OPTIMAL CONTROL APPLICATIONS AND METHODS
Optim. Control Appl. Meth. 2005; **26**:55–83
Published online 29 December 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/oca.750

A new method for suboptimal control of a class of non-linear systems

Ming Xin[‡,§] and S. N. Balakrishnan[*,†,¶]

Department of Mechanical and Aerospace Engineering, University of Missouri-Rolla, 1870 Miner Circle, Rolla, MO 65409-1350, U.S.A.

SUMMARY

In this paper, a new non-linear control synthesis technique (θ–D approximation) is discussed. This approach achieves suboptimal solutions to a class of non-linear optimal control problems characterized by a quadratic cost function and a plant model that is affine in control. An approximate solution to the Hamilton–Jacobi–Bellman (HJB) equation is sought by adding perturbations to the cost function. By manipulating the perturbation terms both semi-global asymptotic stability and suboptimality properties are obtained. The new technique overcomes the large-control-for-large-initial-states problem that occurs in some other Taylor series expansion based methods. Also this method does not require excessive online computations like the recently popular state dependent Riccati equation (SDRE) technique. Furthermore, it provides a closed-form non-linear feedback controller if finite number of terms are taken in the series expansion. A scalar problem and a 2-D benchmark problem are investigated to demonstrate the effectiveness of this new technique. Both stability and convergence proofs are given. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: non-linear systems; optimal control; perturbation methods

1. INTRODUCTION

Numerous techniques exist for the synthesis of control laws for non-linear systems. Optimal control of non-linear dynamics with respect to a mathematical index of performance has also been extensively investigated in the last few decades. One of the difficulties in obtaining optimal solutions for non-linear systems is that optimal feedback control depends on the solution to the Hamilton–Jacobi–Bellman (HJB) equation [1]. The HJB equation is extremely difficult to solve

*Correspondence to: S. N. Balakrishnan, Mechanical and Aerospace Engineering, University of Missouri-Rolla, Rolla, MO 65409-0050, U.S.A.
†E-mail: bala@umr.edu
‡E-mail: xin@umr.edu
§Postdoctoral Research Fellow.
¶Professor.

Contract/grant sponsor: Anteon Corporation

Received 22 October 2003
Revised 20 September 2004

in general rendering optimal control techniques of limited use for non-linear systems. Consequently, a number of papers investigated methods to find suboptimal solutions to non-linear control problems.

One such technique is the power series expansion based method. Al'brekht [2] and Lukes [3] have derived a sufficient condition for obtaining optimal feedback control of a non-linear analytic system and developed a formal recursive procedure to construct a suboptimal control scheme as a power series in states. However, a closed form solution for the recursive procedure was not given. Garrard *et al.* [4] extended the above idea by expanding the optimal cost function as a power series in terms of an artificial variable ε and obtained a recursive solution to the HJB equation. Their modifications simplified the calculations, but the updated technique can only be applied to a certain class of systems in which the non-linearity can be considered as small perturbations. Nishkawa *et al.* [5] proposed a method to determine the coefficients of a series expansion. But the convergence of the series is not guaranteed when the non-linearity is large. In addition, higher-order approximations do not necessarily give better results. Wernli and Cook [6] developed an approach by bringing the original non-linear system into an apparent linearization form. Their suboptimal control involves finding the Taylor series expansion of the solution to a state dependent Riccati equation. But the convergence of this series is not guaranteed and the resulting control law leads to a large control effort when the initial states are large. Garrard [7,8] also formulated another approach that expanded both optimal cost and non-linear dynamics as a power series of the states and used the same idea as before. This is applicable to a wider class of non-linear systems. A drawback of their method is that one has to assume the structure of the optimal cost as a scalar polynomial with undetermined coefficients which contains all possible combinations of products of the elements of the state vector. Therefore, as the system order increases, the complexity of determining these coefficients increases dramatically. Zhang *et al.* [9] and Krikelis *et al.* [10] proposed a better way for calculating the coefficients of the series solution given by Garrard. But it can only guarantee the stability around the origin. The common problem with these methods is that they do not offer asymptotic stability in the large.

Another recently emerging technique that systematically solves the non-linear regulator problem is the state dependent Riccati equation (SDRE) method [11]. By turning the equations of motion into a linear-like structure, this approach permits the designer to employ linear optimal control methods such as the LQR methodology and the H_∞ design technique for the synthesis of non-linear control systems. The SDRE method however, needs online computation of the algebraic Riccati equation at *each sample time*.

The idea of inverse optimal design for non-linear systems has been investigated in many papers [12–17]. Since solving HJB equation is generally very hard for a given plant and cost functional, inverse optimal approach starts with a stabilizing controller for a given plant and then find a meaningful cost functional for which this controller is optimal. Saridis *et al.* [12] developed a technique from this point of view. Given an arbitrarily selected admissible feedback control, a recursive algorithm solving the generalized HJB (GHJB) equation was proposed for sequential improvement of the control law that converges to the optimal solution. Saridis and Wang [13] also extended this theory to stochastic non-linear systems and proposed design procedures using upper and lower bounds of the cost function. But finding an appropriate value function to the GHJB equation is still a very difficult task. Beard *et al.* [14] adopted a Galerkin approximation to solve the GHJB equation. It gave the solution to the above problem. However, to find an admissible control to satisfy all the 10 conditions proposed in this paper is not

an easy task. Pan *et al.* [16] combined the inverse optimal control with the backstepping design to solve the non-linear H_∞ optimal control problem. They employed a concept called non-linear Cholesky factorization for the cost function which can be used in the backstepping design to yield the optimal control. However, the construction of the approximation function is not unique. How to judge different selections is still open. Margaliot and Langholz [17] derived some conditions using Young's inequality under which the closed form solutions to the HJB equation can be achieved. This method can construct a large family of explicitly solvable optimal control problems by using various class K functions (continuous and strictly increasing).

In this paper, we attempt a direct solution to a class of optimal control problems. A new non-linear controller synthesis (θ–D approximation) technique based on approximate solution to the HJB equation is proposed. This approach achieves suboptimal solutions to a class of non-linear optimal control problems characterized by a quadratic cost function and a plant model that is affine in control. By introducing an intermediate variable θ, the optimal cost can be expanded as a power series in terms of θ. The HJB equation is then reduced to a set of recursive algebraic equations and yields a closed-form controller with a finite number of terms. By adding perturbations to the cost function and manipulating them, we are able to ensure convergence of the series and achieve semi-global asymptotic stability. In addition, this technique can overcome the problem of large-control-for-large-initial-states encountered by some other power series expansion based control laws [6]. Tuning the parameters in the perturbation terms also enables us to modulate the system transient performance in a flexible way. The formulation of θ–D approximation method is presented in Section 2. In Section 3 a scalar benchmark problem is studied and a 2-D non-linear regulator problem is investigated by comparing the θ–D technique with the SDRE approach. Conclusions are given in Section 4. Proof of convergence of the cost function expansion and bounds on the error in cost are presented in the appendix.

2. SUBOPTIMAL CONTROL OF A CLASS OF NON-LINEAR SYSTEMS

2.1. Problem statement

In this paper we restrict ourselves to the state feedback control problem for the class of non-linear time-invariant systems described by

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{g}\mathbf{u} \tag{1}$$

where $\mathbf{x} \in \Omega \subset R^n$, $\mathbf{f} \in R^n$, $\mathbf{g} \in R^{n\times m}$, $\mathbf{u} \in U \subset R^m$; $\mathbf{g}$ is a constant matrix.

The objective is to find a controller that minimizes the quadratic cost function, J given by

$$J = \frac{1}{2}\int_0^\infty (\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u})\,\mathrm{d}t \tag{2}$$

where $\mathbf{Q} \in R^{n\times n}$, $\mathbf{R} \in R^{m\times m}$. $\mathbf{Q}$ is a semi-positive definite matrix and $\mathbf{R}$ is a positive definite matrix. To ensure that the optimal control problem is well posed, we make the following assumptions:

(C1) $\mathbf{x} \in \Omega$ and Ω is a compact set including origin as an interior point; $\mathbf{u} \in U$ and U is a compact set.

(C2) $\mathbf{f}(\mathbf{x})$ is continuously differentiable on Ω and $\mathbf{f}(\mathbf{0})=\mathbf{0}$.
(C3) The system (1) is controllable over the compact set Ω.
(C4) The system (1) is zero-state observable through $\boldsymbol{Q}$.

In Lukes [3], it was shown that with these conditions and a quadratic cost function (2), the optimal control problem admits a unique solution near the origin and the solution can be obtained through a continuously differentiable optimal cost satisfying the HJB partial differential equation [1]:

$$V_{\mathbf{x}}^{\mathrm{T}}\mathbf{f}(\mathbf{x}) - \tfrac{1}{2}\, V_{\mathbf{x}}^{\mathrm{T}}\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}V_{\mathbf{x}} + \tfrac{1}{2}\,\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} = 0 \tag{3}$$

where $V_{\mathbf{x}} = \partial V(\mathbf{x})/\partial \mathbf{x}$ and $V(\mathbf{x})$ is the optimal cost, i.e.

$$V(\mathbf{x}) = \min_{\mathbf{u}} \frac{1}{2}\int_0^{\infty} (\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u})\,\mathrm{d}t \tag{4}$$

with $V(\mathbf{x})>0$ and $V(\mathbf{0})=\mathbf{0}$.

The necessary condition for optimality leads to

$$\mathbf{u} = -\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}V_{\mathbf{x}} \tag{5}$$

Note that the controller needs $V_{\mathbf{x}}$. However, the HJB equation is extremely difficult to solve in general, rendering optimal control techniques of limited use for non-linear systems. In the next section, the θ–D suboptimal control formulation to find an approximate solution to the HJB equation is presented.

2.2. θ–D Suboptimal control development

The approximation process is carried out by the addition of a vanishing power series to the cost function and assuming a power series solution to the gradient of the optimal cost V.

Now consider perturbations $\sum_{i=1}^{\infty}\mathbf{D}_i\theta^i$ added to the cost function:

$$J = \frac{1}{2}\int_0^{\infty}\left[\mathbf{x}^{\mathrm{T}}\left(\mathbf{Q} + \sum_{i=1}^{\infty}\mathbf{D}_i\theta^i\right)\mathbf{x} + \mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u}\right]\mathrm{d}t \tag{6}$$

where θ is a scalar and $\mathbf{D}_i$ is a matrix. θ and $\mathbf{D}_i$ are chosen such that $\mathbf{Q} + \sum_{i=1}^{\infty}\mathbf{D}_i\theta^i$ is semi-positive definite.

Write the original state equation as

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{g}\mathbf{u} = \left\{\mathbf{A}_0 + \theta\left[\frac{\mathbf{A}(\mathbf{x})}{\theta}\right]\right\}\mathbf{x} + \mathbf{g}\mathbf{u} \tag{7}$$

where $\mathbf{A}_0$ is a constant matrix such that $(\mathbf{A}_0, \mathbf{g})$ is a stabilizable pair and $[\mathbf{A}_0 + \mathbf{A}(\mathbf{x}), \mathbf{g}]$ is pointwise controllable.

Define

$$\boldsymbol{\lambda} = V_{\mathbf{x}} \tag{8}$$

By using (8) and (6) in HJB equation (3) we have the perturbed HJB equation:

$$\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{f}(\mathbf{x}) - \frac{1}{2}\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\boldsymbol{\lambda} + \frac{1}{2}\mathbf{x}^{\mathrm{T}}\left(\mathbf{Q} + \sum_{i=1}^{\infty}\mathbf{D}_i\theta^i\right)\mathbf{x} = 0 \tag{9}$$

Assume a power series expansion of $\boldsymbol{\lambda}$ in terms of θ

$$\boldsymbol{\lambda} = \sum_{i=0}^{\infty}\mathbf{T}_i\theta^i\mathbf{x} \tag{10}$$

where $\mathbf{T}_i$ are to be determined and assumed to be symmetric.

