Fully Probabilistic Control for Stochastic Nonlinear Control Systems with Input Dependent Noise

Randa Herzallah

Non–linearity and Complexity Research Group
Aston University, United Kingdom
Email:r.herzallah@aston.ac.uk

Abstract

Robust controllers for nonlinear stochastic systems with functional uncertainties can be consistently designed using probabilistic control methods. In this paper a generalised probabilistic controller design for the minimisation of the Kullback–Leibler divergence between the actual joint probability density function (pdf) of the closed loop control system, and an ideal joint pdf is presented emphasising how the uncertainty can be systematically incorporated in the absence of reliable systems models. To achieve this objective all probabilistic models of the system are estimated from process data using mixture density networks (MDNs) where all the parameters of the estimated pdfs are taken to be state and control input dependent. Based on this dependency of the density parameters on the input values, explicit formulations to the construction of optimal generalised probabilistic controllers are obtained through the techniques of dynamic programming and adaptive critic methods. Using the proposed generalised probabilistic controller, the conditional joint pdfs can be made to follow the ideal ones. A simulation example is used to demonstrate the implementation of the algorithm and encouraging results are obtained.

Keywords: Functional uncertainty. Dual heuristic programming. Adaptive critic. Nonlinear stochastic systems. Mixture of Gaussians. Fully probabilistic design.

1. Introduction

Control systems should be designed to run satisfactorily under stochastic uncertain conditions and variations in operating conditions. In highly complex and nonlinear systems it is practically impossible to identify globally valid models. This is not only due to the uncertainty arising from the presence of unknown parameters or functions. It concerns also characteristics variations of the dynamics of the systems, particularly if the systems are time varying or exhibiting a number of distinct modes of behavior during their operation. Robust control strategies in the absence of reliable world models and in the presence of strong uncertainty must be designed for practical control to take place. Among existing robust control strategies, probabilistic control has been proposed to identify satisfactory controllers. Since the control problem is formulated as the minimisation of an objective function with respect to the optimal control law, the probabilistic control problem is transformed to a stochastic optimisation problem.

The conventional theory of stochastic control is particularly suitable for taking into account randomly varying system parameters and designing probabilistic control strategies under uncertain working conditions. However, the general stochastic control problem is intractable to solve and requires an exponential amount of memory and computation time (Astrom, 1970; Werbos, 1992). Therefore, predominantly linear stochastic systems with Gaussian random variables were studied and the performance objectives concerned were confined to be either mean value or variance of the stochastic output (Anderson and Moore, 1971; Niu et al., 2008; Solo, 1990; Everdij and Blom, 1996; Blackmore et al., 2010). For practical control systems where either the system variables or the noise are not Gaussian, the mean and variance of system outputs may not be sufficient to characterise the closed loop system behavior (Fabri and Kadirkamanathan, 2001; Wang and Afshar, 2009; Herzallah, 2012). As a result, several groups of control methods have been developed. Examples of the so far developed groups of control methods are: 1) stochastic distribution control, which is concerned with the problem of designing a controller so that the pdf of the system output follows a

pre-specified ideal distribution (Wang and Afshar, 2009); 2) closed loop pdf control (Kárný, 1996; Kárný and Guy, 2006), where the selected control influences the closed loop description of the system under consideration; and 3) the control of the tracking error pdf (Herzallah and Lowe, 2008).

This paper is concerned with extending the second group of control methods, known as fully probabilistic design (FPD). For stochastic systems with measurable states x_t , the objective of the standard FPD method is to determine the pdf of a randomised optimal control law, u_t described by

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

that minimises the following backward recurrence equation,

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

Here $-\ln(\gamma(x_{t-1}))$ is the expected minimum cost–to–go function and the most complete actual probabilistic description of the closed loop system $h(x_t,u_t \mid x_{t-1})$ factorises

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

by the chain rule (Peterka, 1981) into the model $s(x_t|u_t,x_{t-1})$ describing the system dynamics and the controller description $c(u_t|x_{t-1})$. Similarly,

$${}^{I}h(x_t,u_t \mid x_{t-1}) = {}^{I}s(x_t|u_t,x_{t-1})\,{}^{I}c(u_t|x_{t-1}), \tag{3}$$

is the factorisation of the actual 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.

However, up to now, the solution to the FPD method is developed only for stochastic systems where the involved pdfs in Equations (2) and (3) are estimated using single model approaches such as the parametric models belonging to exponential family (EF) (Barndorff-Nielsen, 1978), or neural network Gaussian models (Herzallah and Kárnaý, 2011). For these stochastic systems, the pdf of optimal controller, $c^*(u_t|x_{t-1})$, minimising the cost–to–go function given in Equation (1) is shown to be generally determined by the following functional recursion,

$$\begin{aligned} c^*(u_t|x_{t-1}) &= \frac{{}^{I}c(u_t|x_{t-1})\exp[-\beta(u_t,x_{t-1})]}{\gamma(x_{t-1})}, \\ \gamma(x_{t-1}) &= \int {}^{I}c(u_t|x_{t-1})\exp[-\beta(u_t,x_{t-1})]du_t, \\ \beta(u_t,x_{t-1}) &= \int s(x_t|u_t,x_{t-1}) \Bigg[\ln\left(\frac{s(x_t|u_t,x_{t-1})}{{}^{I}s(x_t|u_t,x_{t-1})}\right) \\ &\quad - \ln(\gamma(x_t)) \Bigg] dx_t. \end{aligned} \tag{4}$$

Nevertheless, using single model approaches for estimating the involved pdfs is restrictive in many real world applications that are characterised by strong nonlinearity, multimodality and uncertainty. Recently a methodology known as multiple model (a.k.a mixture of experts) adaptive estimation and control has become highly appealing in the identification of unknown and arbitrarily random noise and in modelling the general pdfs of the system dynamics (Murray-Smith and , Eds.; Narendra and Driolet, 2001; Karniel et al., 2001). In this framework, a weighted sum of

local models is used to model the nonlinear dynamics of the system. In effect, every local model is a representation of one particular operation mode. Depending on the problem being handled, a number of modeling techniques are employed to represent the local models including standard neural networks (Park and Sandberg, 1991), statistical mixture models (Titterington et al., 1985; Smídl et al., 2005), and regression type models (Wang et al., 2013) among others. The multiple model approach has been recently exploited in the FPD method for deriving a randomised controller for systems that operate in different operation modes (Kárný et al., 2003).

The method proposed in (Kárný et al., 2003), however, is constrained by its high computational complexity. In principle, the optimal control law can be obtained by solving a recurrence equation analogical to the dynamic programming solution. However, this necessitates sequential evaluations and storage of the expected cost–to–go function subject to the probability density functions of the system dynamics. The implementation and exploitation of the available analytical results of FPD become even more arduous for the majority of practical systems, which are non–linear and non–Gaussian in nature. Most important, the parameters of the probabilistic models involved in the FPD are assumed in (Kárný et al., 2003) to be input and state independent, limiting the resulting control strategies to deterministic certainty–equivalent systems.

An objective of this paper is to generalise the standard FPD method in the context of multiple model techniques by developing a generalised probabilistic controller, minimising the Kullback–Leibler divergence between the actual joint pdf of the closed loop control system, and an ideal joint pdf, when the involved pdfs have their parameters dependent on the input values in the way reflecting the uncertainty of the network dynamics. Mixture density networks (MDNs) from the neural network field (Bishop, 1995; Herzallah, 2012) will be used in this paper to estimate the required complete pdfs, as combinations of Gaussian components, such that their parameters are state and control dependent. Furthermore, to overcome the computational complexity involved in the FPD method, the recently developed probabilistic dual heuristic programming (DHP) adaptive critic methods will be adopted and extended here as well to develop the proposed generalised probabilistic control solution. The proposed method describes a general way of how to understand and incorporate uncertainty in deriving optimal control strategies. It emphasises that when the density function parameters are dependent on the input values, not only should the expected value of the Kullback–Leibler distance be minimised but also the variance of its cost function.

