Fully Probabilistic Control Design in an Adaptive Critic Framework

Randa Herzallah[a], Miroslav Kárný[b]

[a] *Faculty of Engineering Technology, Al-Balsa Applied University, Jordan*

herzallah.r@gmail.com

[b] *Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, school@utia.cas.cz*

Abstract

Optimal stochastic controller pushes the closed-loop behavior as close as possible to the desired one. The fully probabilistic design (FPD) uses probabilistic description of the desired closed loop and minimizes Kullback-Leibler divergence of the closed-loop description to the desired one. Practical exploitation of the fully probabilistic design control theory continues to be hindered by the computational complexities involved in numerically solving the associated stochastic dynamic programming problem. In particular very hard multivariate integration and an approximate interpolation of the involved multivariate functions. This paper proposes a new fully probabilistic control algorithm that uses the adaptive critic methods to circumvent the need for explicitly evaluating the optimal value function, thereby dramatically reducing computational requirements. This is a main contribution of this short paper.

Key words:

stochastic control design, fully probabilistic design, adaptive control, adaptive

1 Introduction

Stochastic control design minimizes an expected cost function with respect to feedback control strategies, e.g. [1,5]. It influences selected characteristics, e.g. noncentral second moments, of the joint probability density function (pdf) of variables occurring in the optimized closed loop. The studied FPD [11,7,12] pushes this joint pdf to the user-specified ideal pdf describing the desired behavior of the closed loop. The FPD has a strong intuitive appeal and provides an explicit minimizing strategy. Although the minimizer can be obtained explicitly, computational requirements of the FPD approach are still intensive. Numerically the FPD approach involves computation of subsequent integrations to minimize an expected cost function subject to the probability density function of the system dynamics. Practical implementation of the FPD approach is difficult because of 1) multivariate integration and curse of dimensionality 2) non Gaussian probability density functions prevent the cost function from being written in a closed analytical form, which consequently does not allow exploitation of the rich available analytical results 3) the FPD approach assumes the existence of perfectly known pdf models of the systems to be controlled, which are rarely available.

The contribution of this paper lies in developing an adaptive critic solution to the FPD problem. The proposed fully probabilistic adaptive critic approach uses a critic network that approximates the derivative of a cost function derived from a Kullback-Leibler distance between the joint probability density function of the closed loop system and an ideal joint probability density func-

tion. The critic network critiques the controller and the outputs of that controller, hence considered as a feedback rather than an open loop controller. The action network provides estimate for the conditional distribution of the optimal control strategy as derived from the FPD either on or off line. In contrast to the original FPD, the proposed adaptive critic solution reduces the computational requirements and does not assume the existence of perfectly known pdf models of the system dynamics to be controlled. As such, more robust control strategy can be derived for real world systems where hypothetical probability measures of the system dynamics are assumed. This paper provides a basis for considering the computational intelligence-based adaptive critic methods along with the existing classical FPD approach for developing a more robust and practically implementable control.

To emphasize, this work uses neural network approximation methods to complement the techniques of conventional stochastic control theory, which are well developed, tested and implemented. This represents the novelty of the new probabilistic adaptive critic framework proposed in this paper: whilst the proposed design is firmly rooted in stochastic control, the needed probabilistic models are handled by stochastic version of neural networks. These are proved to be very effective tools for obtaining probabilistic models of stochastic linear and nonlinear mappings. The new design provides a general solution for stochastic systems subject to random inputs and deterministic systems characterized by functional uncertainty with unknown probability density functions. Hence the contribution of this work to intelligent control stems from the nature of the plant and the environment being considered, which covers functional uncertainty and randomness. These are the typical conditions under which an intelligent controller is expected to operate so as to improve the performance

and autonomy of conventional control schemes.

Throughout, $\equiv$ is defining equality; $f(\cdot|\cdot)$ stands for a probability density function (pdf); the conditioning symbol | is also used as separator in functions that need not to be pdfs; t labels discrete-time moments, $t \in \{1, \ldots, H\}$; $H \leq \infty$ is a given control horizon; $d_t = (x_t, u_t)$ is the data record at time t consisting of an observed vectorial measurable state x_t and of an optional vectorial system input u_t; $d(t)$ stands for the sequence $(d_1, \ldots, d_t)$; integrals are multiple and definite over the integrand domain.

2 Problem Formulation

Assume that the system can be represented by the following nonlinear stochastic model

$$x_t = g(x_{t-1}, u_t, \epsilon_t), \tag{1}$$

where x_t is the measured state vector, u_t is the control input to the system, ϵ_t is a white noise, which has zero mean and covariance P, and g(.) is an unknown nonlinear function that represents the system dynamics. Because of the existence of the noise, only the conditional probability density functions (pdfs) of the future state values can be specified at each instant of time t as follows

$$s(x_t|u_t, x_{t-1}), \tag{2}$$

In general $s(\cdot|\cdot)$ needs not to be known and no assumption is made on whether ϵ_t has a known probability density function.

