43rd IEEE Conference on Decision and Control
December 14-17, 2004
Atlantis, Paradise Island, Bahamas

Effects of Quantization, Saturation, and Sampling Time in Multi Output Systems

R. Sandoval-Rodriguez*, Chaouki T. Abdallah*
Electrical & Computer Engineering Department
MSC01 1100
1 University of New Mexico
Albuquerque, NM 87131-0001
{rsandova,chaouki}@ece.unm.edu

R. H. Byrne
Intelligent Systems and Robotics Center
Sandia National Labs
PO Box 5800
Mail Stop 1005
Albuquerque, NM 87185-1005
rhbyrne@sandia.gov

Abstract— **In this paper we analyze the effects of sampling and quantization in the states of the plant, and the saturation of the control signal, on the behaviors of closed-loop systems. We present graphical and analytical tools to determine the minimum number of bits in the A/D converters, and the maximum sampling time to ensure stability of closed-loop LTI Single Input Multiple Output Systems.**

I. Introduction

The use of continuous-time analysis with infinite precision resolution can lead us to select feedback gains that might not work as expected when implemented using computer control. This is due to finite precision calculations, propagation delays, and processing time in the computer system. By a computer system we refer to a system that may be as simple as a micro-controller with D/A (digital to analog) and A/D (analog to digital) converters, or a complete Personal Computer (PC) with a multipurpose data acquisition board. In either case, we use finite precision arithmetic to represent the gains of the controller in memory, and also to represent both the values of the states as read from the A/D converter, and the control signals sent to the D/A converter. In this paper we focus on the effects of representing the state and control signals with finite resolution, given that modern PCs can store controller gains with resolution much higher than the number of bits in the D/A and A/D converters. For the effects of finite resolution in the controller gains we refer the reader to [2].

In a related framework, modern applications of control systems make use of general purpose communication networks to transmit the signals from the sensors of the plant to the controller, and from the controller to the actuators in the plant. Such communication networks introduce issues such as the division of the messages into finite size packets, propagation time delay of the packets, and lost packets. We lump together some of these issues, and address the problem as a control system using a limited bandwidth communication channel.

The work here follows the setting of [1], [9], and the pioneering work of [3], [12]. In [10] we analyzed the effects of quantization and sampling time in the scalar system case. In this paper we extend the analysis to the multi-output case, in particular we completely analyze these effects on the double integrator and related systems. We focus on the double integrator because many nonlinear systems can be reduced to this form after applying feedback linearization [8].

* The research of both authors is partially supported by NSF-0233205 and ANI- 0312611.

II. Analysis of Systems with Quantized and Saturated States

A continuous-time infinite-precision state feedback controller with control signal

$$u = -[k_1 x_1 + ... + k_n x_n] \tag{1}$$

can asymptotically drive to the origin, any initial condition of an LTI system, provided that the system is controllable (or more generally stabilizable). However, when quantization and saturation are present in both state and control signals, asymptotic convergence to the origin is no longer possible. Instead, we can "contain" [3] the states of the system in the innermost quantization level with the proper selection of the control signal among the permissible values. The control signal may be computed by multiplying the vector of quantized states by the vector gain k, calculated assuming no quantization, and then passing the resulting control signal through the quantizing and saturation blocks. Example 1 shows the initial-state response of both a quantized system and a non-quantized system, when the control signal is calculated as mentioned.

Example 1: Let us consider a double integrator system with state feedback, so that the feedback gain vector $k = [1 \quad 5]$, places both closed-loop eigenvalues in the left half plane. The closed-loop simulation is run in Simulnik®. We used two double integrators, where in the first one we closed the loop trough the gain $-k$, while in the second system we used quantizing and saturation blocks for the states and control signal. The saturation values are $x_{sat} = \pm 10$, and the quantizer has $N = 8$ levels. Figure 1 shows the response of the systems to the initial condition $x_1 = 8.7,\ x_2 = -8.7$, in the phase plane. The non-quantized system converges asymptotically to the origin, while for the quantized system,

the states are 'trapped' in a limit cycle encircling the innermost quantization square (intersection of the innermost quantization levels, in the general case these intersections are rectangles).