To summarise, the aim of this article is to extend and develop the solution of the FPD control problem given in Equation (4) such that the system and control models' uncertainties are taken into consideration when deriving the optimal control law. Compared with the results presented in (Kárný et al., 2003) this article has three distinct features. Firstly, based on using the well developed probabilistic based MDN methods (Herzallah, 2012), the involved pdfs in the FPD method are estimated such that their parameters are dependent on the input values in the way reflecting the uncertainty of the network dynamics. Having achieved this, the solution to the FPD control problem is generalised such that models uncertainty is taken into consideration in the derivation of the optimal control law. Secondly, to minimise the computational complexity involved in the proposed generalised FPD control method, the probabilistic dual heuristic programming (DHP) adaptive critic methods (Herzallah and Kárnaý, 2011) are adopted and extended here . This extended method of probabilistic DHP adaptive critic method is referred to as the generalised probabilistic DHP adaptive critic method. Thirdly, until now the FPD control method has only been demonstrated, theoretically and numerically, on regulation control problems. As opposed to this, the development of the proposed generalised probabilistic control solution in this paper will be demonstrated on the general tracking control problem where the system states are required to follow desired state values. The regulation problem where the objective is to reach a zero state can be considered as a special case of the tracking problem where the desired value of the state is set to zero.

To emphasise, generalised probabilistic controllers proposed in this article provide a pragmatic method for effectively and robustly controlling complex stochastic dynamical systems under uncertain conditions. They take knowledge of uncertainty into consideration in the derivation of the optimal control law. This represents the novelty of the new generalised probabilistic controller proposed in this paper.

The paper is organised in the following way. Section 2 provides some preliminaries, formulates and solves the generalised probabilistic control problem. The solution to the generalised probabilistic DHP adaptive critic is demonstrated in Section 3. Section 4 provides the algorithm of the generalised probabilistic DHP control problem. Numerical example is in Section 5. Section 6 provides concluding remarks.

2. Preliminaries, Problem Statements and Solution

This preparatory section recalls basic elements of modelling the conditional pdf of the system response and formulates and solves the generalised fully probabilistic control method proposed in this paper.

2.1. Model Representation

Consider a stochastic discrete time nonlinear control system as follows

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

where $x_t \in R^n$ is the measured state vector, $u_t \in R^r$ is the control input vector, $f(.) : R^n \longmapsto R^n$ is unknown smooth nonlinear function of the state, and $\epsilon_t \in R^n$ is an independent noise signal. It should be pointed out that ϵ_t is an arbitrary bounded independent random variable where in general f can be not invertible with respect to it.

Since the random forces affecting the state values of the system are arbitrary independent random variables, the pdf of the system states will in general be non–Gaussian density function. Hence in this case, a sum of squares or cross entropy error function for estimating the system states is not expected to yield good prediction. In order to obtain a complete description of the state values, the conditional pdf of the state, x_t conditioned on previous state x_{t-1} and control input values u_t should be modelled. This paper applies MDNs to model the general pdf of the system state from process data. This method for estimating the pdf of the system states falls under the multiple models approach and will be discussed next. The solution to the FPD control problem when the pdfs of the system states are estimated using MDNs is given in Section 2.3.

2.2. Modelling of system states pdf

For a general stochastic system of the form described in Equation (5), MDNs provide a general framework for modelling the complete general pdf of the random state variables, $s(x_t \mid x_{t-1}, u_t)$ (Bishop, 1995; Herzallah, 2012). Here, the pdf of the states, x_t is represented as a linear combination of kernel functions of the form

$$s(x_t \mid x_{t-1}, u_t) = \sum_{j=1}^{M} \alpha_j(x_{t-1}, u_t) f_j(x_t \mid x_{t-1}, u_t), \tag{6}$$

where M is the number of kernels in the mixture model. The parameters $\alpha_j(x_{t-1}, u_t)$ are the prior probabilities of x_t having been generated from the j^{th} component of the mixture. The functions $f_j(x_t \mid x_{t-1}, u_t)$ represent the conditional pdf of x_t for the j^{th} kernel. Various choices of the kernel functions are possible (Bishop, 1995), however, in this article Gaussian kernel functions are considered,

$$f_j(x_t \mid x_{t-1}, u_t) \propto \exp\left\{ - \frac{\| x_t - \hat{x}_{t,j}(x_{t-1}, u_t) \|^2}{2\sigma^2_{t,j}(x_{t-1}, u_t)} \right\}, \tag{7}$$

where $\hat{x}_{t,j}(x_{t-1}, u_t)$, and $\sigma^2_{t,j}(x_{t-1}, u_t)$ represent the centre and the variance respectively of the j^{th} kernel. In Equation (7) the entries of the state vector x_t are assumed to be conditionally independent within each kernel and can be described by a common variance $\sigma^2_{t,j}(x_{t-1}, u_t)$, hence spherical Gaussian kernels. A spherical Gaussian assumption can be relaxed in a very straightforward way, by using a full covariance matrix for each Gaussian kernel. However using full covariance is not necessary, because in principle a Gaussian mixture model with sufficiently many kernels of the type given by Equation (7) can approximate any given density function arbitrarily accurately providing that the mixing coefficients and the Gaussian parameters are correctly chosen.

For any given values of (x_{t-1}, u_t), the mixture model (6) provides a general formalism for modelling sufficiently smooth conditional pdf $s(x_t \mid x_{t-1}, u_t)$. Note that the prior probabilities, the centre, and the variances of the state pdf are taken to be continuous functions of the input variables (x_{t-1}, u_t). These functions are estimated as the outputs of a feed–forward neural network that takes (x_{t-1}, u_t) as input. They represent the set of parameters which govern the system state pdf. This combination of a density model and a feed–forward neural network is represented schematically in Figure 1.

Measured process data, states and applied control inputs, are then used to train the MDN and obtain the parameters of the state pdf. The MDN is then trained to minimise the negative logarithm of the probability density function of the system states by using back–propagation

$$E = -\sum_{q} \ln\left\{ \sum_{j=1}^{M} \alpha_j(x_{t-1}, u_t) f_j(x_t \mid x_{t-1}, u_t) \right\}. \tag{8}$$

Details of the derivatives of the error function given by Equation (8) with respect to the outputs of the networks and constraints of the mixture models can be found in (Herzallah and Lowe, 2004; Bishop, 1995).

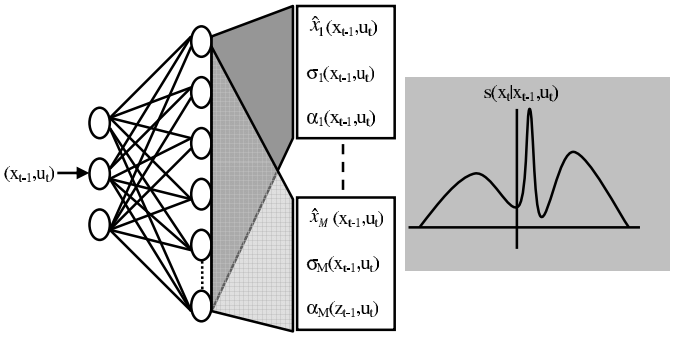

Figure 1: The architecture of the system states Mixture density network

2.3. Control Objective, Problem Formulation and Solution

As discussed in Section 1, within the standard FPD approach the stochastic system described by Equation (5) is controlled by randomised controllers described by conditional pdfs $c(u_t \mid x_{t-1})$. Since the pdf of the system states $s(x_t \mid x_{t-1}, u_t)$ is a general pdf and is estimated using MDNs as a weighted sum of mixture of Gaussians, the pdf of the randomised controller $c(u_t \mid x_{t-1})$ need then to be acquired from the general pdf of the system states. Such learning must be able to divide up the control into appropriate regions which can be recombined to generate the system behavior. This implies that for each behavior captured by a kernel function of the pdf of the system states, a paired control input kernel function must be designed (Herzallah, 2012). As such, the pdf of the control inputs can be approximated as a weighted sum of these control input kernel functions as follows

$$c(u_t \mid x_{t-1}) = \sum_{j=1}^{M} \phi_j(x_{t-1}) g_j(u_t \mid x_{t-1}), \tag{9}$$

where $\phi_j(x_{t-1})$ are the prior probabilities of u_t having been generated from the j^{th} component of the control input mixture, and $g_j(u_t \mid x_{t-1})$ represent the conditional density of u_t for the j^{th} control input kernel.

