On the stability and numerical stability of a model state dependent delay differential equation

Felicia Maria G. Magpantay

Doctor of Philosophy

Department of Mathematics and Statistics

McGill University
Montreal,Quebec

2011-08-15

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

ACKNOWLEDGEMENTS

I would like to thank my Ph.D. supervisor, Professor Antony R. Humphries for his teaching, guidance and support throughout the preparation of this thesis. I would also like to thank Professor Nicola Guglielmi for his help with this endeavour. It has taken many years of school for me to reach this point of my academic life and I am grateful to my teachers, professors and friends in the many institutions that I have attended during those years: the Philippine Science High School scholarship system, the University of the Philippines, Trent University (in particular the Trent International Program), the University of Western Ontario and of course, McGill University. I would especially like to thank Layan El Hajj, Philip Rempel, Ivo Panayotov, Victoria de Quehen and Gabriel Gauthier-Shalom for their companionship and support as we made our way through graduate school. Last but not the least, I would like to thank my family for all my strength and for having raised me to become independent without ever losing sight of home.

This study was funded by the Institut des Sciences Mathématiques, McGill University and NSERC through the research grant of Professor Humphries.

ABSTRACT

In this thesis the following model state dependent delay differential equation is considered,

$$\varepsilon\dot{u}\left(t\right)=\mu u\left(t\right)+\sigma u\left(t-a-cu\left(t\right)\right).$$

For fixed ε, a and c, the analytical stability region of this equation is known and it is the same for both the constant delay ($c=0$) and state dependent delay ($c\neq 0$) cases. Different approaches are used to directly prove stability in parts of this analytic region for the state dependent DDE: first using a Gronwall argument and then using a Lyapunov-Razumikhin method which is a generalisation of the work of Barnea [6] who considered the $\mu=c=0$ case. The parameter regions in which stability is proven by these methods contain the entire delay independent portion of the analytical stability region and parts of the delay dependent portion. These methods are then extended to show the stability of the backward Euler method with linear interpolation applied to the model DDE. Using the Lyapunov-Razumikhin method, stability is proven in larger parameter regions that depend on the stepsize, but always contain the region found for the DDE. Analytic expressions for regions in which general Θ methods are stable were also derived and evaluated numerically. In the last chapter a new scheme for numerically integrating scalar DDEs with multiple state dependent delays is presented. This scheme is based on singularly diagonally implicit Runge-Kutta (SDIRK) methods in order to solve stiff problems such as the equation above with small ε. Due to the nature of SDIRK methods, if there is no overlapping then at each step a set of scalar equations are solved one-by-one using a Newton-bisection algorithm. New continuous extensions which are piecewise polynomial are chosen to accompany the SDIRK scheme so as not to destroy the SDIRK structure in the overlapping cases and to avoid the problem of spiking when there is a sharp change in the numerical solution.

ABRÉGÉ

Dans cette thèse, l'équation différentielle à retard (DDE) modèle d'état dépendant suivante est considérée,

$$\varepsilon \dot{u}\left(t\right) = \mu u\left(t\right) + \sigma u\left(t - a - cu\left(t\right)\right).$$

Pour ε, a et c fixés, la région de stabilité analytique de cette équation est connue et est la même pour le retard constant ($c = 0$) ainsi que pour l'état de retard dépendant ($c \neq 0$). Différentes approches sont utilisées pour prouver directement la stabilité dans certaines parties de cette région analytique pour la DDE d'état dépendant: d'abord en utilisant un argument de Gronwall, puis en utilisant une méthode de Lyapunov-Razumikhin qui est une généralisation du travail de Barnea [6] qui considère le cas $\mu = c = 0$. Les régions de paramètres dans lesquelles la stabilité est prouvée par ces méthodes contiennent la partie entière de retard indépendant de la région de stabilité analytique et certaines parties de la portion de retard dépendant. Ces méthodes sont ensuite étendues pour montrer la stabilité de la méthode d'Euler arrière avec interpolation linéaire appliquée à la DDE modèle. En utilisant la méthode de Lyapunov-Razumikhin, la stabilité est prouvée dans des régions de paramètres plus grandes qui dépendent du pas de discrétisation, mais qui contiennent toujours la région trouvée pour la DDE. Des expressions analytiques pour les régions dans lesquelles les méthodes Θ générales sont stables ont également été tirées et évaluées numériquement. Dans le dernier chapitre d'un nouveau schéma pour intégration numérique des DDE scalaires avec des multiples retards d'état dépendant est présenté. Ce schéma est basé sur des méthodes de Runge-Kutta singulièrement et diagonalement implicites (SDIRK) afin de résoudre des problèmes raides tels que l'équation ci-dessus avec des petites valeurs de ε. En raison de la nature des méthodes SDIRK, s'il n'y a pas de chevauchement, alors à chaque iteration un ensemble d'équations scalaires sont résolues, une par une, en utilisant un algorithme de bissection de Newon. Des nouvelles extensions continues qui sont polynomiales par morceaux sont choisies pour accompagner le schéma SDIRK afin de ne pas détruire la structure SDIRK dans les cas de chevauchement et pour éviter le problème des piques quand il y a un changement brusque de la solution numérique.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure page

CHAPTER 1
Introduction

Ordinary differential equations (ODEs) use the current time and current state of a system to determine the rate of change of the state. The theory behind these equations as well as the numerical methods that are employed to solve them is well-established, a good thing since ODEs prove to be effective in describing many physical phenomena. However, many phenomena require information from the past state of the system as well as the current state. This naturally leads us to consider delay differential equations (DDEs). The additional complication of requiring values from the past results in a much more difficult equation to solve both analytically and numerically. For instance, consider the logistic model (also called the Verhulst-Pearl model) for population dynamics,

$$\dot{u}(t) = au(t)(1 - u(t)), \tag{1.0.1}$$

where $a > 0$. In this model the total carrying capacity of the system is one, and the rate of change of the population is proportional to its current size $u(t)$ multiplied by $1 - u(t)$ which reflects the bottleneck effect due to competition within the population for resources. A modification of this ODE model was presented by Hutchinson [31] in 1948,

$$\dot{u}(t) = au(t)(1 - u(t-1)). \tag{1.0.2}$$

This model takes into account a time lag in the bottleneck effect which may be due to the maturation time of the population, or the recovery time of resources. The solutions of (1.0.1) all monotonically converge to the fixed point at $u = 1$. In contrast, one may find solutions of (1.0.2) that are monotonic for $a \in \left(0, \frac{1}{e}\right)$, exhibit decaying oscillations about $u = 1$ for $a \in \left(\frac{1}{e}, \frac{\pi}{2}\right)$ and approach periodic orbits for $a > \frac{\pi}{2}$ (see Figure 1–1). Wright [60] proved the convergence of all positive solutions to $u = 1$ for $a \in \left[0, \frac{3}{2}\right]$. The convergence for the case $a \in \left(\frac{3}{2}, \frac{\pi}{2}\right)$ is still an open conjecture [54].

Another difference we note between (1.0.1) and (1.0.2) is that the ODE requires an initial value while the DDE requires an initial history function defined over an interval of length one. If we think of these equations as dynamical systems where the solutions describe a flow from a

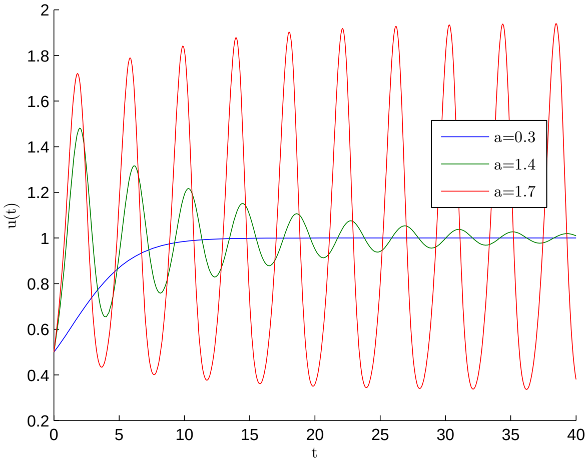


Figure 1–1: Sample trajectories of Hutchinson's equation (1.0.2) for population dynamics

space of initial vectors, then the DDE is a flow from a space of functions, an infinite dimensional vector space. This illustrates the infinite dimensionality of DDE systems, a property of DDEs that is further discussed in texts by Bellman and Cooke [10] and Hale and Verduyn Lunel [29]. Here we just note that this increase in dimensionality led to the interesting changes in dynamics in going from (1.0.1) to (1.0.2). In other DDE systems one may also observe chaotic behavior of solutions even in the scalar case, non-injectivity between initial data and solutions, and possible termination of solutions.

There are now many mathematical models that incorporate time delays. In biology they are used to add immune response time in disease dynamics, and sojourn times in epidemic models [3, 54]. Delays arise in engineering systems from feedback loops and observation lags such as in traffic control [17, 46]. DDEs are now also being used in models in various other fields of science such as chemical kinetics, electrodynamics, optics, ecology, just to name a few [5, 36]. The delays in these models may be constant, time dependent or state dependent. Fewer models use state dependent delays, perhaps due to the difficulties in analysing and numerically simulating such equations. This is the motivation for our work in tackling some of the various complications state dependent DDEs bring to the table. Our results will hopefully encourage more people to utilize the wealth of dynamics and wide potential for practical applications of state dependent DDEs.

1.1 Retarded functional differential equations and delay differential equations

In this thesis the following model state dependent DDE is considered

$$\varepsilon \dot{u}\left(t\right) = \mu u\left(t\right) + \sigma u\left(t - a - cu\left(t\right)\right). \tag{1.1.1}$$

This is the simplest equation one can write with a state dependent delay. The only nonlinearity in this equation comes from the state dependence and from this, interesting dynamics arise that cannot be found in the constant delay case ($c = 0$), much less the ODE case ($\sigma = 0$) [38, 39, 40, 41]. This model problem is a prototype for a general class of DDE problems when the solutions are close to zero. In this thesis the dynamics of (1.1.1) are considered as well as the generalisation of (1.1.1) to N delays,

$$\varepsilon \dot{u}\left(t\right) = -\gamma u\left(t\right) - \sum_{i=1}^{N} \kappa_i u\left(t - a_i - c_i u\left(t\right)\right). \tag{1.1.2}$$

If $N = 1$ then this is just (1.1.1). The change in notation from μ and σ to γ and κ is due to our focus on bounded solutions of the N-delay problem.

In equations (1.1.1) and (1.1.2), the derivative depends on values of the state at discrete delayed times. There are other types of DDEs such as DDEs with distributed delays and neutral DDEs in which the derivative depends on values of the derivative in the past. These equations are covered under the general theory of retarded functional differential equations (RFDEs). Our treatment of RFDEs including the definitions, results on existence and uniqueness is based on the text by Hale and Verduyn Lunel [29].

Let $\mathbb{R}^d$ be the d-dimensional linear vector space over the reals equipped with the Euclidean norm $|\cdot|$. Let $r \geqslant 0$ and $C = C\left([-r, 0], \mathbb{R}^d\right)$ be the Banach space of continuous functions mapping $[-r, 0]$ to $\mathbb{R}^d$ with the supremum norm denoted by $\|\cdot\|$. Let Ω be an open subset of $\mathbb{R} \times C$ and $f : \Omega \to \mathbb{R}^d$. If $u \in C\left([t_0 - r, t_f], \mathbb{R}^d\right)$ then for every $t \in [0, t_f]$ define $u_t \in C$ such that

$$u_t\left(\theta\right) = u\left(t + \theta\right), \quad \theta \in \left[-r, 0\right].$$

Then with $\cdot$ indicating a right hand derivative consider the RFDE

$$\dot{u}(t) = f\left(t, u_t\right). \tag{1.1.3}$$

A solution of (1.1.3) passing through (t_0, ϕ) is a function $u \in C\left([t_0 - r, t_f), \mathbb{R}^d\right)$ for some $t_f > t_0$ such that $(t, u_t) \in \Omega$ and u satisfies (1.1.3) for $t \in [t_0, t_f)$ and $u_{t_0} = \phi$. We write $u = u\,(t_0, \phi)$ when we want to explicitly show the dependence of the solution on the initial time and function.

