp[-(u_t - H_2)^T\Omega_2(u_t - H_2)] \\ \vdots \\ \exp[-(u_t - H_L)^T\Omega_L(u_t - H_L)] \end{bmatrix}^T .\digamma^T\chi^T[h'(x_{t-1}) + g'(x_{t-1})u_t]\mathrm{d}u_t$$
$$= [\chi\digamma]^T h'(x_{t-1}) + [\chi\digamma]^T g'(x_{t-1})H. \tag{29}$$

Finally the fifth term can be calculated by propagating $\lambda[x_t]$ through the stochastic model of (13) back to u_t, and then through the action network, which yields,

$$\int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})\lambda[x_t]\frac{\partial x_t}{\partial u_t}\frac{\partial u_t}{\partial x_{t-1}}\mathrm{d}(x_t, u_t) = \int \exp[-(u_t - \hat{u}_t)^T\Gamma^{-1}(u_t - \hat{u}_t)] \times$$
$$\left\{\int \exp[-(x_t - \hat{x}_t)^T\Sigma^{-1}(x_t - \hat{x}_t)][\chi\phi(x_t)]^T g(x_{t-1})W\psi'(x_{t-1})\mathrm{d}x_t\right\}\mathrm{d}u_t$$
$$= [\chi\digamma]^T g(x_{t-1})W\psi'(x_{t-1}). \tag{30}$$

Adding (20), (21), (22), (29), (30) together, yields the target vector of the critic network,

$$\begin{aligned} \lambda^*[x_{t-1}] &= 2[h(x_{t-1}) + g(x_{t-1})\hat{u}_t]^T\Sigma^{-1}[h'(x_{t-1}) + g'(x_{t-1})\hat{u}_t] \\ &+ 2[h(x_{t-1}) + g(x_{t-1})\hat{u}_t]^T\Sigma^{-1}g(x_{t-1})W\psi'(x_{t-1}) + 2\hat{u}_t^T\Gamma^{-1}W\psi'(x_{t-1}) \\ &+ [\chi\digamma]^T h'(x_{t-1}) + [\chi\digamma]^T g'(x_{t-1})H + [\chi\digamma]^T g(x_{t-1})W\psi'(x_{t-1}). \end{aligned} \tag{31}$$

Equation (31) can then be used to correct the critic network and update its parameters.

Remark: All of the above integrals (implied by (20)– (30)) are Gaussian integrals which can be evaluated using theorems and corollaries provided in [10], chapter 10.

C. Probabilistic Control

The randomized control input can be computed by solving the optimality equation (11). The first term on the right hand side of (11) is the expected value of the partial derivatives of $U(x_t, u_t)$ and x_t with respect to x_t and u_t respectively,

$$\int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})\frac{\partial U(x_t, u_t)}{\partial x_t}\frac{\partial x_t}{\partial u_t}\mathrm{d}(x_t, u_t) = \int \exp[-(u_t - \hat{u}_t)^T\Gamma^{-1}(u_t - \hat{u}_t)] \times$$
$$\left\{\int \exp[-(x_t - \hat{x}_t)^T\Sigma^{-1}(x_t - \hat{x}_t)]2\hat{x}_t^T\Sigma^{-1}g(x_{t-1})\mathrm{d}x_t\right\}\mathrm{d}u_t$$
$$= \int \exp[-(u_t - \hat{u}_t)^T\Gamma^{-1}(u_t - \hat{u}_t)]2[h(x_{t-1}) + g(x_{t-1})u_t]^T\Sigma^{-1}g(x_{t-1})\mathrm{d}u_t$$
$$= 2[h(x_{t-1}) + g(x_{t-1})\hat{u}_t]^T\Sigma^{-1}g(x_{t-1}). \tag{32}$$

The second term requires the evaluation of the expected value of the partial derivatives of $U(x_t, u_t)$ with respect to u_t,

$$\int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})\frac{\partial U(x_t, u_t)}{\partial u_t}\mathrm{d}(x_t, u_t) = \int \exp[-(u_t - \hat{u}_t)^T\Gamma^{-1}(u_t - \hat{u}_t)] \times$$
$$\left\{\int \exp[-(x_t - \hat{x}_t)^T\Sigma^{-1}(x_t - \hat{x}_t)]2\hat{u}_t^T\Gamma^{-1}\mathrm{d}x_t\right\}\mathrm{d}u_t$$
$$= \int \exp[-(u_t - \hat{u}_t)^T\Gamma^{-1}(u_t - \hat{u}_t)]2\hat{u}_t^T\Gamma^{-1}\mathrm{d}u_t$$
$$= 2\hat{u}_t^T\Gamma^{-1}. \tag{33}$$

