Neural Networks 24 (2011) 1128–1135

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

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

[b] Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Czech Republic

ARTICLE INFO

Article history:
Received 12 November 2010
Received in revised form 5 May 2011
Accepted 13 June 2011

Keywords:
Stochastic control design
Fully probabilistic design
Adaptive control
Adaptive critic

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 paper.

1. Introduction

Stochastic control design minimizes an expected cost function with respect to feedback control strategies, e.g. Astrom (1970) and Bertsekas (2001). 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 (Guy & Kárný, 2005; Kárný, 1996; Kárný & Guy, 2006) 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 function. 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

* Corresponding author. Tel.: +962 796734908.
E-mail addresses: herzallah.r@gmail.com (R. Herzallah), school@utia.cas.cz (M. Kárný).

0893-6080/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2011.06.006

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

$$\mathcal{D}\left(f \parallel {}^I f\right) \equiv \int f(D) \ln\left(\frac{f(D)}{{}^I f(D)}\right) \mathrm{d}D. \tag{4}$$

The KLD in (4) has the following key property

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

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 (Peterka, 1981) allows its factorization 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 reemphasize, 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 factorized 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)

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

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

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

Full derivation of (8) is given in the Appendix. Eq. (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 the 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 (Lin, 2011; Lin & Yang, 2008; Liu, Javaherian, Kovalenko, & Huang, 2008; Prokhorov, Santiago, & Wunsch, 1995; Prokhorov & Wunsch, 1997; Si, Barto, Powell, & Wunsch, 2004). 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 (Balakrishnan & Biega, 1996; Han & Balakrishnan, 2002; Kulkarni & KrishnaKumar, 2003). 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 (Werbos, 1992). The critic network is trained forward in time, which reduces computational time and storage requirements in real time control applications. This also