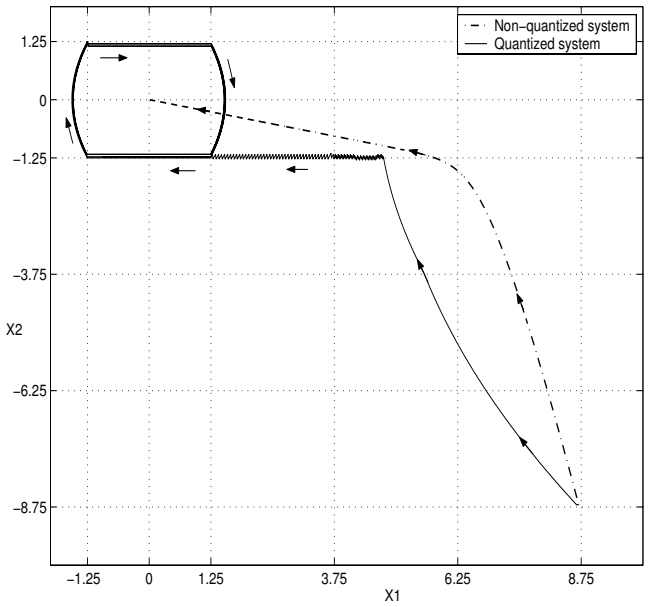

Fig. 1. Initial State Response

We can explain the presence of the limit cycle as follows. From the state-space dynamic equation of the double integrator system

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \tag{2}$$

we can see that the dynamic update of x_1 depends only on the state x_2, while that of x_2 depends only on the control input u as follows

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= u \end{aligned} \tag{3}$$

Figure 2 shows a vector field of the double integrator using quantization in the states. For easy visualization we only show the four innermost quantization levels of each state variable (The innermost quantization level, which corresponds to the 'zero value' is not part of the four levels shown). The horizontal vectors show the contribution of x_2 to the dynamic update of x_1, and changes continuously along the vertical axis despite the quantization. The vertical vectors show the contribution of the control signal u in the dynamic equation of x_2 and remains constant inside each quantization square, in contrast to the dynamic update of x_1. Given that the amplitude of the control signal depends linearly on the states values ($u = -kx$), the innermost quantization square has zero control signal value. The diagonal vectors provide the resulting heading for the solution trajectories passing trough the corresponding quantization squares. Thus, any trajectory entering from the right or below the innermost quantization square, will keep the value of x_2 constant while decreasing the value of x_1 (travelling right to left). On the other hand, any trajectory entering form the left or above the innermost quantization square will keep the value of x_2 constant while increasing the value of x_1 (travelling left to right). Also, from Figures 1 and 2, we see that any trajectory leaving the innermost quantization square from the left and below the x_1 axis, will be forced to describe an arc that crosses the x_1 axis, and re-enters the innermost square from the left and above the x_1 axis. On the other hand, a trajectory leaving the innermost square from the right and above the x_1 axis, will describe an arc that crosses the x_1 axis and re-enters the innermost square from the right and below the x_1 axis. Along the line $k_1x_1 + k_2x_2 =$

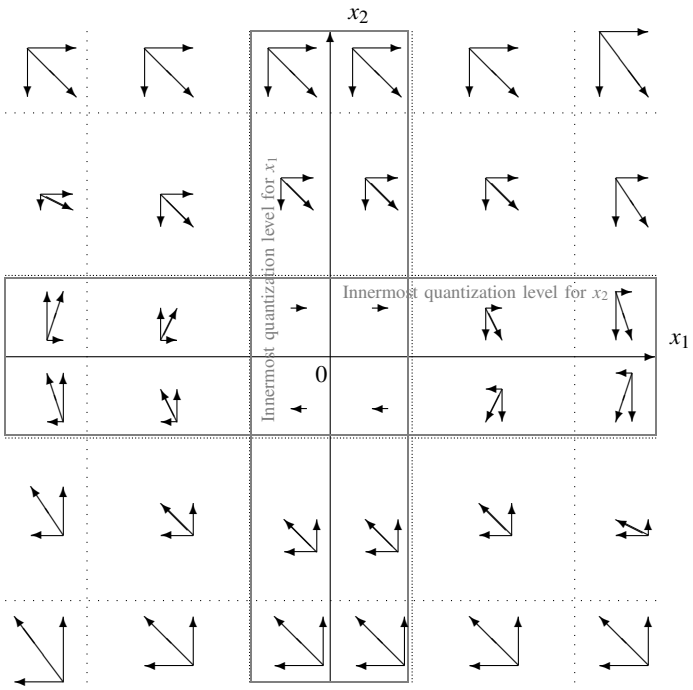

Fig. 2. Vectors field of the double integrator.

