



**HAL**  
open science

## Homographic scheme for Riccati equation

François Dubois, Abdelkader Saïdi

► **To cite this version:**

| François Dubois, Abdelkader Saïdi. Homographic scheme for Riccati equation. 2000. hal-00554484v2

**HAL Id: hal-00554484**

**<https://hal.science/hal-00554484v2>**

Preprint submitted on 8 May 2011

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Homographic scheme for Riccati equation $\square$

FRANÇOIS DUBOIS <sup>ab</sup> AND ABDELKADER SAÏDI <sup>c</sup>

<sup>a</sup> Conservatoire National des Arts et Métiers  
15 rue Marat, F-78 210, Saint Cyr l'Ecole, France.

<sup>b</sup> Applications Scientifiques du Calcul Intensif  
bât. 506, BP 167, F-91 403 Orsay Cedex, France.

<sup>c</sup> Université Louis Pasteur  
Institut de Recherche Mathématique Avancée  
7 rue René Descartes, F-67 084 Strasbourg, France.  
dubois@asci.fr, saidi@math.u-strasbg.fr.

## Abstract

In this paper we present a numerical scheme for the resolution of matrix Riccati equation, usually used in control problems. The scheme is unconditionally stable and the solution is definite positive at each time step of the resolution. We prove the convergence in the scalar case and present several numerical experiments for classical test cases.

**Keywords:** control problems, ordinary differential equations, stability.

**AMS classification:** 34H05, 49K15, 65L20, 93C15.

## Contents

- 1) Introduction
- 2) Scalar Riccati equation
- 3) Matrix Riccati equation
- 4) Numerical experiments
- 5) Conclusion
- 6) Acknowledgments
- 7) References.

---

$\square$  Rapport CNAM-IAT n° 338-2K, 29 août 2000.

Rapport de recherche n° 2000-32 du laboratoire IRMA de l'Université Louis Pasteur à Strasbourg. Edition du 5 mai 2011.

1) INTRODUCTION.

- We study the optimal control of a differential linear system

$$(1.1) \quad \frac{dy}{dt} = Ay + Bv,$$

where the state variable  $y(\bullet)$  belongs to  $\mathbb{R}^n$  and the control variable  $v(\bullet)$  belongs to  $\mathbb{R}^m$ , with  $n$  and  $m$  be given integers :

$$(1.2) \quad y(t) \in \mathbb{R}^n, \quad v(t) \in \mathbb{R}^m.$$

Matrix  $A$  is composed by  $n$  lines and  $n$  columns and matrix  $B$  contains  $n$  lines and  $m$  columns. Both matrices  $A$  and  $B$  are fixed relatively to the evolution in time. Ordinary differential equation (1.1) is associated with an initial condition

$$(1.3) \quad y(0) = y_0$$

with  $y_0$  be given in  $\mathbb{R}^n$ . Moreover the solution of system (1.1) (1.3) is parametrized by the function  $v(\bullet)$  and instead of the short notation  $y(t)$ , we can set more precisely

$$(1.4) \quad y(t) = y(t; y_0, v(\bullet)).$$

The control problem consists of finding the minimum  $u(t)$  of some quadratic functional  $J(\bullet)$  :

$$(1.5) \quad J(u(\bullet)) \leq J(v(\bullet)), \quad \forall v(\bullet).$$

The functional  $J(\bullet)$  depends on the control variable function  $v(\bullet)$ , is additive relatively to the time and represents the coast function. We set classically :

$$(1.6) \quad J(v(\bullet)) = \frac{1}{2} \int_0^T (Qy(t), y(t))dt + \frac{1}{2} \int_0^T (Rv(t), v(t))dt + \frac{1}{2} (Dy(T), y(T)).$$

Functional  $J(\bullet)$  is parametrized by the horizon  $T > 0$ , the symmetric semi-definite positives  $n$  by  $n$  constant matrices  $Q$  and  $D$  :

$$(1.7) \quad (Qy, y) \geq 0, \quad \forall y \in \mathbb{R}^n, \quad y \neq 0,$$

$$(1.8) \quad (Dy, y) \geq 0, \quad \forall y \in \mathbb{R}^n, \quad y \neq 0.$$

and the symmetric definite positive  $m$  by  $m$  constant matrix  $R$  :

$$(1.9) \quad (Ru, u) > 0, \quad \forall u \in \mathbb{R}^m, \quad u \neq 0.$$

- Problem (1.1) (1.3) (1.5) (1.6) is a classical mathematical modeling of linear quadratic loops for dynamical systems in automatics (see *e.g.* Athans, Falb [AF66], Athans, Levine and Levis [ALL67], Kawakernaak-Sivan [KS72], Faure Robin [FR84], Lewis [Le86]). When control function  $v(\bullet)$  is supposed to be squarely integrable ( $v(\bullet) \in L^2(]0, T[, \mathbb{R}^m)$ ) then the control problem (1.1) (1.3) (1.5) (1.6) has a unique solution  $u(t)$  (see for instance Lions [Li68]) :

$$(1.10) \quad u(t) \in L^2(]0, T[, \mathbb{R}^m).$$

When there is no constraint on the control variable the minimum of functional  $J(\bullet)$  is characterized by the condition :

$$(1.11) \quad dJ(u) \bullet w = 0, \quad \forall w \in L^2(]0, T[, \mathbb{R}^m),$$

which is not obvious to derive because  $y(\bullet)$  is a function of  $v(\bullet)$ . We introduce the differential equation (1.1) as a constraint between  $y(\bullet)$  and  $v(\bullet)$  with the associated Lagrange multiplier  $p$ . We set :

$$(1.12) \quad \mathcal{L}(y, v; p) = J(v) - \int_0^T \left( p, \frac{dy}{dt} - Ay - Bv \right) dt$$

and the variation of  $\mathcal{L}(\bullet)$  under an infinitesimal variation  $\delta y(\bullet)$ ,  $\delta v(\bullet)$  and  $\delta p(\bullet)$  of the other variables can be conducted as follow :

$$\begin{aligned} \delta \mathcal{L} &= \int_0^T (Qy, \delta y) dt + \int_0^T (Rv, \delta v) dt + (Dy(T), \delta y(T)) \\ &\quad - \int_0^T \left( \delta p, \frac{dy}{dt} - Ay - Bv \right) dt - \int_0^T \left( p, \frac{d\delta y}{dt} - A \delta y - B \delta v \right) dt \\ &= \int_0^T (Qy, \delta y) dt + \int_0^T (Rv, \delta v) dt + (Dy(T), \delta y(T)) \\ &\quad - \int_0^T \left( \delta p, \frac{dy}{dt} - Ay - Bv \right) dt - [p \delta y]_0^T + \int_0^T \left( \frac{dp}{dt}, \delta y \right) dt \\ &\quad + \int_0^T (A^t p, \delta y) dt + \int_0^T (B^t p, \delta v) dt, \quad \text{and} \\ (1.13) \quad \left\{ \begin{aligned} \delta \mathcal{L} &= \int_0^T \left( \frac{dp}{dt} + Qy + A^t p, \delta y \right) dt + \int_0^T (Rv + B^t p, \delta v) dt \\ &\quad - \int_0^T \left( \delta p, \frac{dy}{dt} - Ay - Bv \right) dt + (Dy(T) - p(T), \delta y(T)) \end{aligned} \right. \end{aligned}$$

because  $\delta y(0) = 0$  when the initial condition (1.3) is always satisfied by the function  $(y + \delta y)(\bullet)$ .

• The research of a minimum for  $J(\bullet)$  (condition (1.11)) can be rewritten under the form of research of a saddle point for Lagrangien  $\mathcal{L}$  and we deduce from (1.13) the evolution equation for the adjoint variable :

$$(1.14) \quad \frac{dp}{dt} + A^t p + Qy = 0,$$

the final condition when  $t = T$ ,

$$(1.15) \quad p(T) = Dy(T)$$

and the optimal command in terms of the adjoint state  $p(\bullet)$  :

$$(1.16) \quad Ru(t) + B^t p(t) = 0.$$

We observe that the differential system (1.1) (1.14) joined with the initial condition (1.3) and the final condition (1.15) is coupled through the initial optimality condition (1.16). In practice, we need a linear feedback function of the state variable  $y(t)$  instead of the adjoint variable  $p(t)$ . Because adjoint state  $p(\bullet)$  linearly depends on state variable  $y(\bullet)$  we set classically :

$$(1.17) \quad p(T) = X(T-t) y(t), \quad 0 \leq t \leq T,$$

with a symmetric  $n$  by  $n$  matrix  $X(\bullet)$ , positive definite for  $t > 0$  (see *e.g.* [KS72] or [Le86]).

$$(1.18) \quad X(t) \text{ is a symmetric } n \times n \text{ definite positive matrix, } t > 0.$$

The final condition (1.15) is realised for each value  $y(T)$ , then we have the following condition :

$$(1.19) \quad X(0) = D,$$

and introducing the representation (1.17) in the differential equation (1.14) and (1.1) we obtain :

$$-\frac{dX}{dt}(T-t) y(t) + X(T-t) [Ay(t) + Bu(t)] + A^t X(T-t) y(t) + Qy(t) = 0.$$

We replace the control  $u(t)$  by its value obtained in relation (1.16) and we deduce :

$$-\frac{dX}{dt}(T-t) + X(T-t) [A + B(-R^{-1}) B^t X(T-t)] + A^t X(T-t) + Qy(t) = 0.$$

This last equation is realised for each state value  $y(t)$ . Replacing  $t$  by  $T-t$  in this equation, we get :

$$(1.20) \quad \frac{dX}{dt} - (XA + A^t X) + XBR^{-1}BX - Q = 0,$$

which defines the Riccati equation associated with the control problem (1.1) (1.3) (1.5) (1.6).

• In this paper we study the numerical approximation of differential system (1.19) (1.20). Recall that datum matrices  $Q$ ,  $D$  and  $K$ , with  $K$  defined according to :

$$(1.21) \quad K = BR^{-1}B^t,$$

are  $n \times n$  symmetric matrices, with  $Q$  and  $D$  semi-definite positive and  $K$  positive definite ; datum matrix  $A$  is an  $n$  by  $n$  matrix without any other

condition and the unknown matrix  $X(t)$  is symmetric. We have the following property (see *e.g.* [Le86]).

**Proposition 1. Positive definiteness of the solution of Riccati equation.**

Let  $K, Q, D, A$  be given  $n \times n$  matrices with  $K, Q, D$  symmetric matrices,  $Q$  and  $D$  positive matrices and  $K$  a definite positive matrix. Let  $X(\bullet)$  be the solution of the Riccati differential equation :

$$(1.22) \quad \frac{dX}{dt} - (XA + A^t X) + XKX - Q = 0$$

with initial condition (1.19). Then  $X(t)$  is well defined for each  $t \geq 0$ , is symmetric and for each  $t > 0$ ,  $X(t)$  is definite positive and tends to a definite positive matrix  $X_\infty$  as  $t$  tends to infinity :

$$(1.23) \quad X(t) \longrightarrow X_\infty \quad \text{if} \quad t \longrightarrow \infty.$$

Matrix  $X_\infty$  is the unique positive symmetric matrix which is solution of the so-called algebraic Riccati equation :

$$(1.24) \quad -(XA + A^t X) + XKX - Q = 0.$$

- As a consequence of this proposition it is useful to simplify the feedback command law (1.16) by the associated limit command obtained by taking  $t \longrightarrow \infty$ , that is :

$$(1.25) \quad v(t) = -R^{-1} B^t X_\infty y(t),$$

and the differential system (1.1) (1.25) is asymptotically stable (see *e.g.* [Le86]). The practical computation of matrix  $X_\infty$  with direct methods is not obvious and we refer *e.g.* to [La79] for a description of the state of the art. If we wish to compute directly a numerical solution of instationnary Riccati equation (1.22), classical methods for ordinary differential equations like *e.g.* the forward Euler method :

$$(1.26) \quad \frac{1}{\Delta t}(X_{j+1} - X_j) + X_j K X_j - (A^t X_j + X_j A) - Q = 0,$$

or Runge Kutta method as we will see in what follows fail to maintain positivity of the iterate  $X_{j+1}$  at the order  $(j + 1)$  :

$$(1.27) \quad (X_{j+1} x, x) > 0, \quad \forall x \in \mathbb{R}^n, \quad x \neq 0,$$

even if  $X_j$  is positive definite if time step  $\Delta t > 0$  is not small enough, see *e.g.* [Sa97]. Moreover, the stability constraint (1.27) is not classical and there is at our knowledge no simple way to determine a priori if time step  $\Delta t$  is compatible or not with condition (1.27).

- In this paper, we propose a method for numerical integration for Riccati equation (1.22) which maintains condition (1.27) for each time step  $\Delta t > 0$ . We detail in second paragraph the simple case of scalar Riccati equation and prove the convergence for this particular case ; under some constraints on parameters, the scheme is monotonous and remains at the order one of precision, as suggested by results of Dieci and Eirola [DE96]. We present the homographic scheme in the general case in section 3 and establish its principal property : for each time step and without explicit constraint on the time step  $\Delta t$ , the numerical scheme defines a symmetric positive definite matrix. We propose and present four numerical test cases in section 4.