Remark 1: Note that in Equation (9) the prior probabilities of the control inputs are taken to be independent of the system states prior probabilities. The control input prior probabilities however, can be fixed to the corresponding probabilities from the system states pdf (Herzallah, 2012) such that the state and control input are tightly coupled. This means that each control input kernel learns to provide an appropriate control input under the context for which its paired kernel of the state pdf most likely produces the state value.

The kernel functions of the inverse controller are again taken to be Gaussian kernel functions,

$$g_j(u_t \mid x_{t-1}) \propto \exp\left\{ -\frac{\| u_t - \hat{u}_{t,j}(x_{t-1}) \|^2}{2\rho^2_{t,j}(x_{t-1})} \right\}, \tag{10}$$

where $\rho^2_{t,j}(x_{t-1})$ and $\hat{u}_{t,j}(x_{t-1})$ represent the variance and centre of the j^{th} kernel of control input respectively. Similar to the system states pdf, the centre, priors, and the variances of the control input pdf are continuous functions

of the input variable x_{t-1}. They are estimated as the outputs of a feed–forward neural network that takes x_{t-1} as input. The centre, variances and priors of the inverse controller MDN represent the set of parameters which govern the inverse controller pdf. Here the target of the controller mixture density network is the optimal control input as calculated in Section 3.3, Equation (34). The mixture density network is then trained to minimize the negative logarithm of the probability density function of the control input u_t by using back–propagation.

Finally the ideal joint pdf ${}^{I}h(x_t, u_t \mid x_{t-1}) = {}^{I}s(x_t \mid x_{t-1}, u_t)\, {}^{I}c(u_t \mid x_{t-1})$ is assumed to be independent of the component kernels of the actual joint pdf $h(x_t, u_t \mid x_{t-1}) = s(x_t \mid x_{t-1}, u_t)c(u_t \mid x_{t-1})$. For example if a tracking control problem is considered, a reasonable form for the ideal state pdf, ${}^{I}s(x_t \mid x_{t-1}, u_t) = {}^{I}f(x_t \mid u_t, x_{t-1})$ would be a Gaussian pdf with its mean value being set to a desired value and variance to the non–reducible variance of innovations of the active component. It specifies where the system state will be of high probability. Similarly, the ideal control input pdf ${}^{I}c(u_t \mid x_{t-1}) = {}^{I}g(u_t \mid x_{t-1})$ can also be assumed to be Gaussian which expresses where the system control input should be.

With the previous definitions of the state pdf given in Equation (6), the control input pdf given in Equation (9), and the assumed Gaussian form of the ideal joint pdf, an approximate solution to the FPD control problem can be developed. The FPD solution depends on the method which is selected for calculating the active kernel component from an MDN. Several methods for calculating the output from the MDN have been proposed in the literature (Bishop, 1995; Herzallah and Lowe, 2004). These methods can be generalised in a straightforward manner to select the active kernel component from the MDN. A method of interest in multimodal control applications is the most probable component from the MDN corresponding to the most probable branch. Taking the most probable branch from the system states pdf, the system states density function can be approximated by a Gaussian density as follows,

$$\begin{aligned} \arg\max_j\{\alpha_j(x_{t-1}, u_t)\} \longrightarrow s(x_t \mid x_{t-1}, u_t) \\ = f_j(x_t \mid x_{t-1}, u_t). \end{aligned} \tag{11}$$

Similarly the control input pdf can be approximated by a Gaussian density as follows,

$$\begin{aligned} \arg\max_j\{\phi_j(x_{t-1})\} \longrightarrow c(u_t \mid x_{t-1}) \\ = g_j(u_t \mid x_{t-1}). \end{aligned} \tag{12}$$

Taking the most probable branch from the system states and control input pdfs as given by Equations (11) and (12) yields the optimal controller described by the following theorem.

Theorem 1 If the system a) selects most probable branch, $g_j(u_t \mid x_{t-1})$ of the pdf of control inputs u_t, b) models the state evolution by most probable branch (11), c) expresses its aims by the ideal system model ${}^{I}f(x_t \mid x_{t-1}, u_t)$ and by an ideal controller ${}^{I}g(u_t \mid x_{t-1})$, then the optimal controller minimising the cost–to–go function (1) is given by

$$\begin{aligned} g_j^*(u_t|x_{t-1}) &= \frac{{}^{I}g(u_t|x_{t-1})\exp[-\beta(u_t, \hat{x}_{t,j}, x_{t-1})]}{\gamma(x_{t-1})}, \\ \gamma(x_{t-1}) &= \int {}^{I}g(u_t|x_{t-1})\exp[-\beta(u_t, \hat{x}_{t,j}, x_{t-1})]du_t, \\ \beta(u_t, \hat{x}_{t,j}, x_{t-1}) &= \int f_j(x_t|u_t, x_{t-1}) \ln\Big(\frac{f_j(x_t|u_t, x_{t-1})}{{}^{I}f(x_t|u_t, x_{t-1})}\Big)dx_t - \ln(\tilde{\gamma}(x_t)), \\ -\ln(\tilde{\gamma}(x_t)) &= -\ln(\gamma(\hat{x}_{t,j})) - \frac{1}{2}\nabla^2_{x_t}\ln(\gamma(x_t))\mid_{\hat{x}_{t,j}} \sigma^2_{t,j}(u_t, x_{t-1}). \end{aligned} \tag{13}$$

***Proof*:** The result is implied by the following sequence of equalities in which we use the definition of cost–to–go function (1), Fubini theorem on multiple integration (Rao, 1987), marginalisation, normalisation and the chain rule of pdfs (Peterka, 1981) together with conditional independence expressed by the assumptions stated above.

Using the most probable branch of state model (11), the ideal distribution of state model ${}^{I}f(x_t \mid x_{t-1}, u_t)$, and the ideal distribution of the controller ${}^{I}g(u_t \mid x_{t-1})$ in Equation (1) and the chain rule yields,

$$\begin{aligned} -\ln(\gamma(x_{t-1})) = \int f_j(x_t \mid u_t, x_{t-1}) g_j(u_t \mid x_{t-1}) \\ \times\left[\ln\left(\frac{f_j(x_t \mid u_t, x_{t-1})}{{}^{I}f(x_t \mid u_t, x_{t-1})}\right) + \ln\left(\frac{g_j(u_t \mid x_{t-1})}{{}^{I}g(u_t \mid x_{t-1})}\right) - \ln(\gamma(x_t))\right] d(x_t, u_t). \end{aligned} \tag{14}$$