0, the control signal u is zero. This may not be significant in the non-quantized case, since the solution trajectory might touch this line just once. But in the case of quantized states, we might have $k_1x_1 = -k_2x_2$ in some quantization squares, leaving those squares with a 'zero' control signal. For the particular case of example 1, the conflicting squares lie in the second and fourth quadrants. We selected $k = [1 \quad 5]$ to make x_2 the dominant state, thus in all squares in the second quadrant the resulting control signal $u = -kx$ is negative pushing the trajectories downward. In the intersection of the innermost quantization level of x_2 with the second and third quadrants, the dominant state is x_1, which makes the resulting control signal positive. Then, any trajectory entering the innermost quantization level of x_2 from the second quadrant will be repelled by the positive control signal u, creating the chattering asymptote shown in Figure

3. A similar condition occurs on the boundary between the innermost quantization level of x_2 and the fourth quadrant, creating the other chattering asymptote.

The way in which the control signal u appears in the

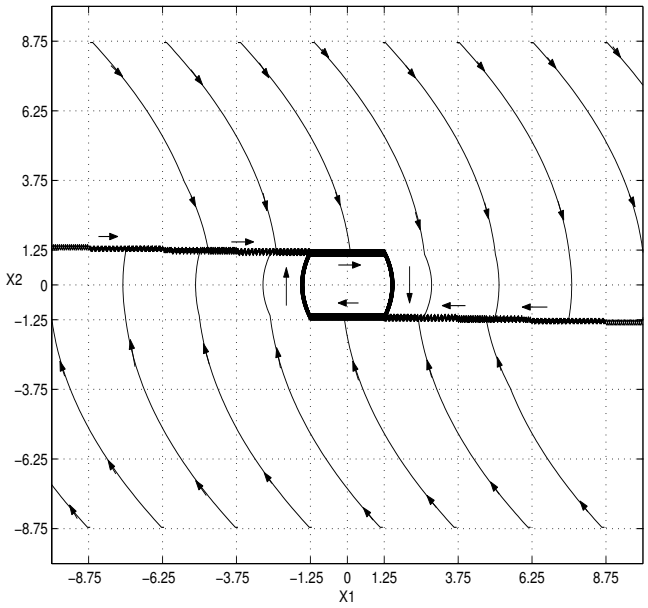

Fig. 3. Phase portrait of the double integrator

dynamic equations of the double integrator makes it easy to stabilize (i.e. to drive it to the innermost quantization levels, and contain it there), and the region of attraction is practically $\mathbb{R}^2$. The author in [11] addresses the case for dimension higher than two. The effects introduced by the quantizing and saturation blocks are not critical, and the region of attraction is not reduced. The slope of the chattering asymptotes can be controlled by the gain k, and may be calculated assuming no quantization. The limit cycle is contained in the innermost quantization square, and tends to become smaller as the number of bits in the A/D and D/A converters increases.

There are however other systems in which the use of quantizing and saturation blocks reduces the region of attraction drastically, and furthermore imposes a minimum number of bits in the A/D and D/A converters to ensure stability. An example of such systems are the decoupled integrators described in equation (8) of [1], and given by:

$$\dot{x} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} x + \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} u \qquad (4)$$