Theorem 1.1.1 (Local existence theorem from Hale and Verduyn Lunel [29])**.** *Suppose $\Omega \subseteq \mathbb{R} \times C$ is open, $f : \Omega \to \mathbb{R}^d$ is continuous, and Lipschitz continuous with respect to its second argument in each compact set in Ω. Then there is a unique solution to* (1.1.3) *through any $(t_0, \phi) \in \Omega$.* □

To observe the dynamics of solutions to RFDEs we need more than the local existence of solutions. A solution u to (1.1.3) on an interval $[t_0 - r, t_f)$ is said to have a continuation v if there (i) is a $\tilde{t}_f > t_f$ such that v is defined on $\left[t_0 - r, \tilde{t}_f\right)$, (ii) v is a solution to (1.1.3) and (iii) v coincides with u on $[t_0, t_f)$. A solution is non-continuable if no continuation exists.

Theorem 1.1.2 (Continuation theorem from Hale and Verduyn Lunel [29])**.** *Suppose $\Omega \subseteq \mathbb{R}\times C$ is open and $f : \Omega \to \mathbb{R}^d$ is completely continuous (f takes closed bounded sets of Ω into bounded sets of $\mathbb{R}^d$). Let u be a noncontinuable solution of* (1.1.3) *on* $[t_0 - r, t_f)$*. Then for any compact set $W \in \Omega$, there is a $t_W \in (t_0, t_f)$ such that $(t, u_t) \notin W$ for $t_W \leqslant t < t_f$.* □

In general, even if f is not locally Lipschitz we assume that f is completely continuous. Otherwise strange behaviors of noncontinuable solutions in finite time intervals may be observed.

RFDE theory covers a wide range of equations, and for this reason the conditions for existence and uniqueness are stricter than necessary for some classes of equations. In particular, the theory is difficult to apply to the class of state dependent DDEs because there is no *a priori* bound on the delay terms. Also, and perhaps more importantly, the existence and uniqueness theorems for RFDEs require Lipschitz continuity with respect to its second argument, a strong condition since the second argument is a function on $[-r, 0]$. For instance, (1.1.1) written as a an RFDE (1.1.3) yields $F\,(t, \phi) = \frac{\mu}{\varepsilon}\phi(0) + \frac{\sigma}{\varepsilon}\phi\,(-a - c\phi(0))$. Such an F is not Lipschitz continuous in $\phi \in C$ so we may not use Theorems 1.1.1 and 1.1.2. Instead we consider the theory developed by Driver [14] for a general class of DDEs with discrete delays.

Consider the general N-delay differential equation with state dependent delays

$$\begin{cases} \dot{u}\,(t) = f\,(t, u\,(t)\,, u\,(\alpha_1\,(t, u(t)))\,, ..., u\,(\alpha_N\,(t, u(t))))\,, & t_0 \leqslant t \leqslant t_f \\ u\,(t) = \varphi\,(t)\,, & t \leqslant t_0. \end{cases} \tag{1.1.4}$$

Let $D \subseteq \mathbb{R} \times \mathbb{R}^{(N+1)d}$ be an open set and let D^* be the projection of D to $\mathbb{R} \times \mathbb{R}^d$ (the space of the first $d+1$ coordinates). Let the function $f \in C\left(D, \mathbb{R}^d\right)$ and $\alpha_i\left(t,u\right) \in C\left(D^*, \mathbb{R}\right)$ for $i = 1, ..., N$. We call $\alpha_i\left(t,u\right)$ the deviated arguments and $\tau_i\left(t,u\right) = t - \alpha_i\left(t,u\right)$ the delay terms. Let $\tau_0 \in [0, \infty]$ and let the deviated arguments be such that $t_0 - \tau_0 \leqslant \alpha_i\left(t,u\right) \leqslant t$ for all $(t,u) \in D^* \cap \{t \geqslant t_0\}$, $i = 1, ..., N$. If $\tau_0 < \infty$ then let $\varphi \in C\left(\left[t_0 - \tau_0, t_0\right], \mathbb{R}^d\right)$ and

$$\left(t_0, \varphi\left(t_0\right)\right), \left(t_0, \varphi\left(\alpha_1\left(t_0, \varphi\left(t_0\right)\right)\right)\right), ..., \left(t_0, \varphi\left(\alpha_N\left(t_0, \varphi\left(t_0\right)\right)\right)\right) \in D^*.$$

A solution to (1.1.4) is a function $u(t) : [t_0 - \tau_0, t_f) \to \mathbb{R}^d$ such that the equations in (1.1.4) are satisfied, $u(t)$ is continuous and $\left(t, u\left(t\right)\right) \in D^*$ for $t \in [t_0, t_f)$. If $\tau_0 = \infty$ then change the intervals $[t_0 - \tau_0, t_f)$ to $(-\infty, t_f)$ everywhere and the same definition of a solution applies.

Theorem 1.1.3 (Local existence theorem, Driver [14])**.** *Let $f\left(t, u, v_1, ..., v_N\right) \in C\left(D, \mathbb{R}^d\right)$ be Lipschitz continuous in u, v_1, ..., v_N in every compact subset of D. Let $\alpha_i\left(t,u\right) \in C\left(D^*, \mathbb{R}\right)$ be Lipschitz continuous in u in every compact subset of D^* for $i = 1, ..., N$. Let $\tau_0 \in [0, \infty]$ be such that $t_0 - \tau_0 \leqslant \alpha_i\left(t,u\right) \leqslant t$ for all $(t,u) \in D^* \cap \{t \geqslant t_0\}$, $i = 1, ..., N$. If $\tau_0 < \infty$ then let $\varphi(t)$ be Lipschitz continuous on $[t_0 - \tau_0, t_0]$. If $\tau_0 = \infty$ then let $\varphi(t)$ be Lipschitz continuous on each finite subinterval of $(-\infty, t_0]$. Then there is a $\delta > 0$ such that a unique solution to* (1.1.4) *exists on $[t_0, t_0 + \delta)$.* □

The Lipschitz continuity condition here is less stringent since it is with respect to vector arguments, unlike in Theorem 1.1.2 where it is with respect to a function argument. It is easy to verify that equations (1.1.1) and (1.1.2) are Lipschitz in this sense.

Theorem 1.1.4 (Extended existence theorem, Driver [14])**.** *Let $f\left(t, u, v_1, ..., v_N\right) \in C\left(D, \mathbb{R}^d\right)$ and $\alpha_i\left(t,u\right) \in C\left(D^*, \mathbb{R}\right)$ for $i = 1, ..., N$. Let $\tau_0 \in [0, \infty]$ be such that $t_0 - \tau_0 \leqslant \alpha_i\left(t,u\right) \leqslant t$ for all $(t,u) \in D^* \cap \{t \geqslant t_0\}$, $i = 1, ..., N$. If $\tau_0 < \infty$ then let $\varphi(t)$ be continuous on $[t_0 - \tau_0, t_0]$. If $\tau_0 = \infty$ then let $\varphi(t)$ be continuous on $(-\infty, t_0]$. Then any solution to* (1.1.4) *can be extended to $[t_0, t_f)$ where $t_0 < t_f \leqslant \infty$. If t_f is finite and cannot be increased then one of the following cases must occur:*

1. $\limsup_{t \to \infty} \left\| u\left(t\right), u\left(\alpha_1\left(t, u(t)\right)\right), ..., u\left(\alpha_N\left(t, u(t)\right)\right) \right\| = \infty$
2. $\left(t, u(t), u\left(\alpha_1\left(t, u(t)\right)\right), ..., u\left(\alpha_N\left(t, u(t)\right)\right)\right)$ *comes arbitrarily close ∂D.* □

Driver notes that for ordinary differential equations, $(t, u(t))$ approaches ∂D, instead of just coming arbitrarily close to it. For general DDEs, stronger conditions are required to obtain this sort of behavior.

1.2 Extending Runge-Kutta methods to solve DDEs

As with ODEs, DDEs generally have to be solved numerically. The standard approach for doing this is to look into existing ODE methods and extend them to solve DDEs. It turns out that different types of extensions work well for some DDE problems but not for all of them. Bellen and Zennaro [8] stress that the integration of different classes of DDEs requires methods that are designed specifically for the relevant class of DDE and cannot be based on a simple adaptation of an ODE method. For instance, if during the integration one encounters a vanishing delay ($\tau\left(t, u(t)\right) = 0$) then an explicit numerical integrator for ODEs becomes implicit for this DDE. Stiff delay equations also lead to new problems that are not covered by the numerical analysis of stiff ODEs. These and other issues that arise in extending ODE methods to solve DDEs are introduced in this section and will be further discussed in Chapters 4 and 5.

We begin with our notation for Runge-Kutta (RK) methods. An s-stage RK method may be represented by its Butcher tableau.

$$\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b^T} \end{array} = \begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}$$

The b_i's are called the weights, and c_i's are called abscissae (for most common methods, $c_i = \sum_{j=1}^{s} a_{ij} \in [0,1]$). Suppose that the ODE to be solved is

$$\begin{cases} \dot{u}(t) = g\left(t, u(t)\right), \quad t_0 \leqslant t \leqslant t_f, \\ u(t_0) = \varphi_0 \end{cases} \tag{1.2.1}$$

where $g : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ and $\varphi_0 \in \mathbb{R}^d$. Given a mesh $\Delta = \{t_n\}_{n=0}^{n_f}$ of discrete time values, the approximation u_n to the solution of (1.2.1) at time t_n is obtained by setting $u_0 = \varphi_0$ and solving

$$\begin{aligned} Y_{n+1}^{(i)} &= u_n + h_{n+1} \sum_{j=1}^{s} a_{ij} g\left(t_{n+1}^{(j)}, Y_{n+1}^{(j)}\right), \qquad i = 1, ..., s \\ u_{n+1} &= u_n + h_{n+1} \sum_{i=1}^{s} b_i g\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}\right), \end{aligned} \tag{1.2.2}$$

where $t_{n+1}^{(i)} = t_n + c_i h_{n+1}$ and $h_{n+1} = t_{n+1} - t_n$.

Now we extend the RK method so that it can be used to solve DDEs. Consider the following DDE with one general state dependent delay

$$\begin{cases} \dot{u}(t) = f\left(t, u\left(t\right), u\left(\alpha\left(t, u(t)\right)\right)\right), & t_0 \leqslant t \leqslant t_f, \\ u(t) = \varphi\left(t\right), & t \leqslant t_0, \end{cases} \tag{1.2.3}$$

where $f : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ and $\varphi : \mathbb{R} \to \mathbb{R}^d$. We use one delay only for simplicity. The extension to multiple delays is straightforward. Assume $\alpha\left(t, u(t)\right) = t - \tau\left(t, u(t)\right) \leqslant t$ for all t. The right-hand-side function requires the value of the solution at this past time which will generally not fall on a mesh point. A natural way to approximate this value is to augment the RK method with a continuous extension that interpolates between mesh values. These methods are called continuous RK (CRK) methods. The interpolant within a time interval $[t_n, t_{n+1}]$ can be constructed by making use of information from the stages of the RK method in the same interval. We follow the standard treatment of doing this by changing the b_i's from constants to polynomial functions $b_i\left(\theta\right)$, $\theta \in [0, 1]$ satisfying

$$b_i(0) = 0, \quad b_i(1) = b_i, \quad i = 1, ..., s$$

Given a mesh $\Delta = \{t_n\}_{n=0}^{n_f}$ of discrete time values, the approximation u_n to the solution of (1.2.3) at time t_n is obtained by setting $u_0 = \varphi\left(t_0\right)$ and solving

$$\begin{aligned} Y_{n+1}^{(i)} &= u_n + h_{n+1} \sum_{j=1}^{s} a_{ij} f\left(t_{n+1}^{(j)}, Y_{n+1}^{(j)}, \tilde{Y}_{n+1}^{(j)}\right), \qquad i = 1, ..., s, \\ u_{n+1} &= u_n + h_{n+1} \sum_{i=1}^{s} b_i f\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}, \tilde{Y}_{n+1}^{(i)}\right), \end{aligned} \tag{1.2.4}$$

where the $\tilde{Y}_{n+1}^{(j)}$ (called the spurious stages) are the values of the continuous extension at time $\alpha\left(t_n^{(j)}, Y_{n+1}^{(j)}\right)$. At time $t = t_n + \theta h_{n+1} \in [t_n, t_{n+1}]$, the continuous extension η is defined by

$$\eta\left(t_n + \theta h_{n+1}\right) = u_n + h_{n+1} \sum_{i=1}^{s} b_i(\theta) f\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}\right). \tag{1.2.5}$$

These types of continuous extensions which only use the existing stage values of the RK method are called interpolants of the first class. It is also possible to create interpolants which interpolate using new stages that are not part of the original RK method. These are called interpolants of the second class and they are sometimes necessary to improve the accuracy of a CRK method.