Now using Fubini theorem, Equation (14) can be recast as follows,

$$-\ln(\gamma(x_{t-1})) = \int g_j(u_t \mid x_{t-1}) \Bigg[\underbrace{\delta(u_t, \hat{x}_{t,j}, x_{t-1}) + \ln\left(\frac{g_j(u_t \mid x_{t-1})}{{}^{I}g(u_t \mid x_{t-1})}\right)}_{\equiv \text{partial cost} \Longrightarrow U(u_t, \hat{x}_{t,j})}$$

$$- \underbrace{< \ln(\gamma(x_t)) >}_{\text{expected optimal cost–to–go}} \Bigg] du_t, \tag{15}$$

where we used,

$$\delta(u_t, \hat{x}_{t,j}, x_{t-1}) = \int f_j(x_t \mid u_t, x_{t-1}) \ln\left(\frac{f_j(x_t \mid u_t, x_{t-1})}{{}^{I}f(x_t \mid u_t, x_{t-1})}\right) dx_t. \tag{16}$$

Given that the system is stabilisable which means finite moments of x_t and since $\ln(\gamma(x_t))$ is continuous bounded and differentiable, the expected value of the cost–to–go function $< \ln(\gamma(x_t)) >$ can be approximated by Taylor series to be given by

$$\begin{aligned} - < \ln(\gamma(x_t)) > &= -\ln(\gamma(\hat{x}_{t,j})) - \frac{1}{2}\nabla^2_{x_t} \ln(\gamma(x_t)) \mid_{\hat{x}_{t,j}} \sigma^2_{t,j}(u_t, x_{t-1}) \\ &= -\ln(\gamma(\hat{x}_{t,j})) + \frac{1}{2}\nabla_{x_t}\lambda[x_t] \mid_{\hat{x}_{t,j}} \sigma^2_{t,j}(u_t, x_{t-1}), \end{aligned} \tag{17}$$

where $\lambda[x_t] = \nabla_{x_t}[-\ln(\gamma(x_t))]$. Substitute Equation (17) into Equation (15) yield,

$$-\ln(\gamma(x_{t-1})) = \int g_j(u_t \mid x_{t-1}) \Bigg[\underbrace{\delta(u_t, \hat{x}_{t,j}, x_{t-1}) + \ln\left(\frac{g_j(u_t \mid x_{t-1})}{{}^{I}g(u_t \mid x_{t-1})}\right)}_{\equiv \text{partial cost} \Longrightarrow U(u_t, \hat{x}_{t,j})}$$

$$- \underbrace{\ln(\gamma(\hat{x}_{t,j})) + \frac{1}{2}\nabla_{x_t}\lambda[x_t] \mid_{\hat{x}_{t,j}} \sigma^2_{t,j}(u_t, x_{t-1})}_{\text{expected optimal cost–to–go}} \Bigg] du_t. \tag{18}$$

Now introducing the following definition

$$\beta(u_t, \hat{x}_{t,j}, x_{t-1}) = \delta(u_t, \hat{x}_{t,j}, x_{t-1}) - \ln(\gamma(\hat{x}_{t,j})) + \frac{1}{2}\nabla_{x_t}\lambda[x_t] \mid_{\hat{x}_{t,j}} \sigma^2_{t,j}(u_t, x_{t-1}), \tag{19}$$

Equation (18) can be rewritten as

$$\begin{aligned} -\ln(\gamma(x_{t-1})) &= \int g_j(u_t|x_{t-1}) \left[\beta(u_t, \hat{x}_{t,j}, x_{t-1}) + \ln\left(\frac{g_j(u_t|x_{t-1})}{{}^{I}g(u_t|x_{t-1})}\right) \right] du_t \\ &= \int g_j(u_t|x_{t-1}) \ln\left(\frac{g_j(u_t|x_{t-1})}{{}^{I}g(u_t|x_{t-1}) \exp(-\beta(u_t, \hat{x}_{t,j}, x_{t-1}))}\right) du_t \\ &= \int g_j(u_t|x_{t-1}) \left[\ln\left(\frac{g_j(u_t|x_{t-1})}{\frac{{}^{I}g(u_t|x_{t-1}) \exp(-\beta(u_t, \hat{x}_{t,j}, x_{t-1}))}{\gamma(x_{t-1})}}\right) - \ln(\gamma(x_{t-1})) \right] du_t. \end{aligned} \tag{20}$$

By adding and subtracting $\ln(\gamma(x_{t-1}))$, the first term in Equation (20) has become a conditional version of the KL divergence. The independence of $\ln(\gamma(x_{t-1}))$ on the optimised $c(u_t|x_{t-1})$ implies that the expression is minimised by the claimed pdf (13). □

Remark 2: The occurrence of the modified expected cost–to–go $[-\ln(\gamma(\hat{x}_{t,j})) + \frac{1}{2}\nabla_{x_t}\lambda[x_t] \mid_{\hat{x}_{t,j}} \sigma^2_{t,j}(u_t, x_{t-1})]$ is the key modification of the standard FPD induced by the dependency of the parameters of the involved pdfs on the input values.

3. Generalised Probabilistic DHP Adaptive Critic Solution

3.1. Kullback–Leibler Performance Function

The development so far assumes no prior knowledge about known pdfs of the system states or control inputs. All required pdfs are assumed to be general density functions and are estimated using the MDN methods from process data as a weighted sum of mixture of Gaussians. The parameters of the required pdfs (means, variances and priors) are taken to be continuous functions of the input variables. This is important, as this dependency of the pdfs parameters on the input values in fact reflects the uncertainty involved in the estimation process and hence facilitates the development of the pragmatic method, discussed in Section 2.3 and further discussed and developed in the current section, for incorporating functional uncertainties in the derivation of the optimal control law.

The optimal control law can be obtained from Equation (13) by implementing the backward dynamic programming approach. However the backward dynamic programming approach is computationally very expensive and grows exponentially with the dimensionality of the state vector. As such, the method adopted here for calculating the optimal control law is based on approximating the derivative of the optimal cost–to–go function defined in Equation (18) using adaptive critic methods (Herzallah and Kárnaý, 2011). Analogues to the method developed in (Herzallah and Kárnaý, 2011), an adaptive critic network to approximate the derivative of the optimal cost–to–go function (18) with respect to the state, $\lambda^*[x_{t-1}] = \nabla_{x_{t-1}}[-\ln(\gamma(x_{t-1}))]$ will be used. An action neural network will also be used to approximate the obtained optimal randomised control inputs. This method is referred to as the generalized probabilistic DHP adaptive critic method.

In this section the generalized probabilistic DHP adaptive critic solution for the uncertain stochastic discrete systems defined in Equation (5) will be developed. The solution method will be derived for the general tracking control problem where the system states are required to follow desired state values. The regulation problem where the objective is to reach a zero state can be considered as a special case of the tracking problem where the desired value of the state is set to zero.

For the general tracking control problem, and following the discussion in Section 2.3, the pdf of the ideal states is taken to be Gaussian with mean equal to the desired state value and variance equal to the non–reducible variance of innovations of the active component,

$$ {}^{I}f(x_t \mid u_t, x_{t-1}) \propto \exp\left\{ -\frac{\| x_t - x_{t,d} \|^2}{2\sigma^2_{t,j}(x_{t-1}, u_t)} \right\}, \tag{21} $$

where $\sigma^n_{t,j}(x_{t-1}, u_t)$ is the variance of the active, j^{th}, component of the state mixture corresponding to the maximal α, and $x_{t,d}$ is the desired state value. This follows because the component of the state mixture capturing the system dynamics at time t is the j^{th} component with maximal α and so it only makes sense to use its innovations to embrace knowledge of uncertainty and use it in the derivation of the optimal control law. Similarly, the ideal pdf of the control input is defined as,

$$ {}^{I}g(u_t \mid x_{t-1}) \propto \exp\left\{ -\frac{\| u_t \|^2}{2\rho^2_{t,j}(x_{t-1})} \right\}, \tag{22} $$

which penalises large control inputs.