The last term is the expected value of the propagation of $\lambda[x_t]$ through the stochastic model of (13) back to u_t,

$$\begin{aligned}
&\int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})\lambda[x_t]\frac{\partial x_t}{\partial u_t}\mathrm{d}(x_t, u_t) = \int \exp[-(u_t - \hat{u}_t)^T\Gamma^{-1}(u_t - \hat{u}_t)] \times \\
&\left\{\int \exp[-(x_t - \hat{x}_t)^T\Sigma^{-1}(x_t - \hat{x}_t)](\chi\phi(x_t))^T g(x_{t-1})\mathrm{d}x_t\right\}\mathrm{d}u_t \\
&= \int \exp[-(u_t - \hat{u}_t)^T\Gamma^{-1}(u_t - \hat{u}_t)] \times \\
&\left\{\int \begin{bmatrix} \exp[-(x_t - \hat{x}_t)^T\Sigma^{-1}(x_t - \hat{x}_t)]\exp[-(x_t - z_1)^T\gamma_1^{-1}(x_t - z_1)] \\ \exp[-(x_t - \hat{x}_t)^T\Sigma^{-1}(x_t - \hat{x}_t)]\exp[-(x_t - z_2)^T\gamma_2^{-1}(x_t - z_2)] \\ \vdots \\ \exp[-(x_t - \hat{x}_t)^T\Sigma^{-1}(x_t - \hat{x}_t)]\exp[-(x_t - z_L)^T\gamma_L^{-1}(x_t - z_L)] \end{bmatrix}^T \chi^T g(x_{t-1})\mathrm{d}x_t\right\}\mathrm{d}u_t \\
&= \int \begin{bmatrix} \exp[-(\hat{x}_t - z_1)^T(\Sigma + \gamma_1)^{-1}(\hat{x}_t - z_1) - (u_t - \hat{u}_t)^T\Gamma^{-1}(u_t - \hat{u}_t)] \\ \exp[-(\hat{x}_t - z_2)^T(\Sigma + \gamma_2)^{-1}(\hat{x}_t - z_2) - (u_t - \hat{u}_t)^T\Gamma^{-1}(u_t - \hat{u}_t)] \\ \vdots \\ \exp[-(\hat{x}_t - z_L)^T(\Sigma + \gamma_L)^{-1}(\hat{x}_t - z_L) - (u_t - \hat{u}_t)^T\Gamma^{-1}(u_t - \hat{u}_t)] \end{bmatrix}^T \chi^T \\
&\times g(x_{t-1})\mathrm{d}u_t \\
&= (\chi\digamma)^T g(x_{t-1}).
\end{aligned} \tag{34}$$

Adding all terms together yields,

$$2[h(x_{t-1}) + g(x_{t-1})\hat{u}_t]^T\Sigma^{-1}g(x_{t-1}) + 2\hat{u}_t^T\Gamma^{-1} + (\chi\digamma)^T g(x_{t-1}) = 0. \tag{35}$$

The solution of this equation cannot be analytically obtained due to the nonlinear nature of $\digamma$. Hence the controller design is a nonlinear optimization problem which generally leads to a numerical solution. The controller network can then be optimized such that the error between optimal control input u_t^* obtained from (35) and estimated control input u_t from the neural network is minimized. The controller can then generates control signals u_t stochastically from a gaussian distribution having a mean $\hat{u}_t$ and a global average covariance Γ calculated as discussed in Section II-C.

IV. Nonlinear Randomized control Algorithm Based on Probabilistic DHP Critic Methods

The exact evaluation of the closed form of optimal controller in FPD methods (7) is nontrivial and computationally very intensive due to its involvement of multivariate integrations. These multivariate integrations are only tractable for the linear Gaussian case where the mean is linear in both state and control values. This motivated the probabilistic DHP adaptive critic approach discussed in Section III. As can be seen from (31), the desired critic value can be calculated from neural network models of the forward dynamics of the system, nonlinear controller, and critic network. Similarly, the control law can be calculated from (35) once the critic network and system dynamic models become available. Although the probabilistic DHP adaptive critic approach involves multiple computation levels, its implementation can be made by means of modular approach constituting of functional modules and algorithmic modules [8], [20], [12]. The key functional modules are the action and critic networks. Algorithmic modules on the other hand include the computation of the desired critic value (31), computation of the optimal control law (35), and the update of the networks parameters. Each module in the adaptive critic can be modified independently from other modules, thus facilitate fast and reliable implementation. The probabilistic DHP adaptive critic approach allows computation of probabilistic control laws from the FPD methods for linear Gaussian systems with less computational complexities, and more important allows the solution to nonlinear and non Gaussian systems.