The objective of the FPD is then to determine a randomized optimal control law described by the conditional pdf

$$c(u_t|x_{t-1}) \tag{3}$$

that minimizes the Kullback-Leibler divergence (KLD) between the actual joint pdf $f(D)$ of the observed data $D = (x(H), u(H))$ and the ideal joint pdf ${}^{I}f(D)$ acting on a set possible Ds and defined as follows

$$\mathcal{D}\left(f||\,{}^{I}f\right) \equiv \int f(D) \ln\left(\frac{f(D)}{{}^{I}f(D)}\right) dD. \tag{4}$$

The KLD in (4) has the following key property

$$\mathcal{D}(f||\,{}^{I}f) \geq 0,\ \mathcal{D}(f||\,{}^{I}f) = 0 \text{ iff } f = {}^{I}f \text{ almost everywhere on } D. \tag{5}$$

The joint pdf $f(D) \equiv f(d(H))$ of the data sequence $D \equiv d(H)$ is the most complete probabilistic description of the (observed) behavior of the closed control loop. The chain rule for pdfs [18] allows its factorisation as follows

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

The first generic factor in (6) is the conditional pdf of the system dynamic given in (2) and the second generic term describes the optional (randomized) causal controller given in (3). To reemphasise, probability density functions of the system dynamics and inverse controller are assumed to be unknown and need to be estimated in this article. The estimation method of these probability density functions will be discussed in Section 3.

The interpretation of the ideal pdf as a result of standard control design implies that it can be factorised in the way mimic to (6) with an "ideal" system

model ${}^{I}s(x_t|u_t, x_{t-1})$ and "ideal" controller ${}^{I}c(u_t|x_{t-1})$ mimic to (2) and (3), respectively

$$ {}^{I}f(D) = \prod_{t=1}^{H} {}^{I}s(x_t|u_t, x_{t-1})\, {}^{I}c(u_t|x_{t-1}). \tag{7} $$

Minimization of (4) with respect to the control input can be obtained recursively by first defining $-\ln(\gamma(x_{t-1}))$ to be the expected minimum cost-to-go function (alternatively called value function) corresponding to (4)

$$ \begin{aligned} -\ln(\gamma(x_{t-1})) &= \min_{\left\{c(u_\tau|x_{\tau-1}\}_{\tau\geq t}^{H}\right\}} \sum_{\tau=t}^{H} \int f(d_t, \ldots, d(H)|x_{t-1}) \\ &\times \ln\left(\frac{s(x_\tau|u_\tau, x_{\tau-1})c(u_\tau|x_{\tau-1})}{{}^{I}s(x_\tau|u_\tau, x_{\tau-1})\, {}^{I}c(u_\tau|x_{\tau-1})}\right) \mathrm{d}(d_t, \ldots, d(H)), \end{aligned} $$

for arbitrary $\tau \in \{1, \ldots, H\}$. Using this definition minimization is then performed recursively to give the following recurrence functional equation

$$ \begin{aligned} -\ln(\gamma(x_{t-1})) &= \min_{c(u_t|x_{t-1})} \int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1}) \\ &\times \Bigg[\underbrace{\ln\left(\frac{s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})}{{}^{I}s(x_t|u_t, x_{t-1})\, {}^{I}c(u_t|x_{t-1})}\right)}_{\equiv\text{partial cost} \Longrightarrow U(x_t, u_t)} - \underbrace{\ln(\gamma(x_t))}_{\text{optimal cost-to-go}} \Bigg] \mathrm{d}(x_t, u_t). \end{aligned} \tag{8} $$

Full derivation of (8) is given in the appendix. Equation (8) constitute the recurrence equation of the dynamic programming solution to the FPD control problem.

The recurrence equation can then be used backward in time to obtain an approximate solution to the exact optimal control history. Here the evaluation of any control action u_t, at time t, involves performing $H - t$ subsequent integrations. Furthermore, the evaluation of the optimal cost-to-go function, $\gamma(x_{t-1})$ involves repeating these subsequent integrations many times. Using stored values of later optimal cost-to-go, the backward propagation is implemented to

evaluate the control strategy, which means very large storage requirements. This backward dynamic programming approach is very expensive computationally for higher dimensional systems. The required expansion of the state and storage of all optimal cost lead to a number of computations that grows exponentially with the number of the state variables, a phenomenon known as the curse of dimensionality.

3 An Adaptive Critic Approach to the Fully Probabilistic Control

In this paper, we seek to avoid the difficulties of the FPD arising from the multivariate integration and curse of dimensionality. This can be achieved by way of the adaptive critic methods derived from the forward dynamic programming approach. They use a critic network to approximate the optimal cost-to-go and an action network to provide prediction for the optimal control policy. The critic methods overcome the curse of dimensionality problem through function approximation while approaching the optimal solution over time. The main objective here is to achieve satisfactory convergence to the optimal or near-optimal solution.

Adaptive critic designs are neural network based designs for optimization that combine concepts of reinforcement learning and approximate dynamic programming [21,20,19,15–17]. They consist of two neural networks, an action network that produces optimal actions and an adaptive critic that approximates the performance of the action network [2,8,13]. Depending on the specific role performed by the key component called the critic, the critic network approximates the optimal cost-to-go function or its derivative and is then trained using recursive equations derived from dynamic programming [22].

The critic network is trained forward in time, which reduces computational time and storage requirements in real time control applications. This also has the advantage of predicting future cost and take prevention a head of time before applying the control signal.

Figure 1 illustrates the FPD adaptive critic design proposed in this paper. It is based on the dual heuristic programming (DHP) scheme of adaptive critic methods. In the proposed FPD adaptive critic, the critic approximates the derivative of the cost function in (8) with respect to the state,

$$\lambda^*[x_{t-1}] = \frac{\partial[-\ln(\gamma(x_t))]}{\partial x_{t-1}}. \tag{9}$$

The proposed architecture contains parametric blocks called the controller or action network, the forward model and the critic network. The action network is responsible for estimating the conditional distribution of control signals, while the critic network provides approximation to the derivative of the optimal cost-to-go function as specified in (9). The forward model can be either a mathematical model or neural network approximation to the conditional distribution of the system dynamics.