Although the stages of an RK method are used to produce the continuous extension (1.2.5), in general η does not actually pass through the stage values ($\eta(t_n + c_i h_{n+1}) \neq Y_{n+1}^{(i)}$). From Bellen and Zennaro [8],

$$\{\eta(t_n + c_i h_{n+1}) = Y_{n+1}^{(i)},\ \forall i = 1, ..., s\} \quad \Leftrightarrow \quad \{b_i(c_j) = a_{ji},\ \forall i, j = 1, ..., s\}.$$

A CRK method that has this property is called *natural*. For examples of RK methods and their continuous extensions, refer to the methods discussed in Chapter 5. Bellen and Zennaro [8] contains a more comprehensive discussion of results on CRK methods.

As mentioned at the beginning of this section, extending RK methods to solve DDEs gives rise to many issues. First there is the problem of overlapping which occurs if for some i, $\alpha\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}\right) \in [t_n, t_{n+1}]$. In this case the equation for the continuous extension (1.2.5) becomes an implicit equation even for explicit RK methods. Also, it appears as if there may be as many as $2s$ unknowns in solving (1.2.4) instead of the usual s unknowns for a fully-implicit method. But notice that it is not actually the values of $Y_{n+1}^{(i)}$ and $\tilde{Y}_{n+1}^{(i)}$ that are required to solve for the update u_{n+1} or even for the continuous extension η. Rather, it is the values of the right hand side function evaluated at these stages that are necessary. Rewriting (1.2.4) in the following K-notation shows that there are only s unknowns even in the overlapping case.

$$\begin{aligned} &K_{n+1}^{(i)} = f\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}, \tilde{Y}_{n+1}^{(i)}\right) = f\left(t_{n+1}^{(j)}, u_n + h_{n+1}\sum_{j=1}^{s} a_{ij} K_{n+1}^{(j)}, \tilde{Y}_{n+1}^{(j)}\right), \qquad i = 1, ..., s, \\ &u_{n+1} = u_n + h_{n+1}\sum_{i=1}^{s} b_i K_{n+1}^{(i)}, \end{aligned} \tag{1.2.6}$$

The spurious stages are given by $\tilde{Y}_{n+1}^{(j)} = \eta\left(\alpha\left(t_{n+1}^{(j)}, u_n + h_{n+1}\sum_{k=1}^{s} a_{ij} K_{n+1}^{(k)}\right)\right)$ and (1.2.5) can be rewritten as

$$\eta\left(t_n + \theta h_{n+1}\right) = u_n + h_{n+1}\sum_{i=1}^{s} b_i(\theta) K_{n+1}^{(i)}. \tag{1.2.7}$$

Theorem 4.3.1 of Bellen and Zennaro [8] guarantees that under certain conditions (essentially uniqueness of the analytic solution and use of a small enough stepsize) solving for the continuous extension, even in the overlapping case, is a well-posed problem.

Another issue that comes up in extending RK methods to solve DDEs is the correct choice of the polynomials $b_i\left(\theta\right)$ in order to preserve the order of the method. Before going into this, here is the definition of the order of an RK method for ODEs.

Definition 1.2.1 (Bellen and Zennaro [8])**.** Recall the ODE system (1.2.1) and consider the local problem

$$\begin{cases} z_{n+1}'(t) = g\left(t, z_{n+1}(t)\right), & t_n \leqslant t \leqslant t_{n+1}, \\ z_{n+1}(t_n) = u_n^*. \end{cases} \tag{1.2.8}$$

An RK method has discrete order p if $p \geqslant 1$ is the largest integer such that, for all C^p-continuous right-hand-side functions $g(t,y)$ in (1.2.1) and for all mesh points,

$$\|z_{n+1}(t_{n+1}) - u_{n+1}\| = O\left(h_{n+1}^{p+1}\right),$$

uniformly with respect to u_n^* in any bounded subset of $\mathbb{R}^d$ and to $n = 0, ..., n_f$. We say that the interpolant η defined by (1.2.5) has uniform order q if $q \geqslant 1$ is the largest integer such that, for all C^q-continuous right-hand-side functions $g(t,y)$ and for all mesh points,

$$\max_{t_n \leqslant t \leqslant t_{n+1}} \|z_{n+1}(t) - \eta(t)\| = O\left(h_{n+1}^{q+1}\right).$$

For an RK method with order p, the global error is order p, that is

$$\|u(t_n) - u_n\| = O\left(h^p\right), \quad n = 0, ..., n_f$$

where $h = \max_{1 \leqslant n \leqslant n_f} h_n$.

Numerical methods for DDEs derived from RK methods with order p are not necessarily of order p when solving DDEs. Two issues that can cause the loss of order are the choice of continuous extension and the improper tracking of discontinuity points. *Discontinuity points* (also called *breaking points*) are values of t for which a derivative of the solution to (1.2.3) becomes discontinuous. Since the solution at time t_0 is given by $\dot{u}\left(t_0\right) = f\left(t_0, \varphi\left(t_0\right), \varphi\left(\alpha\left(t_0, \varphi\left(t_0\right)\right)\right)\right)$ but the history function φ is arbitrary, then in general one may expect that the derivative is discontinuous at t_0. In this case $t = t_0$ is called a 0-level primary discontinuity point and following standard notation it is labeled $\xi_{0,1}$, where the first subscript indicates the level and the second is an index. At the points $t = \xi_{1,j}$ where $\alpha_i(\xi_{1,j}, u(\xi_{1,j})) = t_0$ for some i, the discontinuity in $\dot{u}(t)$ at t_0 will cause a discontinuity in $\ddot{u}(t)$ at $t = \xi_{1,j}$, and these points are called 1-level primary discontinuity points. Similarly at points $t = \xi_{2,k}$ such that $\alpha_i(\xi_{2,k}, u(\xi_{2,k})) = \xi_{1,j}$ for some i and j there can be a discontinuity in $u^{(3)}(t)$. In this way the discontinuity in the first derivative at $t = 0$ propagates to discontinuities in higher derivatives of $u(t)$ at later times. Bellen and Zennaro [8] provides a more complete discussion of breaking points.

Theorem 1.2.2 (Order of a CRK method, Bellen and Zennaro [8], Theorem 6.1.2)**.** *Let $f(t,u,v)$ in* (1.2.3) *be C^p-continuous in $[t_0,t_f]\times\mathbb{R}^d\times\mathbb{R}^d$, the initial function $\varphi(t)$ be C^p-continuous and the delay $\tau(t,u)$ be C^p-continuous in $[t_0,t_f]\times\mathbb{R}^d$. Assume that the discrete mesh $\{t_0,...,t_{n_f}\}$ includes all the discontinuity points of order $\leqslant p$ in $[t_0,t_f]$. If the underlying CRK method has discrete order p and uniform order q, then the method* (1.2.6)-(1.2.7) *has discrete global order and uniform global order $q' = \min\{p,q+1\}$; that is*

$$\max_{1\leqslant n\leqslant n_f}\|u(t_n)-u_n\| = O\left(h^{q'}\right), \quad \max_{t_0\leqslant t\leqslant t_{n_f}}\|u(t)-\eta(t)\| = O\left(h^{q'}\right)$$

where $h = \max_{1\leqslant n\leqslant n_f} h_n$. □

Another issue that comes up in the numerical integration of DDEs is stability failure which will be discussed in Chapter 4. In Chapter 5 a class of RK methods, called singularly diagonally implicit RK methods, is implemented to solve stiff DDE problems in which the problems of overlapping and capturing a discontinuity point in the mesh sometimes occur in tandem. We also deal with the appropriate choice of continuous extension for this method, preferring those that preserve the structure of the SDIRK method to those that preserve the order of the method.

1.3 Literature review

The general theory of RFDEs and DDEs has been developed over the years by many authors. The classical texts are by Bellman and Cooke [10], El'sgol'ts and Norkin [15], Hale [28], Hale and Verduyn Lunel [29] and Kolmanovskii and Myshkis [33]. Driver [14] contains the results on existence, uniqueness and continuation of solutions for DDEs with multiple state dependent delays. Hartung et al. [30] and Walther [58] contain more recent results on the solution manifold of state dependent delay equations.

There have also been many recent texts on both theory and applications of DDEs [3, 17, 36, 54]. Mallet-Paret and Nussbaum [38, 39, 40, 41], Mallet-Paret, Nussbaum and Paraskevopoulos [42] have considered the model equation (1.1.1) and looked into so-called slowly oscillating periodic solutions as $\varepsilon\to 0$. By the classical arguments of El'sgol'ts and Norkin [15], the analytical solution of the $c = 0$ case of (1.1.1) may be written as

$$u\,(t) = \sum_k C_k \exp\left(\lambda_k t\right).$$

Applying this to (1.1.1) with $c = 0$ yields the characteristic equation for the eigenvalues of the DDE. A necessary and sufficient condition for the solutions of this DDE to approach zero as $t \to \infty$ is that all the characteristic roots have negative real parts. The stability region in the (μ, σ) parameter has been derived by Bellman and Cooke [10] for the constant delay case. For the state dependent case, results by Györi and Hartung [25] show that the state dependent DDE is exponentially stable if and only if the trivial solution of the constant delay problem is exponentially stable.

The method of Lyapunov functions for ODEs were first extended to Lyapunov functionals for RFDEs by Krasovskii [34]. For this reason these functionals are sometimes called Lyapunov-Krasovskii functionals. Razumikhin [51] developed the theory on how one might go back from the more difficult Lyapunov functionals for DDEs to Lyapunov functions again. The proof and some applications of this theorem are presented by Barnea in [6] and he also applied it to (1.1.1) with $\mu = c = 0$. Following his method of proof, we generalise his results to arbitrary μ and c. Other papers that state and employ theorems of the Razumikhin-type are Ivanov, Liz and Trofimchuk [32] and Krisztin [35]. A comprehensive discussion of Lyapunov functionals and functions for DDEs is available in the text by Hale and Verduyn Lunel [29] among other texts.

The main reference for the theory of numerical methods for DDEs is the text by Bellen and Zennaro [8]. Baker, Paul and Willé [5] provide an introduction to issues in numerically solving DDEs. More recent results on a more general class of equations (including neutral delay equations) are available in Bellen et al. [9]. The generalization of A-stability for ODE methods to P- and GP-stability for DDE methods was introduced by Barwell [7]. Zennaro [61] showed that any A-stable method is also P-stable for DDEs. The extensions to D-stability were considered by Wiederholt [59] and Guglielmi [21]. The stability of numerical methods for DDEs have also been looked at by Al-Mutib [1], Baker and Paul [4], Calvo and Grande [12], Guglielmi [20, 21], Guglielmi and Hairer [22], Liu and Spijker [37], Maset [43] and Torelli [57].

The problems involved in extending RK methods to solve DDEs are documented in Bellen and Zennaro [8]. These problems include loss of order, loss of stability, discontinuity tracking and the choice of continuous extensions. Guglielmi and Hairer also talk about the problems of discontinuity tracking in [24]. There are several DDE solvers currently available. Particularly relevant to this work is DDE23 by Shampine and Thompson [53] which is a friendly MATLAB solver for non-state dependent DDEs, DDESD by Shampine [52] which is the MATLAB solver for DDEs with state dependent delays and RADAR5 by Guglielmi and Hairer [23] which solves stiff DDEs. Other DDE solvers are DKLAG6 by Corwin, Sarafyan and Thompson [13], DDVERK by Enright and Hayashi [16], DDE_SOLVER by Thompson and Shampine [56] and ARCHI by Paul [49]. The solver that we develop in MATLAB is based on SDIRK methods which are discussed in Hairer and Wanner [26].

1.4 Main results

The main results of this thesis concern aspects of both the theory and numerical analysis of DDEs. Chapter 2 presents results on the properties and some dynamics of a model DDE with N state dependent delays. Equation (1.1.1) gives the $N = 1$ case which is the focus of my work. All parameters in the equation are real numbers with $\varepsilon, a > 0$. A scaling argument shows that there are effectively only two free parameters in the equation because if $c \neq 0$ it is always possible to rescale and set $\varepsilon = a = c = 1$, and if $c = 0$ it is possible to again set $\varepsilon = a = 1$. However all parameters are kept arbitrary in the discussion to allow for limiting cases such as $\varepsilon \to 0$ which yields saw-tooth graphs and other interesting solutions. More complex dynamics can be observed for the $N > 1$ case which is discussed in Section 2.2. The use of a state dependent delay imposes the possibility of the delay becoming an advance and the solution becoming dependent on both past and future states. In Theorems 2.1.4 and 2.2.2 I find conditions on the parameters for which this does not happen for the model equations. Furthermore, in Theorems 2.1.8 and 2.2.4 I derive sufficient conditions for which the deviated arguments become monotonically increasing after a finite time.