## 2) SCALAR RICCATI EQUATION.

- When the unknown is a scalar variable, we write Riccati equation under the following form :

$$(2.1) \quad \frac{dx}{dt} + kx^2 - 2ax - q = 0,$$

with

$$(2.2) \quad k > 0, \quad q \geq 0,$$

and an initial condition :

$$(2.3) \quad x(0) = d, \quad d \geq 0, \quad a^2 + q^2 > 0.$$

We approach the ordinary differential equation (2.1) with a finite difference scheme of the type proposed by Baraille [Ba91] for hypersonic chemical cinetics and independently with the “family method” proposed by Cariolle [Ca79] and studied by Miellou [Mi84]. We suppose that time step  $\Delta t$  is given strictly positive. The idea that we have proposed in [Du93], [DS95] and [DS2k] is to write the approximation  $x_{j+1}$  at time step  $(j+1)\Delta t$  as a rational fraction of  $x_j$  with positive coefficients. We decompose first the real number  $a$  into positive and negative parts :

$$(2.4) \quad a = a^+ - a^-, \quad a^+ \geq 0, \quad a^- \geq 0, \quad a^+ a^- = 0,$$

and factorise the product  $x^2$  into the very simple form :

$$(2.5) \quad (x^2)_{j+1/2} = x_j x_{j+1}.$$

### **Definition 1. Numerical scheme in the scalar case.**

For resolution of the scalar differential equation (2.1), we define our numerical scheme by the following relation :

$$(2.6) \quad \frac{x_{j+1} - x_j}{\Delta t} + k x_j x_{j+1} - 2a^+ x_j + 2a^- x_{j+1} - q = 0.$$

- The scheme (2.6) is implicit because some equation has to be solved in order to compute  $x_{j+1}$  whereas  $x_j$  is supposed to be given. In the case of our scheme this equation is **linear** and the solution  $x_{j+1}$  is directly obtained with scalar scheme (2.6) by the homographic relation :

$$(2.7) \quad x_{j+1} = \frac{(1 + 2a^+ \Delta t) x_j + q \Delta t}{k \Delta t x_j + (1 + 2a^- \Delta t)}.$$

**Proposition 2. Algebraic properties of the scalar homographic scheme.**

Let  $(x_j)_{j \in \mathbb{N}}$  be the sequence defined by initial condition :

$$(2.8) \quad x_0 = x(0) = d$$

and recurrence relation (2.7). Then sequence  $(x_j)_{j \in \mathbb{N}}$  is globally defined and remains positive for each time step.

$$(2.9) \quad x_j \geq 0, \quad \forall j \in \mathbb{N}, \quad \forall \Delta t > 0.$$

- If  $\Delta t > 0$  is chosen such that :

$$(2.10) \quad 1 + 2|a|\Delta t - kq\Delta t^2 \neq 0,$$

then  $(x_j)_{j \in \mathbb{N}}$  converges towards the positive solution  $x^*$  of the “algebraic Riccati equation” :

$$(2.11) \quad kx^2 - 2ax - q = 0$$

and explicitly computed according to the relation

$$(2.12) \quad x^* = \frac{1}{k} (a + \sqrt{a^2 + kq}).$$

- In the exceptional case where  $\Delta t > 0$  is chosen such that (2.10) is not satisfied, then the sequence  $(x_j)_{j \in \mathbb{N}}$  is equal to the constant  $\frac{1 + 2a^+ \Delta t}{k \Delta t}$  for  $j \geq 1$  and the scheme (2.7) cannot be used for the approximation of Riccati equation (2.1).

**Proof of proposition 2.**

- The proof of the relation (2.9) is a consequence of the fact that the recurrence relation (2.7) defines an homographic function  $f$  :

$$(2.13) \quad x_{j+1} = f(x_j)$$

$$(2.14) \quad f(x) = \frac{\alpha x + \beta}{\gamma x + \delta}$$

with positive coefficients  $\alpha, \beta, \gamma, \delta$  :



$$(2.15) \quad \alpha = 1 + 2a^+ \Delta t, \quad \beta = q \Delta t, \quad \gamma = k \Delta t, \quad \delta = 1 + 2a^- \Delta t.$$

Because  $x_0 = d \geq 0$ , it is then clear that  $x_j \geq 0$  for each  $j \geq 0$  and relation (2.9) is established. The homographic function  $f(\bullet)$  is a constant scalar equal to  $\frac{\alpha}{\gamma} = \frac{\beta}{\delta}$  when the determinant :

$$(2.16) \quad \Delta = \alpha\delta - \beta\gamma$$

is equal to zero. It is the case when condition (2.10) is not realised. When (2.10) holds, we have simply :

$$(2.17) \quad \frac{f'(x)}{f(x)} = \frac{\alpha\delta - \beta\gamma}{(\gamma x + \delta)(\alpha x + \beta)}$$

and  $f$  is a bounded monotonic function on the interval  $]0, \infty[$ . Let  $x^*$  be the positive solution of equation :

$$(2.18) \quad f(x) = x.$$

It is elementary to observe that  $x^*$  is given by the relation (2.12). Let  $x_-$  be defined by :

$$(2.19) \quad x_- = -\frac{q}{k x^*} = x^* - 2 \frac{\sqrt{a^2 + kq}}{k},$$

the other root of equation (2.18). We set :

$$(2.20) \quad u_j = \frac{x^* - x_j}{x_j - x_-}.$$

Then we have :

$$(2.21) \quad u_{j+1} = \frac{\frac{\alpha x^* + \beta}{\gamma x^* + \delta} - \frac{\alpha x_j + \beta}{\gamma x_j + \delta}}{\frac{\alpha x_j + \beta}{\gamma x_j + \delta} - \frac{\alpha x_- + \beta}{\gamma x_- + \delta}} = \frac{\gamma x_- + \delta}{\gamma x^* + \delta} \frac{x_j - x^*}{x_- - x_j},$$

if (2.10) holds. Replacing  $x_-$  by its value given from (2.19) and using (2.21), we obtain :

$$(2.22) \quad u_{j+1} = \frac{\delta x^* - \beta}{\alpha x^* + \beta} u_j.$$

The sequence  $(u_j)_{j \in \mathbb{N}}$  is geometric and the ratio  $\frac{u_{j+1}}{u_j}$  has always an absolute value less than 1. Effectively we have on one hand :

$$-(\alpha x^* + \beta) \leq \delta x^* - \beta$$

because  $\alpha + \delta = 2(1 + |a|\Delta t)$  is strictly positive and  $x^*$  is positive and on the other hand :

$$\delta x^* - \beta \leq \alpha x^* + \beta$$

because  $(\alpha - \delta)x^* + 2\beta = 2\Delta t(ax^* + q) = \Delta t(k(x^*)^2 + q)$  which is a positive number. The absolute value of  $\frac{u_{j+1}}{u_j}$  is not exactly equal to 1 because  $x^* > 0$  and  $ax^* + q = x^*(kx^* - a) = x^*\sqrt{a^2 + kq} > 0$  according to (2.10). Then  $u_j$  is converging to zero and  $x_j$  toward  $x^*$  that completes the proof.  $\square$

**Theorem 1. Convergence of the scalar scheme.**

We suppose that the data  $k, a, q$  of Riccati equation satisfy (2.2) and (2.10) and that the datum  $d$  associated with the initial condition (2.3) is relatively closed to  $x^*$ , ie :

$$(2.23) \quad -\frac{1}{k\tau} + \eta \leq d - x^* \leq C,$$

where  $C$  is some given strictly positive constant ( $C > 0$ ),  $x^*$  calculated according to relation (2.12) is the limit in time of the Riccati equation,  $\tau$  is defined from data  $k, a, q$  according to :

$$(2.24) \quad \tau = \frac{1}{2\sqrt{a^2 + kq}},$$

and  $\eta$  is some constant chosen such that

$$(2.25) \quad 0 < \eta < \frac{1}{k\tau}.$$

• We denote by  $x(t; d)$  the solution of differential equation (2.1) with initial condition (2.3). Let  $(x_j(\Delta t; d_\Delta))_{(j \in \mathbb{N})}$  be the solution of the numerical scheme defined at the relation (2.7) and let  $d_\Delta$  be the initial condition :

$$(2.26) \quad x_0(\Delta t; d_\Delta) = d_\Delta.$$

We suppose that the numerical initial condition  $d_\Delta > 0$  satisfies a condition analogous to (2.23) :

$$(2.27) \quad -\frac{1}{k\tau} + \eta \leq d_\Delta - x^* \leq C,$$

with  $C$  and  $\eta > 0$  equal to the constant introduced in (2.23) and satisfying (2.25) .

• Then the approximated value  $(x_j(\Delta t; d_\Delta))_{j \in \mathbb{N}}$  is arbitrarily closed to the exact value  $x(j\Delta t; d)$  for each  $j$  as  $\Delta t \rightarrow 0$  and  $d_\Delta \rightarrow d$ . More precisely, if  $a \neq 0$  we have the following estimate for the error at time equal to  $j\Delta t$  :

$$(2.28) \quad |x(j\Delta t; d) - x_j(\Delta t; d_\Delta)| \leq A(\Delta t + |d - d_\Delta|), \quad \forall j \in \mathbb{N}, 0 < \Delta t \leq B$$

with constants  $A > 0, B > 0$ , depending on data  $k, a, q, \eta$  but independent on time step  $\Delta t > 0$  and iteration  $j$ .

- If  $a = 0$ , the scheme is second order accurate in the following sense :

$$(2.29) \quad |x(j\Delta t; d) - x_j(\Delta t; d_\Delta)| \leq A(\Delta t^2 + |d - d_\Delta|), \quad \forall j \in \mathbb{N}, 0 < \Delta t \leq B$$

with constants  $A$  et  $B$  independent on time step  $\Delta t$  and iteration  $j$ .

**Remark.**

- A direct application of the Lax theorem (see *e.g.* [La74]) for numerical scheme associated to ordinary differential equations is not straightforward because both Riccati equation and the numerical scheme are nonlinear. Our proof is based on the fact that this problem is algebraically relatively simple.

**Lemma 1.**

Let  $d_\Delta$  be some discrete initial condition and  $x(t; d_\Delta)$  be the exact solution of Riccati differential equation (2.1) associated to initial condition (2.26). We set

$$(2.30) \quad y(t; d_\Delta) = x(t; d_\Delta) - x^*.$$

Then we have

$$(2.31) \quad y(t; d_\Delta) = \frac{(d_\Delta - x^*) e^{-t/\tau}}{1 + k\tau(d_\Delta - x^*)(1 - e^{-t/\tau})}.$$

**Proof of Lemma 1.**

- The real function  $\mathbb{R} \ni t \mapsto y(t; d_\Delta) \in \mathbb{R}$  introduced in relation (2.30) satisfies the equation  $\frac{dy}{dt} + ky^2 + \frac{1}{\tau}y = 0$ , that can be rewritten under the form :

$$\frac{dy}{ky^2 + \frac{y}{\tau}} \equiv \tau \left( \frac{dy}{y} - \frac{k\tau}{1 + k\tau y} dy \right) = -dt.$$

After integration, we have :

$$\frac{y(t; d_\Delta)}{1 + k\tau y(t, d_\Delta)} = \frac{y(0; d_\Delta)}{1 + k\tau y(0, d_\Delta)} e^{-t/\tau},$$

giving simply :

$$y(t; d_\Delta) = \frac{(d_\Delta - x^*) e^{-t/\tau}}{1 + k\tau(d_\Delta - x^*)(1 - e^{-t/\tau})}$$

*i.e.* relation (2.31). Then Lemma 1 is established. □

**Lemma 2.**

Let  $d_\Delta$  be some discrete initial condition associated to (2.26) and  $\epsilon_j$  the error between the exact solution  $x(j\Delta t; d_\Delta)$  of the differential equation and the solution of the numerical scheme  $x_j(\Delta t; d_\Delta)$  :

$$(2.32) \quad \epsilon_j = x(j\Delta t; d_\Delta) - x_j(\Delta t; d_\Delta).$$

Let  $x^*$  and  $x_-$  defined in (2.12) and (2.19) be the two fixed points of the homographic function  $f(\bullet)$  introduced in (2.14) and (2.15). We set

$$(2.33) \quad h(\xi) = \frac{x^* + x_- \xi}{1 + \xi}, \quad \xi > -1$$

and we introduce on one hand the function  $t \mapsto z(t)$  relative to the exact solution :

$$(2.34) \quad z(t) = \frac{x(t; d_\Delta) - x^*}{x_- - x(t, d_\Delta)}$$

and on the other hand a new sequence  $u_j$  by the same algebraic relation :

$$(2.35) \quad u_j = \frac{x_j(\Delta t; d_\Delta) - x^*}{x_- - x_j(\Delta t; d_\Delta)}.$$

Then we have the following estimate :

$$(2.36) \quad |\epsilon_j| \leq |h'(z(0))| |u_j - z(j\Delta t)|.$$

**Proof of Lemma 2.**

• We have constructed function  $h(\bullet)$  in order to have  $h(z(\theta)) = z(h(\theta))$ ,  $\forall \theta \in \mathbb{R}$ . Then we have

$$x(t; d_\Delta) = h(z(t)) \quad \text{for each } t > 0$$

$$x_j(\Delta t, d_\Delta) = h(u_j) \quad \text{for each } j \geq 0,$$

with function  $z(\bullet)$  introduced at relation (2.34) and sequence  $u_j$  in (2.35). Then  $\epsilon_j$  can be rewritten with the help of this function  $h(\bullet)$  and we have :

$$(2.37) \quad \epsilon_j = h(z(j\Delta t)) - h(u_j) \leq \sup_{\xi \in [z(j\Delta t), u_j]} |h'(z(\xi))| |z(j\Delta t) - u_j|.$$

• We note that due to (2.22), the sequence  $u_j$  is a geometric converging one, then we have :  $-|u_0| \leq u_j \leq |u_0|$  for each  $j \geq 0$ . Moreover thanks to relation (2.31) of Lemma 1, we have the following calculus :

$$\begin{aligned}
 z(t) &= \frac{y(t; d_\Delta)}{x_- - x^* - y(t; d_\Delta)} \\
 &= \frac{\frac{(d_\Delta - x^*) e^{-t/\tau}}{1 + k\tau(d_\Delta - x^*)(1 - e^{-t/\tau})}}{-\frac{1}{k\tau} - \frac{(d_\Delta - x^*) e^{-t/\tau}}{1 + k\tau(d_\Delta - x^*)(1 - e^{-t/\tau})}} \\
 &= -\frac{k\tau(d_\Delta - x^*)}{1 + k\tau(d_\Delta - x^*)} e^{-t/\tau},
 \end{aligned}$$

thus satisfy clearly :  $-|z(0)| \leq z(j\Delta t) \leq |z(0)|$  ,  $\forall j \geq 0$ . We remark also that :

$$-1 < z(0) = u_0 = -\frac{k\tau(d_\Delta - x^*)}{1 + k\tau(d_\Delta - x^*)}.$$

• Moreover,  $|h'(\xi)|$  is a decreasing function for  $\xi > -1$ . Then by the two points finite difference Taylor formula and the above statements, we deduce from (2.37) the estimate (2.36) and Lemma 2 is established.  $\square$