3.1 Forward Model

The method adopted for estimating the conditional distribution of the system dynamics is based on the method proposed in [9], where neural network models are used to provide a prediction for the conditional expectation of the system output and calculating the variance of its residual error. For the general class of stochastic nonlinear systems given by (1), the forward model, which provides a prediction for the conditional expectation of the state can be

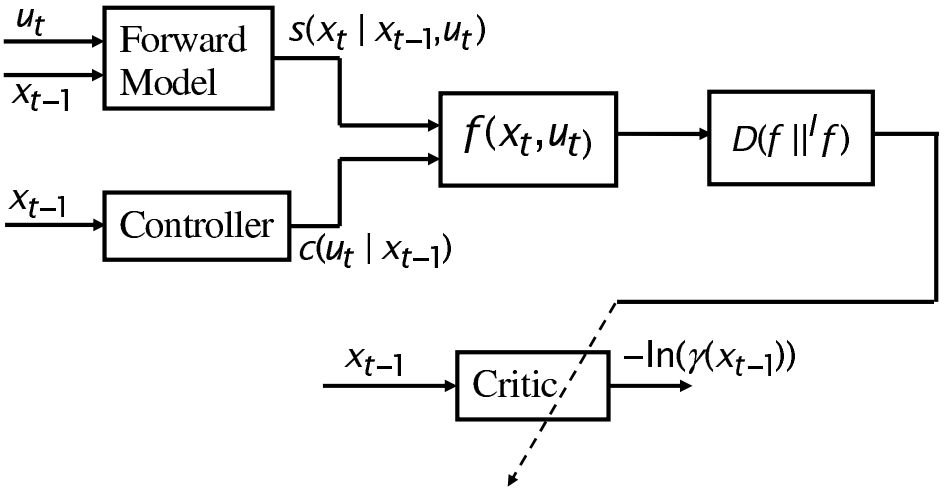


Fig. 1. The architecture of the proposed probabilistic DHP adaptive critic method

either a mathematical or neural approximation of the following general form

$$\hat{x}_t = N_f(x_{t-1}, u_t), \tag{10}$$

determined by a non-linear mapping N_f. The sum of squares error between actual state values x_t and estimated state values $\hat{x}_t$ is used to optimize the parameters of this model. Once the forward model of the plant is identified the following stochastic model can be built

$$x_t = \hat{x}_t + \tilde{\epsilon}_t, \tag{11}$$

where $\tilde{\epsilon}_t$ represents the residual error of the system output, which is shown [9] to be close to Gaussian random noise with zero mean and $\tilde{P}$ covariance. According to the developed stochastic model in (11), the distribution of the state values will be Gaussian distribution with expected means provided by the approximating nonlinear network and a global covariance $\tilde{P}$ given by the residual value of the error between actual and estimated states $E\left((x_t - \hat{x}_t)(x_t - \hat{x}_t)^T\right)$.

3.2 Controller Network

Similar to the forward model, the conditional distribution of control signals is estimated using neural network models, which are used to provide a prediction for the conditional expectation of control. For the proposed probabilistic DHP AC, the optimal control law can be computed from the following optimality equation, which is obtained by deriving (8) with respect to the control input,

$$\frac{\partial[-\ln(\gamma(x_{t-1}))]}{\partial u_t}\Big|_{u_t=u_t^*} = \int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1}) \tag{12}$$
$$\times\Big[\frac{\partial U(x_t,u_t)}{\partial x_t}\frac{\partial x_t}{\partial u_t} + \frac{\partial U(x_{t-1},u_t)}{\partial u_t} + \lambda[x_t]\frac{\partial x_t}{\partial u_t}\Big]\mathrm{d}(x_t,u_t) = 0.$$

Here the controller network can be optimized such that the error between optimal control input u_t^*, obtained from (12) 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. Hence, the controller generates a control signal u_t stochastically from a Gaussian distribution having a mean computed from the controller network and a global covariance matrix equal to the residual error between the output of the network and the optimal control signal, $E\left((u_t^* - u_t)(u_t^* - u_t)^T\right)$.

***Remark 1**:* Although the conditional distribution function of the system output and control signals are assumed to be Gaussian in this paper, expected values of these distributions are estimated using nonlinear neural network models.

***Remark 2**:* The sum of square errors does not require that the distributions of the states or control inputs to be Gaussian. If a sum of square error is used however the quantities, which can be determined are the conditional

expectations of the estimated outputs and the global average covariances of the residual errors.

Remark 3: It would be possible to consider more general estimators of the residual error, but the Gaussian assumption is adequate for the moment. Future work, will consider how this assumption may be relaxed by modelling the noise distribution using a mixture model.