Substitute (10) into the perturbed HJB equation (9) and equate the coefficients of powers of θ to zero to get the following equations:

$$\mathbf{T}_0\mathbf{A}_0 + \mathbf{A}_0^{\mathrm{T}}\mathbf{T}_0 - \mathbf{T}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_0 + \mathbf{Q} = \mathbf{0} \tag{11}$$

$$\mathbf{T}_1(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_0) + (\mathbf{A}_0^{\mathrm{T}} - \mathbf{T}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})\mathbf{T}_1 = -\frac{\mathbf{T}_0\mathbf{A}(\mathbf{x})}{\theta} - \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{x})\mathbf{T}_0}{\theta} - \mathbf{D}_1 \tag{12}$$

$$\mathbf{T}_2(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_0) + (\mathbf{A}_0^{\mathrm{T}} - \mathbf{T}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})\mathbf{T}_2 = -\frac{\mathbf{T}_1\mathbf{A}(\mathbf{x})}{\theta} - \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{x})\mathbf{T}_1}{\theta} + \mathbf{T}_1\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_1 - \mathbf{D}_2 \tag{13}$$

$$\vdots$$

$$\mathbf{T}_n(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_0) + (\mathbf{A}_0^{\mathrm{T}} - \mathbf{T}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})\mathbf{T}_n = -\frac{\mathbf{T}_{n-1}\mathbf{A}(\mathbf{x})}{\theta} - \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{x})\mathbf{T}_{n-1}}{\theta} + \sum_{j=1}^{n-1}\mathbf{T}_j\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_{n-j} - \mathbf{D}_n \tag{14}$$

Since the right-hand side of Equations (11)–(14) involve $\mathbf{x}$ and θ, $\mathbf{T}_i$ is a function of $\mathbf{x}$ and θ. Thus we denote it as $\mathbf{T}_i(x, \theta)$. The expression for control can be obtained in terms of the power series:

$$\mathbf{u} = -\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}V_{\mathbf{x}} = -\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\sum_{i=0}^{\infty}\mathbf{T}_i(\mathbf{x}, \theta)\theta^i\mathbf{x} \tag{15}$$

It is easy to see that Equation (11) is an algebraic Riccati equation. The rest of the equations are Lyapunov equations that are *linear* in $\mathbf{T}_i$. In the rest of this paper, we will call this method the θ–D technique. The algorithm without $\mathbf{D}_i$ term is called the θ approximation. The algorithm in Reference [6] turns out to be the θ approximation although their derivation is different. A major drawback of the θ approximation, however, is that large initial conditions may give rise to large control or lead to instability that will be discussed later in Section 3.

Construct the following expression for $\mathbf{D}_i,\ i = 1, \ldots, n$:

$$\mathbf{D}_1 = k_1\mathrm{e}^{-l_1 t}\left[-\frac{\mathbf{T}_0\mathbf{A}(\mathbf{x})}{\theta} - \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{x})\mathbf{T}_0}{\theta}\right] \tag{16}$$

$$\mathbf{D}_2 = k_2 \mathrm{e}^{-l_2 t}\left[-\frac{\mathbf{T}_1\mathbf{A}(\mathbf{x})}{\theta} - \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{x})\mathbf{T}_1}{\theta} + \mathbf{T}_1\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_1\right] \tag{17}$$

$$\vdots$$

$$\mathbf{D}_n = k_n \mathrm{e}^{-l_n t}\left[-\frac{\mathbf{T}_{n-1}\mathbf{A}(\mathbf{x})}{\theta} - \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{x})\mathbf{T}_{n-1}}{\theta} + \sum_{j=1}^{n-1}\mathbf{T}_j\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_{n-j}\right] \tag{18}$$

where k_i and $l_i > 0, i = 1, \ldots, n$ are scalar adjustable design parameters.

A key difference between the algorithm in Reference [6] and this study is the terms involving $\mathbf{D}_i$. Earlier part of this research led to a version of the θ approximation. The θ approximation leads to Equations (11)–(14) without $\mathbf{D}_i$ terms. In applications though, whenever the initial conditions were large, the transient control effort was huge. After careful analysis, it was found that large control usually results from the state dependent term $\mathbf{A}(\mathbf{x})$ on the right-hand side of Equations (12)–(14). The rise in control magnitudes happens when there are terms in $\mathbf{A}(\mathbf{x})$ which could grow to a high magnitude when $\mathbf{x}$ is large. For example, when $\mathbf{A}(\mathbf{x})$ includes a cubic term, a high initial state would result in a high initial $\mathbf{T}_i$ and would be amplified further into the solution of $\mathbf{T}_{i+1}$. Consequently it leads to large initial control magnitude. This observation led to search for ways to alleviate this effect. It was observed that vanishing perturbations to the cost function provide extra terms on the right-hand side of Equations (12)–(14). The next step in this process was to find expressions that could help with offsetting the problem of large control without compromising the original problem. This could happen only if those extra terms are small and decrease fast with time. So $\mathbf{D}_i$ is chosen as

$$\begin{aligned} &-\frac{\mathbf{T}_{i-1}\mathbf{A}(\mathbf{x})}{\theta} - \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{x})\mathbf{T}_{i-1}}{\theta} + \sum_{j=1}^{i-1}\mathbf{T}_j\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_{i-j} - \mathbf{D}_i \\ &\quad = \varepsilon_i(t)\left[-\frac{\mathbf{T}_{i-1}\mathbf{A}(\mathbf{x})}{\theta} - \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{x})\mathbf{T}_{i-1}}{\theta} + \sum_{j=1}^{i-1}\mathbf{T}_j\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_{i-j}\right], \quad i = 1, \ldots, n \end{aligned} \tag{19}$$

where $\varepsilon_i(t) = 1 - k_i \mathrm{e}^{-l_i t}$ is a small number, i.e. $0 \leqslant \varepsilon_i(t) < 1$. $\varepsilon_i(t)$ can be used to suppress this large value to propagate in Equations (12)–(14). $\varepsilon_i(t)$ is chosen to satisfy some conditions required in the proof of convergence and stability of the θ–D algorithm. On the other hand, the exponential term $\mathrm{e}^{-l_i t}$ with $l_i > 0$ is used to let the perturbation terms in the cost function and HJB equation diminish as time evolves.

Remark 2.1

As long as $\varepsilon_i(t)$ is a small number, we can limit t on a compact time interval $[0, T]$ which is the time period we are interested in. Of course T could be very large. Theoretically no matter how large T is, we can always choose a small enough l_i and a proper k_i to make $\varepsilon_i(t) = 1 - k_i \mathrm{e}^{-l_i t}$ a small number. In applications, selection of (k_i, l_i) is problem dependent since these parameters affect the transient performance.

Remark 2.2

Equations (11)–(18) are sufficient but not necessary conditions for $V_{\mathbf{x}} = \sum_{i=0}^{\infty}\mathbf{T}_i(\mathbf{x}, \theta)\theta^i\mathbf{x}$ to be the solution of the HJB equation. This can be explained by noting that we only assume that the gradient of the optimal cost function is of the form $\sum_{i=0}^{\infty}\mathbf{T}_i(\mathbf{x}, \theta)\theta^i\mathbf{x}$. To recover the optimal

control, an extra condition has to be imposed such that this solution is the gradient of a positive definite function with $V(\mathbf{0})=\mathbf{0}$. This point has been discussed in several papers [18–20]. Here we give a proposition under which such an optimal cost function exists is proposed. This result can also be found in References [18,19].

Proposition[18,19]
Suppose the θ–D approximation (11)–(18) has a positive definite matrix valued solution $\sum_{i=0}^{\infty}\mathbf{T}_i(\mathbf{x},\theta)\theta^i$. If the vector valued function $\boldsymbol{\lambda}(\mathbf{x},\theta)=\sum_{i=0}^{\infty}\mathbf{T}_i(\mathbf{x},\theta)\theta^i\mathbf{x}$ satisfies

$$\frac{\partial\lambda_i}{\partial x_j}(\mathbf{x},\theta)=\frac{\partial\lambda_j}{\partial x_i}(\mathbf{x},\theta) \tag{20}$$

for all $\mathbf{x}\in R^n$ and $i,j=1,2,\ldots,n$. Equation (15) is the global optimal state feedback control for the regulation problem (1) and (2). Moreover the optimal cost is given by

$$V(\mathbf{x})=\mathbf{x}^{\mathrm{T}}\int_0^1\lambda(t\mathbf{x})\,\mathrm{d}t \tag{21}$$

The positive definiteness of $\sum_{i=0}^{\infty}\mathbf{T}_i(\mathbf{x},\theta)\theta^i$ will be proved in the subsequent development. Solutions of $\mathbf{T}_i$'s depend upon the factorization of $\mathbf{f}(\mathbf{x})$, i.e. $\mathbf{A}_0$ and $\mathbf{A}(\mathbf{x})$. It has been shown [20] that finding the optimal factorization of $\mathbf{f}(\mathbf{x})$ such that (20) is satisfied is as hard as solving the HJB equation. Even in some other suboptimal control method like the SDRE [20], it can only be checked numerically. The θ–D method in this paper can be written in closed form as a suboptimal control due to the difficulty in checking this condition (20).

The steps of applying the θ–D method are summarized as follows:

(1) Solve the algebraic Riccati equation (11) to get $\mathbf{T}_0$ once $\mathbf{A}_0$, $\mathbf{Q}$ and $\mathbf{R}$ are determined. Note that the resulting $\mathbf{T}_0$ is a positive definite constant matrix.
(2) Solve Lyapunov equation (12) to get $\mathbf{T}_1(\mathbf{x},\theta)$. Note that it is a linear equation in terms of $\mathbf{T}_1$ and an interesting property of this and the rest of the equations is that the coefficient matrices $\mathbf{A}_0-\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_0$ and $\mathbf{A}_0^{\mathrm{T}}-\mathbf{T}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}$ are constant. Let $\mathbf{A}_{c_0}=\mathbf{A}_0-\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_0$. Through linear algebra, Equation (12) can be brought into a form like $\hat{\mathbf{A}}_0\mathrm{Vec}\,(\mathbf{T}_1)=\mathrm{Vec}\,[\mathbf{Q}_1(\mathbf{x},\theta,t)]$ where $\mathbf{Q}_1(\mathbf{x},\theta,t)$ is the right-hand side of Equation (12); $\mathrm{Vec}\,(\mathbf{M})$ denotes stacking the elements of matrix $\mathbf{M}$ by rows in a vector form; $\hat{\mathbf{A}}_0=\mathbf{I}\otimes\mathbf{A}_{c_0}^{\mathrm{T}}+\mathbf{A}_{c_0}^{\mathrm{T}}\otimes\mathbf{I}$ is a constant matrix and the symbol $\otimes$ denotes the Kronecker product. The resulting solution of $\mathbf{T}_1$ can be written in closed-form as $\mathrm{Vec}\,(\mathbf{T}_1)=\hat{\mathbf{A}}_0^{-1}\mathrm{Vec}[\mathbf{Q}_1(\mathbf{x},\theta,t)]$.
(3) Solve Equations (13) for $\mathbf{T}_2(\mathbf{x},\theta)$ by following the same procedure in step 2. Number of $\mathbf{T}_i$s needed depends on the problem.

As can be seen, closed-form solutions for $\mathbf{T}_2,\ldots,\mathbf{T}_n$ can be obtained with just one constant matrix inverse operation. The expression of $\mathbf{Q}_i(\mathbf{x},\theta,t)$ on the right-hand side of the equations is already known and needs only simple matrix multiplications and additions. This has been used in different practical problems requiring on-line control [21].

The following theorem will show that the convergence of the series expansion of $\sum_{i=0}^{\infty}\mathbf{T}_i(\mathbf{x},\theta)\theta^i$ on a specified compact set Ω can be obtained by choosing appropriate $\mathbf{D}_i$ matrices.

Theorem 2.1

If the following conditions are satisfied:

(i) $\mathbf{x} \in \Omega$, where $\Omega \subset R^n$ is a compact set,
(ii) $(\mathbf{A}_0, \mathbf{g})$ is a controllable pair,
(iii) $\mathbf{A}(\mathbf{x})$ is continuous on Ω and $\|\mathbf{A}(\mathbf{x})\|_2 \neq 0, \ \forall \mathbf{x} \in \Omega$,
(iv) $\lambda_{\max}[(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_0) + (\mathbf{A}_0^{\mathrm{T}} - \mathbf{T}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})] < 0$, where $\lambda_{\max}$ denotes the largest eigenvalue,
(v) $\mathbf{D}_i, \ i = 1, \ldots, n$ are chosen according to Equations (16)–(18),

The series $\sum_{i=0}^{\infty} \mathbf{T}_i(\mathbf{x}, \theta)\theta^i$ produced by the algorithm in Equations (11)–(14) is a pointwise convergent series.