**Lemma 3.**

Let  $\tau$  be defined in (2.24). We introduce  $\alpha$  and  $\beta$  according to the following relations

$$(2.38) \quad \alpha = \frac{1}{2\tau} - |a|$$

$$(2.39) \quad \beta = \frac{1}{2\tau} + |a|$$

and when  $\Delta t > 0$  we define function  $\varphi(\Delta t)$  according to

$$(2.40) \quad \varphi(\Delta t) = 1 + \frac{\tau}{\Delta t} \log \left( \frac{1 - \alpha \Delta t}{1 + \beta \Delta t} \right).$$

Then with notations introduced in Lemma 1 and the following one

$$(2.41) \quad \theta_j = \frac{j\Delta t}{\tau},$$

we have

$$(2.42) \quad u_j - z(j\Delta t) = -z(0) e^{-\theta_j} \left[ 1 - \exp(\theta_j \varphi(\Delta t)) \right].$$

**Proof of Lemma 3.**

• We study the difference that majorate the right hand side of (2.37). From relation (2.21), we know that sequence  $u_j$  is geometric and more precisely :

$$u_j = \left( \frac{\gamma x_- + \delta}{\gamma x^* + \delta} \right)^j z(0)$$

with  $\gamma$  and  $\delta$  computed in (2.15)  $\gamma = k \Delta t$ ,  $\delta = 1 + 2a^- \Delta t$ . We have from equation (2.11) and definition (2.24) of variable  $\tau$  :

$$\begin{aligned} \frac{\gamma x_- + \delta}{\gamma x^* + \delta} &= \frac{k \Delta t \frac{1}{k} \left( a - \frac{1}{2\tau} \right) + 1 + 2a^- \Delta t}{k \Delta t \frac{1}{k} \left( a + \frac{1}{2\tau} \right) + 1 + 2a^- \Delta t} \\ &= \frac{\Delta t \left( a^+ - a^- - \frac{1}{2\tau} \right) + 1 + 2a^- \Delta t}{\Delta t \left( a^+ - a^- + \frac{1}{2\tau} \right) + 1 + 2a^- \Delta t} \\ &= \frac{1 + \Delta t \left( |a| - \frac{1}{2\tau} \right)}{1 + \Delta t \left( |a| + \frac{1}{2\tau} \right)} \\ &= \frac{1 - \alpha \Delta t}{1 + \beta \Delta t}. \end{aligned}$$

We can now write :

$$\begin{aligned} u_j - z(j\Delta t) &= -z(0) \left[ e^{-j\Delta t/\tau} - \left( \frac{1 - \alpha \Delta t}{1 + \beta \Delta t} \right)^j \right] \\ &= -z(0) e^{-j\Delta t/\tau} \left[ 1 - \exp\left( \theta_j + j \log\left( \frac{1 - \alpha \Delta t}{1 + \beta \Delta t} \right) \right) \right] \\ &= -z(0) e^{-\theta_j} \left[ 1 - \exp\left( \theta_j \left( 1 + \frac{\tau}{\Delta t} \log\left( \frac{1 - \alpha \Delta t}{1 + \beta \Delta t} \right) \right) \right) \right]. \end{aligned}$$

and relation (2.42) is proven. Lemma 3 is established.  $\square$

**Lemma 4.**

Let  $\varphi(\bullet)$  be defined by relation (2.40). We suppose that time step  $\Delta t$  satisfies the condition

$$(2.43) \quad 0 < \Delta t \leq \frac{\tau}{2}.$$

Then we have

$$(2.44) \quad \left| \varphi(\Delta t) - \tau \left[ \frac{\beta^2 - \alpha^2}{2} \Delta t - \frac{1}{3}(\alpha^3 + \beta^3) \Delta t^2 \right] \right| \leq \frac{1}{2}(\alpha^4 + \beta^4) \tau \Delta t^3 .$$

**Proof of Lemma 4.**

- We have the following elementary calculus :

$$1 - \xi + \xi^2 - \xi^3 \leq \frac{1}{1 + \xi} \leq 1 - \xi + \xi^2 - \xi^3 + \xi^4 \quad \text{if } |\xi| < 1$$

and we deduce after integration :

$$-\frac{\xi^4}{4} \leq \left[ \log(1 + \xi) - \left( \xi - \frac{\xi^2}{2} + \frac{\xi^3}{3} \right) \right] \leq -\frac{\xi^4}{4} + \frac{\xi^5}{5} \leq \frac{\xi^4}{2} \quad \text{if } |\xi| < \frac{1}{2} .$$

Then we have

$$(2.45) \quad \left| \log(1 + \xi) - \left( \xi - \frac{\xi^2}{2} + \frac{\xi^3}{3} \right) \right| \leq \frac{1}{2} |\xi|^4 \quad \text{if } |\xi| \leq \frac{1}{2}$$

and we use this estimation with  $\xi = -\alpha \Delta t$  and  $\xi = \beta \Delta t$ . We remark that  $|a| \leq \frac{1}{2\tau}$  then  $\beta \leq \frac{1}{\tau}$  and we deduce that when condition (2.43) is satisfied, we have

$$\Delta t \leq \frac{\tau}{2} \leq \frac{1}{2\beta} \leq \frac{1}{2\alpha}, \quad \text{then } \alpha \Delta t \leq \frac{1}{2} \quad \text{and } \beta \Delta t \leq \frac{1}{2} .$$

We deduce from (2.45)

$$\begin{aligned} \left| \log\left(\frac{1 - \alpha \Delta t}{1 + \beta \Delta t}\right) - \left[ (-\alpha - \beta) \Delta t + \frac{\beta^2 - \alpha^2}{2} \Delta t^2 + \frac{1}{3}(-\alpha^3 - \beta^3) \Delta t^3 \right] \right| &\leq \\ &\leq \frac{1}{2}(\alpha^4 + \beta^4) \Delta t^4 . \end{aligned}$$

- As a consequence of the previous inequality, multiplying the above expression by  $\frac{\tau}{\Delta t}$ , we obtain :

$$\begin{aligned} \left| \varphi(\Delta t) - 1 - \frac{\tau}{\Delta t} \left[ (-\alpha - \beta) \Delta t + \frac{\beta^2 - \alpha^2}{2} \Delta t^2 + \frac{1}{3}(-\alpha^3 - \beta^3) \Delta t^3 \right] \right| &\leq \\ &\leq \frac{1}{2}(\alpha^4 + \beta^4) \tau \Delta t^3 , \end{aligned}$$

and due to the fact that  $(-\alpha - \beta) \tau = -1$  we have since  $\tau \leq \frac{1}{\beta} \leq \frac{1}{\alpha}$ ,

$$\varphi(\Delta t) - \left[ \frac{\beta^2 - \alpha^2}{2} \tau \Delta t - \frac{1}{3}(\alpha^3 + \beta^3) \tau \Delta t^2 \right] \leq \frac{1}{2}(\alpha^4 + \beta^4) \tau \Delta t^3 .$$

and relation (2.44) is proven. Lemma 4 is established.  $\square$

**Lemma 5.**

We suppose that  $a \neq 0$ . If time step  $\Delta t$  satisfy

$$(2.46) \quad 0 < \Delta t \leq \inf \left( \frac{\tau}{2}, |a| \tau^2 \right),$$

we have

$$(2.47) \quad |\varphi(\Delta t) - |a| \Delta t| \leq \frac{7}{12} |a| \Delta t.$$

**Proof of Lemma 5.**

• If  $a \neq 0$ , we observe that :  $\beta^2 - \alpha^2 = \frac{2|a|}{\tau} > 0$ , and inequality (2.46) suppose that  $\Delta t$  has been chosen such that  $\Delta t \leq |a| \tau^2$ . Due to previous computations, we have the following set of estimations :

$$\begin{aligned} \left| \left[ \frac{\beta^2 - \alpha^2}{2} \tau \Delta t - \varphi(\Delta t) \right] \right| &\leq \left( \frac{1}{3}(\alpha^3 + \beta^3) \Delta t + \frac{1}{2}(\alpha^4 + \beta^4) \Delta t^2 \right) \tau \Delta t \\ &= \left[ \frac{\Delta t}{3} \left( \frac{1}{4\tau^3} + \frac{3a^2}{\tau} \right) + \frac{\Delta t^2}{2} \left( \frac{1}{8\tau^4} + \frac{3a^2}{\tau^2} + 2a^4 \right) \right] \tau \Delta t \\ &\leq \left( \frac{\Delta t}{3\tau^3} + \frac{\Delta t^2}{2\tau^4} \right) \tau \Delta t \\ &\qquad\qquad\qquad \text{because } |a| \leq \frac{1}{2\tau} \text{ due to (2.24)} \\ &\leq \frac{\Delta t^2}{\tau^2} \left( \frac{1}{3} + \frac{1}{2}|a|\tau \right) \\ &\leq \frac{\Delta t^2}{\tau^2} \left( \frac{1}{3} + \frac{1}{4} \right) = \frac{7}{12} \frac{\Delta t^2}{\tau^2} \\ &\leq \frac{7}{12} |a| \Delta t \end{aligned}$$

and relation (2.47) is proven. □

**Lemma 6.**

We suppose that  $a = 0$ . If time step  $\Delta t$  satisfy condition (2.43), we have

$$(2.48) \quad \left| \varphi(\Delta t) + \frac{1}{12\tau^2} \Delta t^2 \right| \leq \frac{1}{32\tau^2} \Delta t^2$$

**Proof of Lemma 6.**

• If  $a = 0$ , then  $\beta = \alpha = \frac{1}{2\tau}$  and we have simply :

$$\frac{1}{3}(\alpha^3 + \beta^3) \tau \Delta t^2 = \frac{2}{3} \left( \frac{1}{2\tau} \right)^3 \tau \Delta t^2 = \frac{1}{12\tau^2} \Delta t^2$$



$\frac{1}{2}(\alpha^4 + \beta^4) \tau \Delta t^3 = \left(\frac{1}{2\tau}\right)^4 \tau \Delta t^3 = \frac{1}{16\tau^3} \Delta t^3 \leq \frac{1}{32\tau^2} \Delta t^2$  due to the relation (2.43). Then the relation (2.48) is a direct of previous estimation (2.44) established in Lemma 4. That completes the proof of Lemma 6.  $\square$

**Lemma 7.**

We suppose that function  $t \mapsto z(t)$  is defined at relation (2.34) and that the numerical initial condition  $d_\Delta$  satisfies hypothesis (2.27). We denote by  $h(\bullet)$  the expression introduced in (2.33). Then we have

$$(2.49) \quad |z(0)| \leq \frac{1}{\eta} \min\left(C, \frac{1}{k\tau} - \eta\right)$$

$$(2.50) \quad |h'(z(0))| \leq k\tau \left(\min\left(C, \frac{1}{k\tau} - \eta\right)\right)^2.$$

**Proof of Lemma 7.**

• We have, due to relation (2.27) :  $-C \leq x^* - d_\Delta \leq \frac{1}{k\tau} - \eta$ .

Then following (2.19) and (2.24) we deduce :

$$-C - \frac{1}{k\tau} \leq x^* - \frac{1}{k\tau} - d_\Delta = x_- - d_\Delta \leq -\eta \quad \text{and in consequence}$$

$$(2.51) \quad \frac{1}{|x_- - d_\Delta|} \leq \frac{1}{\eta}.$$

From relation (2.34) we have  $z(0) = \frac{d_\Delta - x^*}{x_- - d_\Delta}$  and inequality (2.49) is a direct consequence of (2.51) and of hypothesis (2.27). We derive now the expression (2.33) relatively to variable  $\xi$  and we have easily :

$$|h'(z(0))| \leq \frac{|x_- - x^*|}{(1 + z(0))^2} = \frac{|x_- - d_\Delta|^2}{|x_- - x^*|} \leq k\tau \left(\min\left(C, \frac{1}{k\tau} - \eta\right)\right)^2$$

due to the above expression of  $z(0)$  and relation (2.51). The estimation (2.50) is established and Lemma 7 is proven.  $\square$

**Proof of Theorem 1.**

• We cut the expression inside the absolute value of left hand side of (2.28) into two parts. The first one is the error for the continuous solution of the ordinary differential equation when changing initial data and the second one is to the error  $\epsilon_j$  between the solution of the ordinary differential equation and the discrete solution given by the scheme for the same initial condition :

$$|x(j\Delta t; d) - x_j(\Delta t; d_\Delta)| \leq |x(j\Delta t; d) - x(j\Delta t; d_\Delta)| + |\epsilon_j|$$

and due to definition (2.30),

$$(2.52) \quad |x(j\Delta t; d) - x_j(\Delta t; d_\Delta)| \leq |y(j\Delta t; d) - y(j\Delta t; d_\Delta)| + |\epsilon_j|.$$

• We first study the term  $|y(j\Delta t; d) - y(j\Delta t; d_\Delta)|$  in right hand side of (2.52). We first minorate the absolute value of the denominator of the right hand side of (2.31). Under the hypothesis (2.27), we have :

$$1 + k\tau(d_\Delta - x^*)(1 - e^{-t/\tau}) \geq 1 \quad \text{if } d_\Delta - x^* > 0,$$

and in the other case when  $d_\Delta - x^* \leq 0$  then  $|d_\Delta - x^*| \leq \frac{1}{k\tau} - \eta$  due to (2.27) and we have in consequence

$$1 + k\tau(d_\Delta - x^*)(1 - e^{-t/\tau}) \geq 1 - (1 - k\tau\eta) = k\tau\eta > 0 \quad \text{if } d_\Delta - x^* \leq 0.$$

Therefore the denominator of right hand side of (2.31) is always strictly positive and is in all cases minorated by  $k\tau\eta$  :

$$1 + k\tau(d_\Delta - x^*)(1 - e^{-t/\tau}) \geq k\tau\eta \quad \forall t > 0.$$

In consequence of (2.33) and previous algebra,  $y(t; d)$  and  $y(t; d_\Delta)$  are well defined for each  $t$  such that  $0 \leq t < \infty$  and we have :

$$\begin{aligned} x(j\Delta t; d) - x(j\Delta t; d_\Delta) &= y(j\Delta t; d) - y(j\Delta t; d_\Delta) \\ &= \frac{(d - d_\Delta)e^{-t/\tau}}{(1 + k\tau(d - x^*)(1 - e^{-t/\tau})) (1 + k\tau(d_\Delta - x^*)(1 - e^{-t/\tau}))}, \end{aligned}$$

then :

$$(2.53) \quad |x(j\Delta t, d) - x(j\Delta t, d_\Delta)| \leq \left( \frac{1}{k\tau\eta} \right)^2 |d - d_\Delta|$$

and the estimation of the first term introduced in (2.52) is then controlled .