3.3 Adaptive Critic Network

Given estimation of control law from the control network and the derivative of the critic network $\lambda[x_{t-1}]$, the critic network can be optimized by computing its desired value as follows

$$\lambda^*[x_{t-1}] = \int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})\Big[\frac{\partial U(x_t,u_t)}{\partial x_t}\frac{\partial x_t}{\partial x_{t-1}} + \frac{\partial U(x_t,u_t)}{\partial x_t}\frac{\partial x_t}{\partial u_t}\frac{\partial u_t}{\partial x_{t-1}} + \frac{\partial U(x_t,u_t)}{\partial u_t}\frac{\partial u_t}{\partial x_{t-1}} + \lambda[x_t]\frac{\partial x_t}{\partial x_{t-1}} + \lambda[x_t]\frac{\partial x_t}{\partial u_t}\frac{\partial u_t}{\partial x_{t-1}}\Big]\mathrm{d}(x_t,u_t), \quad (13)$$

which is obtained by deriving (8) with respect to the state. The parameters of the critic network can then be adapted such that the error between the desired value of the critic given in (13) and estimated value from the critic network is minimized.

3.4 Training Algorithm

Assuming dual heuristic programming DHP for training the adaptive critic network, the training process for the adaptive critic network has been known to be a two stage process. The training of the action network, which outputs

the optimal control policy $u[x_{t-1}]$ 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 utility function can be calculated, this in combination with the critic outputs and the system model derivatives, allows the use of (13) 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 (12) solving for the target u_t^*, which is then used to correct the action network. These two steps continue until a predetermined level of convergence is reached.

To reemphasise, adaptive critic designs are based on an algorithm that cycles between a policy improvement routine and an action determination operation. Here the algorithm approximates the optimal control law and the value function $-\ln(\gamma(x_{t-1}))$. The algorithm terminates when control law and value function or its derivative $\lambda[x_{t-1}]$ have converged to the optimal or suboptimal control law and value function or its derivative respectively. The proof of convergence given in [6,10] is directly applicable to the probabilistic adaptive critic design in this paper. A simple example on the convergence of the proposed probabilistic critic network is given in Section 5.

Although adaptive critic algorithm involves multiple computational levels, its implementation can be made by means of modular approach constituting of functional modules and algorithmic modules. The key functional modules are the action and critic networks. Algorithmic modules however include networks parameters updating, solve the optimality Equation (12), and compute the desired value of the critic network (13). Each module in the adaptive critic can

be modified independently from other modules thus algorithmic changes and debugging are performed fast and reliable. The speed of convergence is dependent on the suitability of the initialized control law and value function as well as the optimization algorithm [6,9]. Further discussion on the convergence and speed of convergence of adaptive critic designs can be found in [14]. Moreover, empirical evidence on the convergence of the adaptive critic design can be found in [2,8,13,15].

4 Linear Gaussian Case

Since analytic solution of the fully probabilistic design method can be obtained if all pdfs of the system dynamics and the controller are Gaussian, this section will evaluate the adaptive critic solution and the FPD solution to the linear Gaussian case. In other words, the linear Gaussian systems will serve as a benchmark problem for checking the meaning of the proposed adaptive critic method and illustrating the use of the FPD.

For that purpose the theory developed in the previous section is applied here to a linear-Gaussian state space model, described by the following stochastic equation,

$$\begin{aligned} x_t &= Ax_{t-1} + Bu_t + \omega_t \\ s(x_t \mid x_{t-1}, u_t) &\rightsquigarrow \mathcal{N}_{x_t}(Ax_{t-1} + Bu_t, \Sigma), \end{aligned} \tag{14}$$

where A and B are the state and control matrices respectively, ω_t is the noise of the residual error of the state, and where the covariance of the residual error Σ is estimated as described in Section 3. The system is initially in state x_{t-1} and the aim is to return the system state to the origin. Hence, the recurrence

functional equation defined in (8) is considered as the performance function to be minimized. As a result of the chain rule Equation (8) can be expressed in the form,

$$-\ln(\gamma(x_{t-1})) = \int f(x_t, u_t)\Big[\ln\frac{s(x_t|x_{t-1},u_t)}{{}^{I}s(x_t|u_t,x_{t-1})} + \ln\frac{c(u_t|x_{t-1})}{{}^{I}c(u_t|x_{t-1})} - \ln(\gamma(x_t))\Big]\mathrm{d}(x_t,u_t). \tag{15}$$

The system state distribution $s(.\mid .)$ in Equation (15) is defined in (14). The ideal state distribution is assumed to be,

$${}^{I}s(x_t|u_t,x_{t-1}) = \mathcal{N}_{x_t}(0,\Sigma) \tag{16}$$

It reflects the regulation problem with the realistic aim of reaching the zero state, with a spread being determined by the covariance of the innovation Σ.

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

$$\begin{aligned} u_t &= Cx_{t-1} + \epsilon_t \\ c(u_t|x_{t-1}) &\rightsquigarrow \mathcal{N}_{u_t}(Cx_{t-1},\Gamma), \end{aligned} \tag{17}$$

where C is the matrix of the controller parameters, ϵ_t is the residual error of the control input, and Γ is the covariance of the residual error of control. The distribution of the ideal controller is also assumed to be

$${}^{I}c(u_t|x_{t-1}) = \mathcal{N}_{u_t}(0,\Gamma), \tag{18}$$

where Γ is the innovations in estimating the optimal control input.