Proof

Considering (12) and the selection of $\mathbf{D}_1$ in (16), (12) can be written as

$$\mathbf{T}_1(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_0) + (\mathbf{A}_0^{\mathrm{T}} - \mathbf{T}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})\mathbf{T}_1 = -\varepsilon_1(\mathbf{T}_0\mathbf{A} + \mathbf{A}^{\mathrm{T}}\mathbf{T}_0)\,\frac{1}{\theta} \tag{22}$$

with

$$\varepsilon_1 = 1 - k_1 \mathrm{e}^{-l_1 t} \tag{23}$$

For clarity, from now on, we omit the argument $\mathbf{x}$ in $\mathbf{A}(\mathbf{x})$ and t in $\varepsilon_i(t)$ to simplify the notation. Assume that the solution to the equation

$$\hat{\mathbf{T}}_1(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0) + (\mathbf{A}_0^{\mathrm{T}} - \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})\hat{\mathbf{T}}_1 = -\varepsilon_1(\hat{\mathbf{T}}_0\mathbf{A} + \mathbf{A}^{\mathrm{T}}\hat{\mathbf{T}}_0) \tag{24}$$

is $\hat{\mathbf{T}}_1$ with

$$\hat{\mathbf{T}}_0 = \mathbf{T}_0 \tag{25}$$

By using the linearity of Lyapunov equation (24), the solution to (22) becomes

$$\mathbf{T}_1 = \frac{1}{\theta}\,\hat{\mathbf{T}}_1 \tag{26}$$

Similarly assume that the solution to the equation

$$\hat{\mathbf{T}}_2(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0) + (\mathbf{A}_0^{\mathrm{T}} - \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})\hat{\mathbf{T}}_2 = -\varepsilon_2(\hat{\mathbf{T}}_1\mathbf{A} + \mathbf{A}^{\mathrm{T}}\hat{\mathbf{T}}_1 - \hat{\boldsymbol{T}}_1\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_1) \tag{27}$$

is $\hat{\mathbf{T}}_2$. Then the solution to (13) is

$$\mathbf{T}_2 = \frac{1}{\theta^2}\,\hat{\mathbf{T}}_2 \tag{28}$$

In the same manner,

$$\mathbf{T}_n = \frac{1}{\theta^n}\,\hat{\mathbf{T}}_n \tag{29}$$

where $\hat{\mathbf{T}}_n$ is the solution of

$$\hat{\mathbf{T}}_n(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0) + (\mathbf{A}_0^{\mathrm{T}} - \hat{\mathbf{T}}_0\mathbf{g}\boldsymbol{R}^{-1}\mathbf{g}^{\mathrm{T}})\hat{\mathbf{T}}_n$$

$$= -\varepsilon_n\left[\hat{\mathbf{T}}_{n-1}\mathbf{A}(\mathbf{x}) + \mathbf{A}^{\mathrm{T}}(\mathbf{x})\hat{\mathbf{T}}_{n-1} - \sum_{j=1}^{n-1}\hat{\mathbf{T}}_j\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_{n-j}\right] \tag{30}$$

and

$$\varepsilon_n = 1 - k_n\mathrm{e}^{-l_n t} \tag{31}$$

From (25), (26), (28) and (29) we note that proving the convergence of $\sum_{i=0}^{\infty}\mathbf{T}_i(\mathbf{x},\theta)\theta^i$ is equivalent to proving the convergence of $\sum_{i=0}^{\infty}\hat{\mathbf{T}}_i(\mathbf{x})$ where $\hat{\mathbf{T}}_i$ satisfy Equations (11), (24), (27) and (30). Here we list them for clarity:

$$\hat{\mathbf{T}}_0\mathbf{A}_0 + \mathbf{A}_0^{\mathrm{T}}\hat{\mathbf{T}}_0 - \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0 + \mathbf{Q} = 0 \tag{32}$$

$$\hat{\mathbf{T}}_1(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0) + (\mathbf{A}_0^{\mathrm{T}} - \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})\hat{\mathbf{T}}_1 = -\varepsilon_1(\hat{\mathbf{T}}_0\mathbf{A} + \mathbf{A}^{\mathrm{T}}\hat{\mathbf{T}}_0) \tag{24a}$$

$$\hat{\mathbf{T}}_2(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0) + (\mathbf{A}_0^{\mathrm{T}} - \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})\hat{\mathbf{T}}_2 = -\varepsilon_2(\hat{\mathbf{T}}_1\mathbf{A} + \mathbf{A}^{\mathrm{T}}\hat{\mathbf{T}}_1 - \hat{\boldsymbol{T}}_1\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_1) \tag{27a}$$

$$\vdots$$

$$\hat{\mathbf{T}}_n(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0) + (\mathbf{A}_0^{\mathrm{T}} - \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})\hat{\mathbf{T}}_n$$

$$= -\varepsilon_n\left[\hat{\mathbf{T}}_{n-1}\mathbf{A}(\mathbf{x}) + \mathbf{A}^{\mathrm{T}}(\mathbf{x})\hat{\mathbf{T}}_{n-1} - \sum_{j=1}^{n-1}\hat{\mathbf{T}}_j\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_{n-j}\right] \tag{30a}$$

The objective now is to find a norm bound for each $\hat{\mathbf{T}}_i$ in order to prove the convergence of the series $\sum_{i=1}^{\infty}\hat{\mathbf{T}}_i$. Given a continuous Lyapunov equation

$$\bar{\mathbf{A}}^{\mathrm{T}}\bar{\mathbf{P}} + \bar{\mathbf{P}}\bar{\mathbf{A}} = -\bar{\mathbf{Q}} \tag{33}$$

where $\bar{\mathbf{A}}, \bar{\mathbf{P}}, \bar{\mathbf{Q}} \in R^{n\times n}$, if $\bar{\mathbf{A}}$ is a stable matrix, we have the norm bound for $\bar{\mathbf{P}}$ [22].

$$\|\bar{\mathbf{P}}\|_{\bullet} \leqslant \frac{\|\bar{\mathbf{Q}}\|_{\bullet}}{-\mu_{\bullet}(\bar{\mathbf{A}}^{\mathrm{T}}) - \mu_{\bullet}(\bar{\mathbf{A}})} \tag{34}$$

where $\mu_{\bullet}(\bar{\mathbf{A}})$ is a matrix measure of $\bar{\mathbf{A}}$ induced from $\|\ \|_{\bullet}$. In the case of 2-norm,

$$\mu_2(\bar{\mathbf{A}}) \triangleq \tfrac{1}{2}\lambda_{\max}(\bar{\mathbf{A}} + \bar{\mathbf{A}}^{\mathrm{T}}) \tag{35}$$

In the following, $\|\bullet\|$ is defined as a 2-norm and denote $\mu(\bullet) = \mu_2(\bullet)$.

In Equation (32), $(\mathbf{A}_0, \mathbf{g})$ is a controllable pair and $\mathbf{R}$ is positive definite and $\mathbf{Q}$ is semi-positive definite. Hence, the Riccati equation (32) has a positive definite solution and $(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0)$ is a stable matrix.

From (34) we have

$$\|\hat{\mathbf{T}}_1\| \leqslant \frac{\|\varepsilon_1[\hat{\mathbf{T}}_0\mathbf{A}(\mathbf{x}) + \mathbf{A}^{\mathrm{T}}(\mathbf{x})\hat{\mathbf{T}}_0]\|}{-\mu(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0) - \mu(\mathbf{A}_0^{\mathrm{T}} - \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})} \tag{36}$$

Let

$$C = \frac{1}{-\mu(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0) - \mu(\mathbf{A}_0^{\mathrm{T}} - \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})} \tag{37}$$

Since we assume that $\lambda_{\max}[(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0) + (\mathbf{A}_0^{\mathrm{T}} - \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})] < 0$, we have $C > 0$.

Then

$$\|\hat{\mathbf{T}}_1\| \leqslant C\varepsilon_1\|\hat{\mathbf{T}}_0\mathbf{A} + \mathbf{A}^{\mathrm{T}}\hat{\mathbf{T}}_0\| \tag{38}$$

$$\leqslant C\varepsilon_1[\|\hat{\mathbf{T}}_0\|(\|\mathbf{A}\| + \|\mathbf{A}^{\mathrm{T}}\|)] \tag{39}$$

Since $\mathbf{A}(\mathbf{x})$ is continuous on a compact set Ω, it is bounded on Ω.

Let

$$C_1 = \max_{\mathbf{x}\in\Omega}(\|\mathbf{A}(\mathbf{x})\| + \|\mathbf{A}^{\mathrm{T}}(\mathbf{x})\|) \tag{40}$$

Then

$$\|\hat{\mathbf{T}}_1\| \leqslant \varepsilon_1 C C_1\|\hat{\mathbf{T}}_0\| \tag{41}$$

We can see that

$$\varepsilon_1 \cdot C \cdot C_1 \cdot \|\hat{\mathbf{T}}_0\| = O(\varepsilon_1) \tag{42}$$

since C, C_1 and $\|\hat{\mathbf{T}}_0\|$ are constants.

Since it is assumed that $\|\mathbf{A}(\mathbf{x})\|$ is not zero, $C_1 \neq 0$. If it is zero, the non-linear control problem will reduce to the linear regulator problem for which the solution is nothing but $\mathbf{T}_0$, the solution of the algebraic Riccati equation (11).

For later use, define

$$S_0 = \|\hat{\mathbf{T}}_0\| \tag{43}$$

and

$$S_1 = \varepsilon_1 \cdot C \cdot C_1 \cdot \|\hat{\mathbf{T}}_0\| \tag{44}$$

Then

$$S_1 = O(\varepsilon_1) \tag{45}$$

Therefore, by choosing sufficiently small ε_1, we can always make $S_1/S_0 < \varepsilon_1 C C_1 < 1$.

Consider Equation (27). A norm-bounded inequality for $\hat{\mathbf{T}}_2$ becomes

$$\|\hat{\mathbf{T}}_2\| \leqslant C\varepsilon_2 \|\hat{\mathbf{T}}_1\mathbf{A} + \mathbf{A}^{\mathrm{T}}\hat{\mathbf{T}}_1 - \hat{\mathbf{T}}_1\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_1\| \tag{46}$$

Let

$$C_g = \|\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\| \tag{47}$$

Then

$$\begin{aligned}\|\hat{\mathbf{T}}_2\| &\leqslant C \cdot \varepsilon_2(C_1\|\hat{\mathbf{T}}_1\| + \|\hat{\mathbf{T}}_1\|^2 C_g) \\ &\leqslant C \cdot \varepsilon_2(C_1 \cdot \varepsilon_1 C C_1\|\hat{\mathbf{T}}_0\| + \varepsilon_1^2 C^2 C_1^2\|\hat{\mathbf{T}}_0\|^2 C_g) \\ &= C^2\varepsilon_1\varepsilon_2 C_1^2\|\hat{\mathbf{T}}_0\|(1 + \varepsilon_1 C\|\hat{\mathbf{T}}_0\|C_g)\end{aligned} \tag{48}$$

Let

$$C_2 = \max_{t\in[0,T]}(1 + \varepsilon_1 C\|\hat{\mathbf{T}}_0\|C_g) \tag{49}$$

Then we get

$$\|\hat{\mathbf{T}}_2\| \leqslant \varepsilon_1\varepsilon_2 C^2 \cdot C_1^2 \cdot C_2\|\hat{\mathbf{T}}_0\| \tag{50}$$

Let

$$S_2 = \varepsilon_1\varepsilon_2 C^2 \cdot C_1^2 \cdot C_2\|\hat{\mathbf{T}}_0\| \tag{51}$$

We can see that

$$S_2 = O(\varepsilon_1\varepsilon_2) \tag{52}$$

Note that from Equation (44) we have

$$\frac{S_2}{S_1} = \varepsilon_2 C C_1 C_2 = O(\varepsilon_2) \tag{53}$$

Therefore, if ε_2 is picked sufficiently small, we can make

$$\frac{S_2}{S_1} < 1 \tag{54}$$

For $\hat{\boldsymbol{T}}_3$, we have

$$\begin{aligned}\|\hat{\mathbf{T}}_3\| &\leqslant C\varepsilon_3\|\hat{\mathbf{T}}_2\mathbf{A} + \mathbf{A}^{\mathrm{T}}\hat{\mathbf{T}}_2 - \hat{\mathbf{T}}_1\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_2 - \hat{\mathbf{T}}_2\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_1\| \\ &\leqslant C\varepsilon_3[C_1\|\hat{\mathbf{T}}_2\| + 2C_g\|\hat{\mathbf{T}}_1\|\,\|\hat{\mathbf{T}}_2\|] \\ &\leqslant C\varepsilon_3[\varepsilon_1\varepsilon_2C^2C_1^3C_2\|\hat{\mathbf{T}}_0\| + 2C_g\varepsilon_1^2\varepsilon_2C^3C_1^3C_2\|\hat{\mathbf{T}}_0\|^2] \\ &\leqslant C^3\varepsilon_1\varepsilon_2\varepsilon_3C_1^3C_2\|\hat{\mathbf{T}}_0\|\,[1 + 2C_g\varepsilon_1C\|\hat{\mathbf{T}}_0\|] \end{aligned} \tag{55}$$

Let

$$C_3 = \max_{t\in[0,T]}(1 + 2C_g\varepsilon_1C\|\hat{\mathbf{T}}_0\|) \tag{56}$$

So we have

$$\|\hat{\mathbf{T}}_3\| \leqslant \varepsilon_1\varepsilon_2\varepsilon_3C^3C_1^3C_2C_3\|\hat{\mathbf{T}}_0\| \tag{57}$$

Let

$$S_3 = \varepsilon_1\varepsilon_2\varepsilon_3C^3C_1^3C_2C_3\|\hat{\mathbf{T}}_0\| = O(\varepsilon_1\varepsilon_2\varepsilon_3) \tag{58}$$

According to Equation (51) we have

$$\frac{S_3}{S_2} = \varepsilon_3CC_1C_3 = O(\varepsilon_3) \tag{59}$$

Therefore, if ε_3 is picked sufficiently small, we can make

$$\frac{S_3}{S_2} < 1 \tag{60}$$

In a similar manner we can derive for $\hat{\mathbf{T}}_n$ that

$$\|\hat{\mathbf{T}}_n\| \leqslant (\varepsilon_1\cdots\varepsilon_n)\cdot C^nC_1^nC_2\ldots C_n\|\hat{\mathbf{T}}_0\| \tag{61}$$

and

$$(\varepsilon_1\cdots\varepsilon_n)\cdot C^nC_1^nC_2\ldots C_n\|\hat{\mathbf{T}}_0\| = O(\varepsilon_1\cdots\varepsilon_n) \tag{62}$$

Once the bound for each $\hat{\mathbf{T}}_i$ is determined, the convergence of the series $\sum_{i=1}^{\infty}\hat{\mathbf{T}}_i$ can be obtained. Define a series $\sum_{n=0}^{\infty}S_n$ with S_0 and S_1 defined in (43) and (44) and

$$S_n = (\varepsilon_1\cdots\varepsilon_n)\cdot C^nC_1^nC_2\ldots C_n\|\hat{\mathbf{T}}_0\| \tag{63}$$

Then

$$\frac{S_n}{S_{n-1}} = \varepsilon_n\cdot CC_1C_n = O(\varepsilon_n) \tag{64}$$

By choosing a sufficiently small ε_n such that $\lim_{n\to\infty}\varepsilon_n\cdot CC_1C_n < 1, \sum_{i=0}^{\infty}S_i$ is a convergent series. Since each $\|\hat{\mathbf{T}}_i\| \leqslant S_i$, $\sum_{i=0}^{\infty}\hat{\mathbf{T}}_i$ is also a convergent series. Therefore $\sum_{i=0}^{\infty}\mathbf{T}_i(\mathbf{x},\theta)\theta^i$ is convergent. □

Remark 2.3

The proof of Theorem 2.1 enables us to observe that θ is just an intermediate variable used in a power series expansion. It turns out to be cancelled by the choice of $\mathbf{D}_i$ matrices (see Equations (16)–(18)). $1/\theta^i$ appears linearly in Equations (11)–(14), which is shown in the proof of Theorem 2.1. When $\mathbf{T}_i$ multiplies θ^i in the control, they get cancelled. What come into play eventually are the $\hat{\mathbf{T}}_i$ matrices since $\mathbf{T}_i = \hat{\mathbf{T}}_i/\theta^i$.