Here, the control signal appears in all the dynamic equations with the same amplitude and sign, limiting its action to either increasing or decreasing all the states simultaneously. The presence of the terms $\lambda_1, \lambda_2, ...\lambda_n$ in the system's matrix A imposes constraints on the initial conditions $x_1(0), x_2(0), ...x_n(0)$ when the control signal is bounded by $u = \pm u_{max}$. From equation (4) we can obtain the bounds on the region of attraction for each axis in a similar fashion as described in [10]. Assuming a negative initial condition $x_i(0) < 0$, the lower bound for x_i results from

$$\dot{x}_i(0) = \lambda_i x_i(0) + u_{max} > 0. \qquad (5)$$

Thus, in order to guarantee that the trajectory starts moving towards $x_i = 0$, the lower bound for x_i yields

$$x_i(0) > -\frac{u_{max}}{\lambda_i}. \qquad (6)$$

Similarly, assuming a positive initial condition $x_i(0) > 0$, the upper bound for x_i can be obtained from

$$\dot{x}_i(0) = \lambda_i x_i(0) - u_{max} < 0. \qquad (7)$$

Simplifying, the upper bound results in

$$x_i(0) < \frac{u_{max}}{\lambda_i}. \qquad (8)$$

However, while these bounds guarantee that the trajectory will not grow along the corresponding axis at the beginning of the motion, they do not guarantee that the trajectory will converge to the origin as the control signal is bounded. Therefore, in order to obtain the region of attraction we have to analyze the trajectories generated from the initial conditions. For the particular case of a two-dimensional system, state feedback controllers split the phase plane into two regions along the line

$$-k_1 x_1 - k_2 x_2 = 0. \qquad (9)$$

Initial conditions above this line generate negative control signals, while initial conditions below this line generate positive control signals. Given this, trajectories will be driven to the first or third quadrant and then pulled to the origin. The slope of this 'pulling' asymptote is slightly larger than the slope of the line in equation (9).

With the purpose of relating these results to limited information control [5], [6], [7], we use a binary control scheme. In other words, for initial conditions above the line in equation (9), the control signal is $u = -u_{max}$. And for initial conditions below the line, the control signal is $u = u_{max}$. This can be done by selecting a large gain vector k that places the closed-loop eigenvalues far inside the left half plane. This way the line in equation (9) results in a chattering asymptote, and trajectories hitting this asymptote inside the bounds placed by equations (6) and (8) will slide to the origin. Without loss of generality, the asymptote can be changed to

$$\lambda_1 x_1 - \lambda_2 x_2 = 0 \qquad (10)$$

Given that the vector gain $k = [-\lambda_1 \quad \lambda_2]$ also places the closed-loop eigenvalues in the left half plane. This asymptote intersects the axes at the states' bounds (See fig. 4). Then, a trajectory starting above the chattering asymptote must hit the asymptote before it reaches the lower bound in the x_1 axis, while a trajectory starting below the chattering

asymptote must hit the asymptote before it reaches the upper bound in the x_1 axis. We can construct the 'envelope' of the region of attraction as follows: given an initial condition in $x_1(0) = x_{10}$, find the value of $x_2(0) = x_{20}$ such that the trajectory starting at this point hits the chattering asymptote at its intersection with the axes bounds. Thus for a trajectory above the chattering asymptote and given the initial condition $x_1(0) = x_{10}$, we can find the time at which the trajectory in x_1 reaches $x_1(t) = -\frac{u_{max}}{\lambda_1}$. The solution $x(t)$ for equation (4), applying the negative control signal $u(t) = -u_{max}$, results

$$x_1(t) = e^{\lambda_1 t} x_{10} - \frac{u_{max}}{\lambda_1}\left(e^{\lambda_1 t} - 1\right) \tag{11}$$

$$x_2(t) = e^{\lambda_2 t} x_{20} - \frac{u_{max}}{\lambda_2}\left(e^{\lambda_2 t} - 1\right) \tag{12}$$

Substituting $x_1(t) = -\frac{u_{max}}{\lambda_1}$ in equation (11) and solving for t yields

$$t = \frac{1}{\lambda_1} \ln\left(\frac{2u_{max}}{u_{max} - \lambda_1 x_{10}}\right) \tag{13}$$

substituting equation (13), and $x_2(t) = -\frac{u_{max}}{\lambda_2}$ in equation (12), and solving for x_{20} results

$$x_{20} = \frac{\frac{u_{max}}{\lambda_2}\left(\left(\frac{2u_{max}}{u_{max} - \lambda_1 x_{10}}\right)^{\frac{\lambda_2}{\lambda_1}} - 2\right)}{\left(\frac{2u_{max}}{u_{max} - \lambda_1 x_{10}}\right)^{\frac{\lambda_2}{\lambda_1}}}. \tag{14}$$

Proceeding similarly, we can find the bottom half of the envelope of the region of attraction. Figure 4 shows the envelope of the region of attraction constructed using the equations just derived.