Using Equations (7), (10), (21), and (22) the first two terms in Equation (18) evaluate to

$$ \begin{aligned} \delta(u_t, \hat{x}_{t,j}, x_{t-1}) &= \int f_j(x_t \mid u_t, x_{t-1}) \frac{2\hat{x}_{t,j}(u_t, x_{t-1})x_t - \hat{x}^2_{t,j}(u_t, x_{t-1}) - 2x_t x_{t,d} + x^2_{t,d}}{2\sigma^2_{t,j}} dx_t \\ &= \frac{\{\hat{x}_{t,j}(u_t, x_{t-1}) - x_{t,d}\}^2}{2\sigma^2_{t,j}(u_t, x_{t-1})}, \end{aligned} \tag{23} $$

$$ \ln\left(\frac{g_j(u_t \mid x_{t-1})}{{}^{I}g(u_t \mid x_{t-1})} \right) = \frac{2\hat{u}_{t,j}(x_{t-1})u_t - \hat{u}^2_{t,j}(x_{t-1})}{2\rho^2_{t,j}}. \tag{24} $$

The optimal cost–to–go function defined by the recurrence Equation (18) can then be rewritten as,

$$-\ln(\gamma(x_{t-1})) = \int g_j(u_t \mid x_{t-1}) \Bigg[\underbrace{\frac{\{\hat{x}_{t,j}(u_t, x_{t-1}) - x_{t,d}\}^2}{2\sigma^2_{t,j}(u_t, x_{t-1})} + \frac{2\hat{u}_{t,j}(x_{t-1})u_t - \hat{u}^2_{t,j}(x_{t-1})}{2\rho^2_{t,j}}}_{\equiv \text{partial cost} \Longrightarrow U(u_t, \hat{x}_{t,j})}$$
$$- \underbrace{\Big\{ \ln(\gamma(\hat{x}_{t,j})) - \frac{1}{2}\nabla_{x_t}\lambda[x_t] \mid_{\hat{x}_{t,j}} \sigma^2_{t,j}(u_t, x_{t-1}) \Big\}}_{\text{expected optimal cost–to–go}} \Bigg] d(u_t). \tag{25}$$

This completes the formulation of the optimal cost to go function. The solution to the generalized probabilistic control using the DHP adaptive critic method can now be obtained by calculating the desired value of the critic network, $\lambda^*[x_{t-1}] = \nabla_{x_{t-1}}[-\ln(\gamma(x_{t-1}))]$ and the optimal control inputs. We start first by calculating the desired critic value. The solution method to calculating the optimal control inputs will be given in Section 3.3.

3.2. Desired Value of the Critic Network

Analogous to the fully probabilistic DHP adaptive critic approach, a critic network is used in this paper to approximate the derivative of the optimal cost–to–go function (25) with respect to the states,

$$\begin{aligned}
&\nabla_{x_{t-1}}[-\ln(\gamma(x_{t-1}))] = \int g_j(u_t \mid x_{t-1}) \Big[\nabla_{\hat{x}_{t,j}} U(u_t, \hat{x}_{t,j}) \nabla_{x_{t-1}} \hat{x}_{t,j} \\
&+ \nabla_{\hat{x}_{t,j}} U(u_t, \hat{x}_{t,j}) \nabla_{u_t} \hat{x}_{t,j} \nabla_{x_{t-1}} u_t + \nabla_{\sigma^2_{t,j}} U(u_t, \hat{x}_{t,j}) \nabla_{x_{t-1}} \sigma^2_{t,j} + \nabla_{\sigma^2_{t,j}} U(u_t, \hat{x}_{t,j}) \nabla_{u_t} \sigma^2_{t,j} \nabla_{x_{t-1}} u_t \\
&+ \nabla_{u_t} U(u_t, \hat{x}_{t,j}) \nabla_{x_{t-1}} u_t + \nabla_{\rho^2_{t,j}} U(u_t, \hat{x}_{t,j}) \nabla_{x_{t-1}} \rho^2_{t,j} + \lambda[\hat{x}_{t,j}] \nabla_{x_{t-1}} \hat{x}_{t,j} + \lambda[\hat{x}_{t,j}] \nabla_{u_t} \hat{x}_{t,j} \nabla_{x_{t-1}} u_t \\
&+ \frac{1}{2}\nabla^2_{\hat{x}_{t,j}} \lambda[\hat{x}_{t,j}] \nabla_{x_{t-1}} \hat{x}_{t,j} \sigma^2_{t,j} + \frac{1}{2}\nabla^2_{\hat{x}_{t,j}} \lambda[\hat{x}_{t,j}] \nabla_{u_t} \hat{x}_{t,j} \nabla_{x_{t-1}} u_t \sigma^2_{t,j} + \frac{1}{2}\nabla_{\hat{x}_{t,j}} \lambda[\hat{x}_{t,j}] \nabla_{x_{t-1}} \sigma^2_{t,j} \\
&+ \frac{1}{2}\nabla_{\hat{x}_{t,j}} \lambda[\hat{x}_{t,j}] \nabla_{u_t} \sigma^2_{t,j} \nabla_{x_{t-1}} u_t \Big] d(u_t).
\end{aligned} \tag{26}$$

The solution of this equation cannot be obtained analytically due to the involvement of the integration action and the nonlinear functions of $\nabla U(u_t, \hat{x}_{t,j})$ and $\lambda[\hat{x}_{t,j}]$. A possible approach would be to use Taylor expansion of the nonlinear functions around the mean value of the active component of control inputs and retaining terms up to the second order. For this purpose, denote

$$\begin{aligned}
&\psi(u_t, \hat{x}_{t,j}) = \nabla_{\hat{x}_{t,j}} U(u_t, \hat{x}_{t,j}) \nabla_{x_{t-1}} \hat{x}_{t,j} + \nabla_{\hat{x}_{t,j}} U(u_t, \hat{x}_{t,j}) \nabla_{u_t} \hat{x}_{t,j} \nabla_{x_{t-1}} u_t \\
&+ \nabla_{\sigma^2_{t,j}} U(u_t, \hat{x}_{t,j}) \nabla_{x_{t-1}} \sigma^2_{t,j} + \nabla_{\sigma^2_{t,j}} U(u_t, \hat{x}_{t,j}) \nabla_{u_t} \sigma^2_{t,j} \nabla_{x_{t-1}} u_t + \lambda[\hat{x}_{t,j}] \nabla_{x_{t-1}} \hat{x}_{t,j} \\
&+ \lambda[\hat{x}_{t,j}] \nabla_{u_t} \hat{x}_{t,j} \nabla_{x_{t-1}} u_t + \frac{1}{2}\nabla^2_{\hat{x}_{t,j}} \lambda[\hat{x}_{t,j}] \nabla_{x_{t-1}} \hat{x}_{t,j} \sigma^2_{t,j} + \frac{1}{2}\nabla^2_{\hat{x}^2_t} \lambda[\hat{x}_{t,j}] \nabla_{u_t} \hat{x}_{t,j} \nabla_{x_{t-1}} u_t \sigma^2_{t,j} \\
&+ \frac{1}{2}\nabla_{\hat{x}_{t,j}} \lambda[\hat{x}_{t,j}] \nabla_{x_{t-1}} \sigma^2_{t,j} + \frac{1}{2}\nabla_{\hat{x}_{t,j}} \lambda[\hat{x}_{t,j}] \nabla_{u_t} \sigma^2_{t,j} \nabla_{x_{t-1}} u_t.
\end{aligned} \tag{27}$$

It can be seen that $\psi(u_t, \hat{x}_{t,j})$ is continuous because $\nabla U(u_t, \hat{x}_{t,j})$ and $\lambda[\hat{x}_{t,j}]$ are continuous and it is differentiable with respect to u_t. As a result by Taylor expanding $\psi(u_t, \hat{x}_{t,j})$ around the mean value of the active component of control inputs we get

$$\psi(u_t, \hat{x}_{t,j}) = \psi(\hat{u}_{t,j}, \hat{x}_{t,j}) + \nabla_{u_t}\psi(u_t, \hat{x}_{t,j}) \mid_{\hat{u}_{t,j}} [u_t - \hat{u}_{t,j}] + \frac{1}{2}\nabla^2_{u_t}\psi(u_t, \hat{x}_{t,j}) \mid_{\hat{u}_{t,j}} [u_t - \hat{u}_{t,j}]^2. \tag{28}$$