In Chapter 3 the parameters ε, a and c are fixed and I look for regions in the (μ, σ) parameter space for which the zero solution of (1.1.1) is Lyapunov stable, or asymptotically stable. The $c = 0$ case is well-known and its analytic stability domain $\Sigma_\star = \overset{\Delta}{\Sigma} \cup \overset{w}{\Sigma} \cup \overset{c}{\Sigma}$ can be derived using pole location techniques. This region is given in Definition 3.1.1 and shown in Figure 3–2. The cone $\overset{\Delta}{\Sigma}$ is sometimes called the delay independent region and $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$ the delay dependent region because the size of the latter depends on the delay term a. By the results of Györi and Hartung [25], the local stability domain of the $c \neq 0$ case is the same as the analytic stability domain of the $c = 0$ case. However I still consider different ways to directly prove stability so that discrete versions of these methods may be adapted to prove the stability of numerical solutions to the problem. In $\overset{\Delta}{\Sigma}$ it can be easily shown that the zero solution is asymptotically stable using a contraction argument. In $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$ I use two different approaches.

In Section 3.2 a Gronwall argument is used to directly prove asymptotic stability in part of $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$. This result is given in Theorem 3.2.6. In Sections 3.4.1-3.4.2 a larger region for which the solution is Lyapunov stable is found using a method based on a Lyapunov-Razumikhin theorem for RFDEs. This second approach is a generalisation of the work of Barnea [6] who considered the $\mu = c = 0$ case. These results are summarised in Theorems 3.4.7, 3.4.13 and 3.4.15. Section 3.5 shows measurements of the different stability regions that were found.

Chapter 4 is on the stability properties of CRK methods applied to DDEs. The equations involved are known to be well-posed for problems that satisfy certain Lipschitz and continuity conditions and when the stepsize is small enough [8]. But this stepsize is very small for stiff equations such as (1.1.1) with small ε, and are therefore not practical. As in ODEs, stiff problems are the motivation for studying the stability of numerical methods. We consider P(0), GP(0), D(0) or GD(0)-stability, generalizations of A-stability to DDE methods.

Much of the existing literature on the stability of DDE methods involves the constant delay DDE (1.1.1) with $c = 0$ where the stepsize is a submultiple of the delay term a. Such a choice of stepsize removes the need for interpolation to determine the values of the spurious stages. However this technique cannot be used for time dependent and state dependent problems. In this work I use direct approaches to find a stepsize-independent stability region of the backward Euler method in the (μ, σ) parameter space of (1.1.1) using both constant delays ($c = 0$) and state dependent delays ($c \neq 0$). In Theorem 4.6.3 I prove that if $(\mu, \sigma) \in \Sigma_\star$ then there exists a BE solution to (1.1.1) that converges to zero for all stepsizes $h > a$. For the case $h \in (0, a]$, discrete versions of both the Gronwall and Razumikhin-like arguments are conducted to derive regions in the (μ, σ) parameter space for which the backward Euler solution to (1.1.1) is stable. Using the Razumikhin-like method, stability is proven in larger parameter regions that depend on the stepsize, but always contain the region found for the DDE. These results are presented in Theorems 4.4.6 and 4.5.15. An extension of these results to general Θ methods is presented at the end of this chapter.

Chapter 5 is a discussion on the implementation of some methods to solve stiff, scalar DDEs with multiple state dependent delays. The solver is based on L-stable singularly diagonally implicit Runge-Kutta (SDIRK) methods with a selection of continuous extensions. SDIRK methods are RK methods where the A matrix is lower triangular and the entries in the diagonal all have the same value. These methods are sometime called semi-implicit methods because the stages are solved for in order, one at a time. When using an s-stage SDIRK scheme to numerically integrate scalar ODE problems, at every time step we need to solve s scalar equations one at a time. This is a more tractable problem than solving an s-dimensional system all at once. The advantage is the same when SDIRK schemes are used to solve DDEs without overlapping. In the overlapping case, the need to determine the continuous extension at the current time interval introduces a need to solve all the stages all at once again, a difficult iteration for stiff problems. In order to retain the properties of SDIRK methods new continuous extensions that are piecewise polynomials are used so as not to destroy the SDIRK structure in the overlapping cases.

Some of the issues involved in implementing CRK methods to solve DDEs are also tackled in Chapter 5. I concentrate on the issue of choosing appropriate continuous extensions for the methods. From Theorem 1.2.2, a discrete method of order p needs a continuous extension of at least order $p-1$ to retain its order for solving DDEs. However higher order continuous extensions usually use polynomials of higher degrees and this can lead to “spiking” in the continuous extension. Spiking happens when the solution is undergoing sudden changes and the continuous extension leaves the convex hull of the adjacent mesh values. Figure 5–3 shows an example of a continuous extension spiking and how it affects the solution at later times. This problem may be avoided by choosing continuous extensions that are convex combinations of the mesh values such as continuous extensions based on linear interpolation and piecewise linear interpolation.

CHAPTER 2
Properties of a DDE with N linearly state dependent delays

The properties of a model delay differential equation with N state dependent delays are discussed in this chapter. We begin by looking at $N = 1$, the simplest case, in Section 2.1 and prove existence, uniqueness and boundedness results as well as some properties of the delay term and deviated argument. In Section 2.2 we derive conditions for which these results can be extended to arbitrary $N \geqslant 1$. Finally, in Section 2.3 we look at bounds on the solutions of the model problem and estimate from these bounds where these solutions might bifurcate.

2.1 The model problem: N=1

Consider the following DDE with one state dependent delay

$$\begin{aligned} \varepsilon \dot{u}(t) &= \mu u(t) + \sigma u(t-a-cu(t)), & t \geqslant 0, \\ u(t) &= \varphi(t), & t \leqslant 0, \end{aligned} \tag{2.1.1}$$

where $\varepsilon, a, c, \mu, \sigma \in \mathbb{R}$, and $\varepsilon, a > 0$ and the history function $\varphi(t)$ is a real-valued continuous function. The delay function is $\tau(t, u(t)) = a + cu(t)$ and the deviated argument is $\alpha(t, u(t)) = t - a - cu(t)$. If $a \neq 0$ we can rescale the problem to set $a = 1$. Let $v(t) = u(at)$, then the equation for $v(t)$ is

$$\bar{\varepsilon}\dot{v}(t) = \mu v(t) + \sigma v(t - 1 - \bar{c}v(t)),$$

where $\bar{\varepsilon} = \frac{\varepsilon}{a}$, $\bar{c} = \frac{c}{a}$. If $c \neq 0$, we can also set $c = 1$. Let $w(t) = \bar{c}v(t)$ then we derive

$$\bar{\varepsilon}\dot{w}(t) = \mu w(t) + \sigma w(t - 1 - w(t)).$$

These scaling arguments have been presented by Mallet-Paret and Nussbaum in [41]. Since it is also possible to divide by ε this shows that for $a \neq 0$ and $c \neq 0$, (2.1.1) actually only has two free parameters μ and σ. However we will keep all five parameters arbitrary in our discussion of this equation to limiting cases such as $\varepsilon \to 0$.

To compare the constant delay problem ($c = 0$) and the state dependent problem ($c \neq 0$) consider a Taylor expansion of the solutions about $u = 0$,

$$u(t - a - cu(t)) = u(t-a) + u'(t-a)(-cu(t)) + \frac{u''(t-a)}{2!}(-cu(t))^2 + \dots.$$

Thus,

$$\varepsilon \dot{u}(t) = \mu u(t) + \sigma \left(u(t-a) + u'(t-a)(-cu(t)) + \frac{u''(t-a)}{2!}(-cu(t))^2 + \dots \right).$$

Ignoring the higher order terms leaves the constant delay problem $\varepsilon \dot{u}(t) = \mu u(t) + \sigma u(t-a)$. This suggests that the local stability region of the $c \neq 0$ case is the same as the stability region of the constant delay problem. Numerics and the results of Györi and Hartung [25] confirm that this is so. This is discussed further in Section 3.1.

For $c = 0$, it is easy to see using the method of steps that (2.1.1) is a well-posed problem.

Theorem 2.1.1. *Consider the constant delay problem* (2.1.1) *with $a > 0$, $c = 0$. Let the history function $\varphi(t)$ be continuous over the interval $[-a, 0]$. Then there exists a unique solution $u(t) \in C([-a, \infty), \mathbb{R})$ to the constant delay problem. Furthermore, if $\varphi(t)$ is smooth then $u(t)$ is C^∞ for all $t \geqslant 0$ except at points in the mesh $a\mathbb{N} = \{0, a, 2a, ...\}$. If $\varphi(t)$ is C^p-continuous then $u(t)$ is C^{p+n}-continuous on the interval $(na, (n+1)a]$. In either case, the solution is C^n at points $t = na$, $n \in \mathbb{N}$.*

Proof. The discontinuity in $\dot{u}$ at $t = 0$ propagates to a discontinuity in $u^{(n+1)}$ at $t = n$, $n \in \mathbb{N}$. For existence and uniqueness, use Bellman's method of steps [10]. □

For $c \neq 0$, before looking at existence and uniqueness we must find the conditions for which the delay term does not become an advance. This requires $t - a - cu(t) \leqslant t$ which is equivalent to the condition

$$\begin{cases} u(t) \geqslant -\frac{a}{c}, & \text{if } c > 0, \\ u(t) \leqslant -\frac{a}{c}, & \text{if } c < 0. \end{cases}$$

Thus, if $c > 0$ and $\varphi(0) > -\frac{a}{c}$, or if $c < 0$ and $\varphi(0) < -\frac{a}{c}$ then the delay cannot become an advance. In these cases the local existence and uniqueness result of Driver [14] given in Theorem 1.1.3 states that there is a solution to (2.1.1) for some nonzero time interval. This result applies for any value of μ, σ and ε. For the region $\mu + \sigma < 0$, the following lemma shows that the delay term cannot become an advance.

Lemma 2.1.2. *Let $\varepsilon, a > 0$, $c \neq 0$ and $\mu + \sigma < 0$. If $c > 0$ and $\varphi(0) > -\frac{a}{c}$ then any solution to* (2.1.1) *must satisfy $u(t) > -\frac{a}{c}$ for all $t \geqslant 0$ such that the solution exists. If $c < 0$ and $\varphi(0) < -\frac{a}{c}$ then any solution to* (2.1.1) *must satisfy $u(t) < -\frac{a}{c}$ for all $t \geqslant 0$ such that the solution exists.*

Proof. Let $c > 0$ and $u(0) = \varphi(0) > -\frac{a}{c}$. Suppose there is a first time $t_1 > 0$ such that $u(t_1) = -\frac{a}{c}$. Then since $u(t) > -\frac{a}{c}$ for $t < t_1$ then $\dot{u}(t_1) \leqslant 0$. But $u(t_1 - a - cu(t_1)) = u(t_1) = -\frac{a}{c}$ so

$$\varepsilon \dot{u}(t_1) = \mu\left(-\frac{a}{c}\right) + \sigma\left(-\frac{a}{c}\right) = -\frac{a}{c}(\mu + \sigma) > 0$$

This is a contradiction and it proves that $u(t) > -\frac{a}{c}$ for all $t \geqslant 0$ such that the solution exists. The case for $c < 0$ can be proven similarly. □

Properties of the solution when $\mu < 0$ and $\sigma < 0$.