Define the following matrix,

$$\tilde{E} = A + BC \tag{19}$$

Using (19) in (14), the stochastic equation of the state space model can then be rewritten as follows,

$$x_t = \mu_t + B\epsilon_t + \omega_t = \tilde{E}x_{t-1} + B\epsilon_t + \omega_t. \tag{20}$$

By the backward induction, it will be verified that $-\ln(\gamma(x_t)) = 0.5x_t^T M x_t + \mathbb{Q}_0$ with some positive semidefinite matrix $M \geq 0$ and some constant $\mathbb{Q}_0 \geq 0$. Since it proves convenient to make use of this result almost immediately, but, first the following definition is introduced

$$M = \tilde{E}^T M \tilde{E} + C^T \Gamma^{-1} C + \tilde{E}^T \Sigma^{-1} \tilde{E}. \tag{21}$$

We claim that the optimal performance index satisfies the following condition

$$-\ln(\gamma(x_{t-1})) = 0.5x_{t-1}^T M x_{t-1} + \mathbb{Q}_0, \tag{22}$$

To prove this we simply substitute into (15) and verify that it is satisfied. For the left hand side of the equation we get:

$$-\ln(\gamma(x_{t-1})) = 0.5x_{t-1}^T M x_{t-1} + \mathbb{Q}_0. \tag{23}$$

For the right hand side, we get

$$\begin{aligned}
&= \int \sum_{t=0}^{H-1} f(x_t, u_t) \Big[\ln \frac{s(x_t|x_{t-1}, u_t)}{{}^I s(x_t|u_t, x_{t-1})} + \ln \frac{c(u_t|x_{t-1})}{{}^I c(u_t|x_{t-1})} \\
&+ 0.5x_t^T M x_t + \mathbb{Q}_0 \Big] \mathrm{d}(x_t, u_t) \\
&= \frac{1}{2} x_{t-1}^T C^T \Gamma^{-1} C x_{t-1} + \frac{1}{2} x_{t-1}^T \tilde{E}^T \Sigma^{-1} \tilde{E} x_{t-1} + \frac{1}{2} x_{t-1}^T \tilde{E}^T M \tilde{E} x_{t-1} \\
&+ \frac{1}{2} \mathrm{tr}[M\Sigma] + \mathbb{Q}_0.
\end{aligned} \tag{24}$$

Because M satisfies (21) and because the distributions of the errors are constant with respect to x, this tells that,

$$\int f(x_t,u_t)\Big[\ln\frac{s(x_t|x_{t-1},u_t)}{{}^{I}s(x_t|u_t,x_{t-1})}+\ln\frac{c(u_t|x_{t-1})}{{}^{I}c(u_t|x_{t-1})}-\ln(\gamma(x_t))\Big]\mathrm{d}(x_t,u_t).$$
$$=0.5x_{t-1}^{T}Mx_{t-1}+\mathbb{Q}_0. \tag{25}$$

Comparing (25) and (22), we see that the quadratic nature of the optimal performance index is satisfied.

4.1 Proposed Adaptive Critic in Linear Gaussian Case

In this section, the target of the critic as they would be generated by DHP from the proposed adaptive critic solution will be calculated and compared to that of the correct target values. The correct target values of the critic network can be obtained by deriving (25) with respect to the state, x_{t-1} as follows:

$$\lambda[x_{t-1}]=Mx_{t-1} \tag{26}$$

To calculate the output of the critic network, $\lambda[x_t]$ should firstly be calculated, and then we carry out the calculations implied by (13),

$$\lambda[x_t]=Mx_t \tag{27}$$

The first term on the right hand side of (13), requires calculation of the partial derivatives of $U(x_t,u_t)$ and x_t with respect to x_t and x_{t-1} respectively,

$$\int s(x_t|u_t,x_{t-1})c(u_t|x_{t-1})\frac{\partial U(x_t,u_t)}{\partial x_t}\frac{\partial x_t}{\partial x_{t-1}}\mathrm{d}(x_t,u_t)=\tilde{E}^{T}\Sigma^{-1}Ax_{t-1}$$

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,

$$\int s(x_t|u_t,x_{t-1})c(u_t|x_{t-1})\frac{\partial U(x_t,u_t)}{\partial x_t}\frac{\partial x_t}{\partial u_t}\frac{\partial u_t}{\partial x_{t-1}}\mathrm{d}(x_t,u_t)=\tilde{E}^{T}\Sigma^{-1}BCx_{t-1}$$

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,

$$\int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})\frac{\partial U(x_t, u_t)}{\partial u_t}\frac{\partial u_t}{\partial x_{t-1}}\mathrm{d}(x_t, u_t) = C^T\Gamma^{-1}Cx_{t-1}$$

The fourth term requires propagating $\lambda[x_t]$ through the stochastic model of (14) back to x_t, which yields,

$$-\int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})\lambda[x_t]\frac{\partial x_t}{\partial x_{t-1}}\mathrm{d}(x_t, u_t) = A^T M[A + BC]x_{t-1}$$

Finally the fifth term can be calculated by propagating $\lambda[x_t]$ through the stochastic model of (14) back to u_t, and then through the action network, which yields,

$$-\int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})\lambda[x_t]\frac{\partial x_t}{\partial u_t}\frac{\partial u_t}{\partial x_{t-1}}\mathrm{d}(x_t, u_t) = MB^T C^T[A + BC]x_{t-1}$$

Adding all terms together, yields the target vector of the critic network,

$$\begin{aligned}\lambda[x_{t-1}] = {} & \tilde{E}^T\Sigma^{-1}Ax_{t-1} + \tilde{E}^T\Sigma^{-1}BCx_{t-1} + C^T\Gamma^{-1}Cx_{t-1} \\ & + A^T M[A + BC]x_{t-1} + MB^T C^T[A + BC]x_{t-1}\end{aligned} \tag{28}$$

Using (21) in (28) yields,

$$\lambda[x_{t-1}] = Mx_{t-1} \tag{29}$$

From (29), it can clearly be seen that the target value as estimated by the critic network is equal to the correct critic value. This validates the theoretical development of the proposed DHP adaptive critic solution of the FPD problem proposed in this paper.