By substituting Equation (28) into Equation (26), the derivative of the cost function with respect to the state, $\lambda^*[x_{t-1}]$ can be calculated as follows

$$\lambda^*[x_{t-1}] = \nabla_{x_{t-1}}[-\ln(\gamma(x_{t-1}))] = \int g_j(u_t \mid x_{t-1}) \Big[\psi(\hat{u}_{t,j}, \hat{x}_{t,j}) + \nabla_{u_t}\psi(u_t, \hat{x}_{t,j}) \mid_{\hat{u}_{t,j}} [u_t - \hat{u}_{t,j}]$$
$$+\frac{1}{2}\nabla^2_{u_t}\psi(u_t, \hat{x}_{t,j}) \mid_{\hat{u}_{t,j}} [u_t - \hat{u}_{t,j}]^2 + \nabla_{u_t} U(u_t, \hat{x}_{t,j})\nabla_{x_{t-1}} u_t + \nabla_{\rho^2_{t,j}} U(u_t, \hat{x}_{t,j})\nabla_{x_{t-1}}\rho^2_{t,j}\Big] d(u_t)$$
$$= \frac{\hat{u}_{t,j}}{\rho^2_{t,j}}\nabla_{x_{t-1}} u_t + \frac{\hat{u}^2_{t,j}}{2\rho^4_{t,j}}\nabla_{x_{t-1}}\rho^2_{t,j} + \psi(\hat{u}_{t,j}, \hat{x}_{t,j}) + \frac{1}{2}\nabla^2_{u_t}\psi(u_t, \hat{x}_{t,j}) \mid_{\hat{u}_{t,j}} \rho^2_{t,j}. \tag{29}$$

The parameters of the critic network can then be adapted such that the error between the desired value of the critic given in Equation (29) and estimated value of the critic network is minimised.

3.3. Generalised Controller Design

The generalised probabilistic DHP adaptive critic method computes the optimal control inputs by deriving Equation (25) with respect to the control inputs and setting the derivative equal to zero. This results in the following optimality equation for computing the generalised probabilistic control values:

$$\nabla_{u_t}[-\ln(\gamma(x_{t-1}))] = \int g_j(u_t \mid x_{t-1}) \Big[\nabla_{\hat{x}_{t,j}} U(u_t, \hat{x}_{t,j})\nabla_{u_t}\hat{x}_{t,j} + \nabla_{\sigma^2_{t,j}} U(u_t, \hat{x}_{t,j})\nabla_{u_t}\sigma^2_{t,j}$$
$$+\nabla_{u_t} U(u_t, \hat{x}_{t,j}) + \lambda[\hat{x}_{t,j}]\nabla_{u_t}\hat{x}_{t,j} + \frac{1}{2}\nabla^2_{\hat{x}_{t,j}}\lambda[\hat{x}_{t,j}]\nabla_{u_t}\hat{x}_{t,j}\sigma^2_{t,j}$$
$$+\frac{1}{2}\nabla_{\hat{x}_{t,j}}\lambda[\hat{x}_{t,j}]\nabla_{u_t}\sigma^2_{t,j}\Big] d(u_t) = 0. \tag{30}$$

Similarly the solution of this equation cannot be obtained analytically due to the involvement of the integration action and the nonlinear functions of $\nabla U(u_t, \hat{x}_{t,j})$ and $\lambda[\hat{x}_{t,j}]$. Hence to facilitate the solution of Equation (30), Taylor expansion of the nonlinear functions around the mean value of the active component of control inputs is considered. For this purpose, denote

$$\eta(u_t, \hat{x}_{t,j}) = \nabla_{\hat{x}_{t,j}} U(u_t, \hat{x}_{t,j})\nabla_{u_t}\hat{x}_{t,j} + \nabla_{\sigma^2_{t,j}} U(u_t, \hat{x}_{t,j})\nabla_{u_t}\sigma^2_{t,j} + \lambda[\hat{x}_{t,j}]\nabla_{u_t}\hat{x}_{t,j}$$
$$+\frac{1}{2}\nabla^2_{\hat{x}_{t,j}}\lambda[\hat{x}_{t,j}]\nabla_{u_t}\hat{x}_{t,j}\sigma^2_{t,j} + \frac{1}{2}\nabla_{\hat{x}_{t,j}}\lambda[\hat{x}_{t,j}]\nabla_{u_t}\sigma^2_{t,j}. \tag{31}$$

It can be seen that $\eta(u_t, \hat{x}_{t,j})$ is continuous because $\nabla U(u_t, \hat{x}_{t,j})$ and $\lambda[\hat{x}_{t,j}]$ are continuous and it is differentiable with respect to u_t. As a result by Taylor expanding $\eta(u_t, \hat{x}_{t,j})$ around the mean value of the active component of control inputs we get

$$\eta(u_t, \hat{x}_{t,j}) = \eta(\hat{u}_{t,j}, \hat{x}_{t,j}) + \nabla_{u_t}\eta(u_t, \hat{x}_{t,j}) \mid_{\hat{u}_{t,j}} [u_t - \hat{u}_{t,j}] + \frac{1}{2}\nabla^2_{u_t}\eta(u_t, \hat{x}_{t,j}) \mid_{\hat{u}_{t,j}} [u_t - \hat{u}_{t,j}]^2. \tag{32}$$

By substituting Equation (32) into Equation (30), the optimality equation can be further formulated as

$$\int g_j(u_t \mid x_{t-1}) \Big[\eta(\hat{u}_{t,j}, \hat{x}_{t,j}) + \nabla_{u_t}\eta(u_t, \hat{x}_{t,j}) \mid_{\hat{u}_{t,j}} [u_t - \hat{u}_{t,j}] + \frac{1}{2}\nabla^2_{u_t}\eta(u_t, \hat{x}_{t,j}) \mid_{\hat{u}_{t,j}} [u_t - \hat{u}_{t,j}]^2$$
$$+\frac{\hat{u}_{t,j}}{\rho^2_{t,j}}\Big] d(u_t) = 0. \tag{33}$$

The integral (33) is easily evaluated to give,

$$\eta(\hat{u}_{t,j}, \hat{x}_{t,j}) + \frac{1}{2}\nabla^2_{u_t}\eta(u_t, \hat{x}_{t,j}) \mid_{\hat{u}_{t,j}} \rho^2_{t,j} + \frac{\hat{u}_{t,j}}{\rho^2_{t,j}} = 0. \tag{34}$$

This equation can then be used to calculate optimal control inputs taking into consideration functional uncertainty in this calculation. The control input obtained in Equation (34) can then be used as the target value for estimating the conditional pdf of control inputs as specified by Equation (9). The controller can then generate control inputs u_t stochastically from its pdf as estimated by the MDN by taking the most probable control input for example (Herzallah, 2012).

4. Generalised Nonlinear Probabilistic Control Algorithm

The optimisation method of the generalised DHP adaptive critic method is a two stage process. The first stage is the optimisation of the critic network that predicts the desired critic value as specified by Equation (29). As can be seen from this equation the desired critic value can be calculated from the density model of the states of the system, density model of nonlinear controller, and the critic network model. The second stage is the optimisation of the control input network. The control input can be calculated using Equation (34) once the critic network and system state density models become available. The implementation of this two stage optimisation method can be performed efficiently by utilising the modular approach constituting of functional modules and algorithmic modules (Ferrari and Stengel, 2004; Lendaris et al., 2002; Herzallah, 2007). The main functional modules are the action and critic networks. Algorithmic modules on the other hand include the computation of the desired critic value (29), computation of the optimal control law (34), and the update of the networks' parameters. Each of these modules can be modified independently from other modules, thus facilitating fast and reliable implementation.