Lemma 2.1.2 gives one bound to the solution when $\mu+\sigma < 0$. The entire stability region of (2.1.1) actually satisfies $\mu+\sigma < 0$ as will be discussed in Chapter 3. For $c > 0$, $\mu < 0$ and $\sigma < 0$ we can do better and get a lower and upper bound on the solution to show global existence of the solution. The global existence of the solution to (2.1.1) when $\mu < 0$ and $\sigma < 0$ has recently been presented by Mallet-Paret and Nussbaum in [41]. However, these results are still derived and proven here because they are later extended to the case of multiple state dependent delays in Section 2.2. Define

$$L_0 = -\frac{a}{c}, \qquad M_0 = \frac{a\sigma}{c\mu}, \qquad \tau_0 = a + cM_0. \tag{2.1.2}$$

Lemma 2.1.3. *Let $\varepsilon, a, c > 0$, $\mu < 0$ and $\sigma < 0$. Let the history function $\varphi(t)$ be continuous with $\varphi(t) \in (L_0, M_0)$ for $t \in [-\tau_0, 0]$. If $u(t)$ is a solution to* (2.1.1) *then $u(t) \in (L_0, M_0)$ for all t for which the solution exists.*

Proof. Since $c > 0$ and $\mu + \sigma < 0$ then by Lemma 2.1.2 $u(t) > L_0$ for all t. Now suppose that there exists $t_1 > 0$ such that $u(t_1) = M_0$ and $u(t) < M_0$ for $t < t_1$. This implies that $\dot{u}(t_1) \geqslant 0$. However,

$$\varepsilon \dot{u}(t_1) = \mu M_0 + \sigma u(t_1 - a - cu(t_1)) < \mu M_0 + \sigma L_0 = 0$$

which is a contradiction. Thus $u(t) \in (L_0, M_0)$ for all t. □

Theorem 2.1.4. *Let $\varepsilon, a, c > 0$, $\mu < 0$ and $\sigma < 0$. Let the history function $\varphi(t) : [-\tau_0, 0] \to (L_0, M_0)$ be Lipschitz continuous with $\varphi(t) \in (L_0, M_0)$ for $t \in [-\tau_0, 0]$. Then there exists a unique solution $u \in C^1([0,\infty), (L_0, M_0))$ to* (2.1.1)*. In particular, $u(t) \in (L_0, M_0)$ for all $t \geqslant 0$.*

Proof. Here we employ the local and extended existence theory for a general delay differential equations with multiple state dependent delays by Driver [14] which are given in Chapter 1 as Theorems 1.1.3 and 1.1.4. Define $D = \mathbb{R} \times (-L, M)^2$ and $D^* = \mathbb{R} \times (-L, M)$ where $L > L_0$ and $M > M_0$. Consider the DDE

$$\begin{aligned} \dot{u}(t) &= f(t, u(t), u(\alpha(t, u(t)))), \quad & t \geqslant 0 \\ u(t) &= \varphi(t), & t \leqslant 0 \end{aligned} \tag{2.1.3}$$

where $f(t, u, v) = \frac{\mu}{\varepsilon}u + \frac{\sigma}{\varepsilon}v$, $\alpha(t, u(t)) = \min\{t, t - a - cu(t)\}$ and $\bar{\tau}_0 = a + cM$. It is easy to see that f is continuous with respect to all of its variables and Lipschitz continuous with respect to u and v in the domain D. Also, $\alpha(t, u)$ is Lipschitz with respect to its u argument and satisfies $-\bar{\tau}_0 \leqslant \alpha(t, u) \leqslant t$ in $D^* \cap \{t \geqslant 0\}$. Since the history function $\varphi(t)$ is given to be Lipschitz continuous in $[-\tau_0, 0]$ then by Theorem 1.1.3 we have local existence and uniqueness of the solution to (2.1.3).

If $u(t) \in C([-\tau, 0], (L_0, M_0))$ then Lemma 2.1.3 states that $u(t)$ remains inside the interval (L_0, M_0). In this case, the system written above is equivalent to (2.1.1) and the history function needs only to be defined and Lipschitz in the interval $[-\tau_0, 0]$. Hence we have local existence and uniqueness of the solution to (2.1.1).

To prove that the solution exists for $t \in [0, \infty)$ we apply Theorem 1.1.4 to (2.1.3). The theorem states that given the same conditions as for existence and uniqueness, the solution can be extended to $[0, \beta)$ where $0 < \beta \leqslant \infty$. If β is finite and cannot be increased then as $t \to \beta$ either the solution goes off to ∞ or $(t, u(t), u(\alpha(t, u(t))))$ comes arbitrarily close to the boundary of D. But Lemma 2.1.3 states that $u(t)$ must stay inside (L_0, M_0) so neither case is possible. Thus we must have $\beta = \infty$ and this completes the proof of the continuation of the unique solution to $[0, \infty)$. □

Lemma 2.1.5. *Let $\varepsilon, a, c > 0$, $\mu < 0$ and $\sigma < 0$ and let $u(t)$ be a solution to (2.1.1) with $u(t) \in (L_0, M_0)$ for $t \in [-\tau_0, \infty)$. Then $u(t)$ must be behave in one of the following manners:*

(A) There exists $\bar{T}$ such that $u(t) \downarrow 0$ for $t > \bar{T}$

(B) There exists $\bar{T}$ such that $u(t) \uparrow 0$ for $t > \bar{T}$

(C) For every $T > 0$ there exists $T_1, T_2 > T$ such that the solution attains a positive local maximum at T_1 and a negative local minimum at T_2.

Proof. From Lemma 2.1.3, for all $t > 0$, $u(t) \in (L_0, M_0)$ and

$$t - \tau_0 \leqslant \alpha(t, u(t)) \leqslant t \tag{2.1.4}$$

Suppose there exists a time T such that $u(t) \geqslant 0$ for all $t > T$. Then for $t > \bar{T} = T + \tau_0$, we have $\dot{u}(t) \leqslant 0$ because of (2.1.4). This means there is some $\bar{u} \in [0, M_0)$ such that $u(t) \downarrow \bar{u}$ when $t > \bar{T}$. Again because of (2.1.4) we must also have $u(t - a - cu(t)) \to \bar{u}$. This means $\dot{u}(t) \to \frac{\mu+\sigma}{\varepsilon}\bar{u}$ which is only possible if $\bar{u} = 0$. Thus in this case, we have the behavior in (A).

If there exists a time T such that $u(t) \leqslant 0$ for all $t > T$ then using a similar argument we must have the behavior in (B). If there is no time T such either $u(t) \geqslant 0$ or $u(t) \leqslant 0$ holds for $t \geqslant T$ then the solution must be always be changing signs which yields the behavior (C). □

In Theorem 2.1.6, we derive a uniform bound $\bar{\tau}$ such that delay term satisfies $\tau(t, u(t)) \geqslant \bar{\tau} > 0$. Using this bound the global existence and uniqueness of the solutions to (2.1.1) could have been proven using the method of steps. This bound is still useful because in a numerical integration of this equation, taking the stepsize to be smaller than $\bar{\tau}$ allows us to avoid overlapping.

Theorem 2.1.6. *Let $\varepsilon, a, c > 0$, $\mu < 0$ and $\sigma < 0$ and let $u(t)$ be a solution to* (2.1.1) *with $u(t) \in (L_0, M_0)$ for $t \in [-\tau_0, \infty)$. Define*

$$\bar{\tau} = \frac{a\left(1 + \frac{\mu}{\sigma}\right)}{1 + \frac{\mu}{\sigma} - \frac{a\mu^2}{\sigma\varepsilon} - \frac{a\sigma^2}{\mu\varepsilon}}, \qquad \bar{L} = -\frac{a}{c} + \frac{\bar{\tau}}{c},$$

and define $h(v) \in C\left(\left(-\frac{a}{c}, 0\right), \mathbb{R}\right)$ as in (2.1.8)*. Then $h(v)$ has a unique zero $v^* \in \left(-\frac{a}{c}, 0\right)$ and $v^* \geqslant \bar{L} > -\frac{a}{c}$. Define also,*

$$\begin{aligned} \tilde{\tau} &= \min\{a + c\varphi(0), \bar{\tau}\}, & \tilde{L} &= \min\{\varphi(0), \bar{L}\}, \\ \tau^* &= \min\{a + c\varphi(0), a + cv^*\}, & L^* &= \min\{\varphi(0), v^*\}. \end{aligned}$$

Then $\tau(t) \geqslant \tau^* \geqslant \tilde{\tau} > 0$ and $u(t) \geqslant L^* \geqslant \tilde{L} > -\frac{a}{c}$ for all $t \geqslant 0$.

Remark 2.1.7. This theorem shows that the solution can be bounded away from $-\frac{a}{c}$ and this bounds the delay term away from zero. Any local minima of the solution cannot cross the root v^* of $h(v)$. The bound $\bar{L}$ is less strict but it has an explicit expression that depends on the model parameters.

Proof. Suppose the solution $u\left(t\right)$ never attains a negative minimum. Then it is easy to show that $\tau\left(t\right) \geqslant \min\left\{a + c\varphi(0), a\right\} \geqslant \tilde{\tau} > 0$ and $u\left(t\right) \geqslant \min\left\{\varphi(0), 0\right\} \geqslant \tilde{L} > -\frac{a}{c}$ for all $t \geqslant 0$. So suppose instead that the solution attains a negative minimum at $u\left(t\right) = v_0 < 0$. Then $\dot{u}\left(t\right) = 0$ and

$$u\left(t - a - cv_0\right) = -\frac{\mu}{\sigma}v_0. \tag{2.1.5}$$

Since the requirements of Lemma 2.1.3 are satisfied then $u\left(t\right) \in \left(L_0, M_0\right)$ for all $t \geqslant -\tau_0$. This bound is used in the following Gronwall argument.

$$\begin{aligned}
\varepsilon\dot{u}\left(t\right) - \mu u\left(t\right) &\geqslant \sigma M_0 = \frac{a\sigma^2}{c\mu} \\
\frac{d}{dt}\left(u\left(t\right)e^{-\frac{\mu}{\varepsilon}t}\right) &\geqslant \frac{a\sigma^2}{c\mu\varepsilon}e^{-\frac{\mu}{\varepsilon}t} \\
u\left(s\right)e^{-\frac{\mu}{\varepsilon}s}\Big|_{s=t-a-cu(t)}^{t} &\geqslant -\frac{a\sigma^2}{c\mu^2}e^{-\frac{\mu}{\varepsilon}t}\Big|_{s=t-a-cu(t)}^{t} \\
u\left(t\right)e^{-\frac{\mu}{\varepsilon}t} - u\left(t - a - cu\left(t\right)\right)e^{-\frac{\mu}{\varepsilon}(t-a-cu(t))} &\geqslant -\frac{a\sigma^2}{c\mu^2}\left(e^{-\frac{\mu}{\varepsilon}t} - e^{-\frac{\mu}{\varepsilon}(t-a-cu(t))}\right)
\end{aligned} \tag{2.1.6}$$

Set $u(t) = v_0$ and substitute for $u\left(t - a - cv_0\right)$ from (2.1.5) in (2.1.6),

$$v_0e^{-\frac{\mu}{\varepsilon}t} + \frac{\mu}{\sigma}v_0e^{-\frac{\mu}{\varepsilon}(t-a-cv_0)} \geqslant -\frac{a\sigma^2}{c\mu^2}\left(e^{-\frac{\mu}{\varepsilon}t} - e^{-\frac{\mu}{\varepsilon}(t-a-cv_0)}\right),$$

$$\left(1 + \frac{\mu}{\sigma}e^{\frac{\mu}{\varepsilon}(a+cv_0)}\right)v_0 + \frac{a\sigma^2}{c\mu^2}\left(1 - e^{-\frac{\mu}{\varepsilon}(a+cv_0)}\right) \geqslant 0. \tag{2.1.7}$$

Define the following function

$$h\left(v\right) = \left(1 + \frac{\mu}{\sigma}e^{\frac{\mu}{\varepsilon}(a+cv)}\right)v + \frac{a\sigma^2}{c\mu^2}\left(1 - e^{-\frac{\mu}{\varepsilon}(a+cv)}\right). \tag{2.1.8}$$

Then we must have $h\left(v_0\right) \geqslant 0$. Consider the behavior of this function.

$$h'\left(v\right) = 1 + \left(\frac{\mu}{\sigma} + \frac{\mu^2c}{\sigma\varepsilon}v - \frac{a\sigma^2}{\mu\varepsilon}\right)e^{\frac{\mu}{\varepsilon}(a+cv)},$$

$$h''\left(v\right) = -\left(-\frac{\mu^2c}{\sigma\varepsilon} + \frac{a\sigma^2c}{\varepsilon^2} - \frac{\mu^2c}{\sigma\varepsilon}\left(1 + \frac{\mu c}{\varepsilon}v\right)\right)e^{\frac{\mu}{\varepsilon}(a+cv)}.$$

Easily, $h'\left(v\right) > 0$ and $h''\left(v\right) < 0$ when $v < 0$. Also,

$$h\left(-\frac{a}{c}\right) = \left(1 + \frac{\mu}{\sigma}e^0\right)\left(-\frac{a}{c}\right) + \frac{a\sigma^2}{c\mu^2}\left(1 - e^0\right) = -\frac{a}{c}\left(1 + \frac{\mu}{\sigma}\right) < 0.$$