4.2 Action Network in Linear Gaussian Case

For linear Gaussian systems, the parameters of the action network C in Equation (17) can be obtained for a corresponding value function $\lambda[x_t]$ by carrying out calculations implied by Equation (12). The first term on the right hand side of (12), requires calculation of the partial derivatives of $U(x_t, u_t)$ and x_t with respect to x_t and u_t respectively,

$$\begin{aligned}&\int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})\frac{\partial U(x_t, u_t)}{\partial x_t}\frac{\partial x_t}{\partial u_t}\mathrm{d}(x_t, u_t)\\&= B^T\Sigma^{-1}Ax_{t-1} + B^T\Sigma^{-1}BCx_{t-1}\end{aligned}$$

For the second term partial derivatives of $U(x_t, u_t)$ with respect to u_t need to be calculated

$$\int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})\frac{\partial U(x_t, u_t)}{\partial u_t}\mathrm{d}(x_t, u_t) = \Gamma^{-1}Cx_{t-1}$$

The last term can be calculated by propagating $\lambda[x_t]$ through the stochastic model of (14) back to u_t, which yields

$$\int s(x_t|u_t, x_{t-1})c(u_t|x_{t-1})\lambda[x_t]\frac{\partial x_t}{\partial u_t}\mathrm{d}(x_t, u_t) = B^TMAx_{t-1} + B^TMBCx_{t-1}$$

Adding all terms together and solving for the parameters of the action network yields,

$$C = -(B^TMB + B^T\Sigma^{-1}B + \Gamma^{-1})^{-1}(B^TMA + B^T\Sigma^{-1}A). \tag{30}$$

The controller can then generates control signals u_t stochastically from a gaussian distribution having a mean $\hat{C}x_{t-1}$ and a global average covariance Γ calculated as discussed in Section 3. Here $\hat{C}$ is the estimate of the optimal controller

parameters C obtained from the neural network model.

5 Simulation Example

In order to illustrate the validity of the theoretical development of the proposed probabilistic adaptive critic, the theory developed in Section 3 is applied here to a single input 2–outputs control problem described by the following stochastic equation

$$\mathbf{x}_{t+1} = \mathbf{G}\mathbf{x}_t + \mathbf{H}\mathbf{u}_{t+1} + \mathbf{w}_{t+1}, \tag{31}$$

where

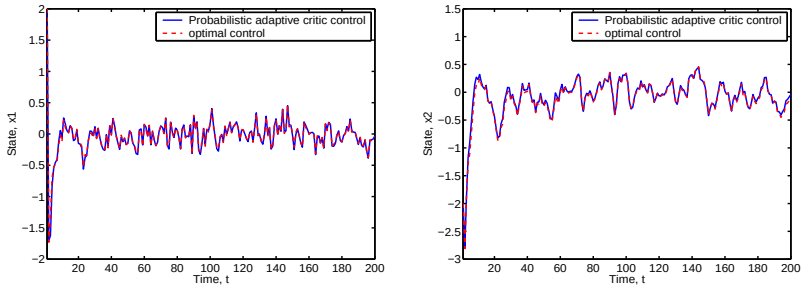
$$\mathbf{G} = \begin{bmatrix} 0 & 0 \\ -0.5 & 1 \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad E[\mathbf{w}_{t+1}\mathbf{w}_{t+1}^T] = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}.$$

The plant is initially, at time $t = 0$, in state $\mathbf{x}_0 = [2; -2]$, and the aim is to return the plant state to the origin, or a state close to the origin. As a first step in the solution, the Gaussian probability density function of the stochastic model described by Equation (31) is estimated, as discussed in Section 3, using two generalized linear networks to provide predictions for the expected values of states x^1, and x^2, and a global diagonal covariance matrix $\Sigma = [0.0098, 0.0101]$. Another generalized linear network is used to provide prediction for the expected value of control signal. The variance of the controller is taken to be, $\Gamma = [0.01]$. Next, the critic networks were taken to be multi-layer perceptron neural networks with five 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.

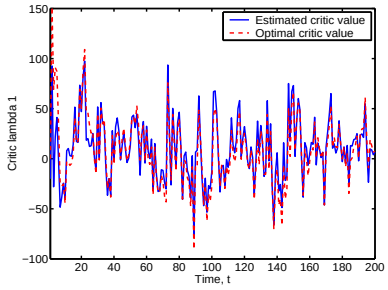
The target values of the critic networks are calculated as specified in Equation (13) for random values of x_t until the critic network converges. The action network is then trained from the converged critic networks, which are used in solving Equation (12) for control. After the action network converges, the critic networks are again trained (by adapting weights of the previously converged critic) using the outputs of the trained action network. This process is repeated until action and critic networks converge.

For comparison purposes, two experiments were conducted to demonstrate the optimal control using policy method of dynamic programming and the proposed probabilistic adaptive critic approach. The numerical results are shown in Figure 2. To demonstrate the validity of the proposed probabilistic adaptive critic approach we have presented plots of the proposed probabilistic adaptive critic states and optimal states histories in Figures 2(a), and (b). In each of the cases we can see that the optimal state values (as obtained from the policy method of dynamic programming) and the proposed probabilistic adaptive critic solutions are superimposed on each other. Figures 2(c), and (d) show the estimated critic values from the proposed probabilistic adaptive critic approach and the optimal critic values . It can be seen that the estimated critic values converge to the optimal critic values. The optimal control history and the adaptive critic control history are presented in Figure 2(e). This figure shows that the estimated control is always close to the optimal control.