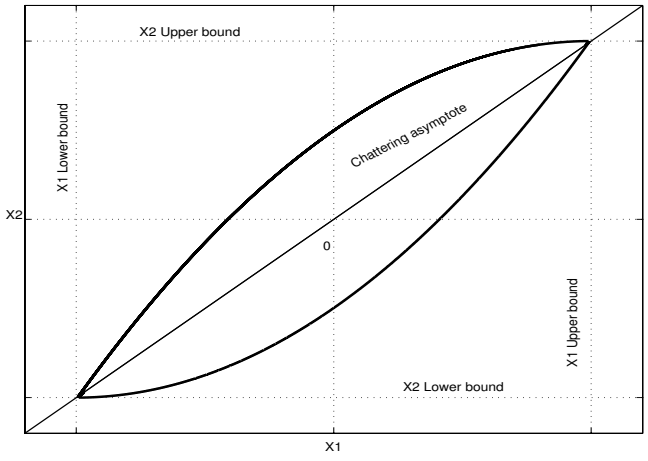

Fig. 4. Region of attraction and chattering asymptote

Let us study next the effect of quantizing; as a particular case let us consider the following system

$$\dot{x} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} x + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u. \tag{15}$$

Using eight-level quantizers in the states, and a control signal bounded by $u = \pm 10$, a state feedback controller with vector gain $k = [-4 \quad 8]$ practically works as a binary control, with the exception that several quantization squares represent a zero control signal. The quantization codes for those squares satisfies equation (10). These quantization squares are $(-10,-5)$, $(-5,-2.5)$, $(0,0)$, $(5,2.5)$, and $(10,5)$. Figure 5 shows the phase plane sectioned by the quantization squares, with each square showing its control signal value such that an upward pointing arrow indicates $u = +10$, while a downward pointing arrow indicates $u = -10$. In contrast to the non-quantized system, there is no

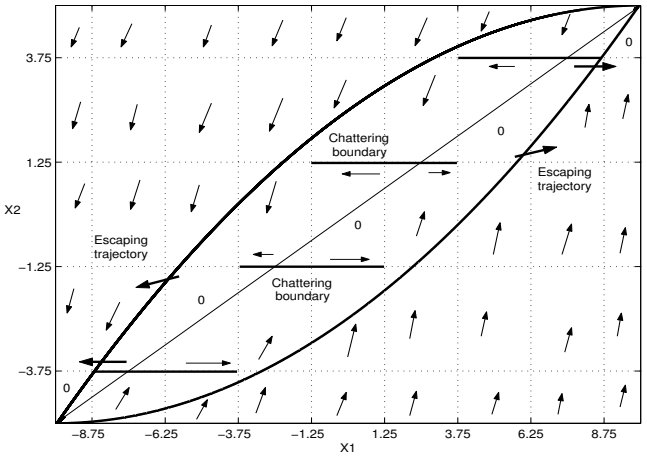

Fig. 5. Quantized phase plane, control signals, region of attraction, chattering boundaries and escaping trajectories

chattering asymptote, and in its place there are several chattering boundaries between squares of opposite signs control signals. The ticker lines show the chattering boundaries and the sliding direction is indicated by the arrows. The diagonal line

$$4x_1 - 8x_2 = 0 \tag{16}$$

divides the chattering boundaries, creating opposite sliding directions. Given the absence of control signal in some squares, trajectories inside those squares and close to the envelope can escape from the region of attraction given the 'inertia' effect of the initial conditions, as indicated in Figure 5. In order to correct this problem we assigned a negative control signal $u = -10$ to the quantization squares with codes $(-10,-5)$ and $(-5,-2.5)$; and a positive control signal $u = 10$ to the quantization squares with codes $(5,2.5)$ and $(10,5)$. No control law in the form $u = k_1 x_1 + k_2 x_2$ can assign these arbitrary values, and we must use a two-dimensional selection function to obtain the control signals. These selection functions allow us to assign arbitrary control signal values to each quantization square. Figure 6 shows a family of trajectories, and the limit cycle to which they converge.