Thus $h\left(-\frac{a}{c}\right) < 0$ and is strictly increasing and concave down for $v \in \left(-\frac{a}{c}, 0\right)$. The function must have a root $v^* < 0$ because $h(v) > 0$ for all $v \geqslant 0$. Since at the negative minimum v_0 has to satisfy $h(v_0) \geqslant 0$ then this lower bound on the solution should occur at a positive distance from $-\frac{a}{c}$. We can find an explicit lower bound on the distance of the root v^* to $-\frac{a}{c}$. Since $h(v)$ is strictly increasing and concave down $\forall v \in \left(-\frac{a}{c}, 0\right)$ then $v^* \geqslant v^{**}$ where v^{**}, the root of the tangent line to $h(v)$ at the point $\left(-\frac{a}{c}, -\frac{a}{c}\left(1+\frac{\mu}{\sigma}\right)\right)$. This is illustrated in Figure 2–1. If we let $v^{**} = \bar{L} = -\frac{a}{c} + \frac{\bar{\tau}}{c}$ (where $\bar{L} < 0$ and $\bar{\tau} > 0$) then using the tangent line,

$$\frac{0 - \left(-\frac{a}{c}\left(1+\frac{\mu}{\sigma}\right)\right)}{\frac{\bar{\tau}}{c}} = h'\left(-\frac{a}{c}\right) = 1 + \frac{\mu}{\sigma} - \frac{a\mu^2}{\sigma\varepsilon c} - \frac{a\sigma^2}{\mu\varepsilon}$$

$$\Rightarrow \bar{\tau} = \frac{a\left(1+\frac{\mu}{\sigma}\right)}{1 + \frac{\mu}{\sigma} - \frac{a\mu^2}{\sigma\varepsilon} - \frac{a\sigma^2}{\mu\varepsilon}}$$

Thus, if $\varphi(0) \geqslant \bar{L}$ then the solution never crosses below $\bar{L}$. If $\varphi(0) < \bar{L}$ then the solution has to start off increasing ($\dot{u}(0)^+ > 0$) otherwise it would have to reach a minimum (because the solution is bounded) below $\bar{L}$ which is a contradiction. If the solution reaches a minimum for any $t > 0$ then this minimum cannot be lower than $\bar{L}$. Thus, $u(t) \geqslant \tilde{L} = \min\left\{\varphi(0), \bar{L}\right\} > -\frac{a}{c}$ for $t \geqslant 0$. The result $\tau(t) \geqslant \tilde{\tau} > 0$ and $u(t) \geqslant \tilde{L} > -\frac{a}{c}$ follows easily. □

We now consider the behaviour of the delay term and deviated argument when $\mu < 0$ and $\sigma < 0$. In Theorem 2.1.8 we show that $\alpha(t, u(t))$ is eventually monotonically increasing. This result and its proof was also recently presented by Mallet-Paret and Nussbaum in [41]. We still state and prove this result here since we extend this in Theorem 2.2.4 to a special case of the N-delay equation. Let the following notation denote the derivative of $\alpha(t, u(t))$ with respect to time

$$\dot{\alpha}(t, u(t)) = \frac{d}{dt}\alpha(t, u(t))$$

Theorem 2.1.8. *Let $\varepsilon, a, c > 0$, $\mu < 0$ and $\sigma < 0$. Let $u(t)$ be a solution to (2.1.1) with $u(t) \in (L_0, M_0)$ for $t \in [-\tau_0, \infty)$. Then there exists T, $0 \leqslant T \leqslant \tau_0$ such that $u(t) \in C^2([T,\infty), \mathbb{R})$, $\alpha(t, u(t)) \in C^2([T,\infty), [0,+\infty))$ and $\dot{\alpha}(t, u(t)) > 0$ for all $t > T$.*

Proof. Generally, $\varepsilon\dot{\varphi}(0) \neq \mu\varphi(0) + \sigma\varphi(-a - c\varphi(0))$ for arbitrary φ so there is usually a discontinuity in $\dot{u}(t)$ at $t = 0$. We also have

$$\varepsilon\ddot{u}(t) = \mu\dot{u}(t) + \sigma\dot{u}(\alpha(t, u(t)))(1 - c\dot{u}(t))$$

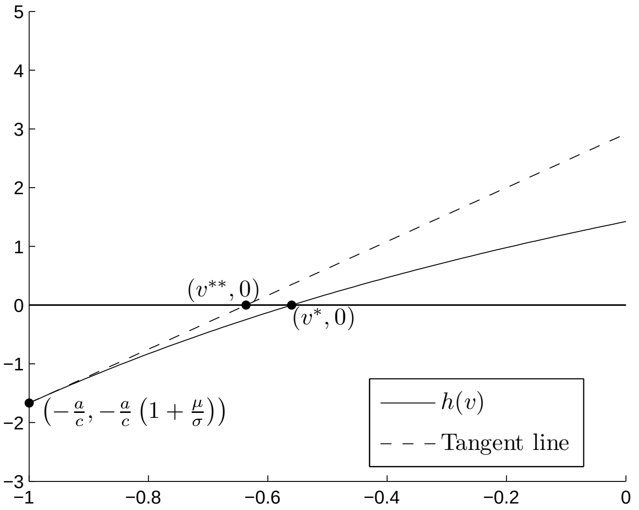


Figure 2–1: Sample plot of $h\left(v\right)$ for $\varepsilon=a=c=1$, $\mu=-1$ and $\sigma=-1.5$. Properties of the function h dictate that the root v^{*} of h would provide a lower bound for the minimum v_0. We can also find an explicit expression for $v^{**}\in\left(-\frac{a}{c},v^{*}\right]$, in terms of the equation parameters.

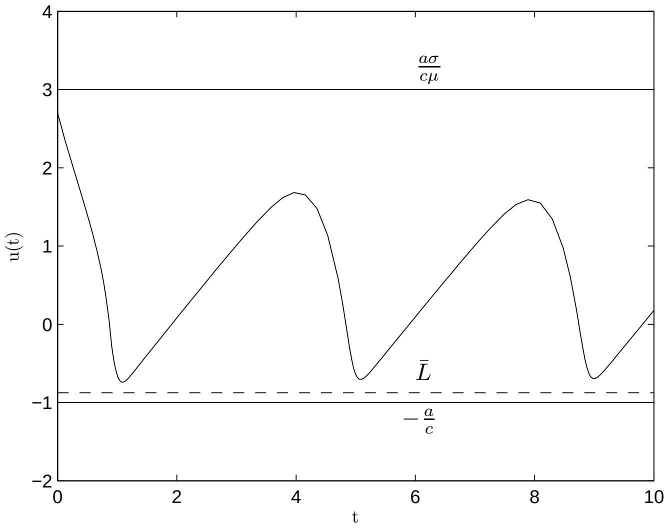


Figure 2–2: Sample plot showing the root $\bar{L}$ of $h\left(v\right)$ located at a positive distance above $\frac{a}{c}$. The solution is uniformly bounded away from $-\frac{a}{c}$ and so the delay term is bounded away from zero.

So there also exists a discontinuity in $\ddot{u}\,(t)$ when $\alpha\,(t, u(t)) = 0$. Define

$$T = \sup_t \{\alpha\,(t, u(t)) = 0\}$$

Then $T \in [0, \tau_0]$ by Lemma 2.1.3. When restricted to $t \geqslant T$ we have $u\,(t) \in C^2\,([T, \infty)\,, \mathbb{R})$.

$$\dot{\alpha}\,(t, u(t)) = 1 - c\dot{u}\,(t) = 1 - \frac{c\mu}{\varepsilon} u\,(t) - \frac{c\sigma}{\varepsilon} u\,(\alpha\,(t, u(t))) \tag{2.1.9}$$

$$\ddot{\alpha}\,(t, u(t)) = -c\ddot{u}\,(t) = -\frac{c\mu}{\varepsilon}\dot{u}\,(t) - \frac{c\sigma}{\varepsilon}\dot{u}\,(\alpha\,(t, u(t)))\,\dot{\alpha}\,(t, u(t)) \tag{2.1.10}$$

By definition of T we have $\alpha\,(t, u(t)) \geqslant 0$ for $t > T$. Thus when restricted to $t \geqslant T$ we have $\alpha\,(t, u(t)) \in C^2\,([T, \infty)\,, [0, +\infty))$.

As a result of the definition of T then either (i) $\dot{\alpha}\,(T, u(T)) > 0$ or (ii) $\dot{\alpha}\,(T, u(T)) = 0$ holds. In the case of (i) then by continuity there exists $\delta > 0$ such that $\dot{\alpha}\,(t, u(t)) > 0$ for all $t \in [T, T + \delta)$. If (ii) holds then $\ddot{\alpha}\,(T, u(T)) = -\frac{c\mu}{\varepsilon}\dot{u}\,(T)$. But since $0 = \dot{\alpha}\,(T, u(T)) = 1 - c\dot{u}\,(T)$ then $\dot{u}\,(T) = \frac{1}{c}$. Thus $\ddot{\alpha}(T, u(T)) = -\frac{\mu}{\varepsilon} > 0$ and in this case there exists a $\delta > 0$ such that $\dot{\alpha}\,(t, u(t)) > 0$ for all $t \in (T, T + \delta)$. Thus in either case, we have $\delta > 0$ such that $\dot{\alpha}\,(t, u(t)) > 0$ for all $t \in (T, T + \delta)$.

Now the problem is to show that it is possible to let $\delta \to \infty$. Suppose not, then there exists $T_1 = T + \delta$ such that $\dot{\alpha}\,(T_1, u(T_1)) = 0$ and $\dot{\alpha}\,(t, u(t)) > 0$ for all $t \in (T, T_1)$. That means $\ddot{\alpha}\,(T_1, u(T_1)) \leqslant 0$. However from (2.1.10), $\ddot{\alpha}\,(T_1, u(T_1)) = -\frac{c\mu}{\varepsilon}\dot{u}\,(T_1) = -\frac{\mu}{\varepsilon} > 0$. This is a contradiction and thus there is no such $T_1 > T$. □

2.2 The model N-Delay problem

Let $N \in \mathbb{Z}$, $N \geqslant 1$. Consider the model DDE with N state dependent delays.

$$\begin{aligned} &\varepsilon\dot{u}\,(t) = -\gamma u\,(t) - \sum_{i=1}^{N} \kappa_i u\,(t - a_i - c_i u\,(t)), && t \geqslant 0, \\ &u\,(t) = \varphi\,(t)\,, && t \leqslant 0, \end{aligned} \tag{2.2.1}$$

where $\varepsilon, a_i, c_i, \gamma, \kappa_i \geqslant 0$ for $i = 1, ..., N$. If we set $N = 1$ we get back our 1-delay equation in (2.1.1) with $a = a_1$, $c = c_1$, $\mu = -\gamma$ and $\sigma = \kappa_1$. We would like to extend the results we had in the previous section with $\mu < 0$ and $\sigma < 0$. In later chapters it will be useful to keep the notation we had in Section 2.1 for the 1-delay equation. However, for the N-delay equation we change the notation as this will be more convenient for the results that we will consider for these problems.

Without loss of generality assume

$$0 > -\frac{a_1}{c_1} \geqslant -\frac{a_2}{c_2} \geqslant ... \geqslant \frac{a_N}{c_N}. \tag{2.2.2}$$

For the N-delay equation, define

$$L_0 = -\frac{a_1}{c_1}, \qquad M_0 = \frac{a_1}{c_1\gamma}\sum_{i=1}^{N}\kappa_i, \qquad \tau_0 = \max_{j=1,\dots,N}\left\{a_j + c_j M_0\right\}. \tag{2.2.3}$$

The boundedness, existence and uniqueness results for the N-delay equation parallel Lemma 2.1.3 and Theorem 2.1.4 for the 1-delay problem provided condition (2.2.4) holds. Note that this condition just becomes $\gamma > 0$ if we set $N = 1$.

Lemma 2.2.1. *Let $\varepsilon, a_i, c_i > 0$, $\gamma_i > 0$ and $\kappa_i > 0$ for $i = 1, ..., N$. Let the history function $\varphi(t)$ be continuous with $\varphi(t) \in (L_0, M_0)$ for $t \in [-\tau_0, 0]$. Let*

$$\gamma > \sum_{i=2}^{N}\kappa_i, \tag{2.2.4}$$

and (2.2.2) *hold. Then if $u(t)$ is a solution to* (2.2.1) *then $u(t) \in (L_0, M_0)$ for all $t \geqslant 0$.*

Proof. Suppose there exists $t_1 > 0$ such that $u(t) \in (L_0, M_0)$ for all $t < t_1$ and $u(t_1) = L_0 = -\frac{a}{c}$. Then $u(t) > -\frac{a}{c}$ for $t < t_1$. This implies that $\dot{u}(t_1) \leqslant 0$. However, $\alpha_1(t_1, u(t_1)) = t_1 - a_1 - c_1 L_0 = t_1$. Also, $u(\alpha_i(t_1, u(t_1))) < M_0$ for $i = 2, ..., N$ so

$$\begin{aligned}\varepsilon\dot{u}(t_1) &= -\gamma u(t_1) - \sum_{i=1}^{N}\kappa_i u(\alpha_i(t_1, u(t_1))),\\ &\geqslant (\gamma + \kappa_1)\frac{a_1}{c_1} - M_0\sum_{i=2}^{N}\kappa_i = \frac{a_1}{c_1\gamma}\left[(\kappa_1 + \gamma)\gamma - \Big(\kappa_1 + \sum_{i=2}^{N}\kappa_i\Big)\sum_{i=2}^{N}\kappa_i\right].\end{aligned}$$

Using (2.2.4) we see that the last term must be positive. Since $\varepsilon > 0$ we get $\dot{u}(t_1) > 0$ which is a contradiction.