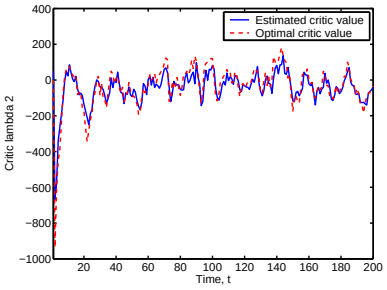


(a)

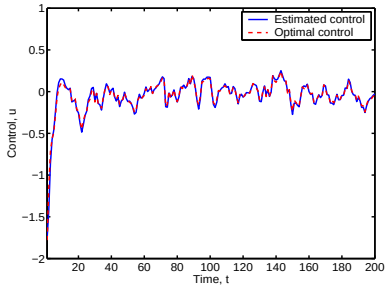
(b)



(c)



(d)



(e)

Fig. 2. Controlled multi dimensional stochastic system: (a) Optimal and critic estimated values for state 1. (b) Optimal and critic estimated values for state 2. (c) Critic output 1. (d) Critic output 2. (e) Optimal and critic values for control.

6 Conclusion

A new approximate neural network based solution of the FPD problem has been proposed. The proposed solution uses the adaptive critic method to ap-

proximate the FPD problem. It reduces computational requirements of the original FPD problem and does not assume the existence of perfectly known probability pdf models of the system dynamics to be controlled. As such, more robust control strategy can be derived for real world systems where hypothetical probability measures of the system dynamics are assumed. This paper provides a basis for considering the computational intelligence-based adaptive critic methods along with the existing classical FPD approach for developing a more robust and practically implementable control.

Moreover, the proposed formulation of the FPD in an adaptive critic framework allows exploitation of functional uncertainty when deriving the optimal control law. Functional uncertainty are usually state and control dependent. If functional uncertainty are made available for the proposed approach, an adaptive critic controller can provide superior performance with minimum a priori knowledge of system dynamics. In future work, we will discuss how to estimate and exploit functional uncertainty in a fully probabilistic adaptive critic design approach.

7 Appendix

7.1 Justification of the Fully Probabilistic Design

Stochastic control design orders controllers according to the expected cost assigned to them [3]. Generally, the optimal controller is designed such as to minimize a performance index J of the following form

$$J = \int \tilde{\mathcal{U}}(L(D), D) f(D) \, \mathrm{d}D. \tag{32}$$

The loss function $L(D)$ with real, possibly infinite, values orders a posteriori closed-loop behaviors D, i.e., it quantifies preference among them and expresses compromise between generic multiple-objectives of the control design. It extends partial preference ordering, which the designer has among possible behaviors. It has to be isotonic with respect to this ordering but its choice is far from being unique. Usually, unnecessary requirements (like smoothness, additivity etc.) are added in order to reduce this freedom.

The function $\tilde{\mathcal{U}}$ shapes additionally the loss so that the prior risk attitude of the designer can be respected. Again, $\tilde{\mathcal{U}}$ is far from being unique and has to meet just a few requirements, like isotonicity with respect to L values, measurability and a sort of smoothness.

The risk-attitude-expressing the role of $\tilde{\mathcal{U}}$ justifies the following assumption

$$\tilde{\mathcal{U}}(L(D), D) \equiv \mathcal{U}(L(D), f(D)). \tag{33}$$

The assumption (33) says that the risk-related modification of the loss is the same for equally probable behaviors. It is acceptable and resembles a sort of likelihood principle [4].

Assume that a randomized controller can be found for a given conditional pdf of system dynamics and any pair of L, $\tilde{\mathcal{U}}$, such that it minimizes the performance index J. The interconnection of the randomized controller with the pdf of the system is given by the ideal joint pdf, ${}^{I}f(D)$. Then, the following proposition can then be introduced,

Proposition 1 (Justification of the FPD) *Let i) (33) holds; ii)* $\mathcal{U}(\cdot,\cdot)$ *has continuous partial derivative with respect to the second argument for almost all*

D; iii) the minimum of J is reached for $f(D) = {}^{I}f(D)$*; iv)* $\mathcal{U}(L(D), {}^{I}f(D)) =$ *constant for almost all D.*

Then, J as given in (32) is an increasing affine transformation of the KLD, $\mathcal{D}\left(f(D)||\,{}^{I}f(D)\right)$.

Proof: Let us denote, $f = f(D)$, ${}^{I}f = {}^{I}f(D)$ and let $f = {}^{I}f + \varepsilon\delta$ be a variation of the optimal pdf ${}^{I}f$ minimizing the performance index J. Here $\varepsilon \geq 0$ and $\delta = \delta(D)$ is an arbitrary function, which has zero integral over $\S D$ and which guarantees that $f \geq 0$. Substituting (33) into (32) and setting the derivative of J with respect to $f(D)$ to zero yields the necessary condition for ${}^{I}f$ to be a minimizer. Thus,

$$\int \delta(D) \left[\mathcal{U}(L(D), {}^{I}f) + {}^{I}f \frac{\partial}{\partial z}\mathcal{U}(L(D), z)_{z={}^{I}f(D)}\right] \mathrm{d}D = 0 \underbrace{\Rightarrow}_{\int \delta D \,\mathrm{d}D=0 \;and\; iv)}$$

$${}^{I}f \frac{\partial \mathcal{U}(L(D), z)}{\partial z}_{|z={}^{I}f(D)} = A > 0 \Rightarrow \mathcal{U}(Z(D), f(D)) = A \ln(f(D)) + \mathcal{B}(D),$$

for some function $\mathcal{B}(D)$. By applying again the assumption *iv*), we get the claimed result. □

Neither the adopted "likelihood principle" nor smoothness represent a significant restriction of generality. The restriction degree caused by the "uniformity" requirement *iv*) is not clear yet.