Lemma 2.1

If the five conditions (i)–(v) in Theorem 2.1 are satisfied, the series $\sum_{i=0}^{\infty} \mathbf{T}_i(\mathbf{x}, \theta)\theta^i$ is positive definite.

Proof

Rewrite (32) as

$$\hat{\mathbf{T}}_0(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0) + (\mathbf{A}_0^{\mathrm{T}} - \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})\hat{\mathbf{T}}_0 = -\mathbf{Q} - \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0 \tag{65}$$

Now, we can obtain the following equality by manipulating a few equations as (65) + (24) + (27) + (30)

$$\begin{aligned}&(\hat{\mathbf{T}}_0 + \hat{\mathbf{T}}_1 + \hat{\mathbf{T}}_2 + \cdots + \hat{\mathbf{T}}_n)(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0) + (\mathbf{A}_0^{\mathrm{T}} - \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})(\hat{\mathbf{T}}_0 + \hat{\mathbf{T}}_1 + \hat{\mathbf{T}}_2 + \cdots + \hat{\mathbf{T}}_n)\\ &= -\mathbf{Q} - \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0 - \sum_{k=1}^{n} \varepsilon_k(\hat{\mathbf{T}}_{k-1}\mathbf{A} + \mathbf{A}^{\mathrm{T}}\hat{\mathbf{T}}_{k-1}) + \sum_{k=2}^{n} \varepsilon_k \sum_{j=1}^{k-1} \hat{\mathbf{T}}_j\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_{k-j}\end{aligned} \tag{66}$$

Convergence of the series $\sum_{i=0}^{\infty} \hat{\mathbf{T}}_i$ has been proved in Theorem 2.1. Assume that $\mathbf{T}_\mathbf{s} = \sum_{i=0}^{\infty} \hat{\mathbf{T}}_i$. When $n \to \infty$, Equation (66) becomes

$$\begin{aligned}&\mathbf{T}_s(\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0) + (\mathbf{A}_0^{\mathrm{T}} - \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}})\mathbf{T}_s\\ &= -\mathbf{Q} - \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0 - \sum_{k=1}^{n\to\infty} \varepsilon_k(\hat{\mathbf{T}}_{k-1}\mathbf{A} + \mathbf{A}^{\mathrm{T}}\hat{\mathbf{T}}_{k-1}) + \sum_{k=2}^{n\to\infty} \varepsilon_k \sum_{j=1}^{k-1} \hat{\mathbf{T}}_j\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_{k-j}\end{aligned} \tag{67}$$

To prove the positive definiteness of $\mathbf{T}_\mathbf{s}$, we only need to prove that the right-hand side of Equation (67) is negative definite since $\mathbf{A}_0 - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0$ is a Hurwitz matrix. Showing the right-hand side of Equation (67) negative definite is equivalent to showing

$$\mathbf{x}^{\mathrm{T}}\left[\mathbf{Q} + \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0 + \sum_{k=1}^{n\to\infty} \varepsilon_k(\hat{\mathbf{T}}_{k-1}\mathbf{A} + \mathbf{A}^{\mathrm{T}}\hat{\mathbf{T}}_{k-1}) - \sum_{k=2}^{n\to\infty} \varepsilon_k \sum_{j=1}^{k-1} \hat{\mathbf{T}}_j\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_{k-j}\right]\mathbf{x} > 0 \tag{68}$$

Consider inequality (68) term by term. For the first two terms denoted by t_{12}, we have

$$t_{12} = \mathbf{x}^{\mathbf{T}}(\mathbf{Q} + \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0)\mathbf{x} \geqslant C_{\lambda_1}\|\mathbf{x}\|^2 \tag{69}$$

where $C_{\lambda_1} = \lambda_{\min}(\mathbf{Q} + \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0) = \text{const}$. Since $\mathbf{Q} \geqslant 0,\ \mathbf{R} > 0,\ C_{\lambda_1} > 0$.

For the third term denoted by t_3, we have

$$t_3 = \mathbf{x}^{\mathrm{T}}\left[\sum_{k=1}^{n\to\infty} \varepsilon_k(\hat{\mathbf{T}}_{k-1}\mathbf{A} + \mathbf{A}^{\mathrm{T}}\hat{\mathbf{T}}_{k-1})\right]\mathbf{x} \geqslant \mathbf{x}^{\mathrm{T}}\left[\varepsilon_m \sum_{k=1}^{n\to\infty} \hat{\mathbf{T}}_{k-1}(\mathbf{A} + \mathbf{A})\right]\mathbf{x}$$

where

$$\varepsilon_m = \begin{cases} \min_{t\in[0,T]}\{\varepsilon_1, \ldots \varepsilon_n, \ldots\} & \text{if } t_3 > 0 \\ \max_{t\in[0,T]}\{\varepsilon_1, \ldots \varepsilon_n, \ldots\} & \text{if } t_3 < 0 \end{cases} \quad \text{and} \quad [0, T] \text{ is a compact time interval}$$

Since it is assumed that $\mathbf{A}(\mathbf{x})$ is continuous in a compact set Ω, $\mathbf{A}(\mathbf{x})$ is bounded on Ω.

Then

$$t_3 \geqslant \mathbf{x}^{\mathrm{T}}[\varepsilon_m \mathbf{T}_s(\mathbf{A} + \mathbf{A})]\mathbf{x} \geqslant \varepsilon_m C_{\lambda_2}\|\mathbf{x}\|^2 \tag{70}$$

where $C_{\lambda_2} = \lambda_{\min}[\mathbf{T}_s(\mathbf{A} + \mathbf{A})] = \text{constant}$.

For the last term denoted by t_4, we have

$$t_4 = \mathbf{x}^{\mathrm{T}}\left[\sum_{k=2}^{n\to\infty} \varepsilon_k \sum_{j=1}^{k-1} \hat{\mathbf{T}}_j \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_{k-j}\right]\mathbf{x} \leqslant \|\mathbf{x}\|^2 \sum_{k=2}^{n\to\infty} \varepsilon_k \sum_{j=1}^{\infty} \|\hat{\mathbf{T}}_j\| \, \|\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\| \sum_{j=1}^{n\to\infty} \|\hat{\mathbf{T}}_{n-j}\|$$

Assume that $C_{\lambda_3} = \sum_{j=1}^{\infty} \|\hat{\mathbf{T}}_j\| \, \|\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\| \sum_{j=1}^{n\to\infty} \|\hat{\mathbf{T}}_{n-j}\| = \text{constant}$, since $\sum_{j=1}^{\infty} \|\hat{\mathbf{T}}_j\|$ is a convergent series.

Then $t_4 \leqslant \sum_{k=2}^{n\to\infty} \varepsilon_k C_{\lambda_3}\|\mathbf{x}\|^2$.

Since $0 \leqslant \varepsilon_k < 1$ $(k = 1 \ldots n, \ldots)$ are design parameters, we can always choose small enough ε_k such that $\sum_{k=2}^{n\to\infty} \varepsilon_k$ is a convergent series. Assume that $\varepsilon_c = \sum_{k=2}^{n\to\infty} \varepsilon_k$ and we have

$$t_4 \leqslant \varepsilon_c C_{\lambda_3}\|\mathbf{x}\|^2 \tag{71}$$

Combining inequalities (69)–(71), we have

$$\mathbf{x}^{\mathrm{T}}\left[\mathbf{Q} + \hat{\mathbf{T}}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_0 + \sum_{k=1}^{n\to\infty} \varepsilon_k(\hat{\mathbf{T}}_{k-1}\mathbf{A} + \mathbf{A}^{\mathrm{T}}\hat{\mathbf{T}}_{k-1}) - \sum_{k=2}^{n\to\infty} \varepsilon_k \sum_{j=1}^{k-1} \hat{\mathbf{T}}_j\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_{n-j}\right]\mathbf{x}$$
$$\geqslant \|\mathbf{x}\|^2(C_{\lambda_1} + \varepsilon_m C_{\lambda_2} - \varepsilon_c C_{\lambda_3}) \tag{72}$$

Hence, by choosing small enough ε_m and ε_c, the right-hand side of inequality (72) can be made positive definite and equivalently the right-hand side of Equation (67) is negative definite. Therefore $\mathbf{T}_s = \sum_{i=0}^{\infty} \hat{\mathbf{T}}_i(\mathbf{x})$ or $\sum_{i=0}^{\infty} \mathbf{T}_i(\mathbf{x}, \theta)\theta^i$ is positive definite. □

Based upon the above claims, we can prove that θ–D approximation technique can achieve semi-global asymptotic stability.

Theorem 2.2

If the five conditions in Theorem 2.1 are satisfied and (vi) $\mathbf{D}_i$ are chosen such that $\mathbf{Q} + \sum_{i=1}^{\infty} \mathbf{D}_i\theta^i$ is semi-positive definite, then the closed-loop feedback control system obtained by control law $\mathbf{u} = -\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}} \sum_{i=0}^{\infty} \mathbf{T}_i(\mathbf{x}, \theta)\theta^i\mathbf{x}$ is semi-globally asymptotically stable.

Proof

Let us choose a Lyapunov candidate function

$$L(\mathbf{x}) = \frac{1}{2}\,\mathbf{x}^{\mathrm{T}} \sum_{i=0}^{\infty} \hat{\mathbf{T}}_i(\mathbf{x})\mathbf{x} \tag{73}$$

In Lemma 2.1, it has been shown that $\sum_{i=0}^{\infty} \hat{\mathbf{T}}_i(\mathbf{x})$ is positive definite. So $L(\mathbf{x}) > 0$.

Now,

$$\begin{aligned}\frac{\mathrm{d}L(\mathbf{x})}{\mathrm{d}t} &= \left[\frac{\partial L(\mathbf{x})}{\partial \mathbf{x}}\right]^{\mathrm{T}} \dot{\mathbf{x}} = \left[\frac{\partial L(\mathbf{x})}{\partial \mathbf{x}}\right]^{\mathrm{T}} [\mathbf{f}(\mathbf{x}) + \mathbf{g}\mathbf{u}] \\ &= \left[\mathbf{x}^{\mathrm{T}} \sum_{i=0}^{\infty} \hat{\mathbf{T}}_i + \frac{1}{2}\,\mathbf{x}^{\mathrm{T}} \sum_{i=1}^{\infty} \frac{\partial \hat{\mathbf{T}}_i}{\partial \mathbf{x}}\,\mathbf{x}\right][\mathbf{f}(\mathbf{x}) + \mathbf{g}\mathbf{u}]\end{aligned} \tag{74}$$

In the remaining proof the argument $\mathbf{x}$ in $\mathbf{f}(\mathbf{x})$ will be omitted to simplify the notations.