Given that the region of attraction for the x_2 axis is bounded by $\frac{\pm u_{max}}{\lambda_2} = \pm 5$, we are wasting four levels of quantization in this axis. In order to optimize the resolution

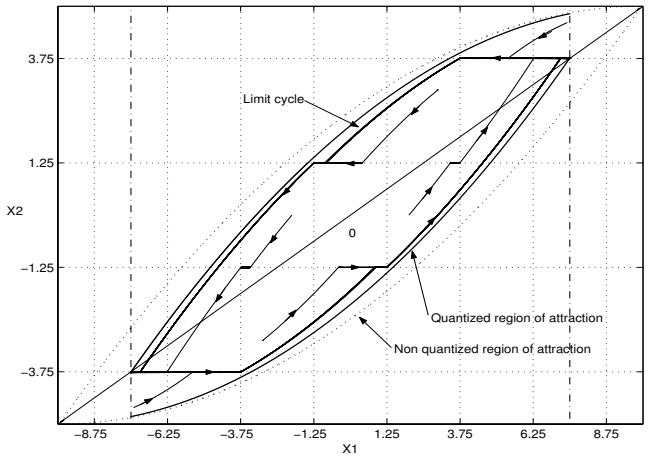

Fig. 6. Family of solutions and the stable limit cycle.

we can set the maximum signals (voltage references) in the A/D converter to these bounding values (± 5), see [3]. With this change, we can reduce the size of the limit cycle, but we can not eliminate it completely due to the unstable nature of the system. Increasing the number of bits in the A/D converters will also reduce the size of the limit cycle, given that the size of the quantization levels is being reduced. However, the number of quantization levels on which the limit cycle is jumping might change (increase or decrease), due to the change in the shape of the quantization rectangles. Proper assignment of the control signal for each quantization rectangle is thus important, in trying to reduce the size of the limit cycle. The presence of limit cycles in quantized open-loop unstable systems seems to be 'natural', but they will also appear in quantized open-loop stable systems when we try to drive the states to points other than an equilibrium point.

III. Analysis of the Effect of Sampling Time in Quantized Systems

In this section we analyze the effect of sampling time in systems with quantizing and saturation blocks. As presented in [10], holding a control signal too long may cause the trajectory to cross over quantization rectangles with opposite sign control signals (intended to create chattering asymptotes), and to eventually leave the region of attraction. Thus, in order to have a trajectory behave as intended, we must guarantee that the trajectory is mapped, at least once, into one of the quantization rectangles adjacent to the quantization rectangle where it was last mapped. For this purpose, we decompose a trajectory into its axial components, then calculate the time it takes each component to cross a quantization level, then take the minimum time of both components. In the general case it is not easy to decompose the trajectory into its components, but this may be done for the systems in equations (2) and (4). Let us consider first the case of the double integrator. The solution of the system in equation (2) for an initial condition on the border of the j^{th} quantization square is given by

$$x_1(t) = x_1(0) + x_2(0)t + \frac{u_j t^2}{2} \quad (17)$$

$$x_2(t) = x_2(0) + u_j t. \quad (18)$$

The final position for each axis is one quantization rectangle towards the innermost quantization level, as shown next

$$x_1(t_1) = x_1(0) - sign(x_1(0)) \left[\frac{2u_{max}}{N} \right] \quad (19)$$

$$x_2(t_2) = x_2(0) - sign(x_2(0)) \left[\frac{2u_{max}}{N} \right] \quad (20)$$

Substituting equations (19) and (20) in equations (17) and (18), respectively, and solving for t_i, yields

$$t_1 = -\frac{x_2(0)}{u_j} \pm \frac{\sqrt{x_2^2(0) - 2u_j sign(x_1(0)) \left[\frac{2u_{max}}{N} \right]}}{u_j} \quad (21)$$

$$t_2 = -\frac{sign(x_2(0)) \left[\frac{2u_{max}}{N} \right]}{u_j}. \quad (22)$$

We also consider the solution for $x_1(t)$ in the innermost square, where there is no control signal applied and the value of x_2 is constant. The minimum time that x_1 takes to cross this square is for $x_2 = \pm \frac{u_{max}}{N}$, assuming $x_2 = \frac{u_{max}}{N}$, the equation of the solution results