The description below is appropriate for direct application to nonlinear stochastic control problems involving functional uncertainties of the form stated in Section 2.1.

1. Estimate the general pdf of the system states described by Equation (5) as discussed in Section 2.2.
2. Design and initialise the weights of the density network of control input.
3. Design and initialise the weights of the critic network.
4. Approximate the state and control density functions by Gaussian densities corresponding to maximal α and ϕ respectively.
5. Specify the ideal density functions of both the state and control inputs.
6. Calculate the desired value of the critic network using Equation (29).
7. 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.
8. Use the output of the converged critic in Equation (34) and solve for the optimal control input.
9. Use this value to update the parameters of the control input density network.
10. Repeat Steps $4-9$ until an acceptable performance is reached.

To summarise, the generalised 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 are approximated by the MDN controller and critic network respectively. The algorithm terminates when an acceptable performance is achieved.

5. Simulation Example

This section demonstrates the generalised probabilistic control on a nonlinear stochastic control system described by the following dynamical equation

$$x_t = (0.5 - 0.02\epsilon_t(1.2 + \tan^{-1}u_t))x_{t-1} + 0.2u_t, \tag{35}$$

where ϵ_t denotes a noise sequence sampled from a mixture of Gaussians with the following mean, μ_{ϵ_t}, and covariance Σ_{ϵ_t}

$$\mu_{\epsilon_t} = \begin{bmatrix} 0 & 2 \end{bmatrix}, \quad \Sigma_{\epsilon_t} = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.1 \end{bmatrix}.$$

The particular system considered here was used in (Yue and Wang, 2003), to illustrate theoretical developments for minimum entropy control.

The control objective is to make the system state, x_t follows a desired state value, $x_{t,d} = 1$. Thus, the generalised probabilistic DHP critic that minimises the cost function (25) is used to design the generalised probabilistic control input. Following the steps in Section 4, the pdf of the system state described by Equation (35) is firstly estimated using MDN as discussed in Section 2.2 and given by Equation (7). As Equation (7) states, the prior probabilities, the centre and noise variances of this state pdf are estimated to be continuous functions of the input variables (x_{t-1}, u_t). Accordingly, the estimated pdf of the system state identifies the input dependent noise of the system states. The MDN of the state pdf has two inputs, x_{t-1} and u_t; and one output x_t. Using the cross validation method the best optimal structure for estimating the pdf of the system state is found to be an MDN with one kernel function and seven hidden neurons. The weight parameters of the pdf of the system state are then kept fixed during the critic and control input network training.

The control is then initiated by another MDN with one kernel function and seven hidden neurons. Similar to the pdf of the system states, the parameters of the pdf of controller MDN are also estimated to be continuous functions of the input variable x_{t-1}. The dependency of these parameters of the MDN of the system states (in particular of the noise variances of the pdf of the system states) and the controller MDN (represented by the dependency of the noise variances of the controller pdf) are taken into consideration in the derivation of the optimal control law as demonstrated by Equations (29) and (34). The control input network has two inputs x_{t-1} and $x_{t,d}$ and one output, u_t. A rough initialisation for the parameters of the control input MDN network was obtained using an off–line training method. Next, the critic network is also taken to be an MDN with one kernel function and nine hidden neurons. The critic network has two inputs x_{t-1} and $x_{t-1,d}$ and one output $\nabla_{x_{t-1}}[-\ln(\gamma(.))]$. The parameters of the critic mixture density network are initialised randomly from a zero mean, isotropic Gaussian, with unit variance scaled by the fan–in of the output units.

Training of both critic and control input networks is carried out on–line using the desired state value, $x_{t,d} = 1$ as given above. The target values of the critic network is then calculated using Equation (29) and the critic training is performed using scaled conjugate gradient method. During training of the critic network, the weights of the control input network are kept fixed. The output from the trained critic is then used in Equation (34) solving for optimal control values and the control network is then trained using same training method as that of the critic network. During training of the control input network, the weights of the critic network are kept fixed. After the control input network trained, the critic network is trained again (by adapting the weights of the previously trained critic) using the outputs of the trained action network.

The control quality of the controller designed in the above manner is then compared with the standard probabilistic DHP adaptive critic technique (Herzallah and Kárnaý, 2011). Here all of the state, control input and critic networks are taken to be standard neural networks. For fair comparison, same noise sequence, initial conditions, and desired state training signal were used during the implementation of each control method.

The system state and desired state and the control input as obtained from the standard and generalised probabilistic DHP adaptive critic controllers are shown in Figure 2. The figure shows that a large transient error results from the standard probabilistic DHP adaptive critic method. This is expected since in the standard probabilistic critic method all of the estimated models of the system state, control input and critic networks are estimated as deterministic models using standard neural networks. This means models uncertainty is not considered in calculating the optimal control input. The proposed generalised probabilistic adaptive critic method, on the other hand, shows no overshoot in the transient period and ensures minimal overshoot since models uncertainty, represented by the dependency of the noise variances of the states and controller pdfs, is taken into consideration in the obtained control input.

6. Conclusions

This paper formalised the design of generalised probabilistic control for stochastic nonlinear and uncertain systems. An MDN is used to estimate the general pdf of the system dynamics representing the actual system response and the results are used to determine subsequent control inputs. In this probabilistic formalism the control objective is to design a generalised randomised control input that controls the joint pdf of the closed loop behavior of the controlled system and makes it as close as possible to an ideal joint pdf. A Kullback–Leibler divergence is used to measure the