Since $V_{\mathbf{x}} = \sum_{i=0}^{\infty} \hat{\mathbf{T}}_i\mathbf{x}$ satisfies the HJB equation

$$V_{\mathbf{x}}^{\mathrm{T}}[\mathbf{f} + \mathbf{g}\mathbf{u}] + \frac{1}{2}\,\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u} + \frac{1}{2}\,\mathbf{x}^{\mathrm{T}}\left(\mathbf{Q} + \sum_{i=1}^{\infty} \mathbf{D}_i\theta^i\right)\mathbf{x} = 0$$

The above equation can be rearranged to get

$$\mathbf{x}^{\mathrm{T}} \sum_{i=0}^{\infty} \hat{\mathbf{T}}_i[\mathbf{f} + \mathbf{g}\mathbf{u}] = -\frac{1}{2}\,\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} - \frac{1}{2}\,\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u} - \frac{1}{2}\,\mathbf{x}^{\mathrm{T}} \sum_{i=1}^{\infty} \mathbf{D}_i\theta^i\mathbf{x} \tag{75}$$

Substitution of Equation (75) into Equation (74) leads to

$$\frac{\mathrm{d}L(\mathbf{x})}{\mathrm{d}t} = -\frac{1}{2}\,\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} - \frac{1}{2}\,\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u} - \frac{1}{2}\,\mathbf{x}^{\mathrm{T}} \sum_{i=1}^{\infty} \mathbf{D}_i\theta^i\mathbf{x} + \frac{1}{2}\,\mathbf{x}^{\mathrm{T}} \sum_{i=1}^{\infty} \frac{\partial \hat{\mathbf{T}}_i}{\partial \mathbf{x}}\,\mathbf{x}[\mathbf{f} + \mathbf{g}\mathbf{u}] \tag{76}$$

Since $\mathbf{Q} + \sum_{i=1}^{\infty} \mathbf{D}_i\theta^i$ is semi-positive definite and $\mathbf{R}$ is positive definite,

$$-\frac{1}{2}\,\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} - \frac{1}{2}\,\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u} - \frac{1}{2}\,\mathbf{x}^{\mathrm{T}} \sum_{i=1}^{\infty} \mathbf{D}_i\theta^i\mathbf{x} < 0 \tag{77}$$

Substitution of $\mathbf{u} = -\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}} \sum_{i=0}^{\infty} \mathbf{T}_i\theta^i\mathbf{x} = -\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}} \sum_{i=0}^{\infty} \hat{\mathbf{T}}_i\mathbf{x}$ into the left-hand side of (77) leads to

$$-\frac{1}{2}\,\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} - \frac{1}{2}\,\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u} - \frac{1}{2}\mathbf{x}^{\mathrm{T}} \sum_{i=1}^{\infty} \mathbf{D}_i\theta^i\mathbf{x} = -\frac{1}{2}\,\mathbf{x}^{\mathrm{T}}\left[\mathbf{Q} + \sum_{i=0}^{\infty} \hat{\mathbf{T}}_i\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}} \sum_{i=0}^{\infty} \hat{\mathbf{T}}_i + \sum_{i=1}^{\infty} \mathbf{D}_i\theta^i\right]\mathbf{x} \tag{78}$$

By using Courant–Fischer theorem [23], we have

$$-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\left[\mathbf{Q}+\sum_{i=0}^{\infty}\hat{\mathbf{T}}_i\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\sum_{i=0}^{\infty}\hat{\mathbf{T}}_i+\sum_{i=1}^{\infty}\mathbf{D}_i\theta^i\right]\mathbf{x}$$
$$\leqslant -\frac{1}{2}\lambda_{\min}\left[\mathbf{Q}+\sum_{i=0}^{\infty}\hat{\mathbf{T}}_i\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\sum_{i=0}^{\infty}\hat{\mathbf{T}}_i+\sum_{i=1}^{\infty}\mathbf{D}_i\theta^i\right]\|\mathbf{x}\|_2^2 \tag{79}$$

Then

$$\frac{\mathrm{d}L(\mathbf{x})}{\mathrm{d}t}\leqslant -\frac{1}{2}\lambda_{\min}\left[\mathbf{Q}+\sum_{i=0}^{\infty}\hat{\mathbf{T}}_i\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\sum_{i=0}^{\infty}\hat{\mathbf{T}}_i+\sum_{i=1}^{\infty}\mathbf{D}_i\theta^i\right]\|\mathbf{x}\|_2^2+\frac{1}{2}\mathbf{x}^{\mathrm{T}}\sum_{i=1}^{\infty}\frac{\partial\hat{\mathbf{T}}_i}{\partial\mathbf{x}}\mathbf{x}[\mathbf{f}+\mathbf{g}\mathbf{u}]$$
$$\leqslant -\frac{1}{2}C_\lambda\|\mathbf{x}\|_2^2+\frac{1}{2}\|\mathbf{x}\|_2^2\left\|\sum_{i=1}^{\infty}\frac{\partial\hat{\mathbf{T}}_i}{\partial\mathbf{x}}\mathbf{x}\right\|_2\left\|\mathbf{A}_0+\mathbf{A}-\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\sum_{i=0}^{\infty}\hat{\mathbf{T}}_i\right\|_2 \tag{80}$$

where

$$C_\lambda=\lambda_{\min}\left[\mathbf{Q}+\sum_{i=0}^{\infty}\hat{\mathbf{T}}_i\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\sum_{i=0}^{\infty}\hat{\mathbf{T}}_i+\sum_{i=1}^{\infty}\mathbf{D}_i\theta^i\right]>0 \tag{81}$$

Thus

$$\frac{\mathrm{d}L(\mathbf{x})}{\mathrm{d}t}\leqslant -\frac{1}{2}\|\mathbf{x}\|_2^2\left[C_\lambda-\left\|\sum_{i=1}^{\infty}\frac{\partial\hat{\mathbf{T}}_i}{\partial\mathbf{x}}\mathbf{x}\right\|_2\left\|\mathbf{A}_0+\mathbf{A}-\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\sum_{i=0}^{\infty}\hat{\mathbf{T}}_i\right\|_2\right] \tag{82}$$

According to the linearity property of Equations (24), (27) and (30), $\hat{\mathbf{T}}_i$ $(i=1,\ldots,n)$ can always be written in the form of

$$\hat{\mathbf{T}}_i=\varepsilon_i\bar{\mathbf{T}}_i(\mathbf{x}) \tag{83}$$

where $\bar{\mathbf{T}}_i(\mathbf{x})$ is the solution of Equations (24), (27) and (30) without ε_i, and

$$\frac{\partial\hat{\mathbf{T}}_i}{\partial\mathbf{x}}\mathbf{x}=\varepsilon_i\frac{\partial\bar{\mathbf{T}}_i}{\partial\mathbf{x}}\mathbf{x} \tag{84}$$

Thus by choosing small enough $\varepsilon_i, i=1\cdots\infty$, one can always make

$$C_\lambda-\left\|\sum_{i=1}^{\infty}\frac{\partial\hat{\mathbf{T}}_i}{\partial\mathbf{x}}\mathbf{x}\right\|_2\left\|\mathbf{A}_0+\mathbf{A}-\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\sum_{i=0}^{\infty}\hat{\mathbf{T}}_i\right\|_2$$
$$=C_\lambda-\left\|\sum_{i=1}^{\infty}\varepsilon_i\frac{\partial\bar{\mathbf{T}}_i}{\partial\mathbf{x}}\mathbf{x}\right\|_2\left\|\mathbf{A}_0+\mathbf{A}-\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\sum_{i=0}^{\infty}\hat{\mathbf{T}}_i\right\|_2>0 \tag{85}$$

Hence, $\mathrm{d}L(\mathbf{x})/\mathrm{d}t<0$. Note that as long as $\mathbf{x}$ lies in a compact set with $\mathbf{A}(\mathbf{x})$ bounded, one can always choose a set of ε_i such that $\mathrm{d}L(\mathbf{x})/\mathrm{d}t<0$. Therefore the underlying system is semi-globally asymptotically stable. □

Since the θ–D method is based on a power series expansion, two claims are made about the convergence of the cost function expansion and the estimation of the error in the cost. Proofs of these claims are given in the appendix.

Claim 2.1

Suppose that the conditions in Theorems 2.1–2.2 are satisfied. The cost function of the θ–D approximation method is a convergent series, i.e. $J = \frac{1}{2} \sum_{n=0}^{\infty} \theta^n J_n$, where

$$J_n = \int_0^{\infty} \left[2\mathbf{x}^{\mathrm{T}} \mathbf{T}_0 \mathbf{g} \mathbf{R}^{-1} \mathbf{g}^{\mathrm{T}} \mathbf{T}_n \mathbf{x} + \mathbf{x}^{\mathrm{T}} \sum_{k=1}^{n-1} \mathbf{T}_k \mathbf{g} \mathbf{R}^{-1} \mathbf{g}^{\mathrm{T}} \mathbf{T}_{n-k} \mathbf{x} \right] \mathrm{dt}$$

Claim 2.2

Suppose that the conditions in Theorems 2.1–2.2 are satisfied. The error in cost if only m terms in $\mathbf{u} = -\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}} \sum_{n=0}^{\infty} \mathbf{T}_n(\mathbf{x}, \theta)\theta^n \mathbf{x}$ are used satisfies $\|J - J^m\| = O(\varepsilon_k \varepsilon_{N-k})$ where $k \geqslant 2m + 1$ and N is a sufficient large number than k and J^m is the cost associated with the truncated $\mathbf{u}$.

2.3. *A systematic way of determining k_i and l_i parameters in the $\mathbf{D}_i$ matrices [21]*

As seen in Section 2.2, $\mathbf{D}_i$ matrices play an important role in the θ–D method. k_i and l_i in $\mathbf{D}_i$ are design parameters. They are essential to ensure the convergence of the power series of $V_{\mathbf{x}}$ and the stability of the closed-loop system. They also offer the flexibility to adjust the system transient performance. Values of k_i and l_i are problem dependent. While some problems require the perturbation $\mathbf{D}_i$ to approach zero quickly [21], some may need them to diminish slowly to ensure satisfactory transient response [24].

A systematic way to select k_i and l_i is based on the fact that the θ–D approach gives an approximate closed-form solution to the state dependent Riccati equation:

$$\mathbf{F}^{\mathrm{T}}(\mathbf{x})\mathbf{P}(\mathbf{x}) + \mathbf{P}(\mathbf{x})\mathbf{F}(\mathbf{x}) - \mathbf{P}(\mathbf{x})\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{P}(\mathbf{x}) + \mathbf{Q} = \boldsymbol{0} \tag{86}$$

where $\mathbf{F}(\mathbf{x}) = \mathbf{A}_0 + \mathbf{A}(\mathbf{x})$.

To see this, assume that

$$V_{\mathbf{x}} = \mathbf{P}(\mathbf{x})\mathbf{x} \tag{87}$$

and $\mathbf{P}(\mathbf{x})$ is symmetric. Substituting Equation (87) into the HJB equation (3) and writing non-linear $\mathbf{f}(\mathbf{x})$ in a linear like structure $\mathbf{f}(\mathbf{x}) = \mathbf{F}(\mathbf{x})\mathbf{x}$ leads to the state dependent Riccati equation (86). Compared with Equation (86), the θ–D method solves a perturbed HJB equation (9) and assumes the power series expansion of

$$V_{\mathbf{x}} = \sum_{i=0}^{\infty} \mathbf{T}_i(\mathbf{x}, \theta)\theta^i \mathbf{x} \tag{88}$$

It can be predicted that the θ–D solution is close to the solution to Equation (86), i.e.

$$\mathbf{P}(\mathbf{x}) \approx \sum_{i=0}^{\infty} \mathbf{T}_i(\mathbf{x}, \theta)\theta^i \tag{89}$$

An efficient procedure for finding the (k_i, l_i) can be determined as follows:

A controller based on $\mathbf{P}(\mathbf{x})$, the solution to Equation (86), is used to generate a state trajectory and then the maximum singular value of $\mathbf{P}(\mathbf{x})$, i.e. $\sigma_{\max}[\mathbf{P}(\mathbf{x})]$, is computed at each state point. Similarly, the (k_i,l_i) parameters determine $\sum_{i=0}^{\infty} \mathbf{T}_i(\mathbf{x}, \theta)\theta^i$ and its associated $\sigma_{\max}[\sum_{i=0}^{\infty} \mathbf{T}_i(\mathbf{x}, \theta)\theta^i]$. Curve fits are then applied to $\sigma_{\max}[\mathbf{P}(\mathbf{x})]$ and $\sigma_{\max}[\sum_{i=0}^{\infty} \mathbf{T}_i(\mathbf{x}, \theta)\theta^i]$. The (k_i,l_i) are selected to minimize the difference between these singular value histories in a least-square sense. The (k_i,l_i) can all be determined in one least-square run off-line.

3. ILLUSTRATIVE EXAMPLES

3.1. A scalar benchmark problem

Consider a simple scalar non-linear regulator problem [25].

Minimize

$$J = \frac{1}{2}\int_0^{\infty} (x^2 + u^2)\, \mathrm{d}t \tag{90}$$

with respect to x and u subject to the constraint

$$\dot{x} = x - x^3 + u \tag{91}$$

This benchmark problem is usually used to illustrate the potential pitfalls of feedback linearization control method [26]. A feedback linearizing controller for this problem is

$$u_{fl} = x^3 - 2x \tag{92}$$

that cancels the beneficial non-linearity term $-x^3$. For large x it requires large control activity that can cause instability in the presence of actuator saturation or uncertainties.