• We now study the error  $\epsilon_j$ . Due to lemmas 2 and 3 and in particular relations (2.36) and (2.42), we have

$$(2.54) \quad |\epsilon_j| \leq |z(0)| |h'(z(0))| e^{-\theta_j} \left| 1 - \exp(\theta_j \varphi(\Delta t)) \right|.$$

(i) If  $a = 0$ , then  $\varphi(\Delta t)$  is negative due to relation (2.48) and  $\theta_j = \frac{j\Delta t}{\tau}$  remains positive. We deduce that in this case

$$\left| 1 - \exp(\theta_j \varphi(\Delta t)) \right| \leq 1 - \exp(\theta_j \varphi(\Delta t)) \leq |\theta_j \varphi(\Delta t)| = \theta_j |\varphi(\Delta t)|$$

and in consequence of (2.54) and (2.48),

$$\begin{aligned}
 |\epsilon_j| &\leq |z(0)| |h'(z(0))| \left( e^{-\theta_j} \theta_j \right) \left( \frac{1}{12} + \frac{1}{32} \right) \frac{\Delta t^2}{\tau^2} \\
 &\leq \frac{11}{96} |z(0)| |h'(z(0))| \sup_{\theta \geq 0} \left( \theta e^{-\theta} \right) \frac{\Delta t^2}{\tau^2} \\
 &\leq \frac{11}{96} \frac{1}{\eta} \min \left( C, \frac{1}{k\tau} - \eta \right) k\tau \left( \min \left( C, \frac{1}{k\tau} - \eta \right) \right)^2 \frac{1}{e} \frac{\Delta t^2}{\tau^2} \\
 &\leq \frac{11}{96e} \frac{k}{\tau\eta} \left( \min \left( C, \frac{1}{k\tau} - \eta \right) \right)^3 \Delta t^2
 \end{aligned}$$

and relation (2.29) is established with

$$(2.55) \quad A = \inf \left( \left( \frac{1}{k\tau\eta} \right)^2, \frac{11}{96e} \frac{k}{\tau\eta} \left( \min \left( C, \frac{1}{k\tau} - \eta \right) \right)^3 \right)$$

$$(2.56) \quad B = \frac{\tau}{2}.$$

(ii) If  $a \neq 0$ , then  $\varphi(\Delta t)$  is positive due to relation (2.47) if  $\Delta t$  satisfies condition (2.46). We suppose moreover that time step satisfies also the condition

$$(2.57) \quad \Delta t \leq \frac{6}{19} \frac{1}{|a|}$$

and due to (2.47), we have

$$(2.58) \quad \varphi(\Delta t) \leq \left( 1 + \frac{7}{12} \right) |a| \Delta t \leq \frac{1}{2}.$$

In order to majorate the expression  $e^{-\theta_j} \left| 1 - \exp(\theta_j \varphi(\Delta t)) \right|$  we distinguish between two cases. On one hand, when  $\theta_j \varphi(\Delta t) \leq 1$ , we have by convexity of the exponential function over the interval  $[0, 1]$  :

$$0 \leq e^{\theta_j \varphi(\Delta t)} - 1 \leq (e - 1) \theta_j \varphi(\Delta t) \quad \text{and we deduce}$$

$$e^{-\theta_j} \left| 1 - \exp(\theta_j \varphi(\Delta t)) \right| \leq (e - 1) [\theta_j e^{-\theta_j}] \varphi(\Delta t) \leq \frac{19}{12} \frac{e - 1}{e} |a| \Delta t.$$

The previous inequality and estimation (2.54) show that under hypotheses (2.46) and (2.57) concerning the time step, we have

$$\begin{aligned}
 |\epsilon_j| &\leq \frac{19}{12} \frac{e - 1}{e} \frac{1}{\eta} \min \left( C, \frac{1}{k\tau} - \eta \right) k\tau \left( \min \left( C, \frac{1}{k\tau} - \eta \right) \right)^2 |a| \Delta t \\
 &\leq \frac{19}{12} \frac{e - 1}{e} \frac{|a| k\tau}{\eta} \left( \min \left( C, \frac{1}{k\tau} - \eta \right) \right)^3 \Delta t \quad \text{due to (2.49) and (2.50)}
 \end{aligned}$$

$$\leq \frac{19}{24} \frac{e-1}{e} \frac{k}{\eta} \left( \min \left( C, \frac{1}{k\tau} - \eta \right) \right)^3 \Delta t \quad \text{due to (2.24)}$$

and

$$(2.59) \quad |\epsilon_j| \leq \frac{19}{12} \frac{1}{e} \frac{k}{\eta} \left( \min \left( C, \frac{1}{k\tau} - \eta \right) \right)^3 \Delta t \quad \text{when } \theta_j \varphi(\Delta t) \leq 1.$$

On the other hand when  $\theta_j \varphi(\Delta t) \geq 1$ , we have :

$$\begin{aligned} e^{-\theta_j} \left| 1 - \exp(\theta_j \varphi(\Delta t)) \right| &= \\ &= e^{-\theta_j/2} e^{\theta_j (\varphi(\Delta t) - 1/2)} \left| 1 - \exp(-\theta_j \varphi(\Delta t)) \right| \\ &\leq e^{-\theta_j/2} \left| 1 - \exp(-\theta_j \varphi(\Delta t)) \right| \quad \text{due to (2.58)} \\ &\leq e^{-\theta_j/2} \theta_j \varphi(\Delta t) \quad \text{because } \theta_j \text{ and } \varphi(\Delta t) \text{ are both positive} \\ &\leq 2 \sup_{\theta \geq 0} \left( \theta e^{-\theta} \right) \varphi(\Delta t) \leq \frac{19}{12} \frac{2}{e} |a| \Delta t \end{aligned}$$

thanks to relation (2.58). Following inequality (2.54), we obtain in this second case

$$\begin{aligned} |\epsilon_j| &\leq \frac{k\tau}{\eta} \left( \min \left( C, \frac{1}{k\tau} - \eta \right) \right)^3 \frac{19}{6e} |a| \Delta t \\ (2.60) \quad |\epsilon_j| &\leq \frac{19}{12e} \frac{k}{\eta} \left( \min \left( C, \frac{1}{k\tau} - \eta \right) \right)^3 \Delta t \quad \text{when } \theta_j \varphi(\Delta t) \geq 1 \end{aligned}$$

and relation (2.29) is proved for this case with

$$(2.61) \quad A = \inf \left( \left( \frac{1}{k\tau\eta} \right)^2, \frac{19}{12e} \frac{k}{\eta} \left( \min \left( C, \frac{1}{k\tau} - \eta \right) \right)^3 \right)$$

$$(2.62) \quad B = \inf \left( \frac{\tau}{2}, |a| \tau^2, \frac{6}{19} \frac{1}{|a|} \right)$$

which ends the proof of Theorem 1. □

### 3) MATRIX RICCATI EQUATION.

• In order to define a numerical scheme to solve the Riccati differential equation (1.22) with initial condition (1.19) we first introduce a real number  $\mu$ , which is chosen positive and such that the real matrix  $[\mu I - (A + A^t)]$  is definite positive :

$$(3.1) \quad \mu > 0, \quad \frac{1}{2}(\mu x, x) - (Ax, x) > 0, \quad \forall x \neq 0.$$

Then we define a definite positive matrix  $M$  wich depends on strictly positive scalar  $\mu$  and matrix  $A$  :

$$(3.2) \quad M = \frac{1}{2}\mu I - A.$$

The numerical scheme is then defined in analogy with relation (2.6). We have the following decomposition :

$$(3.3) \quad A = A^+ - A^-,$$

with

$$(3.4) \quad A^+ = \mu I, \quad A^- = M, \quad \mu > 0, \quad M \text{ positive definite .}$$

Taking as an explicit part the positive contribution  $A^+$  of the decomposition (3.3) of matrix  $A$  and in the implicit part the negative contribution  $A^- = M$  of the decomposition (3.3), we get the following **harmonic scheme** :

$$(3.5) \quad \left\{ \begin{array}{l} \frac{1}{\Delta t}(X_{j+1} - X_j) + \frac{1}{2}(X_j K X_{j+1} + X_{j+1} K X_j) + \\ \qquad \qquad \qquad + (M^t X_{j+1} + X_{j+1} M) = \mu X_j + Q. \end{array} \right.$$

The numerical solution  $X_{j+1}$  given by the scheme at time step  $(j+1)\Delta t$  is then defined as a solution of Lyapunov matrix equation with matrix  $X$  as unknown :

$$(3.6) \quad S_j^t X + X S_j = Y_j.$$

with :

$$(3.7) \quad S_j = \frac{1}{2}I + \frac{\Delta t}{2}K X_j + \Delta t M,$$

and :

$$(3.8) \quad Y_j = X_j + \mu \Delta t X_j + \Delta t Q.$$

We note that  $S_j$  is a (non necessarily symmetric) positive matrix and that  $Y_j$  is a symmetric definite positive matrix if it is the case for  $X_j$ . In all cases, matrix  $Y_j$  is a symmetric positive matrix.

## Definition 2. Symmetric matrices.

Let  $n$  be an integer greater or equal to 1. We define by  $\mathcal{S}_n(\mathbb{R})$ , (respectively  $\mathcal{S}_n^+(\mathbb{R})$ ,  $\mathcal{S}_n^{+*}(\mathbb{R})$ ) the linear space (respectively the closed cone, the open cone) of symmetric-matrices (respectively symmetric positive and symmetric definite positive matrices) ; we have

$$(3.9) \quad (x, Sy) = (Sx, y), \quad \forall x, y \in \mathbb{R}^n, \quad \forall S \in \mathcal{S}_n(\mathbb{R}),$$

$$(3.10) \quad (x, Sx) \geq 0, \quad \forall x \in \mathbb{R}^n, \quad \forall S \in \mathcal{S}_n^+(\mathbb{R}),$$

$$(3.11) \quad (x, Sx) > 0, \quad \forall x \in \mathbb{R}^n, x \neq 0, \quad \forall S \in \mathcal{S}_n^{+*}(\mathbb{R}).$$

The following inclusions  $\mathcal{S}_n^{+*}(\mathbb{R}) \subset \mathcal{S}_n^+(\mathbb{R}) \subset \mathcal{S}_n(\mathbb{R})$  are natural.

**Proposition 3. Property of the Lyapunov equation.**

• Let  $S$  be a matrix which is not necessary symmetric, such that the associated quadratic form:  $\mathbb{R}^n \ni x \mapsto (x, Sx) \in \mathbb{R}$ , is strictly positive :

$$(3.12) \quad S + S^t \in \mathcal{S}_n^{+*}(\mathbb{R}).$$

Then the application  $\varphi_S$  defined by :

$$(3.13) \quad \mathcal{S}_n(\mathbb{R}) \ni X \mapsto \varphi_S(X) = S^t X + X S \in \mathcal{S}_n(\mathbb{R}),$$

is a one to one bijective application on the space  $\mathcal{S}_n(\mathbb{R})$  of real symmetric matrices of order  $n$ .

• Moreover, if matrix  $\varphi_S(X)$  is positive (respectively definite positive) then the matrix  $X$  is also positive (respectively definite positive) :

$$(\varphi_S(X) \in \mathcal{S}_n^+(\mathbb{R}) \implies X \in \mathcal{S}_n^+(\mathbb{R})) \text{ and } (\varphi_S(X) \in \mathcal{S}_n^{+*}(\mathbb{R}) \implies X \in \mathcal{S}_n^{+*}(\mathbb{R})).$$

**Proof of proposition 3.**

• We first observe that  $\varphi_S$  is a linear map. In the case where  $S$  is a symmetric matrix, we can immediatly deduce from (3.12) that  $S$  is definite positive. Let  $X$  be a matrix such that :

$$(3.14) \quad \varphi_S(X) = 0.$$

We prove that  $X = 0$ , ie that  $\ker \varphi_S = \{0\}$ . Let  $x$  be an eigenvector of matrix  $S$  associated to eigenvalue  $\lambda$  :

$$(3.15) \quad Sx = \lambda x, x \neq 0.$$

We deduce from relations (3.14) and (3.15) and the fact that matrix  $S$  is supposed to be symmetric the equality

$$(3.16) \quad S(Xx) = -\lambda(Xx).$$

According to the previous hypothesis the negative scalar  $-\lambda$  cannot be an eigenvalue of  $S$ , then the vector  $(Xx)$  must be equal to zero. This is true for each eigenvalue of matrix  $S$ , wich prove the property in this case, because  $\varphi_S$  is also an endomorphism from  $\mathcal{S}_n(\mathbb{R})$  to  $\mathcal{S}_n(\mathbb{R})$ .