The description below is appropriate for direct application to nonlinear control problems of the form stated in Section II.

1) Estimate the pdf of the stochastic model described by (3) as discussed in Section II.
2) Design and initialize the weights of the action network.
3) Design and initialize the weights of the critic network.
4) Calculate the desired value of the critic network using equation (31).
5) Use the difference between the desired value of the critic network as calculated from the previous step and the output of the critic to update the parameters of the critic network until it converges.
6) Use the output of the converged critic in (35) and solve for the optimal control law.
7) Use this value to update the parameters of the action network.
8) Calculate the global covariance matrix from the residual error between the output of the action network and optimal control law as calculated in step 6 and update the conditional distribution of the optimal controller.
9) Repeat Steps $4-6$ until an acceptable performance is reached.

The flow chart of implementing the probabilistic DHP adaptive critic method for nonlinear systems is given in Figure 1. To summarize, the probabilistic DHP adaptive critic training algorithm cycles between a policy improvement routine and a control input determination operation, where the optimal control law and the derivative of the value function, $\lambda[x_{t-1}]$ are approximated by the action and critic networks respectively. The algorithm terminates when control law and the derivative of the value function have converged to the optimal or suboptimal control law and derivative value function respectively. The proof of convergence given in [16], [8] is directly applicable to the probabilistic adaptive critic design in this paper. A nonlinear control problem example to demonstrate the convergence of the proposed probabilistic critic network is given in Section V. Further discussion on the convergence and speed of convergence of adaptive critic designs can be found in [20]. Moreover, empirical evidence on the convergence of the adaptive critic design can be found in [3], [11], [19], [21].

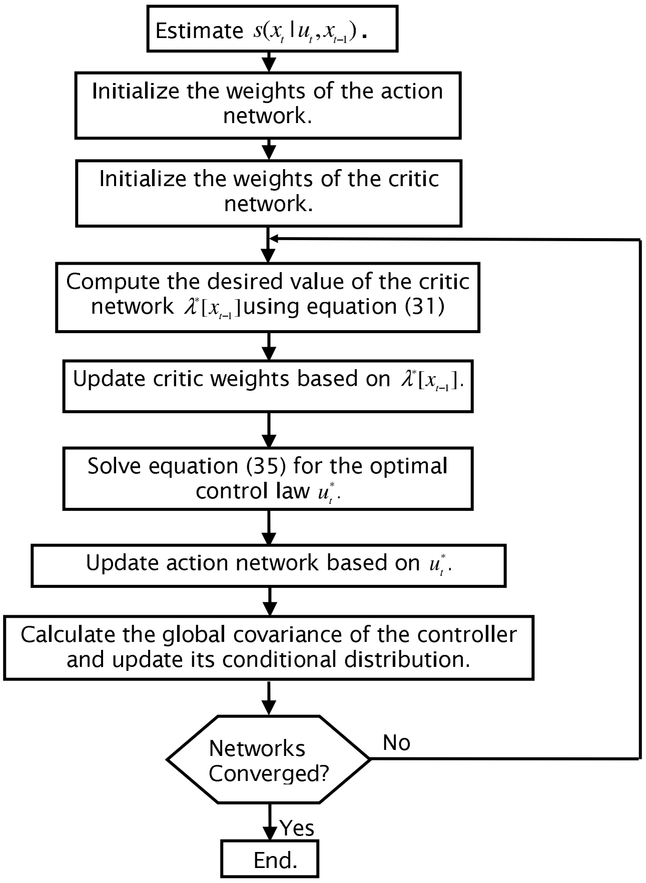

Estimate $s(x_t|u_t,x_{t-1})$.
Initialize the weights of the action network.
Initialize the weights of the critic network.
Compute the desired value of the critic network $\lambda^*[x_{t-1}]$using equation (31)
Update critic weights based on $\lambda^*[x_{t-1}]$.
Solve equation (35) for the optimal control law u_t^*.
Update action network based on u_t^*.
Calculate the global covariance of the controller and update its conditional distribution.
Networks Converged?
No
Yes
End.

