RELATIVE ERROR CONTROL IN QUANTIZATION BASED INTEGRATION

ERNESTO KOFMAN†

†*CIFASIS–CONICET. Laboratorio de Sistemas Dinámicos FCEIA - UNR. Riobamba 245 bis - (2000) Rosario*

***Abstract*— This paper introduces a method to achieve reltive error control in Quantized State System (QSS) methods. Based on the use of logarithmic quantization, the proposed methodology solves the problem of quantum selection.**

***Keywords*— Quantization Based Integration, Continuous System Simulation.**

I. INTRODUCTION

Numerical integration of ordinary differential equations (ODEs) is a topic of permanent research and development. Based on classic methods like Euler, Runge–Kutta and Adams and impulsed with the development of modern and fast computers, several variable–step and implicit ODE solver methods were introduced (Hairer *et al.*, 1993; Hairer and Wanner, 1991; Cellier and Kofman, 2006).

Simultaneously, different software simulation tools implementing those modern methods have been developed. Matlab/Simulink (Shampine and Reichelt, 1997) and Dymola (Elmqvist *et al.*, 1995) can be mentioned among the most popular and efficient general purpose ODE simulation packages.

In spite of the several differences between the mentioned ODE solvers, all of them share a property: they are based on time discretization. This is, they give a solution obtained from a difference equation system, i.e. a discrete–time model.

A completely different approach started to develop since the end of the 90's, where time discretization is replaced by state variables quantization. As a result, the simulation models are not discrete time but discrete event systems. The origin of this idea can be found in the definition of Quantized Systems (Zeigler *et al.*, 2000).

This idea was then reformulated with the addition of hysteresis –to avoid the appearance of infinitely fast oscillations– and formalized as the Quantized State Systems (QSS) method for ODE integration in (Kofman and Junco, 2001). This was followed by the definition of the second order QSS2 method (Kofman, 2002), the third order QSS3 method (Kofman, 2006), a first order Backward QSS method (BQSS) for stiff systems (Migoni *et al.*, 2007), and a first order Centered QSS for marginally stable systems.

The QSS–methods show some important advantages with respect to classic discrete time methods in the integration of discontinuous ODEs (Kofman, 2004), sparsity exploitation (Kofman, 2002), explicit integration of stiff and marginally stable systems (Migoni *et al.*, 2007), absolute stability, and the existence of a global error bound (Cellier and Kofman, 2006).

One of the major drawbacks of the QSS methods is the need of choosing a quantization parameter (called quantum) for each state variable, as the efficience and accuracy of the simulation depends strongly on this choice. The problem is also related to the fact that the methods intrinsically control the absolute error instead of the relative error as classic variable step methods do.

This work shows that the use of time varying quantization, proportional to the magnitude of each state variable (i.e., logarithmic quantization), leads to an intrinsic relative error control in the QSS methods. Moreover, it will be shown that the relative error is approximately proportional to the constant factor that relates the quantum with the state magnitude. This property will permit selecting directly the relative tolerance as a global property of the simulation (as it is done in discrete time variable step methods).

The paper is organized as follows. After introducing some notation, Section II presents the principles of quantization based integration and the QSS methods. Then, Section III introduces the main result (i.e., the relationship between logarithmic quantization and relative error control) and Section IV apply these results to two simulation examples.

A. Notation and Preliminaries

In the sequel, $|M| \triangleq \{|m_{i,j}|\}$, $\mathbb{R}\mathrm{e}(M) \triangleq \{\mathbb{R}\mathrm{e}(m_{i,j})\}$ and $\mathbb{I}\mathrm{m}(M) \triangleq \{\mathbb{I}\mathrm{m}(m_{i,j})\}$ denote the elementwise magnitude, real part and imaginary part, respectively, of a (possibly complex) matrix or vector M. Also, $x \leq y$ ($x < y$) denotes the set of componentwise (strict) inequalities between the components of the real vectors x and y, and similarly for $x \geq y$ ($x > y$). According to these definitions, it is easy to show that

$$|x+y| \leq |x| + |y|, \quad |M\,x| \leq |M| \cdot |x|, \qquad (1)$$

whenever $x, y \in \mathbb{C}^n$ and $M \in \mathbb{C}^{m \times n}$.