• In the case where  $S$  is not symmetric we suppose first that  $S$  is composed into blocks of Jordan type. We first study the case where  $S = \Lambda$  is composed of exactly one Jordan block associated to eigenvalue  $\lambda$  :

$$(3.17) \quad \Lambda = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda & 1 \\ 0 & 0 & \cdots & 0 & \lambda \end{pmatrix}$$

and according to the hypothesis (3.12), we have :

$$(3.18) \quad \operatorname{Re} \lambda > 0.$$

Let  $X_{i,j}$  be an element of the matrix  $X$ , we have from the following relation :

$$(3.19) \quad (\Lambda^t X + X \Lambda)_{i,j} = 2\lambda X_{i,j} + X_{i-1,j} + X_{i,j-1} \quad \text{if } i \geq 2 \text{ and } j \geq 2,$$

and :

$$(3.20) \quad (\Lambda^t X + X \Lambda)_{1,1} = 2\lambda X_{1,1}$$

$$(3.21) \quad (\Lambda^t X + X \Lambda)_{1,j} = 2\lambda X_{1,j} + X_{1,j-1} \quad j \geq 2.$$

Because  $\lambda \neq 0$ ,  $X_{1,1} = 0$  from (3.20), then  $X_{1,j} = 0$  if  $j \geq 2$  from (3.21). By induction,  $X_{i,j} = 0$  taking into account (3.13).

- When matrix  $S$  is composed by a family of Jordan blocks of the previous type, *i.e.*  $S = \Lambda = (\operatorname{diag} \Lambda_j)$  where  $1 \leq j \leq p$  and  $\Lambda_j$  is a Jordan block of the type (3.17) and of order  $n_j$

$$(3.22) \quad \Lambda = \begin{pmatrix} \Lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \Lambda_2 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \Lambda_{p-1} & 0 \\ 0 & 0 & \cdots & 0 & \Lambda_p \end{pmatrix},$$

we decompose the matrix  $X$  into blocks  $X_{i,j}$  of order  $n_i \times n_j$  :

$$(3.23) \quad X = \begin{pmatrix} X_{1,1} & X_{1,2} & X_{1,3} & \cdots & X_{1,p} \\ X_{2,1} & X_{2,2} & X_{2,3} & \cdots & X_{2,p} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ X_{p-1,1} & \cdots & \cdots & X_{p-1,p-1} & X_{p-1,p} \\ X_{p,1} & X_{p,2} & \cdots & X_{p,p-1} & X_{p,p} \end{pmatrix}.$$

Then the block number  $(i, j)$  of the expression  $S^t X + X S$  is equal to  $\Lambda_i^t X_{i,j} + X_{i,j} \Lambda_j$  and for  $i \neq j$  we have to prove that the nondiagonal matrix  $X_{i,j}$  is identically null.

- We establish that if  $\Lambda$  is a Jordan block of order  $n$  of the type (3.17) satisfying inequality (3.18), if  $M$  is a second Jordan block of order  $m$  of the previous type,

$$(3.24) \quad M = \begin{pmatrix} \mu & 1 & 0 & \cdots & 0 \\ 0 & \mu & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \mu & 1 \\ 0 & 0 & \cdots & 0 & \mu \end{pmatrix}$$

such that an inequality analogous to (3.18) is valid :

$$(3.25) \quad \operatorname{Re} \mu > 0.$$

and if  $X$  is a real matrix of order  $n \times m$  chosen such that

$$(3.26) \quad \Lambda^t X + X M = 0,$$

then matrix  $X$  is identically equal to zero. As in relations (3.19)-(3.21), we have :

$$(3.27) \quad (\Lambda^t X + X M)_{1,1} = \lambda X_{1,1} + \mu X_{1,1}$$

$$(3.28) \quad (\Lambda^t X + X M)_{1,j} = \lambda X_{1,j} + \mu X_{1,j} + X_{1,j-1} \quad j \geq 2$$

$$(3.29) \quad (\Lambda^t X + X M)_{i,1} = \lambda X_{i,1} + \mu X_{i,1} + X_{i-1,1} \quad i \geq 2$$

$$(3.30) \quad (\Lambda^t X + X M)_{i,j} = \lambda X_{i,j} + \mu X_{i,j} + X_{i-1,j} + X_{i,j-1} \quad i, j \geq 2.$$

Then due to (3.18), (3.25) and (3.27) we have  $X_{1,1} = 0$ . By induction on  $j$  and according to relation (3.28) we have  $X_{1,j} = 0$ . By induction on  $i$  and due to relation (3.29) we have analogously  $X_{i,1} = 0$ . Finally relation (3.30) prove by induction that  $X_{i,j} = 0$  when  $i$  and  $j \geq 2$ , and matrix  $X$  is identically null.

- When matrix  $S = \Lambda = (\operatorname{diag} \Lambda_j)$  is composed by a family of Jordan blocks of the previous type, the nondiagonal blocks  $X_{i,j}$  of decomposition (3.23) are null due to the previous point. Moreover, the diagonal matrices  $X_{i,i}$  are also identically null due to the property established at relations (3.17) to (3.21). Then the proposition is established when matrix  $S = \Lambda$  is composed by Jordan blocks as in relation (3.22).

- In the general case where real matrix  $S$  satisfy relation (3.14) there exists always a complex matrix  $Q$  such that :

$$(3.31) \quad S = Q^{-1} \Lambda Q,$$

where matrix  $\Lambda$  has a bloc Jordan form given *e.g.* by the right hand side of the relation (3.22). We have also the following elementary calculus :

$$\begin{aligned} \varphi_S(X) &= S^t X + X S \\ &= Q^t \Lambda^t Q^{-t} X + X Q^{-1} \Lambda Q \\ &= Q^t \Lambda^t (Q^{-t} X Q^{-1}) Q + Q^t (Q^{-t} X Q^{-1}) \Lambda Q \\ &= Q^t (\Lambda^t Y + Y \Lambda) Q \end{aligned}$$



with  $Y = Q^{-t}XQ^{-1}$ . Because the matrix  $Q$  is inversible and  $\varphi_S(X)$  is equal to zero, the matrix  $(\Lambda^t Y + Y\Lambda)$  is equal to zero. Matrix  $\Lambda$  is a diagonal bloc Jordan form, and from the previous points we deduce that  $Y$  is equal to zero, then  $X = 0$  and the proof of first assertion of proposition 3 is established in the general case.

- We suppose now that the matrix  $\varphi_S(X)$  is symmetric and positive, that is :  $\varphi_S(X) \in \mathcal{S}_n^+(\mathbb{R})$ . Let  $x$  be an eigenvector of matrix  $X$  associated to the real eigenvalue  $\lambda$  :

$$(3.32) \quad Xx = \lambda x, \quad x \neq 0.$$

From the definition of application  $\varphi_S$ , we have the following relation :

$$(3.33) \quad (x, \varphi_S(X)x) = 2\lambda(x, Sx)$$

and from hypothesis on  $\varphi_S(X)$  the left hand side of relation (3.33) is positive. The expression  $(x, Sx)$  is also strictly positive since relation (3.8) holds. We deduce that the number  $\lambda$  is positive because  $X \in \mathcal{S}_n(\mathbb{R})$  has an orthogonal decomposition in eigenvector spaces.

- If matrix  $\varphi_S(X)$  is symmetric and positive definite ( $\varphi_S(X) \in \mathcal{S}_n^{+*}(\mathbb{R})$ ), we introduce eigenvector  $x$  of matrix  $X$  as previously (relation (3.32)). Then relation (3.33) remains true and the left hand side of this relation is strictly positive. Then the eigenvalue  $\lambda$  of matrix  $X$  is strictly positive and proposition 3 is proven.  $\square$

- The numerical scheme has been written as an equation of unknown  $X = X_{j+1}$  that takes the form :

$$(3.34) \quad \varphi_{S_j}(X) = Y_j.$$

with  $\varphi_{S_j}$  given by a relation of the type (3.13) with the help of matrix  $S_j$  defined in (3.7) and  $Y_j$  in (3.8). Then we have the following proposition.

#### **Proposition 4.**

**Homographic scheme computes a definite positive matrix.**

- The matrix  $X_j$  defined by numerical scheme (3.5) with the initial condition

$$(3.35) \quad X_0 = 0$$

is positive for each time step  $\Delta t > 0$  :

$$(3.36) \quad X_j \in \mathcal{S}_n^+(\mathbb{R}), \quad \forall j \geq 0.$$

- If there exists some integer  $m$  such that  $X_m$  belongs to the open cone  $\mathcal{S}_n^{+*}(\mathbb{R})$ , then matrix  $X_{m+j}$  belongs to the open cone  $\mathcal{S}_n^{+*}(\mathbb{R})$  for each  $j$  :

$$(3.37) \quad \left( \exists m \in \mathbb{N}, X_m \in \mathcal{S}_n^{+*}(\mathbb{R}) \right) \implies \left( \forall j \geq 0, X_{m+j} \in \mathcal{S}_n^{+*}(\mathbb{R}) \right).$$

#### Proof of proposition 4.

- First we have  $Y_0 = \Delta t Q$  and  $S_0 = \frac{1}{2}I + \Delta t M$ , then  $X_1$  is a symmetric positive matrix ( $X \in \mathcal{S}_n^+(\mathbb{R})$ ) according to proposition 3 since matrix  $S_0$  is symmetric positive and  $M$  has been chosen such that

$$(3.38) \quad M + M^t \text{ is positive definite.}$$

The end of the first point follows by induction.

- If real symmetric positive definite matrix  $X_{j+1}$  is given, relation (3.8) clearly indicates that matrix  $Y_j$  is symmetric positive definite and matrix  $S_j$  introduced at relation (3.7) has a symmetric part which is positive definite if we verify that the following matrix

$$(3.39) \quad K X_j + X_j K$$

is positive. But this property is a consequence of the following. On one hand, matrix  $K$  is positive definite and map  $X \mapsto K X + X K$  transforms the closed cone of positive symmetric matrices onto himself. On the other hand, matrix  $X_j$  is symmetric positive definite then  $K X_j + X_j K \in \mathcal{S}_n^{+*}(\mathbb{R})$  due to proposition 3. Proposition 4 is established.  $\square$

- We have defined a numerical scheme for solving in an approximate way the Riccati equation (1.20) with the help of relation (3.5) and the initial condition

$$(3.40) \quad X_0 = 0$$

naturally associated with initial condition (1.19). Recall that time step  $\Delta t$  is **not** limited by any stability condition : matrix  $X_j$  is always symmetric positive and positive definite if matrix  $Q$  is symmetric positive definite. Moreover the equation (3.5) that allows the computation of  $X_{j+1}$  from data is a **linear** equation whose unknown is a symmetric matrix. But  $X_{j+1}$  is a nonlinear (homographic !) function of previous iteration matrix  $X_j$ .

- We now study the convergence of the iterate matrix  $X_j$  as long as discrete time  $j \Delta t$  tends to infinity. We know from Proposition 1 that the solution of the differential Riccati equation tends to the solution of stationary Riccati equation :

$$(3.41) \quad X_\infty K X_\infty - (A^t X_\infty + X_\infty A) - Q = 0.$$

We first study the monotonicity of our numerical scheme. Recall first that if  $A$  and  $B$  are two real symmetric matrices, the condition

$$(3.42) \quad A \leq B$$

and respectively the condition

$$(3.43) \quad A < B$$

means that matrix  $B - A$  is positive ( $B - A \in \mathcal{S}_n^+(\mathbb{R})$ )

$$(3.44) \quad (x, (B - A)x) \geq 0, \quad \forall x \in \mathbb{R}^n,$$

and respectively that matrix  $B - A$  is positive definite ( $B - A \in \mathcal{S}_n^{+*}(\mathbb{R})$ ) :

$$(3.45) \quad (x, (B - A)x) > 0, \quad \forall x \in \mathbb{R}^n.$$

**Proposition 5. Monotonicity.**

• Under the two conditions :

$$(3.46) \quad Q \text{ is a definite positive symmetric matrix}$$

$$(3.47) \quad \frac{1}{2} (KX_\infty + X_\infty K) < \left(\mu + \frac{1}{\Delta t}\right) I,$$

the scheme (3.5) is monotone and we have more precisely :

$$(3.48) \quad \left(0 \leq X_j \leq X_\infty\right) \implies \left(0 \leq X_j \leq X_{j+1} \leq X_\infty\right).$$

**Proof of proposition 5.**

• We know from Lewis [Le86] that for symmetric definite positive matrix  $K$  and symmetric positive matrix  $Q$ , the algebraic Riccati equation (3.41) has a unique symmetric positive solution  $X_\infty$ . Moreover, matrix  $X_\infty$  is positive definite if matrix  $Q$  is positive definite.

• We first establish that matrix  $\Theta \equiv X_\infty - X_{j+1}$  is positive if the left hand side of implication (3.48) is satisfied. We subtract relation (3.41) from numerical scheme (3.5), observe that

$$\begin{aligned} X_j K X_{j+1} - X_\infty K X_\infty &= X_j K (X_{j+1} - X_\infty) + (X_j - X_\infty) K X_\infty \\ &= (X_j - X_\infty) K X_\infty - X_j K \Theta \end{aligned}$$

and that

$$\begin{aligned} X_{j+1} K X_j - X_\infty K X_\infty &= (X_{j+1} - X_\infty) K X_j + X_\infty K (X_j - X_\infty) \\ &= X_\infty K (X_j - X_\infty) - \Theta K X_j \end{aligned}$$

then we obtain the following equation for  $\Theta$  :

$$(3.49) \quad \left\{ \begin{array}{l} \varphi_{\Sigma_1}(\Theta) \equiv \frac{1}{\Delta t} \Theta + \frac{1}{2} (X_j K \Theta + \Theta K X_j) + (M^t \Theta + \Theta M) = \\ = \frac{1}{\Delta t} (X_\infty - X_j) + \mu (X_\infty - X_j) + \frac{1}{2} [(X_j - X_\infty) K X_\infty \\ + X_\infty K (X_j - X_\infty)] \equiv \varphi_{\Sigma_2}(X_\infty - X_j) \end{array} \right.$$

with

$$\begin{aligned} \Sigma_1 &= \frac{1}{2\Delta t} I + \frac{1}{2} K X_j + M \\ \Sigma_2 &= \frac{1}{2} \left( \mu + \frac{1}{\Delta t} \right) - \frac{1}{2} K X_\infty. \end{aligned}$$

The matrix  $\Sigma_1$  admits a positive definite symmetric part  $\Sigma_1 + \Sigma_1^t$  because it is the case for matrix  $I$ . Moreover, since matrix  $X_\infty - X_j$  is symmetric positive by hypothesis, it is sufficient to establish that the symmetric matrix  $\Sigma_2 + \Sigma_2^t$  is positive definite and to apply the proposition 3. This last property is exactly expressed by hypothesis (3.47) and the first point is proven.