Fig. 1. Implementation of the probabilistic DHP adaptive critic for nonlinear systems.

V. Simulation Example

This section demonstrates the probabilistic DHP adaptive critic method on nonlinear stochastic control system. The nonlinear dynamical system is described by the following equation

$$x_t = sin(x_{t-1}) + cos(3 * x_{t-1}) + (2 + cos(x_{t-1}))u_t + \varepsilon_t. \tag{36}$$

The unknown nonlinear dynamics are given by,

$$\begin{aligned} \tilde{h}(x_{t-1}) &= sin(x_{t-1}) + cos(3 * x_{t-1}), \\ \tilde{g}(x_{t-1}) &= (2 + cos(x_{t-1})). \end{aligned}$$

The noise ε_t is assumed to be sampled from a Gaussian distribution, $N(0, 0.01)$. This system has been used in [7] to illustrate theoretical developments for suboptimal dual adaptive control.

The plant is initially, at time $t = 0$, in state x_0, and the aim is to return the plant state to the origin, or a state close to the origin. Thus, a probabilistic DHP critic network that minimizes the cost function (4) is used to design the optimal probabilistic controller and derive optimal control inputs. As a first step in the solution the stochastic model of the forward dynamics of the plant described by (36) is estimated by a Gaussian density function as discussed in Section II-B. Two RBF networks with 15 and 6 Gaussian basis functions are used to estimate the two nonlinear functions $\tilde{h}(x_{t-1})$ and $\tilde{g}(x_{t-1})$ respectively. Hence, the mean of the forward probabilistic model of Gaussian density function is given by $\hat{x}_t = h(x_{t-1}) + g(x_{t-1})u_t$ and the its global variance, $\sigma^2 = 0.0098$ is computed from the residual error of the system dynamics. The weight parameters of the forward

Gaussian probabilistic model are then kept fixed during the critic and action network training. To achieve the control objective of attaining a zero state, the distribution of the ideal state of the system dynamics is taken to be ${}^{I}s(x_t|u_t, x_{t-1}) = \mathcal{N}_{x_t}(0, 0.0098)$. Similarly, the ideal controller distribution is assumed to be ${}^{I}c(u_t|x_{t-1}) = \mathcal{N}_{u_t}(0, 0.01)$.

The control is then initiated by another RBF network with six neurons in the hidden layer for random values of x_t, taken uniformly from the interval $[-4, 4]$. Next, the critic network was also taken to be an RBF neural network with six neurons in the hidden layer. The parameters of the controller and the adaptive critic networks are initialized randomly from a zero mean, isotropic Gaussian, with unit variance scaled by the fan-in of the output units. The target values of the critic network is then calculated using (31) for the specified range of x_t and the critic training is carried out using scaled conjugate gradient method until the weights of the network have converged. The termination criterion of the training process is set to $\mid \triangle F(\chi) \mid < 0.001$ and $\mid \triangle(\chi) \mid < 0.001$ (both must be satisfied) within the limit of 10000 iterations. Here $\mid \triangle F(\chi) \mid$ and $\mid \triangle(\chi) \mid$ is the absolute difference of the error function (defined as the sum of the squares of the errors between target and desired output values) and absolute difference of connection weights of the network between two successive iterations respectively. During training of the critic network, the weights of the action network are kept fixed. The output from the converged critic is then used in (35) solving for optimal control values and the action network is then trained for the same number of iterations, termination criterion and training method as that of the critic network. During training of the action network, the weights of the critic network are kept fixed. After the action network converged, the critic network is trained again (by adapting weights of the previously converged critic) using the outputs of the converged action network. The training of the critic and the action networks are alternated for 3 cycles after which all adaptation is halted and the controller network's ability to return the plant state to the origin is tested.

The control quality of the controller designed in the above manner is then compared with the conventional DHP adaptive critic technique [31]. The same forward neural network model as that used in the above probabilistic design method is used in the conventional DHP critic to represent the forward dynamics of the system. However, only the deterministic forward model represented by the sum of two RBF networks, $\hat{x}_t = h(x_{t-1}) + g(x_{t-1})u_t$ is required for implementation of the conventional DHP adaptive critic method. Moreover, For fair comparison, same noise sequence, initial conditions, and same structure of critic and control neural networks were used during the implementation of each control method.