7.2 Stochastic Principle of Optimality

For the FPD control problem the expected value of the KLD should be minimized

$$\mathcal{D}\left(f||\,{}^{I}f\right) = E\Big\{\sum_{\tau=t}^{H}\int f(d_\tau|x_{\tau-1}) \times \ln\left(\frac{s(x_\tau|u_\tau,x_{\tau-1})c(u_\tau|x_{\tau-1})}{{}^{I}s(x_\tau|u_\tau,x_{\tau-1})\,{}^{I}c(u_\tau|x_{\tau-1})}\right)\,\mathrm{d}(d_\tau)\Big\}.$$

The stochastic principle of optimality can then be derived by firstly defining the minimum expected cost function as follows

$$\begin{aligned}-\ln(\gamma(x_{t-1})) = &\min_{\left\{c(u_\tau|x_{\tau-1}\right\}_{\tau\geq t}^{H}\}} E\Big\{\sum_{\tau=t}^{H}\int f(d_\tau|x_{\tau-1})\\ &\times \ln\left(\frac{s(x_\tau|u_\tau,x_{\tau-1})c(u_\tau|x_{\tau-1})}{{}^{I}s(x_\tau|u_\tau,x_{\tau-1})\,{}^{I}c(u_\tau|x_{\tau-1})}\right)\,\mathrm{d}(d_\tau)\Big\}.\end{aligned}$$

By splitting the summation into two parts

$$\begin{aligned}&-\ln(\gamma(x_{t-1}))\\ &= \min_{c(u_t|x_{t-1})}\Bigg[\min_{(c(u_\tau|x_{\tau-1}))_{\tau=t+1}^{H}}\Big\{\int f(d_1,\ldots,d_t|x_{t-1})\\ &\times \ln\left(\frac{s(x_t|u_t,x_{t-1})c(u_t|x_{t-1})}{{}^{I}s(x_t|u_t,x_{t-1})\,{}^{I}c(u_t|x_{t-1})}\right)\mathrm{d}(d_1,\ldots,d_t)\\ &+\sum_{\tau=t+1}^{H}\int f(d_\tau|x_{\tau-1})\times\ln\left(\frac{s(x_\tau|u_\tau,x_{\tau-1})c(u_\tau|x_{\tau-1})}{{}^{I}s(x_\tau|u_\tau,x_{\tau-1})\,{}^{I}c(u_\tau|x_{\tau-1})}\right)\mathrm{d}(d_\tau)\Big\}\Bigg].\end{aligned}$$

When respecting dependence on the optimised control, we get

$$\begin{aligned}-\ln(\gamma(x_{t-1})) = &\min_{c(u_t|x_{t-1})}\Bigg[\int f(d_1,\ldots,d_t|x_{t-1})\Big\{\ln\left(\frac{s(x_t|u_t,x_{t-1})c(u_t|x_{t-1})}{{}^{I}s(x_t|u_t,x_{t-1})\,{}^{I}c(u_t|x_{t-1})}\right)\\ &-\ln(\gamma(x_t))\Big\}\,\mathrm{d}(d_1,\ldots,d_t)\Bigg]\end{aligned}$$

References

[1] K.J. Astrom. *Introduction to Stochastic Control*. Academic Press, New York, 1970.

[2] 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.

[3] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 1985.

[4] J. M. Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7(3):686–690, 1979.

[5] D.P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Nashua, US, 2001. 2nd edition.

[6] 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.

[7] T. V. Guy and M. Kárný. Stationary fully probabilistic control design. In J. Filipe, J. A. Cetto, and J. L. Ferrier, editors, *Proceedings of the Second International Conference on Informatics in Control, Automation and Robotics*, pages 109–112, Barcelona, September 2005. INSTICC.

[8] 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.

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

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

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

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

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

[14] 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.

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

[16] Wei-Song Lin and Ping-Chieh Yang. Adaptive critic motion control design of autonomous wheeled mobile robot by dual heuristic programming. *Automatica*, 44:2716–2723, August 2008.

[17] Derong Liu, Hossein Javaherian, Olesia Kovalenko, and Ting Huang. Adaptive critic learning techniques for engine torque and air-fule ratio control. *IEEE Transactions on Systems, Man, and Cybernetics*, 38(4):988–993, 2008.

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

[19] D. V. Prokhorov, R. A. Santiago, and D. C. Wunsch II. Adaptive critic designs: A case study of neurocontrol. *Neural Networks*, 8(9):1367–1372, 1995.

[20] D. V. Prokhorov and D. C. Wunsch. Adaptive critic designs. *IEEE Transactions on Neural Networks*, 8(5):997–1007, September 1997.

[21] J. Si, A. Barto, W. Powell, and D. C. Wunsch. *Handbook of Learning and Approximate Dynamic Programming*. Wiley, New York, N.Y., 2004.

[22] 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.