• We consider now the matrix  $S_\infty$  defined by :

$$(3.50) \quad S_\infty = \frac{1}{2} (X_\infty K + K X_\infty) - (A^t + A).$$

The matrix  $S_\infty$  is symmetric and we have the following calculus :

$$(3.51) \quad \begin{aligned} S_\infty &= \frac{1}{2} \left[ (X_\infty K X_\infty) X_\infty^{-1} + X_\infty^{-1} (X_\infty K X_\infty) \right] \\ &\quad - \left[ X_\infty^{-1} (X_\infty A) + (A^t X_\infty) X_\infty^{-1} \right] \\ S_\infty &= \varphi_{\Sigma_3}(X_\infty^{-1}) \end{aligned}$$

with  $\Sigma_3 = \frac{1}{2} X_\infty K X_\infty - A^t X_\infty$ . Then due to relation (3.41) we have  $\Sigma_3 + \Sigma_3^t = Q > 0$  due to hypothesis (3.46). Then the proposition 3 joined with (3.51) and the fact that matrix  $X_\infty^{-1}$  is symmetric positive definite establish that matrix  $S_\infty$  is symmetric definite positive.

• We establish that under the same hypothesis (3.48), the matrix  $Z \equiv X_{j+1} - X_j$  is positive. We start from the numerical scheme (3.5) and replace matrix  $Q$  by its value obtained from relation (3.41). It comes :

$$\begin{aligned} \varphi_{\Sigma_1}(Z) &\equiv \frac{1}{\Delta t} Z + \frac{1}{2} (X_j K Z + Z K X_j) + (M^t Z + Z M) = \\ &= \mu X_j + Q - X_j K X_j - (M^t X_j + X_j M) \end{aligned}$$

$$\begin{aligned}
 &= A^t X_j + X_j A + Q - X_j K X_j \\
 &= X_\infty K X_\infty - X_j K X_j - [A^t (X_\infty - X_j) + (X_\infty - X_j) A] \\
 &= \frac{1}{2} \left[ X_\infty K (X_\infty - X_j) + (X_\infty - X_j) K X_\infty + X_j K (X_\infty - X_j) + \right. \\
 &\quad \left. + (X_\infty - X_j) K X_j \right] - [A^t (X_\infty - X_j) + (X_\infty - X_j) A] \\
 &= \varphi_{\Sigma_4} (X_\infty - X_j) \quad \text{with} \quad \Sigma_4 = \frac{1}{2} (K X_\infty + K X_j) - A.
 \end{aligned}$$

The matrix  $\Sigma_4 + \Sigma_4^t = \frac{1}{2} (X_j K + K X_j) + S_\infty$  is positive definite due to the previous point ; in consequence matrix  $\varphi_{\Sigma_4} (X_\infty - X_j)$  is symmetric positive. The end of the proof is a consequence of proposition 3 and of the fact that the matrix  $\Sigma_1 = \frac{1}{2} I + \frac{1}{2} K X_j + M$  has clearly a symmetric part  $\Sigma_1 + \Sigma_1^t$  which is positive definite.  $\square$

**Proposition 6. Convergence when discrete time tends to infinity.**

We suppose that the data  $K, A, Q$  of Riccati equation (1.20) and parameters  $\mu$  and  $\Delta t$  of harmonic scheme (3.5) satisfy the conditions (3.1), (3.46) and (3.47). Let  $X_j$  and  $X_\infty$  be the solution of scheme (3.5) at discrete time  $j \Delta t$  and the symmetric definite positive matrix solution of the so-called algebraic Riccati equation (3.41). Then  $X_j$  tends to  $X_\infty$  when  $j$  tends to infinity :

$$(3.52) \quad X_j \longrightarrow X_\infty.$$

**Proof of proposition 6.**

• Let  $\mathcal{E}_k$  be the Grassmannian manifold composed by all the linear subspaces of space  $\mathbb{R}^n$  with dimension exactly equal to  $k$  :

$$(3.53) \quad \mathcal{E}_k = \{ W, W \text{ subspace of } \mathbb{R}^n, \dim W = k \}.$$

Then we have the classical characterization of the  $k^o$  eigenvalue of symmetric matrix  $A$  with the so-called inf-sup condition (see *e.g.* Lascaux and Théodor [LT86]) :

$$(3.54) \quad \mu_k = \inf_{W \in \mathcal{E}_k} \sup_{v \in W} \frac{(Av, v)}{(v, v)}.$$

Let  $\lambda_j^k$  be the  $k^o$  eigenvalue of matrix  $X_j$ ,  $\lambda_\infty^k$  the  $k^o$  eigenvalue  $\mu_k$  ( $\mu_1 \leq \mu_2 \leq \dots \leq \mu_n$ ) of matrix  $X_\infty$  and  $W$  a fixed subspace of  $\mathbb{R}^n$  of dimension exactly equal to  $k$ . We deduce from inequality (3.48) of Proposition 5 :

$$\frac{(X_j v, v)}{(v, v)} \leq \frac{(X_{j+1} v, v)}{(v, v)} \leq \frac{(X_\infty v, v)}{(v, v)}, \quad \forall v \neq 0, v \in W.$$

Then when subspace  $W$  is arbitrarily given in set  $\mathcal{E}_k$  we have :

$$\sup_{v \in W} \left[ \frac{(X_j v, v)}{(v, v)} \right] \leq \sup_{v \in W} \left[ \frac{(X_{j+1} v, v)}{(v, v)} \right] \leq \sup_{v \in W} \left[ \frac{(X_\infty v, v)}{(v, v)} \right], \quad \forall W \in \mathcal{E}_k$$

and taking the infimum bound of previous line as subspace  $W$  belongs to Grassmannian manifold  $\mathcal{E}_k$  we deduce, thanks to (3.54)

$$(3.55) \quad \lambda_j^k \leq \lambda_{j+1}^k \leq \lambda_\infty^k.$$

By monotonicity, eigenvalue  $\lambda_j^k$  is converging towards some scalar  $\mu^k$  as  $j$  tends to infinity :

$$(3.56) \quad \lambda_j^k \longrightarrow \mu^k \quad \text{as } j \longrightarrow \infty, \quad 1 \leq k \leq n.$$

• Consider now the unitary eigenvector  $v_j^k$  of matrix  $X_j$  associated with eigenvalue  $\lambda_j^k$  :

$$(3.57) \quad X_j v_j^k = \lambda_j^k v_j^k, \quad \|v_j^k\| = 1, \quad 1 \leq k \leq n, \quad j \geq 0.$$

Because  $X_j$  is a symmetric matrix, the family of eigenvectors  $(v_j^k)_{1 \leq k \leq n}$  is orthonormal and defines an orthogonal operator  $\rho_j$  of space  $\mathbb{R}^n$  defined as acting on the canonical basis  $(e_j)_{1 \leq k \leq n}$  of space  $\mathbb{R}^n$  by the conditions

$$(3.58) \quad \rho_j \cdot e_k = v_j^k, \quad 1 \leq k \leq n, \quad j \geq 0.$$

Rotation  $\rho_j$  belongs to the compact group  $O(n)$  of orthogonal linear transformations of space  $\mathbb{R}^n$ . Then after extraction of a convergent subsequence  $\rho'_j$  of the initial sequence  $(\rho_j)_{j \geq 0}$  we know that there exists an orthogonal mapping  $\rho_\infty \in O(n)$  such that

$$(3.59) \quad \rho'_j \longrightarrow \rho_\infty \quad \text{as } j \longrightarrow +\infty.$$

We introduce the family  $(w_\infty^k)_{1 \leq k \leq n}$  of vectors in  $\mathbb{R}^n$  by the conditions

$$(3.60) \quad w_\infty^k = \rho_\infty \cdot e_k, \quad 1 \leq k \leq n.$$

It constitutes an orthogonal basis of space  $\mathbb{R}^n$  and for each integer  $k$  ( $1 \leq k \leq n$ ), the extracted sequence of vectors  $v_j^k$  is converging towards vector  $w_\infty^k$  :

$$(3.61) \quad v_j^k \longrightarrow w_\infty^k, \quad 1 \leq k \leq n, \quad j \longrightarrow +\infty.$$

• We introduce the symmetric positive definite operator  $Y_\infty$  by the conditions

$$(3.62) \quad Y_\infty \cdot w_\infty^k = \mu^k w_\infty^k, \quad 1 \leq k \leq n.$$

We study now the convergence of the subsequence of matrices  $X_j'$  towards  $Y_\infty$ . We first remark that the sequence of matrices  $(X_j)_{j \geq 0}$  is bounded in space  $\mathcal{S}_n(\mathbb{R})$  :

$$(3.63) \quad \| X_j \| \leq \lambda_\infty^n, \quad \forall j \in \mathbb{N}$$

and we have also the following set of identities :

$$\begin{aligned} (X_j - Y_\infty) w_\infty^k &= X_j (w_\infty^k - v_j^k) - \lambda_\infty^k w_\infty^k + \lambda_j^k v_j^k \\ &= X_j (w_\infty^k - v_j^k) - \lambda_\infty^k (w_\infty^k - v_j^k) + (\lambda_j^k - \lambda_\infty^k) v_j^k. \end{aligned}$$

So we deduce, due also to (3.63) and (3.57) :

$$(3.64) \quad \| (X_j - Y_\infty) w_\infty^k \| \leq \lambda_\infty^n \| w_\infty^k - v_j^k \| + \lambda_\infty^k \| w_\infty^k - v_j^k \| + | \lambda_j^k - \lambda_\infty^k |$$

and according of the convergence (3.61) of subsequences  $v_j^k$  and (3.56) of sequences  $\lambda_j^k$  as index  $j$  tends to infinity, we get from estimation (3.64) :

$$(3.65) \quad X_j' - Y_\infty \longrightarrow 0, \quad j \longrightarrow +\infty.$$

• Due to the definition (3.5) of the numerical scheme, sub-sequence  $X_j'$  necessarily converges to the unique positive definite matrix  $X_\infty$  of the algebraic Riccati equation and in consequence we have necessarily

$$(3.66) \quad Y_\infty = X_\infty.$$

We deduce that for any arbitrary subsequence of the family  $(X_j)_{j \geq 0}$  there exists an extracted sub-subsequence converging towards  $X_\infty$  as  $j$  tends to infinity. Then the entire family  $(X_j)_{j \geq 0}$  converges towards  $X_\infty$  as  $j$  tends to infinity and the property is established.  $\square$

#### 4) NUMERICAL EXPERIMENTS.

##### 4.1) Square root function.

• The first example studied is the resolution of the equation :

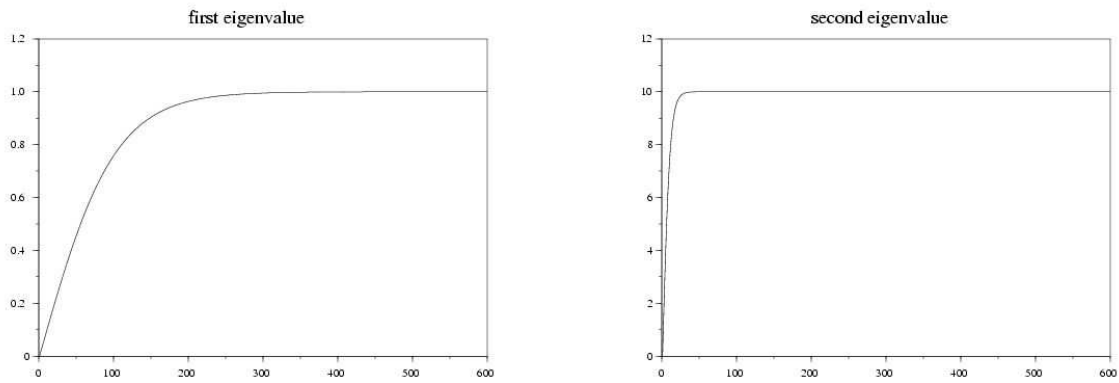
$$(4.1) \quad \frac{dX}{dt} + X^2 - Q = 0, \quad X(0) = 0$$

with  $n = 2$ ,  $A = 0$ ,  $K = I$  and matrix  $Q$  equal to

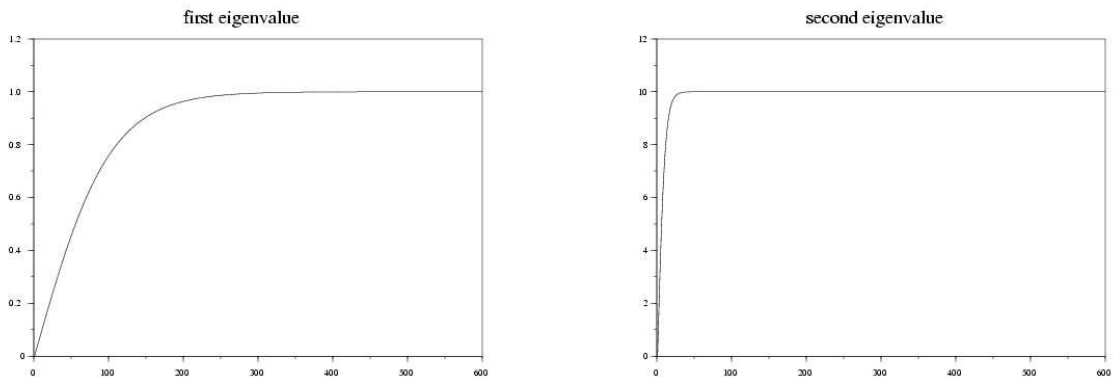
$$(4.2) \quad Q = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 100 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}.$$

• We have tested our numerical scheme for fixed value  $\Delta t = 1/100$  and different values of parameter  $\mu$  :  $\mu = 0.1, 10^{-6}, 10^{+6}$ . For small values of parameter  $\mu$ , the behaviour of the scheme does not change between  $\mu = 0.1$  and  $\mu = 10^{-6}$ . Figures 1 to 4 show the evolution with time of the eigenvalues of matrix  $X_j$  and the convergence is achieved to the square root of matrix  $Q$ . For large value of parameter  $\mu$  ( $\mu = 10^{+6}$ ), we loose completely consistency of the scheme (see figures 5 and 6).