The ability of the obtained controllers from the conventional and probabilistic DHP adaptive critic methods in returning the system state from its initial value which is taken to be $x_0 = 2$ to the origin is then tested. Note that this initial condition was chosen arbitrarily. During testing, the optimal control signal, u_t is generated from the controller network at each time instant t based on previous state value x_{t-1}. This optimal control signal is then forwarded to the plant of Equation (36) and the system output, x_t is measured. Figure 2 shows histories of the probabilistic and conventional DHP adaptive critic states and control efforts. Both of the probabilistic and conventional critic methods managed to regularize the state of the system around zero as required. The probabilistic DHP critic method, however, ensures minimal overshoot compared to the conventional DHP critic method. This is expected since in the conventional critic method both the critic and controller network parameters have been tuned based on the assumption of the existence of an accurate deterministic forward model. In the probabilistic critic method on the other hand, optimal control law is derived such that the distance between the joint pdf of the closed loop system and an ideal joint pdf is minimized. This allows considering system's uncertainty and improves the performance of the derived optimal control law.

VI. Conclusion

In this paper, the solution to the probabilistic DHP adaptive critic method and randomized control input design have been addressed for a class of dynamic stochastic nonlinear systems. Using RBF neural networks to approximate the conditional expectations of the pdfs and unknown nonlinear models, a randomized control strategy has been developed which minimizes the discrepancy between the joint pdf of the closed loop system and a predetermined ideal joint pdf. It has been shown that the controller design is a nonlinear optimization problem and hence need to be solved numerically. A simulated example is used to illustrate the use of the randomized control algorithm as derived from the probabilistic critic and compared against its counter part of conventional critic design methods. The probabilistic design of critic networks showed minimum overshoot compared to the conventional critic design method.

References

[1] B. D. O. Anderson and J. B. Moore. *Linear Optimal Control.* Prentice Hall, Englewood Cliffs, NJ, 1971.

[2] K. J. Astrom. *Introduction to Stochastic Control Theory.* New York: Academic, 1970.

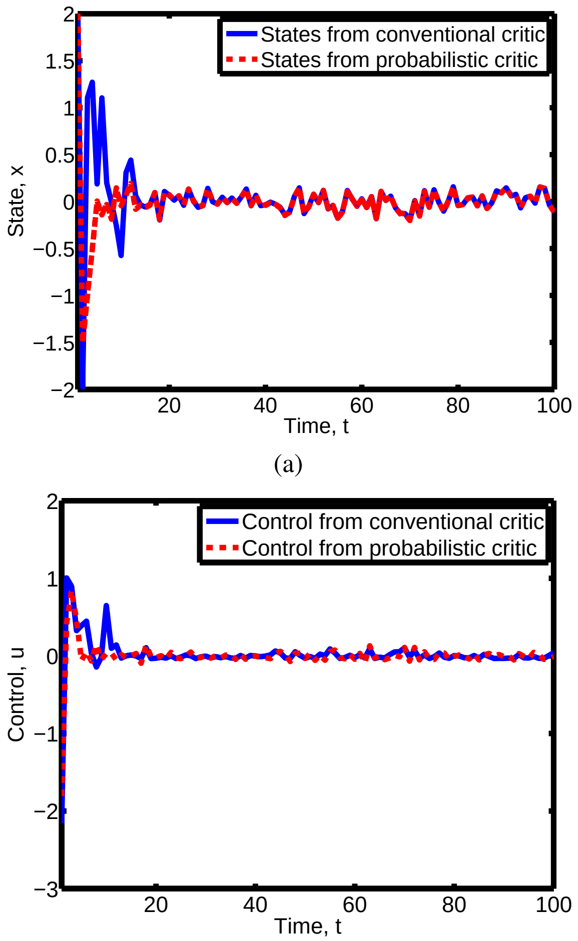
[3] S. N. Balakrishnan and V. Biega. Adaptive-critic-based neural networks for aircraft optimal control. *Journal of Guidance, Control, and Dynamics*, 19(4):893–898, July-August 1996.

[4] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, New York, N.Y., 1995.

[5] L. Blackmore, M. Ono, A. Bektassov, and B. C. Williams. A probabilistic particle-control approximation of chance-constrained stochastic predictive control. *IEEE Transactions on Robotics*, 26(3):502–517, 2010.