Freeman and Kokotovic [25] solved the HJB equation to obtain the optimal control for this problem. Bellman and Bucy [27] also described a general method for solving such problems. The expression of the optimal control is given by

$$u_{\text{opt}} = -(x - x^3) - x\sqrt{x^4 - 2x^2 + 2} \tag{93}$$

Employing the $\theta - D$ technique, we use the linear factorization:

$$A_0 = 1, \quad A(x) = -x^2, \quad g = 1 \tag{94}$$

with $Q = 1$ and $R = 1$.

Following the algorithm in Equations (11)–(14) we can get $T_0 = 1 + \sqrt{2}$

$$T_1 = -\frac{1}{2\sqrt{2}}\left[2\frac{x^2\cdot(1+\sqrt{2})}{\theta} - D_1\right]$$

$$T_2 = -\frac{1}{2\sqrt{2}}\left\{2x^2\cdot\frac{-1}{2\sqrt{2}}\left[2\frac{x^2\cdot(1+\sqrt{2})}{\theta^2} - D_1\right] + \frac{1}{8}\left[2\frac{x^2\cdot(1+\sqrt{2})}{\theta} - D_1\right]^2 - D_2\right\}$$

If there is no D_1 and D_2, T_1 and T_2 becomes

$$T_1 = -\frac{x^2\cdot(1+\sqrt{2})}{\sqrt{2}\theta}, \quad T_2 = \frac{x^4}{4\sqrt{2}\theta^2}$$

As can be seen, the x^2 terms from $A(x)$ is reflected into the solution of T_1 and amplified further into T_2.

Figure 1 shows the state and control responses when the initial states are 1 using only the θ approximation without the D_i terms. We can see that it works as good as the optimal solution. However, Figure 2 shows the responses when $x_0 = 10$. The control response is a zoomed plot in Figure 2. Notice that the initial control level is of the order of 10^4. This phenomenon also happens in the feedback linearization method. In contrast, the new method developed in this paper does not have such problem. Figure 3 demonstrates the results when D_1 and D_2 terms are incorporated as

$$D_1 = 0.98\mathrm{e}^{-2t}\left[-\frac{T_0A(x)}{\theta} - \frac{A^{\mathrm{T}}(x)T_0}{\theta}\right] \tag{95}$$

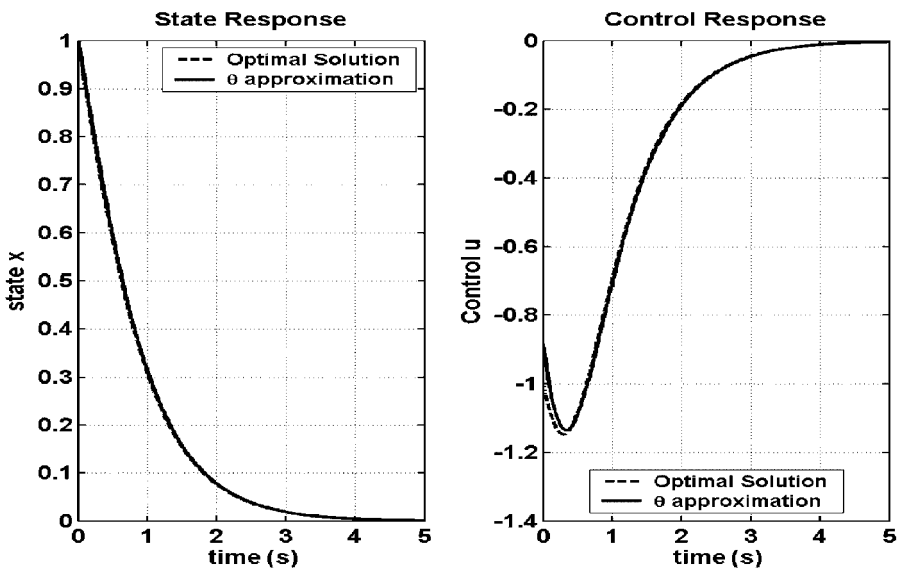

Figure 1. Scalar problem: $x_0 = [1,1]$ without $\mathbf{D}_i$ term.

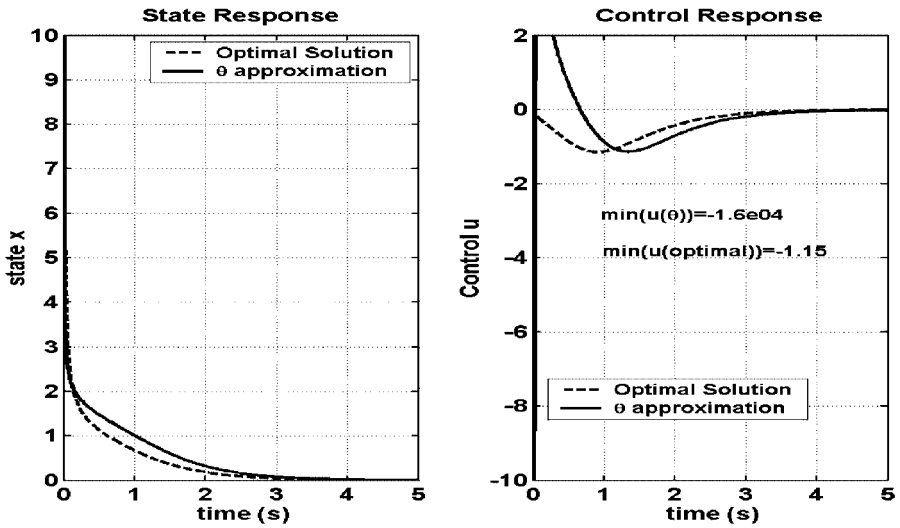

Figure 2. Scalar problem: $x_0 = [10,10]$ without $\mathbf{D}_i$ term.

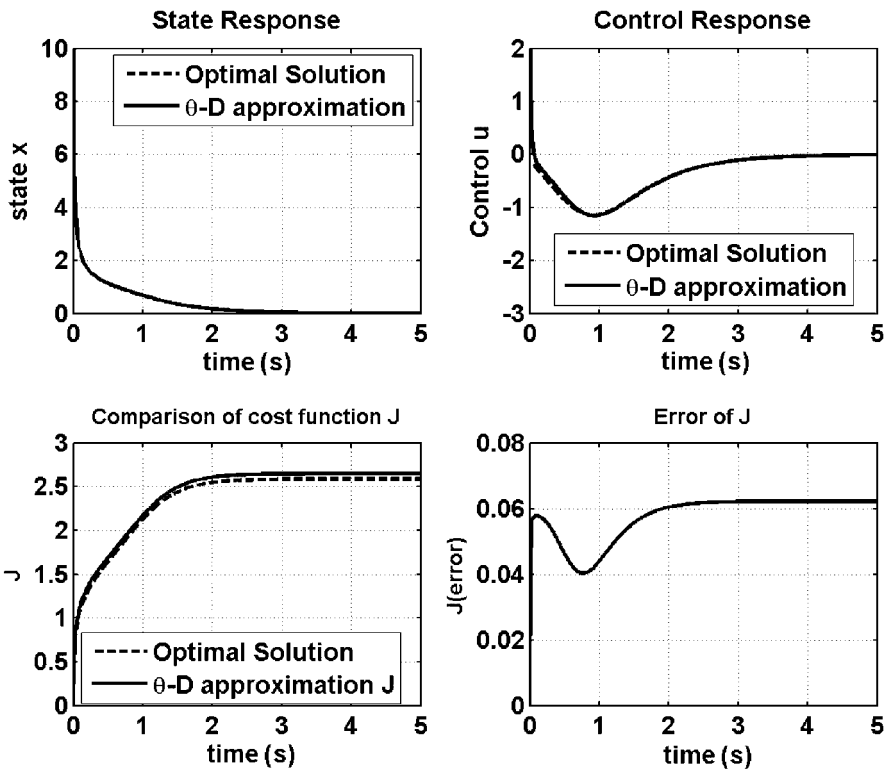

Figure 3. Scalar problem: $x_0 = [10,10]$ with $\mathbf{D}_i$ term included.

$$D_2 = 0.98\mathrm{e}^{-0.9t}\left[-\frac{T_1 A(x)}{\theta} - \frac{A^{\mathrm{T}}(x)T_1}{\theta} + T_1 g R^{-1} g^{\mathrm{T}} T_1\right] \tag{96}$$

The parameters in D_1 and D_2 terms are chosen using the method described in Section 2.3. The numerical experiment with these parameters also shows that the system performance is not sensitive to the variations around these selected values. Note that in the scalar case, Equation (86) gives the exact solution to the optimal control problem [11]. As shown in Figure 3, the maximum control level drops down to less than 2. The comparison of cost functions between the optimal solution and the θ–D approximation is shown in Figure 3. The error in the cost is very small. Note that the comparison is based on the same cost function (90).

Remark 3.1
As a summary, the construction of $\mathbf{D}_i$ in (16)–(18) serves three functions. The first is to suppress the large control if it happens. The second is to provide an appropriate ε_i to guarantee the convergence of power series expansion $\sum_{i=0}^{\infty} \mathbf{T}_i(\mathbf{x}, \theta)\theta^i$ and stability of the closed loop system. The third function is to allow flexibility to modulate the system transient performance by tuning the parameters k_i and l_i in the $\mathbf{D}_i$.

3.2. A 2-D benchmark problem [11]

Finding a control $\mathbf{u} = [u_1\ u_2]^{\mathrm{T}}$to minimize the cost function:

$$J = \frac{1}{2}\int_0^{\infty} \left\{ \mathbf{x}^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{x} + \mathbf{u}^{\mathrm{T}} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \mathbf{u} \right\} \mathrm{d}t \tag{97}$$

with a system described by

$$\dot{x}_1 = x_1 - x_1^3 + x_2 + u_1 \tag{98}$$

$$\dot{x}_2 = x_1 + x_1^2 x_2 - x_2 + u_2 \tag{99}$$

For this problem, $\mathbf{f}(\mathbf{x})$ is factorized as

$$\mathbf{A}_0 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} -x_1^2 & 0 \\ 0 & x_1^2 \end{bmatrix} \tag{100}$$

For analysis, we will compare the θ–D approximation technique with the SDRE method since there is no analytical optimal solution. Figure 4 shows the state and control responses when initial condition is [1,1] (using only θ approximation without $\mathbf{D}_i$ terms). The θ approximation is close to the SDRE solution. Figure 5 presents the results when θ–D controller is applied. We only use the first three terms, e.g. up to $\mathbf{T}_2$ terms in the $\boldsymbol{\lambda}$ expansion (good enough for approximation in this problem as well as some others that we have solved [21,24]). $\mathbf{D}_1$ and $\mathbf{D}_2$ terms are selected using the approach described in Section 2.3 as

$$\mathbf{D}_1 = \mathrm{diag}\{1 \times \mathrm{e}^{-210t}, 1 \times \mathrm{e}^{-60t}\}\left[-\frac{\mathbf{T}_0 \mathbf{A}(x)}{\theta} - \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{x})\mathbf{T}_0}{\theta}\right] \tag{101}$$

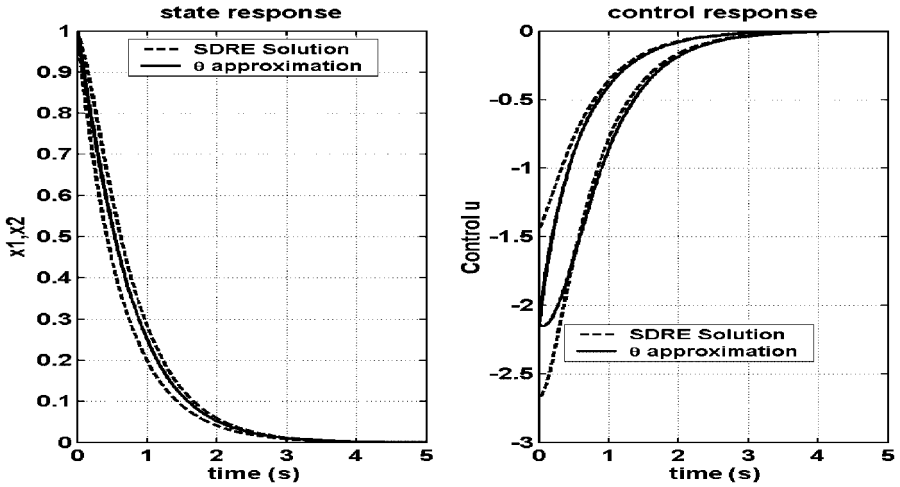

Figure 4. 2-D vector problem: $x_0=[1,1]$ without $\mathbf{D}_i$ term.