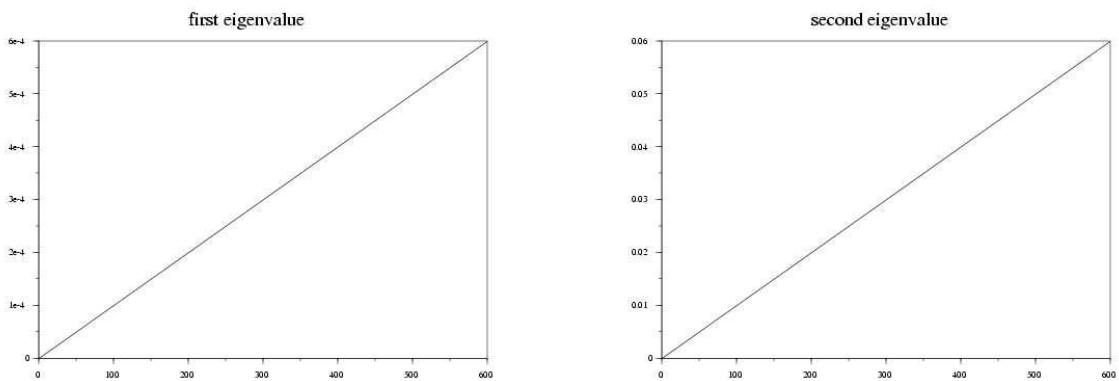
## HOMOGRAPHIC SCHEME FOR RICCATI EQUATION



**Figures 1 and 2.** Square root function test.  
Two first eigenvalues of numerical solution ( $\mu = 0.1$ ).



**Figures 3 and 4.** Square root function test.  
Two first eigenvalues of numerical solution ( $\mu = 10^{-6}$ ).



**Figures 5 and 6.** Square root function test.  
Two first eigenvalues of numerical solution ( $\mu = 10^{+6}$ ).



#### 4.2) Harmonic oscillator.

• The second exemple is the classical harmonic oscillator. Dynamical system  $y(t)$  is governed by the second order differential equation with command  $v(t)$  :

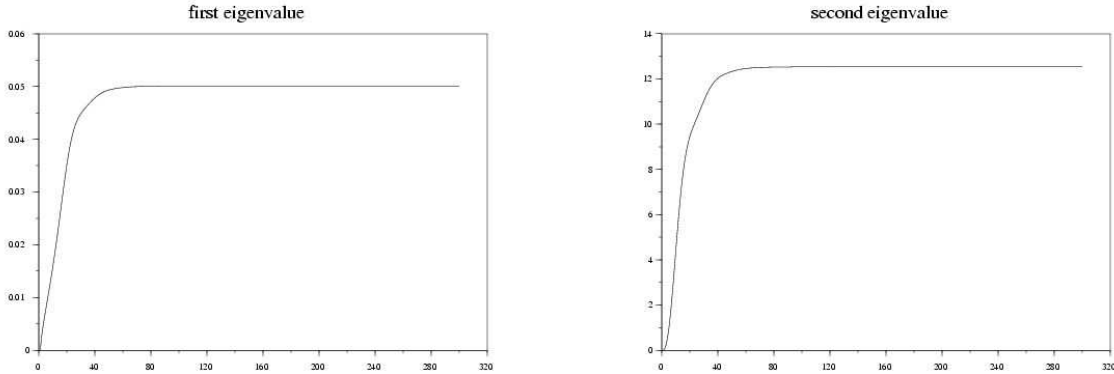
$$(4.3) \quad \frac{d^2 y(t)}{dt^2} + 2\delta \frac{dy(t)}{dt} + \omega^2 y(t) = b v(t).$$

This equation is written as a first order system of differential equations :

$$(4.4) \quad Y = \begin{pmatrix} y(t) \\ \frac{dy(t)}{dt} \end{pmatrix}, \quad \frac{dY}{dt} = \begin{pmatrix} 0 & 1 \\ -\omega^2 & -2\delta \end{pmatrix} Y(t) + \begin{pmatrix} 0 \\ b v(t) \end{pmatrix}.$$

The parameters  $R$ ,  $Q$ ,  $\omega$ ,  $\delta$  and  $b$  of the ordinary differential equation (4.4) and the cost function (1.6) are given by :

$$(4.5) \quad R = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}, \quad Q = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}, \quad \omega = \sqrt{250}, \quad \delta = 0, \quad b = 1.$$

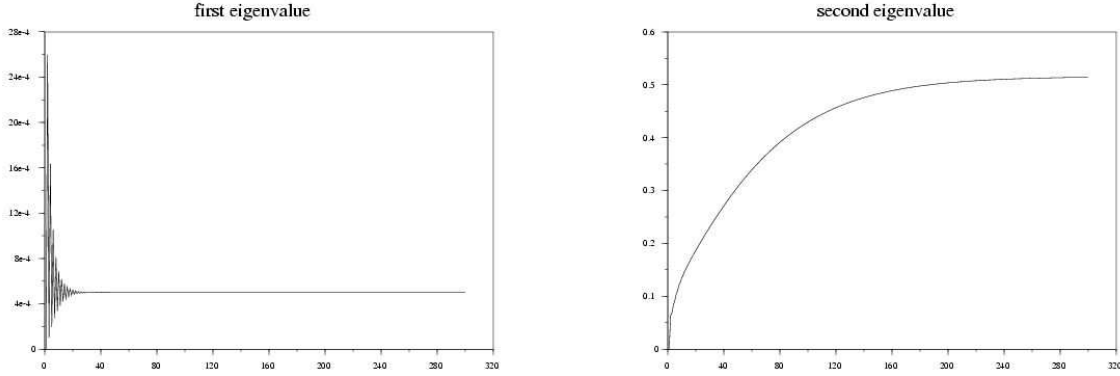


**Figures 7 and 8.** Harmonic oscillator.

Two first eigenvalues of numerical solution ( $\mu = 0.1$ ,  $\alpha = 0.01$ ,  $\Delta t = 0.01$ ).

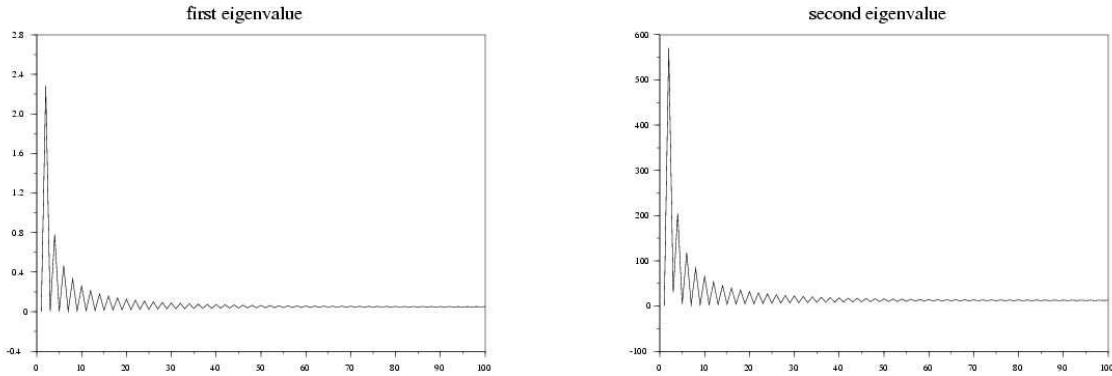
• In this case, we have tested the stability of the scheme for fixed value of parameter  $\mu$  ( $\mu = 0.1$ ) and different values of time step  $\Delta t$ . We have chosen three sets of parameters :  $\alpha = \Delta t = 1/100$  (reference experiment, figures 7 and 8),  $\alpha = 10^{-6}$ ,  $\Delta t = 1/100$  (very small value for  $\alpha$ , figures 9 and 10) and  $\alpha = 1/100$ ,  $\Delta t = 100$  (too large value for time step, figures 11 and 12). Note that for the last set of parameters, classical explicit schemes fail to give any answer. As in previous test case, we have represented the two eigenvalues of discrete matrix solution  $X_j$  as time is increasing. On reference experiment (figures 7 and 8), we have convergence of the solution to the solution of algebraic Riccati equation. If control parameter  $\alpha$  is chosen too small, the first eigenvalue of Riccati matrix oscillates during the first time steps but reach finally the correct

values of limit matrix, the solution of algebraic Riccati equation. If time step is too large, we still have stability but we loose also monotonicity. Nevertheless, convergence is achieved as in previous case.



**Figures 9 and 10.** Harmonic oscillator.

Two first eigenvalues of numerical solution ( $\mu = 0.1$ ,  $\alpha = 10^{-6}$ ,  $\Delta t = 0.01$ ).



**Figures 11 and 12.** Harmonic oscillator.

Two first eigenvalues of numerical solution ( $\mu = 0.1$ ,  $\alpha = 0.01$ ,  $\Delta t = 100$ ).

### 4.3) String of high speed vehicles.

• This example has been considered by Athans, Levine and Levis [ALL67] in modelling position and velocity control for a string of high speed vehicles. Let  $N$  be some integer and

$$(4.6) \quad n = 2N - 1$$

be the order of the given matrices  $A_N$ ,  $K_N$  and  $Q_N$ . The matrices  $A_N$ ,  $K_N$  and  $Q_N$  admit the following structure :

$$(4.7) \quad A_N = \begin{pmatrix} a_{11} & a_{12} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & a_{23} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{N-2N-2} & a_{N-2N-1} & 0 \\ 0 & \cdots & & 0 & a_{N-1N-1} & -1 \\ 0 & \cdots & \cdots & 0 & 0 & -1 \end{pmatrix},$$

$$(4.8) \quad K_N = \text{diag} (1, 0, 1, 0, \cdots, 1, 0, 1, 0, \cdots, 1, 0, 1)$$

$$(4.9) \quad Q_N = \text{diag} (10, 0, 10, 0, \cdots, 10, 0, 10, 0, \cdots, 10, 0, 10).$$

The unknown positive definite matrix  $X_N$  satisfies the algebraic Riccati equation

$$(4.10) \quad X_N K_N X_N - (A_N^t X_N + X_N A_N) - Q_N = 0.$$

The solution of this equation is detailed on 'figure' 13 with 10 significative decimals. The first six decimals are absolutely identical to the ones published by Laub [La79].

columns 1 to 3 :

$$\begin{array}{lll} +1.3630206938E + 00 & +2.6172154724E + 00 & -7.0542734123E - 01 \\ +2.6172154724E + 00 & +7.5925521955E + 00 & -1.6803557707E + 00 \\ -7.0542734123E - 01 & -1.6803557707E + 00 & +1.7747816032E + 00 \\ +9.3685970173E - 01 & +1.4752196951E + 00 & +2.1577096313E + 00 \\ -2.9366643189E - 01 & -4.5950584109E - 01 & -6.0913599887E - 01 \\ +4.7735386064E - 01 & +6.6514730720E - 01 & +6.7071749345E - 01 \\ -1.9737508953E - 01 & -2.6614220828E - 01 & -2.6284317351E - 01 \\ +2.1121165236E - 01 & +2.8065373289E - 01 & +2.6614220828E - 01 \\ -1.6655183115E - 01 & -2.1121165236E - 01 & -1.9737508953E - 01 \end{array}$$

columns 4 to 6 :

$$\begin{array}{lll} +9.3685970173E - 01 & -2.9366643189E - 01 & +4.7735386064E - 01 \\ +1.4752196951E + 00 & -4.5950584109E - 01 & +6.6514730720E - 01 \\ +2.1577096313E + 00 & -6.0913599887E - 01 & +6.7071749345E - 01 \\ +8.2576995027E + 00 & -1.9464979789E + 00 & +1.7558734280E + 00 \\ -1.9464979789E + 00 & +1.8056048615E + 00 & +1.9464979789E + 00 \\ +1.7558734280E + 00 & +1.9464979789E + 00 & +8.2576995027E + 00 \\ -6.7071749345E - 01 & -6.0913599887E - 01 & -2.1577096313E + 00 \\ +6.6514730720E - 01 & +4.5950584109E - 01 & +1.4752196951E + 00 \\ -4.7735386064E - 01 & -2.9366643189E - 01 & -9.3685970173E - 01 \end{array}$$

columns 7 to 9 :

$$\begin{array}{lll}
 -1.9737508953E - 01 & +2.1121165236E - 01 & -1.6655183115E - 01 \\
 -2.6614220828E - 01 & +2.8065373289E - 01 & -2.1121165236E - 01 \\
 -2.6284317351E - 01 & +2.6614220828E - 01 & -1.9737508953E - 01 \\
 -6.7071749345E - 01 & +6.6514730720E - 01 & -4.7735386064E - 01 \\
 -6.0913599887E - 01 & +4.5950584109E - 01 & -2.9366643189E - 01 \\
 -2.1577096313E + 00 & +1.4752196951E + 00 & -9.3685970173E - 01 \\
 +1.7747816032E + 00 & +1.6803557707E + 00 & -7.0542734123E - 01 \\
 +1.6803557707E + 00 & +7.5925521955E + 00 & -2.6172154724E + 00 \\
 -7.0542734123E - 01 & -2.6172154724E + 00 & +1.3630206938E + 00
 \end{array}$$

**Figures 13.** String of high speed vehicles. Matrix X is 9 by 9.

Numerical solution of stationary Riccati equation (4.10).

The parameters  $\mu = 0.1$  and  $\Delta t = 0.1$  have been used in homographic scheme.