If instead $u(t_1) = M_0$ and $u(t) < M_0$ for $t < t_0$ then $\dot{u}(t_1) \geqslant 0$. But $u(t) > -\frac{a_1}{c_1}$ for $t < t_0$ implies that

$$\varepsilon\dot{u}(t_1) < -\gamma M_0 + \frac{a_1}{c_1}\sum_{i=1}^{N}\kappa_i = 0.$$

This is a contradiction. □

Theorem 2.2.2. *Let $\varepsilon, a_i, c_i > 0$, $\gamma_i > 0$ and $\kappa_i > 0$ for $i = 1, ..., N$. Let the history function $\varphi(t)$ be Lipschitz continuous with $\varphi(t) \in (L_0, M_0)$ for $t \in [-\tau_0, 0]$. Let (2.2.2) and (2.2.4) be satisfied. Then there exists a unique solution $u \in C^1([0,\infty), (L_0, M_0))$ to (2.2.1). In particular, $u(t) \in (L_0, M_0)$ for all $t \geqslant 0$.*

Proof. This proof requires Theorems 1.1.3 and 1.1.4, the existence theory for a general DDE with multiple state dependent delays by Driver [14]. Define $D = \mathbb{R} \times (-L, M)^{N+1}$ and $D^* = \mathbb{R} \times (L, M)$ where $L < L_0$ and $M > M_0$. Consider the system

$$\begin{aligned} &\dot{u}(t) = f(t, u(t), u(\alpha_1(t, u(t))), ..., u(\alpha_N(t, u(t)))), && t \geqslant 0 \\ &u(t) = \varphi(t), && t \leqslant 0 \end{aligned} \tag{2.2.5}$$

$$f(t, u, v_1, ..., v_N) = -\frac{\gamma}{\varepsilon} u - \sum_{i=1}^{N} \frac{\kappa_i}{\varepsilon} v_i,$$

$$\alpha_i(t, u(t)) = \min\{t, t - a_i - cu(t)\}, \qquad \tau_0 = \max_{j=1,...,N} \{a_j + c_j M\}$$

As in the proof of Theorem 2.1.4, we use Theorem 1.1.3 to prove the local existence and uniqueness of (2.2.5) and then use the boundedness result in Lemma 2.2.1 to show that in this case, (2.2.5) is equivalent to (2.2.1). Then using Theorem 1.1.4 the existence, uniqueness and boundedness result is extended to $[0, \infty)$. □

Now that we have global existence of the solutions we can look at the smoothness of these solutions. Recall the discussion in Section 2.1 of discontinuity points. The solution $u(t)$ is continuously differentiable for $t > 0$, but in general $\lim_{t \to 0^-} \dot{u}(t) \neq \lim_{t \to 0^+} \dot{u}(t)$. In this case $t = 0$ is a discontinuity point due to a discontinuity in the first derivative. This discontinuity at $t = 0$ propagates to discontinuities in higher derivatives of $u(t)$ at later times. Due to the state dependency of the delay, it is not possible to solve for the location of these points without the actual solution. However the bounds from Lemma 2.2.1 yields an upper bound to the location of the discontinuity points. Since the solution must remain bounded inside (L_0, M_0) then the delay term must satisfy $t - \alpha_i(t, u(t)) = a_i + c_i u(t) \in (a_i + c_i L_0, a_i + c_i M_0)$. Right away this implies that $\alpha_i(t, u(t)) \to +\infty$ as $t \to \infty$. These bounds also dictate that if an n-level discontinuity point $\xi_{n,j}$ gives rise to the $(n+1)$-level point $\xi_{n+1,k}$ then we must have $\xi_{n+1,k} \leqslant \xi_{n,j} + \tau$. Starting from the 0-level point this implies that the n-level primary discontinuities $\xi_{n,k} \leqslant n\tau$ for all k. Thus $u(t) \in C^{n+1}$ for all $t \geqslant n\tau$.

Next consider a special case of the N-delay equation. Let $c = c_1 = c_2 = ... = c_N$. Since we have (2.2.2) then $0 < a_1 \leqslant a_2 \leqslant ... \leqslant a_N$.

Lemma 2.2.3. *Let $\varepsilon, a_i, c_i > 0$, $\gamma_i > 0$ and $\kappa_i > 0$ for $i = 1, ..., N$. Let the history function $\varphi(t)$ be continuous with $\varphi(t) \in (L_0, M_0)$ for $t \in [-\tau_0, 0]$. Let $c = c_1 = ... = c_N$, (2.2.2) and (2.2.4) be satisfied. If $u(t)$ is a solution to (2.2.1) then the deviated arguments $\alpha_i(t, u(t))$ must satisfy the following for all $t > 0$.*

$$t - a_i - \frac{a_1}{\gamma} \sum_{j=1}^{N} \kappa_j \leqslant \alpha_i(t, u(t)) = t - a_i - cu(t) \leqslant t$$

$$\alpha_1(t, u(t)) - \alpha_i(t, u(t)) = a_i - a_1$$

Also, where the following derivatives are defined, we have

$$\alpha_1^{(j)}(t, u(t)) = \alpha_2^{(j)}(t, u(t)) = ... = \alpha_N^{(j)}(t, u(t)), \quad \forall j = 1, 2, ...$$

Proof. The first result easily follows from the boundedness of the solutions. The second result comes from $\alpha_1(t, u(t)) - \alpha_i(t, u(t)) = (t - a_1 - cu(t)) - (t - a_i - cu(t)) = a_i - a_1$. As for the last result,

$$\dot{\alpha}_i(t, u(t)) = 1 - c\dot{u}(t),$$

$$\alpha_i^{(j)}(t, u(t)) = -cu^{(j)}(t).$$

These expressions are all equal for $i = 1, ..., N$. □

So the arguments α_i in this case are separated by constant distances with the α_N term being the furthest back. From the ordering of the a_i's, the upper bound on all the delay terms is $\tau_0 = a_N + cM_0$. Theorem 2.2.4 shows that in this special case the α_i must eventually become monotonically increasing. The proof of this is based on Theorem 2.1.8 for $N = 1$ which is based on a proof by Mallet-Paret and Nussbaum [41]. As in Section 2.1 we write the derivative of the deviated arguments as

$$\dot{\alpha}_i(t, u(t)) = \frac{d}{dt}\alpha_i(t, u(t)).$$

Theorem 2.2.4. *Let $\varepsilon, a_i, c_i > 0$, $\gamma_i > 0$ and $\kappa_i > 0$ for $i = 1, ..., N$. Let the history function $\varphi(t)$ be continuous with $\varphi(t) \in (L_0, M_0)$ for $t \in [-\tau_0, 0]$. Let $c = c_1 = ... = c_N$, (2.2.2) and (2.2.4) be satisfied. If $u(t)$ is a solution to (2.2.1) then there exists T, $0 \leqslant T \leqslant \tau$*

such that $u(t) \in C^2([T,\infty),\mathbb{R})$ and for $i = 1,...,N$, $\alpha_i(t) \in C^2([T,\infty),[T-\tau_0,+\infty))$ and $\dot{\alpha}_i(t,u(t)) > 0$ for all $t > T$.

Proof. The point $t = 0$ is a breaking point as $\dot{u}(t)$ is generally discontinuous at this point. There is also a discontinuity in $\ddot{u}(t)$ when any of the deviated arguments $\alpha_i(t,u(t)) = 0$. Define

$$T = \sup_t \{\alpha_i(t) = 0\}$$

Then $T \in [0,\tau_0]$ and $\alpha_i(t,u(t)) > 0$ for $t > T$. Thus, we must have $u(t) \in C^2([T,\infty),\mathbb{R})$. Then also, $\alpha_i \in C^2([T,\infty),[T-\tau_0,+\infty))$.

Since $\dot{\alpha}_i(t,u(t)) = 1 - c\dot{u}(t)$ for $i = 1,...,N$ then it will be enough to show that $\dot{\alpha}_N(t) > 0$ for $t > T$.

$$\dot{\alpha}_N(t,u(t)) = 1 - c\dot{u}(t) = 1 + \frac{c\gamma}{\varepsilon}u(t) + \frac{c}{\varepsilon}\sum_{i=1}^{n}\kappa_i u(\alpha_i(t,u(t))) \tag{2.2.6}$$

$$\ddot{\alpha}_N(t,u(t)) = -c\ddot{u}(t) = \frac{c\gamma}{\varepsilon}\dot{u}(t) + \frac{c}{\varepsilon}\sum_{i=1}^{N}\kappa_i\dot{u}(\alpha_i(t,u(t)))\,\dot{\alpha}_N(t,u(t)) \tag{2.2.7}$$

where the last line is because the derivatives of all deviated arguments are equal.

As a result of the definition of T then either (i) $\dot{\alpha}_N(T,u(T)) > 0$ or (ii) $\dot{\alpha}_N(T,u(T)) = 0$ holds. If we have case (i) then by continuity there exists $\delta > 0$ such that $\dot{\alpha}_N(t,u(t)) > 0$ for all $t \in [T,T+\delta)$. If (ii) holds then $\ddot{\alpha}_N(T,u(T)) = \frac{c\gamma}{\varepsilon}\dot{u}(T)$. But in this case $\dot{u}(T) = \frac{1}{c}$ so $\ddot{\alpha}_N(T,u(T)) = \frac{\gamma}{\varepsilon} > 0$. Then there exists $\delta > 0$ such that $\dot{\alpha}_N(t,u(t)) > 0$ for all $t \in (T,T+\delta)$. Thus in either case, we have $\delta > 0$ such that $\dot{\alpha}_N(t,u(t)) > 0$ for all $t \in (T,T+\delta)$.

Now the problem is to show that it is possible to let $\delta \to \infty$. Suppose not. Then there exists a finite $T_1 = T + \delta$ such that $\dot{\alpha}_N(T_1,u(T_1)) = 0$ and $\dot{\alpha}_N(t,u(t)) > 0$ for all $t \in (T,T_1)$. At such a point $\ddot{\alpha}_N(T_1,u(T_1)) \leqslant 0$. However from (2.2.7), $\ddot{\alpha}_N(T_1,u(T_1)) = \frac{\gamma c}{\varepsilon}\dot{u}(T_1) = \frac{\gamma}{\varepsilon} > 0$. This is a contradiction and thus there is no such T_1. □

2.3 Bounds on the solutions of the model problems

In this section we look for bounds on the solutions to the 1-delay equation (2.1.1) and the general N-delay equation (2.2.1). It will be convenient to continue using the γ and κ notation from the previous section. So in this section when considering the 1-delay equation we set $\gamma = -\mu < 0$ and $\kappa = -\sigma < 0$ and write

$$\varepsilon\dot{u}(t) = -\gamma u(t) - \kappa u(t - a - cu(t)). \tag{2.3.1}$$

Bounds on the amplitudes of solutions to the 1-delay equation

Define the following function

$$H_1(v,w) = \left(1 + \frac{\gamma}{\kappa} e^{-\frac{\gamma}{\varepsilon}(a+cv)}\right) v + \frac{\kappa w}{\gamma}\left(1 - e^{-\frac{\gamma}{\varepsilon}(a+cv)}\right). \tag{2.3.2}$$

Recall $h(v)$ from (2.1.8) in Theorem 2.1.6. This is actually a special case of $H_1(v,w)$, with $h(v) = H_1\left(v, \frac{a\kappa}{c\gamma}\right)$. We consider some properties of $H(v,w)$ in Lemma 2.3.1.