$$x_1(t) = x_1(0) + \frac{u_{max}}{N} t \quad (23)$$

substituting $x_1(0) = -\frac{u_{max}}{N}$ and $x_1(t) = \frac{u_{max}}{N}$, and solving for t results

$$t = \frac{\frac{2u_{max}}{N}}{\frac{u_{max}}{N}} = 2. \quad (24)$$

Solving for t_i in equations (21) and (22), applying the control signal u_j in each quantization rectangle, the upper bound in the sampling time t_s, is then given by

$$t_s < \min_{i,j} t_i(u_j) \quad (25)$$

Using the same data as in example 1, the upper bound for t_s is

$$t_s < \frac{1}{8} \quad (26)$$

Using $t_s = \frac{1}{10} = 100\, msec$, and running again the Simulink® program for the double integrator of example 1, results in the trajectory shown in Figure 7. We can see from Figure 7 that, despite the large oscillations in the sliding asymptote, the trajectory is contained inside the innermost quantization square.

Now for the case of the system in equation (4), we have that the component on the i^{th} axis of the solution for an initial condition starting at the border of the j^{th} quantization square is given by

$$x_i(t) = e^{\lambda_i t} x_i(0) + \frac{u_j}{\lambda_i} \left(e^{\lambda_i t} - 1 \right) \quad (27)$$

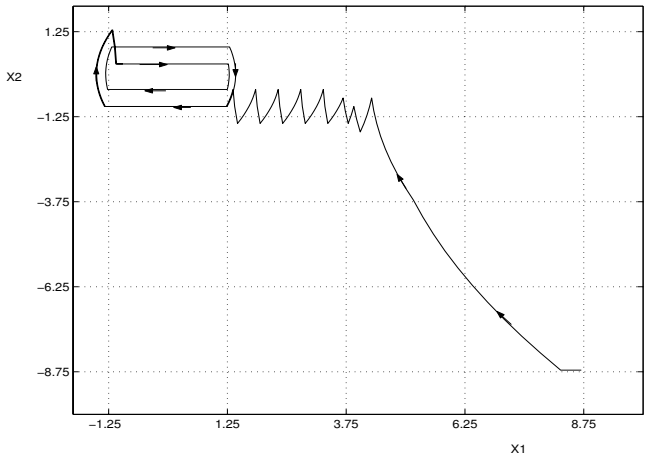

Fig. 7. Solution of the quantized double integrator with sampling time near the maximum.

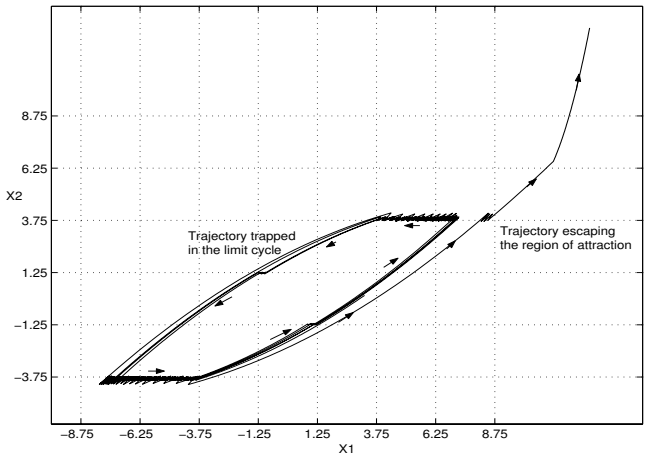

Fig. 8. Solution of the quantized decoupled integrators with sampling times above and below the maximum.

where $x_i(t)$ is given by

$$x_i(t) = x_i(0) + sign(u_j)\left[\frac{2u_{max}}{\lambda_i N}\right] \tag{28}$$

solving for the time

$$t = \frac{1}{\lambda_i}\ln\left(\frac{x_i(0) + sign(u_j)\left[\frac{2u_{max}}{\lambda_i N}\right] + \frac{u_j}{\lambda_i}}{x_i(0) + \frac{u_j}{\lambda_i}}\right) \tag{29}$$