[6] M. H. C. Everdij and H. A. P. Blom. Embedding adaptive JLQG into LQ martingale control with a completely observable stochastic control matrix. *IEEE Transactions on Automatic Control*, 41(3):424–430, 1996.

[7] S. G. Fabri and V. Kadirkamanathan. *Functional Adaptive Control: An Intelligent Systems Approach.* Springer-Verlag, February 2001.



(a)

(b)

Fig. 2. Nonlinear stochastic system: (a) Probabilistic and conventional critic estimated values for the state (b) Probabilistic and conventional critic values for control.

[8] Silvia. Ferrari and Robert. F. Stengel. Model based adaptive critic designs. In Jennie Si, Andrew G. Barto, Warren Buckler Powell, and Don Wunsch, editors, *Handbook of Learning and Approximate Dynamic Programming*, chapter 3, pages 64–94. Institute of Electrical and Electronics Engineers, Inc, Canada, 2004.

[9] N. M. Filatov and H. Unbehausen. Adaptive preditive control policy for nonlinear stochastic systems. *IEEE Transactions on Automatic Control*, 40(11):1943–1949, 1995.

[10] Franklin A. Graybill. *Matrices with Applications in Statistics*. Wadsworth International Group, 1983.

[11] D. Han and S. N. Balakrishnan. State-constrained agile missile control with adaptive critic based neural networks. *IEEE Transactions on Control Systems Technology*, 10(4):481–489, 2002.

[12] R. Herzallah. Adaptive critic methods for stochastic systems with input-dependent noise. *Automatica*, 43:1355–1362, August 2007.

[13] R. Herzallah and D. Lowe. A Bayesian perspective on stochastic neuro control. *IEEE Transactions on Neural Networks*, 19(5):914–924, May 2008.

[14] Randa Herzallah. Probabilistic control for uncertain systems. *Dynamic Systems, Measurement, and Control*, 134(2):021018, 2012.

[15] Randa. Herzallah and Miroslav. Káarnáy. Fully probabilistic control design in an adaptive critic framework. *Neural Networks*, 24(11):1128–1135, 2011.

[16] R. Howard. *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, Massachusetts, London, England., 1960.

[17] M. Kárný and T. V. Guy. Fully probabilistic control design. *Systems & Control Letters*, 55(4):259–265, 2006.

[18] Miroslav Kárný. Towards fully probabilistic control design. *Automatica*, 32(12):1719–1722, 1996.

[19] Nilesh V. Kulkarni and K. KrishnaKumar. Intelligent engine control using an adaptive critic. *IEEE Transactions on Control Systems Technology*, 11(2):164–173, 2003.

[20] George G. Lendaris, Roberto A. Santiago, and Michael S. Carrol. Proposed framework for applying adaptive critics in real−time realm. In *Proceedings of the 2002 International Joint Conference on Neural Networks, IJCNN'02*, pages 1796–1801, Honolulu, HI , USA, 2002.

[21] Chuan-Kai Lin. Radial basis function neural network-based adaptive critic control of induction motors. *Applied Soft Computing*, 2011. Article in Press.

[22] S. Meyn and P. Caines. A new approach to stochastic adaptive control. *IEEE Transactions on Automatic Control*, 32(3):220–226, 1987.

[23] Yugang Niu, D. W.C. Ho, and Xingyu Wang. Robust H_∞ control for nonlinear stochastic systems: A sliding-mode approach. *IEEE Transactions on Automatic Control*, 53(7):1695–1701, 2008.

[24] V. Peterka. Bayesian system identification. In P. Eykhoff, editor, *Trends and Progress in System Identification*, pages 239–304. Pergamon Press, Oxford, 1981.

[25] V. Solo. Stochastic adaptive control and martingale limit theory. *IEEE Transactions on Automatic Control*, 35(1):66–71, 1990.

[26] H. Van Trees and K. Bell. A Bayesian approach to problems in stochastic estimation and control. In *Bayesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking*, pages 601–607. Wiley-IEEE Press, 2007.

[27] H. Wang. Robust control of the output probabolity density functions for multivariable stochastic systems. *IEEE Transactions on Automatic Control*, 44:21032107, 1999.

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

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

[30] Hong Wang and Puya Afshar. ILC-based fixed-structure controller design for output PDF shaping in stochastic systems using LMI techniques. *IEEE Transactions on Automatic Control*, 54(4):760–773, 2009.

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