#### 4.4) Control of the wave equation.

- The fourth example is the control of the wave equation in one space dimension

$$(4.11) \quad \frac{\partial^2 y}{\partial t^2} - c^2 \frac{\partial^2 y}{\partial x^2} = \sum_{i=1}^m \gamma_i(x) u_i(t), \quad x \in ]0, L[$$

with homogeneous Dirichlet boundary conditions

$$(4.12) \quad y(t, 0) = y(t, L) = 0.$$

For solving problem (4.11)-(4.12), we use a spectral decomposition on the eigenmodes  $\Phi_j(x)$  that are solution of the stationary problem :

$$(4.13) \quad -\frac{\partial^2 \Phi_j(x)}{\partial x^2} = \lambda_j \Phi_j(x), \quad x \in ]0, L[$$

$$(4.14) \quad \Phi_j(0) = \Phi_j(L) = 0$$

and are classically explicit by

$$(4.15) \quad \Phi_j(x) = \sqrt{2} \sin\left(\frac{j\pi x}{L}\right), \quad x \in ]0, L[$$

$$(4.16) \quad \lambda_j = \frac{j^2 \pi^2}{L^2}, \quad j = 1, 2, 3 \dots$$

- In practice we restrict to an approximation with the  $N$  first modes :

$$(4.17) \quad y(t, x) = \sum_{j=1}^N y_j(t) \Phi_j(x), \quad t \geq 0, \quad x \in ]0, L[,$$

and with such a spectral approximation, problem (4.11)-(4.12) is projected with  $L^2$  scalar product  $\langle \bullet, \bullet \rangle$

$$(4.18) \quad \langle u, v \rangle = \frac{1}{L} \int_0^L u(t) v(t) dt$$

and the discrete formulation stands as

$$(4.19) \quad \frac{d^2 y_j}{dt^2} + c^2 \lambda_j^2 y_j(t) = \sum_{i=1}^m u_i(t) \langle \gamma_i(\bullet), \Phi_j(\bullet) \rangle, \quad 1 \leq j \leq N.$$

We reduce this discrete differential system to a first order one by setting

$$(4.20) \quad Y(t) = \left( y_1, y_2, \dots, y_N, \frac{dy_1}{dt}, \frac{dy_2}{dt}, \dots, \frac{dy_N}{dt} \right)^t \in \mathbb{R}^{2N}$$

$$(4.21) \quad U(t) = \left( u_1(t), u_2(t), \dots, u_m(t) \right)^t \in \mathbb{R}^n$$

$$(4.22) \quad \Lambda = c^2 \text{diag}(\lambda_1, \dots, \lambda_N)$$

$$(4.23) \quad C_{ji} = \langle \gamma_i(\bullet), \Phi_j(\bullet) \rangle, \quad 1 \leq i \leq n, \quad 1 \leq j \leq N$$

and system (4.19) can be re-written as

$$(4.24) \quad \frac{d}{dt} Y(t) = \begin{pmatrix} 0 & \mathbf{I} \\ -\Lambda & 0 \end{pmatrix} Y(t) + \begin{pmatrix} 0 \\ C \end{pmatrix} U(t).$$

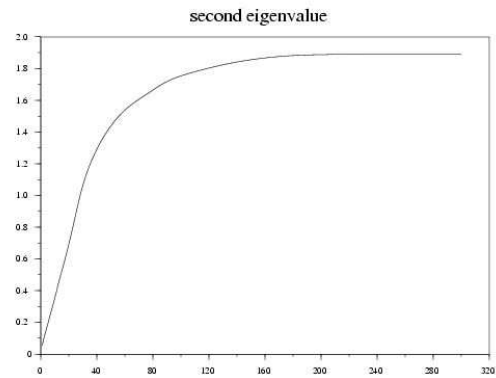
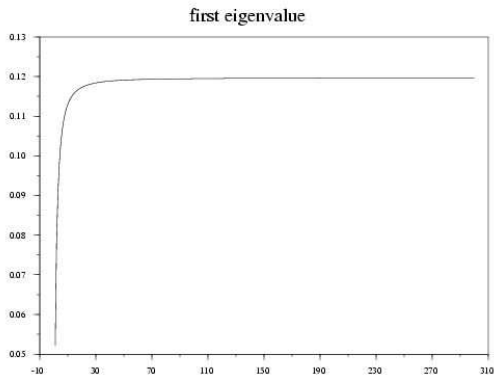
The matrices  $R$  and  $Q$  associated to the definition of the cost function  $J(\bullet)$  (relation (1.6)) are of order  $n$  and  $2N$  respectively. We have chosen the following simple form parameterized by  $\alpha = 1$  and  $\beta = 10$  in this particular test case :

$$(4.25) \quad R = \text{diag}(\alpha, \dots, \alpha)$$

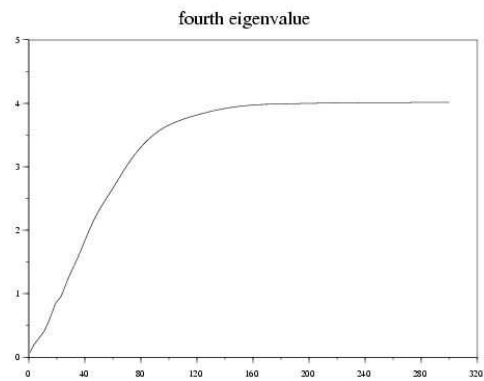
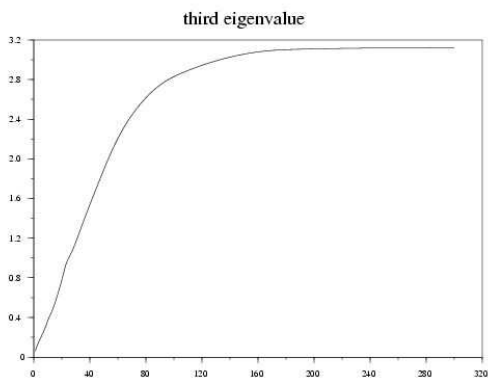
$$(4.26) \quad Q = \text{diag}(\beta, \dots, \beta).$$

• We have tested the scheme for matrices of order  $n = 2N = 10$  and for time step  $\Delta t = 0.01$  and a parameter  $\mu = 0.001$ . We represent on Figures 14 to 23 the ten different eigenvalues of Riccati matrix  $X(t)$  and for these parameters the scheme is stable as for as the explicit two stages Runge-Kutta scheme is unstable.

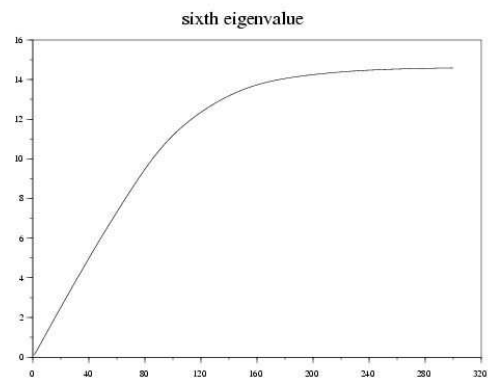
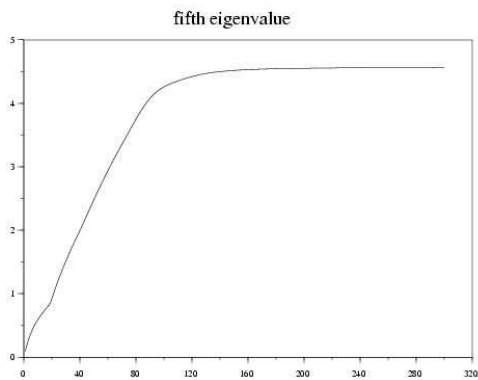
## HOMOGRAPHIC SCHEME FOR RICCATI EQUATION



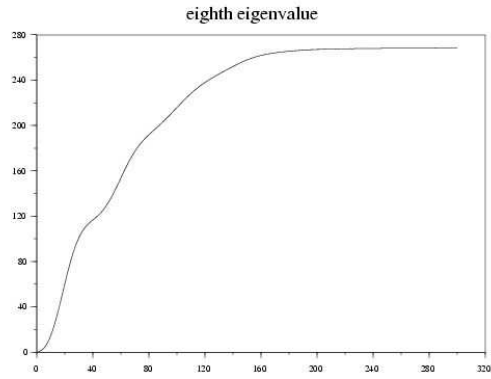
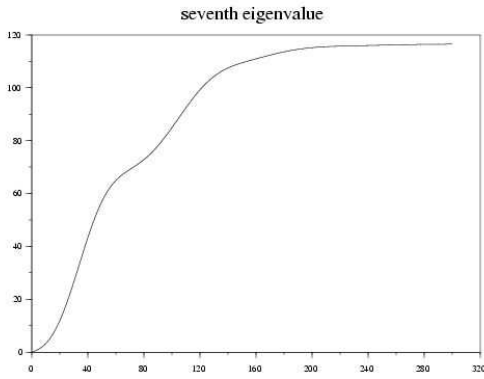
**Figures 14 and 15.** Wave control test.  
Two first eigenvalues of numerical solution.



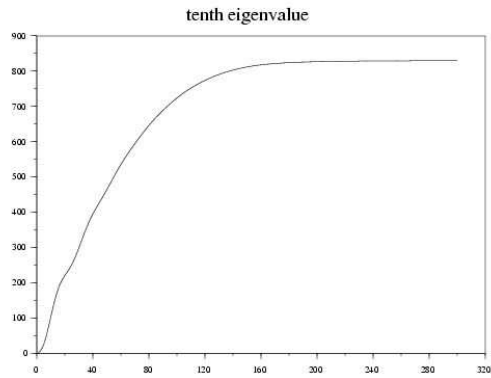
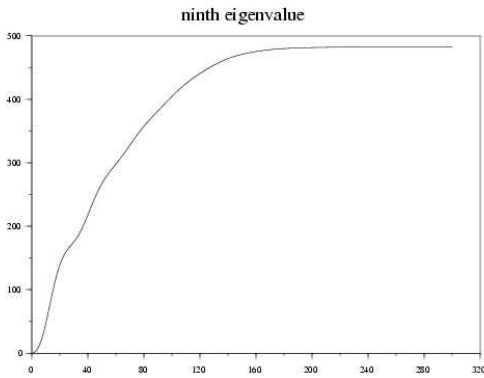
**Figures 16 and 17.** Wave control test.  
Third and fourth eigenvalues of numerical solution.



**Figures 18 and 19.** Wave control test. Fifth and sixth eigenvalues.



**Figures 20 and 21.** Wave control test. Seventh and eighth eigenvalues.



**Figures 22 and 23.** Wave control test. Ninth and tenth eigenvalues.

## 5) CONCLUSION.

- We have proposed an harmonic scheme for the resolution of the matrix Riccati equation. The scheme is implicit, unconditionnaly stable, needs to use one scalar parameter and to solve a linear system of equations for each time step. This scheme is convergent in the scalar case. In the matrix case, harmonic scheme has good monotonicity properties and discrete solution tends to the positive solution of algebraic Ricatti equation as discrete time tends to infinity. We have computed first test cases of matrix square root, harmonic ocsillator, string of vehicles and discretized wave equation where classical explicit schemes fail to give a definite positive discrete solution. Our first numerical experiments show stability and robustness when various parameters have large variations. We plan to develop this work in two directions : first prove the convergence of the harmonic scheme in the case of Ricatti matrix equation and second construct a multistep version in order to achieve second order accuracy.

6) ACKNOWLEDGMENTS.

- The authors thank Marius Tucsnak for helpfull comments on a preliminary draft of this report.

7) REFERENCES.

- [AF66] M. Athans, P.L. Faulb. *Optimal Control. An Introduction to the theory and Its Applications*, Mc Graw-Hill, New York, 1966.
- [ALL67] M. Athans, W.S. Levine and A. Levis. A system for the optimal and suboptimal position and velocity control for a string of high speed of vehicles, in *Proc. 5th Int. Analogue Compytation Meeting*, Lausanne, Switzerland, september 1967.
- [Ba91] R. Baraille. Développement de schémas numériques adaptés à l'hydrodynamique, *Thèse de l'Université Bordeaux 1*, décembre 1991.
- [Ca79] D. Cariolle. Modèle unidimensionnel de chimie de l'ozone, Internal note, *Etablissement d'Etudes et de Recherches Météorologiques*, Paris 1979.
- [DE96] L. Dieci, T. Eirola. Preserving monotonicity in the numerical solution of Riccati differential equations, *Numer. Math.*, vol. 74, p. 35-47, 1996.
- [Du93] F. Dubois. Un schéma implicite non linéairement inconditionnellement stable pour l'équation de Riccati, *unpublished manuscript*, april 1993.
- [DS95] F. Dubois, A. Saïdi. Un schéma implicite non linéairement inconditionnellement stable pour l'équation de Riccati, *Congrès d'Analyse Numérique*, Super Besse, France, may 1995.
- [DS2k] F. Dubois, A. Saïdi. Unconditionally stable scheme for Riccati equation, in *Control of systems governed by partial differential equations*, F. Conrad and M. Tucsnak Editors, *ESAIM: Proceedings*, vol. 8, p. 39-52, 2000.
- [FR84] P. Faurre, M. Robin. *Eléments d'automatique*, Dunod, Paris, 1984.
- [KS72] H. Kawakernaak, R. Sivan. *Linear optimal control systems*, Wiley, 1972.
- [La74] J.D. Lambert. *Computational methods in ordinary differential equations*, J. Wiley & Sons, 1974.
- [La79] A.J. Laub. A Schur Method for Solving Algebraic Riccati Equations, *IEEE Trans. Aut. Control*, vol. AC-24, p. 913-921, 1979.
- [Le86] F.L. Lewis. *Optimal Control*, J. Wiley-Interscience, New York, 1986.
- [Li68] J.L. Lions. *Contrôle optimal des systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.



- [LT86] P. Lascaux, R. Théodor. *Analyse numérique matricielle appliquée à l'art de l'ingénieur*, Masson, Paris, 1986.
- [Mi84] J.C. Miellou. Existence globale pour une classe de systèmes paraboliques semi-linéaires modélisant le problème de la stratosphère : la méthode de la fonction agrégée, *C.R. Acad. Sci., Paris, Serie I*, t. 299, p.723-726, 1984.
- [Sa97] A. Saïdi. Analyse mathématique et numérique de modèles de structures intelligentes et de leur contrôle, *Thèse de l'Université Pierre et Marie Curie*, France, avril 1997.