Then, the sampling time t_s must satisfy

$$t_s < \min_{j,i} \quad \ln\left(\frac{x_i(0) + sign(u_j)\left[\frac{2u_{max}}{\lambda_i N}\right] + \frac{u_j}{\lambda_i}}{x_i(0) + \frac{u_j}{\lambda_i}}\right)^{\frac{1}{\lambda_i}} \tag{30}$$

However, this upper bound for the sampling time does not account for the case when the next mapping of the trajectory is in a contiguous square but outside the region of attraction. Then, in order to avoid a trajectory being mapped outside the region of attraction, we have to consider the minimum distance between a chattering boundary and the envelope of the region of attraction. For the particular case of the system in equation (15), and from Figure 5, we can see that this minimum distance appears between the intersection $(-3.75, -3.75)$ and the envelope of the region of attraction passing just below. Then, a trajectory starting at $(-3.8, -3.7)$, with control signal $u = -10$, will cross the envelope in approximately $t_e = 20\,msec$. Figure 8 shows the trajectories for the initial condition $(-3.8, -3.7)$, with sampling times $t_s = 20\,msec$ and $t_s = 10\,msec$. We can see that the trajectory with sampling time $t_s = 10\,msec$ is trapped on the limit cycle, while the trajectory with sampling time $t_s = 20\,msec$ escapes from the region of attraction.

IV. Conclusions

We have exposed the effects of quantizing and saturation blocks in the states and control signal, along with the effects of sampling time. We showed the presence of limit cycles in open-loop unstable systems. The use of selection functions avoids state conflicts while assigning control signals, and allow us to arbitrary assign the value of the control signal in a conflicting quantization square. We have presented lower bounds for the number of bits required in the A/D converters, and presented upper bounds in the sampling times. We have finally presented a graphical method to obtain the envelope of the region of attraction.

References

[1] J. Baillieul, "Feedback Coding for Information-based Control: Operating Near the Data-Rate Limit", 41st IEEE Conference on Decision and Control, Las Vegas, Nevada, USA, December 2002.

[2] B. Bamieh, "Intersample and Finite Wordlength Effects in Sampled-Data Problems", IEEE Transactions on Automatic Control, Vol. 48, No. 4, April 2003.

[3] R. W. Brockett, D. Liberzon, "Quantized Feedback Stabilization of Linear Systems", IEEE Transactions on Automatic Control, Vol. 45, No. 7, July 2000.

[4] D. F. Delchamps, "Stabilizing a Linear System with Quantized State Feedback", IEEE Transactions on Automatic Control, Vol. 35, No. 8, August 1990.

[5] N. Elia, S. K. Mitter, "Stabilization of Linear Systems with Limited Information", IEEE Transactions on Automatic Control, Vol. 46, No. 9, September 2001.

[6] A. Hassibi, S. P. Boyd, J. P. How, "Control of Asynchronous Dynamical Systems with Rate Constraints on Events", 38th IEEE Conference on Decision and Control, Phoenix, Arizona, USA, December 1999.

[7] H. Ishii, B. A. Francis, "Stabilization with Control Networks", Automatica. April 2002, pp. 1745-1751.

[8] A. Isidori, "Nonlinear Control Systems", Springer-Verlag, 3rd Edition, 1997.

[9] K. Li, J. Baillieul, "The Appropriate Quantization for Digital Finite Communication Bandwidth (DFCB) Control", 42nd IEEE Conference on Decision and Control, Maui, Hawaii, USA, December 2003.

[10] R. Sandoval R., C. T. Abdallah, "On the Effects of Quantization and Sampling in LTI Systems", 12^{th} Mediterranean Conference on Control and Automation, Kusadasi, Aydin, Turkey. June 2004.

[11] A. R. Teel, "Global stabilization and restricted tracking for multiple integrators with bounded controls", Systems & Control Letters, 18(1992) 165-171.

[12] W. S. Wong, R. W. Brockett, "Systems with Finite Communication Bandwidth Constraints II: Stabilization with Limited Information Feedback", IEEE Transactions on Automatic Control, 44, pp. 1049-53.