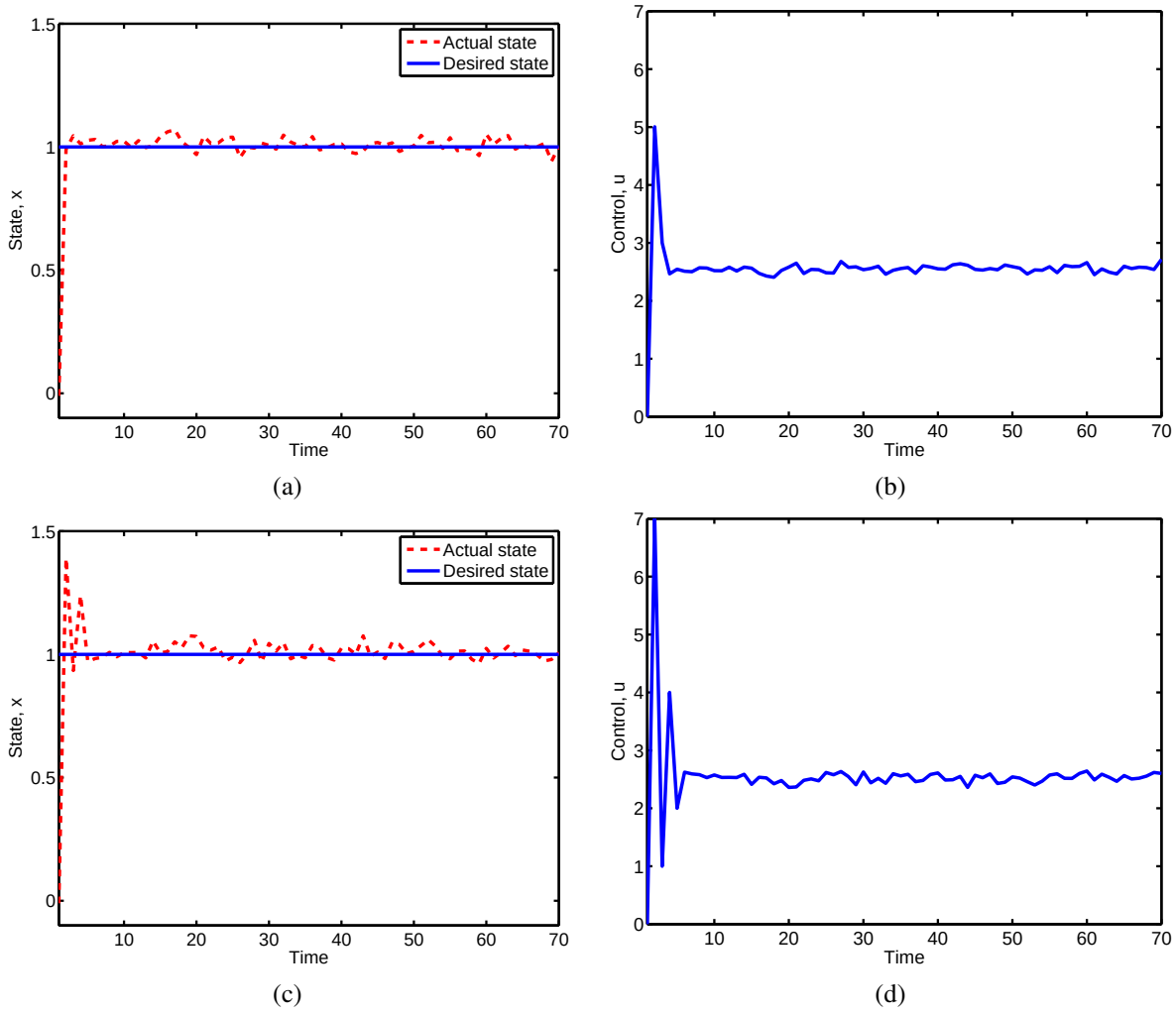

Figure 2: Nonlinear stochastic system: (a) Actual and desired state values from the generalised probabilistic DHP adaptive critic (with input-dependent noise), Section 3 (b) Generalised probabilistic control inputs from the generalised DHP adaptive critic (with input-dependent noise) (c) Actual and desired state values from the standard probabilistic DHP adaptive critic (with input-independent noise), Section 1 (Herzallah and Kárnaý, 2011) (d) Probabilistic control inputs from the standard probabilistic DHP adaptive critic (with input-independent noise).

discrepancy between the actual and the ideal probability density functions. Generalised optimal probabilistic control laws are developed using dynamic programming and adaptive critic methods. Adaptive critic methods reduce computational requirements of the fully probabilistic design control problem and allow the development of more robust and practically implementable control.

Since the pdf of the system states is a general pdf and is estimated using MDNs as a weighted sum of mixture of Gaussians, the pdf of the randomised controller is acquired from the general pdf of the system states. The randomised controller is then obtained by dividing up the control into appropriate regions which can be recombined to generate the desired system response. This development can, therefore, learn the multiple control input models as well as how to select the randomised control input for a given operating condition. Moreover, the proposed formulation of the FPD in an adaptive critic framework and MDNs for estimating the general pdfs of the system response and randomised control inputs, allows exploitation of functional uncertainty when deriving the optimal control law. This is accomplished by utilising the dependency of the probability density functions of the system models on the input values as estimated from the MDN. The proposed generalised probabilistic DHP adaptive critic method has demonstrated that when the density function parameters are dependent on the input values, not only the expected value of the Kullback–Leibler distance that should be minimised but also the variance of its cost function.

A nonlinear stochastic dynamical system has been numerically solved to demonstrate the efficiency of the proposed method. An on–line adaptation for the generalised probabilistic DHP critic is considered and compared to that of the standard probabilistic DHP critic. The proposed generalised probabilistic controller improves the performance of the controlled system and ensures minimal overshoot.

Finally to re–emphasize, the proposed generalised probabilistic critic method derives optimal control inputs such that the distance between the joint pdf of the closed loop system and an ideal joint pdf is minimised. Moreover, the parameters of the joint pdf are estimated such that they are dependent on the state and control input values. This accounts for the systems' uncertainty and improves the performance of the derived generalised probabilistic optimal control law.

Anderson, B. D. O., Moore, J. B., 1971. Linear Optimal Control. Prentice Hall, Englewood Cliffs, NJ.

Astrom, K. J., 1970. Introduction to Stochastic Control Theory. New York: Academic.

Barndorff-Nielsen, O., 1978. Information and exponential families in statistical theory. Wiley, New York.

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

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

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

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

Ferrari, S., Stengel, R. F., 2004. Model based adaptive critic designs. In: Si, J., Barto, A. G., Powell, W. B., Wunsch, D. (Eds.), Handbook of Learning and Approximate Dynamic Programming. Institute of Electrical and Electronics Engineers, Inc, Canada, Ch. 3, pp. 64–94.

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

Herzallah, R., 2012. Probabilistic control for uncertain systems. Dynamic Systems, Measurement, and Control 134 (2), 021018.

Herzallah, R., Kárnaý, M., 2011. Fully probabilistic control design in an adaptive critic framework. Neural Networks 24 (11), 1128–1135.

Herzallah, R., Lowe, D., 2004. A mixture density network approach to modelling and exploiting uncertainty in nonlinear control problems. Engineering Applications of Artificial Intelligence 17, 145–158.

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

Karniel, A., Meir, R., Inbar, G. F., 2001. Polyhedral mixture of linear experts for many-to-one mapping inversion and multiple controllers. Neurocomputing 37, 31–49.

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

Kárný, M., Böhm, J., Guy, T. V., Nedoma, P., 2003. Mixture-based adaptive probabilistic control. International Journal of Adaptive Control and Signal Processing 17, 1–13.

Kárný, M., Guy, T. V., 2006. Fully probabilistic control design. Systems & Control Letters 55 (4), 259–265.

Lendaris, G. G., Santiago, R. A., Carrol, M. S., 2002. Proposed framework for applying adaptive critics in real−time realm. In: Proceedings of the 2002 International Joint Conference on Neural Networks, IJCNN'02. Honolulu, HI , USA, pp. 1796–1801.

Murray-Smith, R., (Eds.), T. A. J., 1997. Multiple Model Approaches to Modelling and Control. Taylor and Francis.

Narendra, K., Driolet, O., 2001. Stochastic adaptive control using multiple models for improved performance in the presence of random disturbances. International Journal of Adaptive Control and Signal Processing 15, 287–318.

Niu, Y., Ho, D. W., Wang, X., 2008. Robust H_∞ control for nonlinear stochastic systems: A sliding-mode approach. IEEE Transactions on Automatic Control 53 (7), 1695–1701.

Park, J., Sandberg, I., 1991. Universal approximation using radial basis function networks. Neural Computation 3, 246–257.

Peterka, V., 1981. Bayesian system identification. In: Eykhoff, P. (Ed.), Trends and Progress in System Identification. Pergamon Press, Oxford, pp. 239–304.

Rao, C., 1987. Linear method of statistical inference and their applications. Academia, Prague, in Czech.

Smídl, V. ., Quinn, A., Kárný, M., Guy, T. V., 2005. Robust estimation of autoregressive processes using a mixture-based filter-bank. Systems and Control Letters 54, 315–323.

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

Titterington, D., Smith, A., Makov, U., 1985. Statistical Analysis of Finite Mixtures. John Wiley and Sons.

Wang, G., Qian, L., Guo, Z., 2013. Continuous tool wear prediction based on gaussian mixture regression model. International Journal of Advanced Manufacturing Technology 66, 1921–1929.

Wang, H., Afshar, P., 2009. 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.

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

Yue, H., Wang, H., 2003. Minimum entropy control of closed-loop tracking errors for dynamic stochastic systems. IEEE Transactions on Automatic Control 48 (1), 118–122.