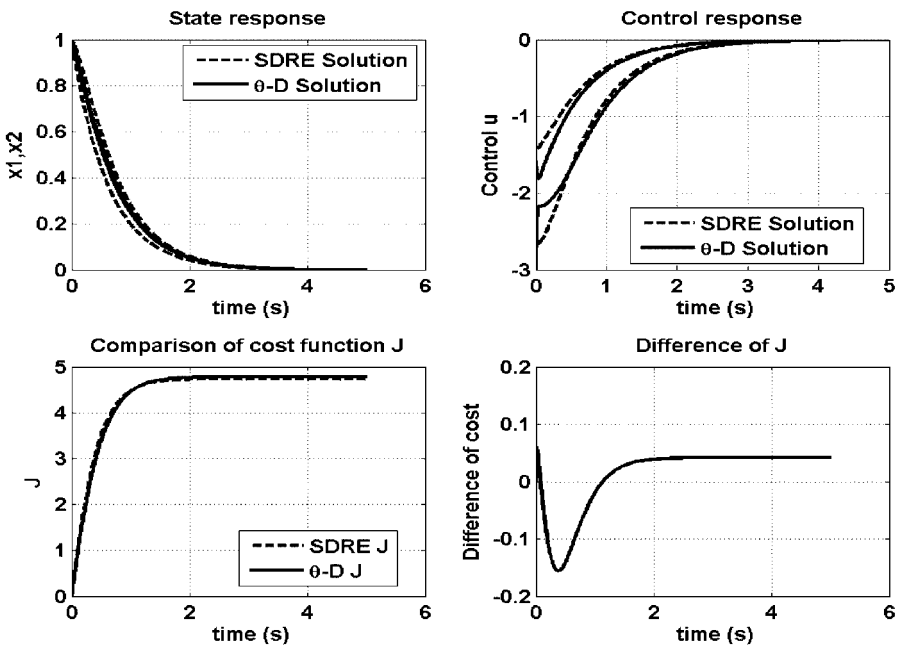

Figure 5. 2-D vector problem: $x_0=[1,1]$ with $\mathbf{D}_i$ term included.

$$\mathbf{D}_2 = \text{diag}\{1 \times e^{-2100t}, 1 \times e^{-1t}\}\left[-\frac{\mathbf{T}_1\mathbf{A}(\mathbf{x})}{\theta} - \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{x})\mathbf{T}_1}{\theta} + \mathbf{T}_1\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_1\right] \tag{102}$$

As can be seen in Figure 5, the θ–D result is also close to the SDRE solution. Figure 5 also gives the cost function comparison which shows that the costs produced by these two method are very close.

Figure 6 shows the results when the initial states are [10,10] using the θ approximation. The control response is a zoomed plot in Figure 6. The maximum magnitude of the control is 10^4 while SDRE control level is less than 80.

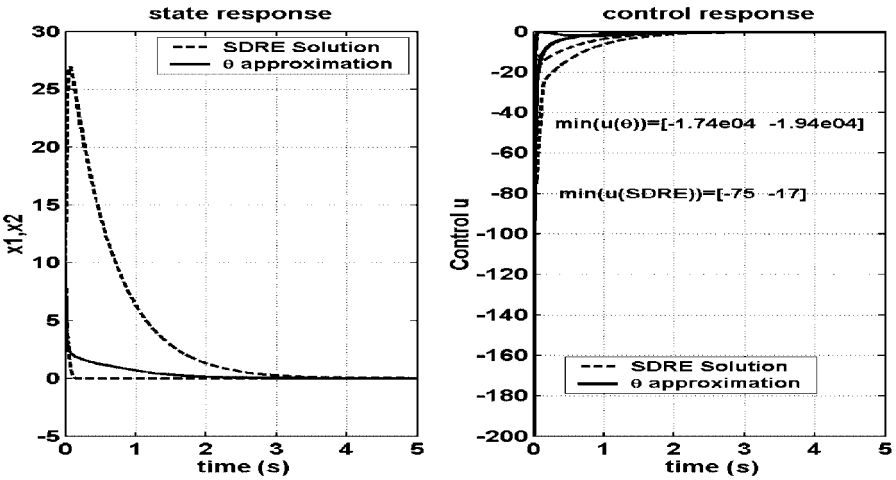

Figure 6. 2-D vector problem: $x_0 = [10,10]$ without $\mathbf{D}_i$ terms.

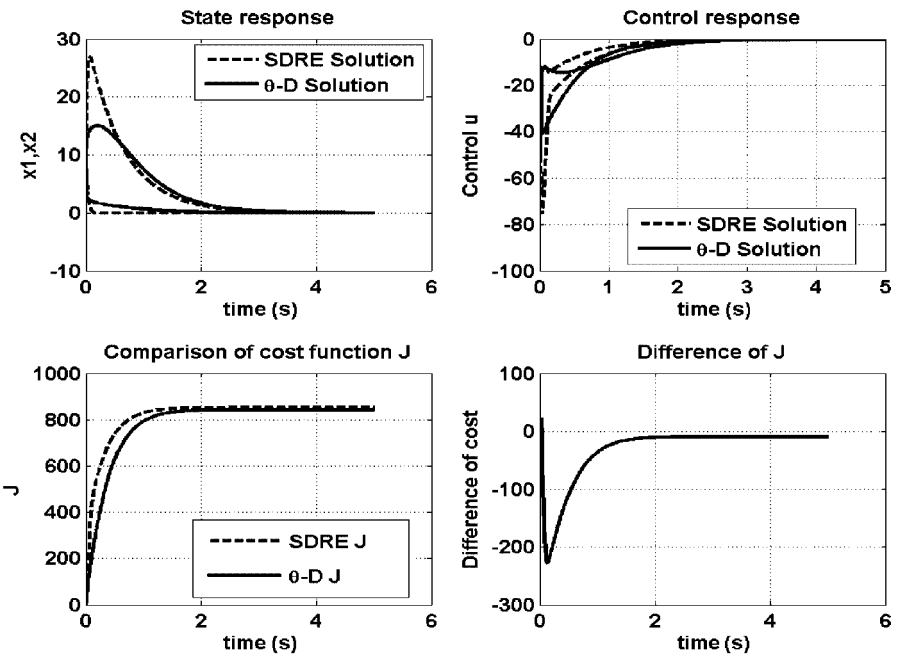

Figure 7. 2-D vector problem: $x_0 = [10,10]$ with $\mathbf{D}_i$ term included.

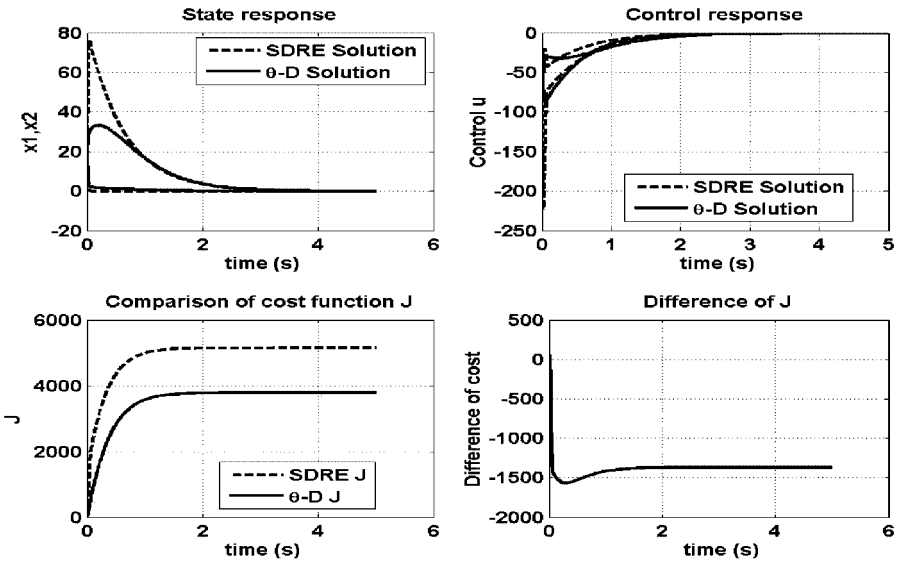

Figure 8. 2-D vector problem: $x_0 = [20,20]$ with $\mathbf{D}_i$ term included.

Figure 7 demonstrates the effect of $\mathbf{D}_i$ terms (θ–D approximation). The maximum magnitude of control level is reduced to about -40. The costs by both methods are close while the θ–D method's is less. Note that the comparison is based on the same cost function (97).

Another advantage of this method is that once we find appropriate parameters of the $\mathbf{D}_i$ matrix, they are not sensitive to the variation of the initial states. In Figure 8, we keep the same parameters in (101) and (102) and show the responses under different large initial states. As can be seen, the θ–D method is not sensitive to the variation of $\mathbf{x}_0$. Compared to the SDRE method, θ–D approach keeps smaller initial control and lower cost. These parameters in the $\mathbf{D}_i$ matrix can be adjusted *off-line* to achieve the desired performance.

As far as implementation is concerned, the θ–D method gives a suboptimal closed-form solution. When implemented online, this method involves only two 2×2 matrix multiplications and three 2×2 matrix additions if we take three terms. However, in comparison, SDRE needs computation of the 2×2 algebraic Riccati equation at *each* sample state. The number of computations will become significant if we want to solve higher-order problems with the SDRE method. Some successful applications of the θ–D method to higher order problems can be referred to References [21] and [24].

4. CONCLUSIONS

In this paper, a new suboptimal non-linear control technique was proposed. This technique can be used to solve a class of non-linear optimal control problems where the model is affine in control and the cost to be minimized is a quadratic cost function. The recursive algorithm results in a closed-form non-linear feedback controller which does not need intensive on-line computations compared to the SDRE technique. In addition, the large control problem encountered in other Talyor series expansion methods was overcome by manipulating the

perturbation terms appropriately. Two illustrative benchmark examples demonstrated the good results and favourable comparisons with other methods. This feedback technique is applicable to a broad class of engineering problems.

APPENDIX A

A.1. Proof of Claim 2.1

In the θ–D approximation, it is assumed that $V_{\mathbf{x}} = \sum_{n=0}^{\infty} \mathbf{T}_n \theta^n \mathbf{x}$ in HJB equation and the suboptimal control

$$\mathbf{u} = -\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}} V_{\mathbf{x}} = -\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}} \sum_{n=0}^{\infty} \mathbf{T}_n \theta^n \mathbf{x} \tag{A1}$$

The cost function is

$$J = \frac{1}{2}\int_0^{\infty} [\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u}]\, \mathrm{d}t \tag{A2}$$

When suboptimal control (A1) is substituted into (A2), J would also be a power series in terms of θ. Now it is needed to prove the convergence of this series.

$$J = \frac{1}{2}\int_0^{\infty} \left[\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{x}^{\mathrm{T}} \sum_{n=0}^{\infty} \mathbf{T}_n \theta^n \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}} \sum_{n=0}^{\infty} \mathbf{T}_n \theta^n \mathbf{x}\right] \mathrm{d}t \tag{A3}$$

The nth term in this power series of J in terms of θ is

$$\begin{aligned} \theta^n J_n &= \theta^n \int_0^{\infty} \left[2\mathbf{x}^{\mathrm{T}}\mathbf{T}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_n\mathbf{x} + \mathbf{x}^{\mathrm{T}} \sum_{k=1}^{n-1} \mathbf{T}_k \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_{n-k}\mathbf{x}\right] \mathrm{d}t \\ &= \theta^n \int_0^{\infty} \mathbf{x}^{\mathrm{T}} \left(2\mathbf{T}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_n + \sum_{k=1}^{n-1} \mathbf{T}_k \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_{n-k}\right)\mathbf{x}\, \mathrm{d}t \end{aligned} \tag{A4}$$

Then

$$\begin{aligned} \theta^n ||J_n|| &\leqslant \theta^n \int_0^{\infty} ||\mathbf{x}||^2 \left|\left|2\mathbf{T}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_n + \sum_{k=1}^{n-1} \mathbf{T}_k \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_{n-k}\right|\right| \mathrm{d}t \\ &\leqslant \theta^n \int_0^{\infty} ||\mathbf{x}||^2 \left(||2\mathbf{T}_0\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}||\, ||\mathbf{T}_n|| + \sum_{k=1}^{n-1} ||\mathbf{T}_k||\, ||\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}||\, ||\mathbf{T}_{n-k}||\right) \mathrm{d}t \end{aligned} \tag{A5}$$

Recall that $||\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}|| = C_g$ (constant). Then

$$\theta^n ||J_n|| \leqslant \theta^n \int_0^{\infty} ||\mathbf{x}||^2 \left(2C_g||\mathbf{T}_0|| \cdot ||\mathbf{T}_n|| + C_g \sum_{k=1}^{n-1} ||\mathbf{T}_k|| \cdot ||\mathbf{T}_{n-k}||\right) \mathrm{d}t \tag{A6}$$

According to (61), it can be obtained

$$||\mathbf{T}_n|| \leqslant \frac{\varepsilon_1 \cdots \varepsilon_n C^n C_1^n C_2 \ldots C_n ||\mathbf{T}_0||}{\theta^n} \tag{A7}$$

Then

$$\begin{aligned}
\theta^n ||J_n|| &\leqslant \theta^n \int_0^\infty ||\mathbf{x}||^2 \left(2C_g ||\mathbf{T}_0||^2 \cdot \frac{\varepsilon_1 \cdots \varepsilon_n C^n C_1^n C_2 \cdots C_n}{\theta^n} \right. \\
&\quad \left. + C_g \sum_{k=1}^{n-1} \frac{(\varepsilon_1 \cdots \varepsilon_k C^k C_1^k C_2 \cdots C_k) \cdot (\varepsilon_1 \cdots \varepsilon_{n-k} C^{n-k} C_1^{n-k} C_2 \cdots C_{n-k})}{\theta^n} ||\mathbf{T}_0||^2 \right) \mathrm{d}t \\
&\leqslant \theta^n \int_0^\infty \frac{||\mathbf{x}||^2 \cdot C_g \cdot ||\mathbf{T}_0||^2 C^n C_1^n}{\theta^n} \left[2(C_2 \cdots C_n)(\varepsilon_1 \cdots \varepsilon_n) \right. \\
&\quad \left. + \sum_{k=1}^{n-1} (C_2 \cdots C_k)(C_2 \cdots C_{n-k})(\varepsilon_1 \cdots \varepsilon_k)(\varepsilon_1 \cdots \varepsilon_{n-k}) \right] \mathrm{d}t
\end{aligned} \tag{A8}$$

Since the stability of θ–D approximation has been proved, the closed loop system can be written as

$$\dot{\mathbf{x}} = \underbrace{\left[\mathbf{A}_0 + \mathbf{A}(\mathbf{x}) - \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}} \sum_{i=0}^{\infty} \mathbf{T}_i(\mathbf{x}, \theta)\theta^i \right]}_{k(\mathbf{x},\theta)} \mathbf{x} \tag{A9}$$

where $k(\mathbf{x}, \theta)$ should be a pointwise Hurwitz matrix.

Thus

$$||\mathbf{x}|| \leqslant C_0 ||\mathbf{x}_0|| \mathrm{e}^{-\lambda t} \tag{A10}$$

where

$$0 > -\lambda \geqslant \lambda_{\max}[k(\mathbf{x}, \theta)] \tag{A11}$$

Then

$$||\mathbf{x}||^2 \leqslant C_0^2 ||\mathbf{x}_0||^2 \mathrm{e}^{-2\lambda t} \tag{A12}$$

and

$$\theta^n J_n \leqslant \theta^n \cdot \frac{C_g ||\mathbf{T}_0||^2 C^n C_1^n M_n(\varepsilon)}{\theta^n} \int_0^\infty C_0^2 ||\mathbf{x}_0||^2 \mathrm{e}^{-2\lambda t} \, \mathrm{d}t \tag{A13}$$

where

$$M_n(\varepsilon) = \left[2(C_2 \cdots C_n)(\varepsilon_1 \cdots \varepsilon_n) + \sum_{k=1}^{n-1} (C_2 \cdots C_k)(C_2 \cdots C_{n-k})(\varepsilon_1 \cdots \varepsilon_k)(\varepsilon_1 \cdots \varepsilon_{n-k}) \right] \tag{A14}$$

Thus

$$\theta^n J_n \leqslant \frac{C_0^2 \|\mathbf{x}_0\|^2}{2\lambda} C_g \|\mathbf{T}_0\|^2 C^n C_1^n \cdot M_n(\varepsilon) \tag{A15}$$

From (A14) we have

$$M_n(\varepsilon) = O(\varepsilon_1 \cdots \varepsilon_n) \tag{A16}$$

or

$$M_n(\varepsilon) < C_\varepsilon \cdot (\varepsilon_1 \cdots \varepsilon_n) \tag{A17}$$

where C_ε is a constant.

Then we have

$$\theta^n J_n \leqslant \frac{C_0^2 \|\mathbf{x}_0\|^2}{2\lambda} \cdot C_g \cdot \|\mathbf{T}_0\|^2 \cdot C^n C_1^n \cdot C_\varepsilon \cdot (\varepsilon_1 \cdots \varepsilon_n) \tag{A18}$$

Define a series $\sum_{n=0}^{\infty} U_n$ with

$$U_n = \frac{C_0^2 \|\mathbf{x}_0\|^2}{2\lambda} \cdot C_g \cdot \|\mathbf{T}_0\|^2 \cdot C^n C_1^n \cdot C_\varepsilon \cdot (\varepsilon_1 \cdots \varepsilon_n) \tag{A19}$$

Thus

$$\frac{U_n}{U_{n-1}} = C \cdot C_1 \varepsilon_n \tag{A20}$$

By choosing sufficiently small ε_n such that $\lim_{n\to\infty} (U_n/U_{n-1}) < 1$, we can make $\sum_{n=0}^{\infty} U_n$ a convergent series. Since each $\theta^n J_n \leqslant U_n$, $\sum_{n=0}^{\infty} \theta^n J_n$ is also a convergent series. □

A.2. Proof of Claim 2.2

Consider (A3)

$$J = \frac{1}{2} \int_0^\infty \left(\mathbf{x}^{\mathrm{T}} \mathbf{Q} \mathbf{x} + \mathbf{x}^{\mathrm{T}} \sum_{n=0}^{\infty} \mathbf{T}_n \theta^n \mathbf{g} \mathbf{R}^{-1} \mathbf{g}^{\mathrm{T}} \sum_{n=0}^{\infty} \mathbf{T}_n \theta^n \mathbf{x} \right) \mathrm{d}t \tag{A3}$$

Now assume that m terms in $\mathbf{u}$ are used, e.g.

$$\mathbf{u}^m = \sum_{n=0}^{m} \mathbf{T}_n \theta^n \mathbf{x} \tag{A21}$$

Denote the cost function associated with $\mathbf{u}^m$ as $J^m(\mathbf{x}, \mathbf{u}^m)$.

Then

$$J^m = \frac{1}{2}\int_0^\infty \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{x}^{\mathrm{T}} \sum_{n=0}^{m} \mathbf{T}_n \theta^n \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}} \sum_{n=0}^{m} \mathbf{T}_n \theta^n \mathbf{x} \right) \mathrm{d}t \tag{A22}$$

Comparing Equation (A22) and (A3), the error in the cost becomes

$$E_J = J - J^m = \frac{1}{2}\int_0^\infty \mathbf{x}^{\mathrm{T}} \sum_{k=2m+1}^{N\to\infty} \mathbf{T}_k \theta^k \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\mathbf{T}_{N-k}\theta^{N-k}\mathbf{x}\, \mathrm{d}t \tag{A23}$$

and

$$\|E_J\| \leqslant \frac{1}{2}\int_0^\infty \|\mathbf{x}\|^2 \sum_{k=2m+1}^{N\to\infty} \|\hat{\mathbf{T}}_k \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_{N-k}\|\, \mathrm{d}t \tag{A24}$$

From (A12) the following inequalities are obtained:

$$\begin{aligned} \|E_J\| &\leqslant \frac{1}{2}\sum_{k=2m+1}^{N\to\infty} \|\hat{\mathbf{T}}_k \mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\hat{\mathbf{T}}_{N-k}\| \int_0^\infty C_0^2 \|\mathbf{x}_0\|^2 \mathrm{e}^{-2\lambda t}\, \mathrm{d}t \\ &\leqslant \frac{C_0^2\|\mathbf{x}_0\|_2^2}{4\lambda} \sum_{k=2m+1}^{N\to\infty} \|\hat{\mathbf{T}}_k\|\, \|\mathbf{g}\mathbf{R}^{-1}\mathbf{g}^{\mathrm{T}}\|\, \|\hat{\mathbf{T}}_{N-k}\| \\ &\leqslant \frac{C_0^2\|\mathbf{x}_0\|^2 C_g}{4\lambda} \sum_{k=2m+1}^{N\to\infty} \varepsilon_k \varepsilon_{N-k} \cdot \|\bar{\mathbf{T}}_k\|\, \|\bar{\mathbf{T}}_{N-k}\| \end{aligned} \tag{A25}$$

where $\bar{\mathbf{T}}_k$ is defined in Equation (83). Since $\|\bar{\mathbf{T}}_k(\mathbf{x})\|$ is bounded on the compact set Ω, let's assume that

$$M_k = \max_{\mathbf{x}\in\Omega} \|\bar{\mathbf{T}}_k(\mathbf{x})\|\, \|\bar{\mathbf{T}}_{N-k}(\mathbf{x})\| \quad \text{where } k = 1, \ldots, N \tag{A26}$$

Thus

$$\|E_J\| \leqslant \frac{C_0^2\|\mathbf{x}_0\|^2 C_g M_k}{4\lambda} \sum_{k=2m+1}^{N\to\infty} \varepsilon_k \varepsilon_{N-k} \tag{A27}$$

Therefore, the error in the cost satisfies $\|E_J\| = O(\varepsilon_k \varepsilon_{N-k})$ for $k \geqslant 2m+1$ and sufficiently small ε_k. □

ACKNOWLEDGEMENTS

Grant from Anteon Corporation in support of this study for Naval Surface Warfare Center is gratefully acknowledged.

REFERENCES

1. Bryson AE, Ho Y-C. *Applied Optimal Control*. Hemisphere Publishing Corporation: New York, 1975.
2. Al'Brekht EG. On the optimal stabilization of non-linear systems. *Journal of Applied Mathematics and Mechanics* 1961; **25**:1254–1266.
3. Lukes DL. Optimal regulation of non-linear dynamical systems. *SIAM Journal of Control and Optimization* 1969; **7**:75–100.
4. Garrard WL, McClamroch NH, Clark LG. An approach to suboptimal feedback control of non-linear systems. *International Journal of Control* 1967; **5**:425–435.
5. Nishikawa Y, Sannomiya N, Itakura H. A method for suboptimal design of non-linear feedback systems. *Automatica* 1971; **7**:703–712.
6. Wernli A, Cook G. Suboptimal control for the non-linear Quadratic Regulator Problem. *Automatica* 1975; **11**:75–84.
7. Garrard WL. Design of non-linear automatic flight control systems. *Automatica* 1977; **13**:497–505.
8. Garrard WL, Enns DF, Snell SA. Non-linear feedback control of highly manoeuverable aircraft. *International Journal of Control* 1992; **56**:799–812.
9. Zhang YL, Gao JC, Zhou CH. Optimal regulation of non-linear systems. *International Journal of Control* 1989; **50**:993–1000.
10. Krikelis NJ, Kiriakidis KI. Optimal feedback control of non-linear systems. *International Journal of Systems Science* 1992; **3**:2141–2153.
11. Cloutier JR, D'Souza CN, Mracek CP. Non-linear regulation and non-linear H_∞ control via the state-dependent Riccati equation technique. *Proceedings of the 1st International Conference on Non-linear Problems in Aviation and Aerospace*, Daytona Beach: FL, May 1996.
12. Saridis GN, Lee CG. An approximation theory of optimal control for trainable manipulators. *IEEE Transactions on Systems, Man, and Cybernetics* 1979; **9**:152–159.
13. Saridis GN, Wang FY. Suboptimal control of non-linear stochastic systems. *Control Theory and Advanced Technology* 1994; **10**(Part 1):847–871.
14. Beard RW, Saridis GN, Wen JT. Galerkin approximations of the generalized Hamilton–Jacobi–Bellman equation. *Automatica* 1997; **33**:2156–2177.
15. Sepulchre R, Jankovic M, Kokotovic PV. *Constructive Non-linear Control*. Springer: New York, 1997.
16. Pan Z, Ezal K, Krener AJ, Kokotovic PV. Backstepping design with local optimality matching. *IEEE Transactions on Automatic Control* 2001; **46**:1014–1027.
17. Margaliot M, Langholz G. Some non-linear optimal control problems with closed-form solutions. *International Journal of Robust and Non-linear Control* 2001; **11**:1365–1374.
18. Huang Y, Lu WM. Non-linear optimal control: alternatives to Hamilton–Jacobi equation. *Proceedings of the 35th Conference on Decision and Control*, Kobe, Japan, 1996.
19. Lu WM, Doyle J. H_∞ Control of non-linear systems: a convex characterization. *IEEE Transactions on Automatic Control* 1995; **40**:1668–1675.
20. Doyle J, Huang Y, Primbs J, Freeman R, Murray R, Packard A, Krstic M. Non-linear control: comparisons and case studies. *Notes from the Non-linear Control Workshop conducted at the American Control Conference*, Alburquerque, NM, June 1997.
21. Xin M, Balakrishnan SN, Stansbery DT, Ohlmeyer EJ. Non-linear missile autopilot design with theta-D technique. *Journal of Guidance, Control and Dynamics* 2004; **27**:406–417.
22. Mori T, Derese IA. A brief summary of the bounds on the solution of the algebraic matrix equations in control theory. *International Journal of Control* 1984; **39**:247–256.
23. Horn RA, Johnson CR. *Matrix Analysis*. Cambridge University Press: Cambridge, 1991.
24. Drake D, Xin M, Balakrishnan SN. A new non-linear control technique for ascent phase of reusable launch vehicles. *AIAA Journal of Guidance, Control and Dynamics* 2004; **27**:938–948.
25. Freeman RA, Kokotovic PV. Optimal non-linear controllers for feedback linearizable systems. *Workshop on Robust Control via Variable Structure and Lyapunov Technique*, Benevento, Italy, September 1994.
26. Slotine J-JE, Li W. *Applied Non-linear Control*. Prentice-Hall: Englewood Cliffs, NJ, 1991.
27. Bellman R, Bucy R. Asymptotic Control Theory. *Journal of SIAM Control* 1964; **2**:11–18.