Lemma 2.3.1. *Let ε, a, c, γ and $\kappa > 0$.*

1. *For every fixed $w > 0$ there is a $v^*(w) < 0$ such that $H_1(v,w)$ is negative if $v < v^*$, zero if $v = v^*$ and positive if $v \in (v^*, 0]$.*
2. *For all $w > 0$, $v^*(w) \in \left(-\frac{a}{c}, 0\right)$.*
3. *If $w_1 > w_2 > 0$ then $v^*(w_1) < v^*(w_2) < 0$.*
4. *For every fixed $w < 0$ there is a $v^*(w) > 0$ such that $H_1(v,w)$ is negative if $v \in [0, v^*)$, zero if $v = v^*$ and positive if $v > v^*$.*
5. *For all $w < 0$, $v^*(w) \in \left(0, -\frac{\kappa}{\gamma} w\right)$.*
6. *If $w_1 < w_2 < 0$ then $v^*(w_1) > v_*(w_2) > 0$.*

Proof. The first derivatives of $H_1(v,w)$ are given by

$$\frac{\partial}{\partial v} H_1(v,w) = 1 + \frac{\gamma}{\kappa}\left(1 - \frac{c\gamma}{\varepsilon} v + \frac{c\kappa^2}{\varepsilon\gamma} w\right) e^{-\frac{\gamma}{\varepsilon}(a+cv)},$$

$$\frac{\partial}{\partial w} H_1(v,w) = \frac{\kappa}{\gamma}\left(1 - e^{-\frac{\gamma}{\varepsilon}(a+cv)}\right).$$

Let $w > 0$. Then $H_1(0,w) > 0$, $H_1\left(-\frac{a}{c}, w\right) < 0$ and $\frac{\partial}{\partial v} H_1(v,w) > 0$ for all $v < 0$. Then $H_1(v,w)$ is monotonically increasing for all $v < 0$ and changes sign in the interval $\left(-\frac{a}{c}, 0\right)$. Properties 1 and 2 easily follow. Also, $\frac{\partial}{\partial w} H_1(v,w) > 0$ for all $v \in \left(-\frac{a}{c}, 0\right)$. This leads to property 3.

Now let $w < 0$. Then $H_1(0,w) < 0$. Let $W_0(x)$ be the restricted Lambert W function $W_0 \in C\left(\left[-\frac{1}{e}, \infty\right), [-1, \infty)\right)$. This section of the Lambert W function is injective and real-valued. Any solution v to $\frac{\partial}{\partial v} H_1(v,w) = 0$ must be given by $v = \frac{\varepsilon}{c\gamma}\left(1 - W_0\left(-\frac{\kappa}{\gamma} e^{1 + \frac{a\gamma}{\varepsilon} + \frac{c\kappa^2}{\varepsilon\gamma} w}\right)\right) + \frac{\kappa^2}{\gamma^2} w$. Thus there can be at most one point for which $\frac{\partial}{\partial v} H_1(v,w)$ changes signs. Since $\lim_{v\to\infty} H_1(v,w) = 1$ then either the derivative is always positive or it starts off negative and then becomes positive. Either way, since $H_1(0,w) < 0$, there can only be one positive solution v to $H(v,w) = 0$. This proves property 4.

If $v > -\frac{\kappa}{\gamma} w$ then $H_1(v,w) > 0$. Thus the positive root to $H_1(v,w) = 0$ must satisfy $v \in \left(0, -\frac{\kappa}{\gamma} w\right)$. This proves property 5. Since $\frac{\partial}{\partial w} H_1(v,w) > 0$ for all $v > 0$ then property 3 follows. □

Recall the definitions of L_0, M_0 and τ_0 in (2.1.2). Suppose that the requirements of Lemma 2.1.3 are satisfied. Then the unique solution to (2.3.1) is bounded inside the interval (L_0, M_0). Following the same Gronwall argument as for (2.1.6) in the proof of Theorem 2.1.6, we get the following relationship

$$u(t) - u(t-a-cu(t))\, e^{-\frac{\gamma}{\varepsilon}(a+cu(t))} \geqslant -\frac{\kappa M_0}{\gamma}\left(1 - e^{-\frac{\gamma}{\varepsilon}(a+cu(t))}\right).$$

Suppose that we have behaviour (C) in Lemma 2.1.5. If the solution attains a local minimum at $u(t) = v$ then $\dot{u}(t) = 0$ and $u(t-a-cu(t)) = -\frac{\gamma}{\kappa} v$. Applying this to the equation yields

$$\left(1 + \frac{\gamma}{\kappa} e^{-\frac{\gamma}{\varepsilon}(a+cv)}\right) v + \frac{\kappa M_0}{\gamma}\left(1 - e^{-\frac{\gamma}{\varepsilon}(a+cv)}\right) \geqslant 0. \tag{2.3.3}$$

This can be written as $H_1(v, M_0) \geqslant 0$. Let $L_1 = v_*(M_0)$. Then by Lemma 2.3.1, $v \geqslant L_1$. Consider what this means. Within every time interval of length τ_0, either the solution must have attained an extremum or it crossed zero ($u(t)$ and $\dot{u}(t)$ cannot have the same sign for longer than an interval of length τ_0). In either case, we must have $u(t) \geqslant v \geqslant L_1$ for $t \geqslant \tau_0$. After another time interval of τ_0, this will be the lower bound on the relevant history of the solution. Using the same arguments we can find a new upper bound on the solution by finding a bound on any possible local maximum within this time interval. Suppose the local maximum occurs at $u(t) = v$ and using the same Gronwall argument to derive

$$\left(1 + \frac{\gamma}{\kappa} e^{-\frac{\gamma}{\varepsilon}(a+cv)}\right) v + \frac{\kappa L_1}{\gamma}\left(1 - e^{-\frac{\gamma}{\varepsilon}(a+cv)}\right) \leqslant 0.$$

This equation can be written as $H_1(v, L_1) \leqslant 0$. Let $M_1 = v^*(L_1)$. Then by Lemma 2.3.1, $v \leqslant M_1$. Then we have a new positive upper bound on $u(t)$ for $t \geqslant 2\tau_0$. Starting with the original bounds L_0 and M_0 and continuing on in this manner we get that after every interval of length $2\tau_0$, we alternate between obtaining a new lower bound on the solution L_k and a new upper bound M_k that can be computed using the following relationships

$$H_1(L_{k+1}, M_k) = 0, \qquad H_1(M_{k+1}, L_{k+1}) = 0. \tag{2.3.4}$$

From properties 2, 3, 5 and 6 in Lemma 2.3.1, $0 \geqslant L_{k+1} \geqslant L_k \geqslant -\frac{a}{c}$ and $0 \leqslant M_{k+1} \leqslant M_k \leqslant \frac{a\kappa}{c\gamma}$. Hence we have proven Theorem 2.3.2.

Theorem 2.3.2. *Let $\varepsilon, a, c > 0$, $\gamma > 0$ and $\kappa > 0$. Let the history function $\varphi(t)$ be continuous with $\varphi(t) \in (L_0, M_0)$ for $t \in [-\tau_0, 0]$. Define the sequence of bounds $\{L_k\}$ and $\{M_k\}$ using the relationship* (2.3.4)*. Let $u(t)$ be the solution to* (2.3.1)*. Then either $u(t) \to 0$ monotonically or $u(t) \in \left[L_{k(t)}, M_{k(t)}\right]$ where $k(t) = \left\lfloor \frac{t+\tau_0}{4\tau_0} \right\rfloor$ for all $t \geqslant 0$.* □

In the limit as $k \to \infty$, the interval $[L_\infty, M_\infty]$ which bounds the solution as $t \to \infty$ can be calculated using the following equations

$$H_1(L_\infty, M_\infty) = 0, \qquad H_1(M_\infty, L_\infty) = 0. \tag{2.3.5}$$

In Figure 2–3 the bounds are plotted as a function of κ for fixed ε, a, c and γ. In Figure 2–4 sample solutions are plotted with the bounds we have derived. These plots show that the bounds are not tight, but they provide us with better bounds on the solution than (L_0, M_0) and they may be used to find parameter regions for which the fixed point at zero is stable, and a necessary condition for a Hopf bifurcation to occur.

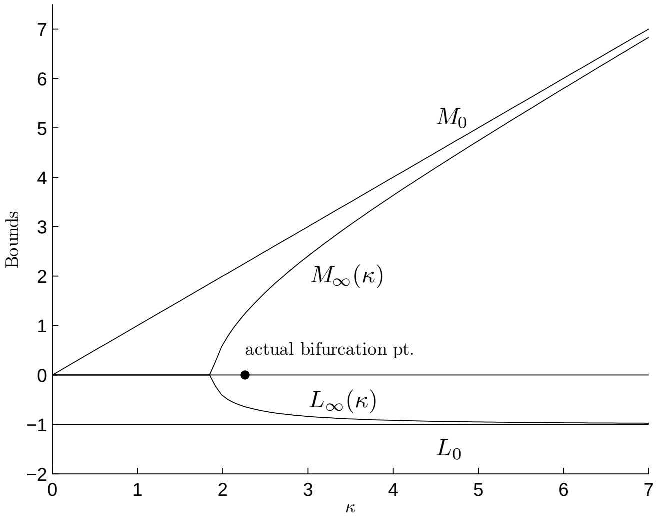
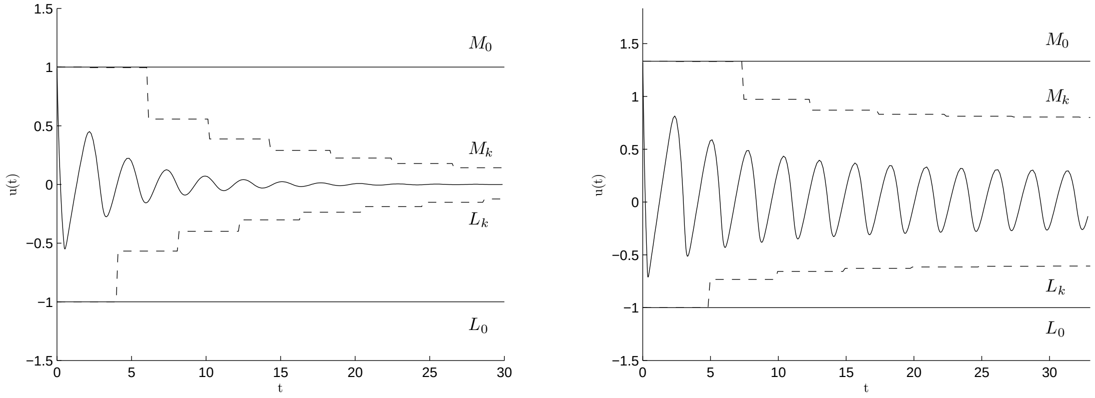


Figure 2–3: Bounds on the solution of (2.3.1) as a function of κ, for $\varepsilon = a = c = \gamma = 1$. The bounds L_∞ and M_∞ are given by (2.3.5). The zero solution of the 1-delay problem loses stability after the point for which the bounds are no longer zero.



(a) $\varepsilon = a = c = 1$, $\gamma = 3$, $\kappa = 3$

(b) $\varepsilon = a = c = 1$, $\gamma = 3$, $\kappa = 4$

Figure 2–4: Sample plots of solutions to (2.3.1). The bounds L_k and M_k on the solution are determined by the iteration (2.3.4).

Bounds on the amplitudes of solutions to the N-delay equation

Let us now consider the N-delay equation,

$$\varepsilon \dot{u}\,(t) = -\gamma u\,(t) - \sum_{i=1}^{N} \kappa_i u\,(t - a_i - c_i u\,(t)). \tag{2.3.6}$$

Here we derive bounds of solutions to (2.3.6) similar to those we have just derived for (2.3.1).

Recall the definition of L_0, M_0 and τ_0 in (2.2.3). Suppose that the requirements of Lemma 2.2.1 are satisfied. Then we have initial bounds (L_0, M_0) on the solution and

$$\dot{u}\,(s) + \frac{\gamma}{\varepsilon} u\,(s) \geqslant -\frac{M_0}{\varepsilon} \sum_{i=1}^{N} \kappa_i,$$
$$\left(u\,(s)\, e^{\frac{\gamma}{\varepsilon} s}\right)' \geqslant -\frac{M_0 e^{\frac{\gamma}{\varepsilon} s}}{\varepsilon} \sum_{i=1}^{N} \kappa_i.$$

Using a Gronwall inequality, integrating from $s = t - a_j - c_j u\,(t)$ to $s = t$ yields

$$u\,(t)\, e^{\frac{\gamma}{\varepsilon} t} - u\,(t - a_j - c_j u\,(t))\, e^{\frac{\gamma}{\varepsilon}(t - a_j - c_j u(t))} \geqslant -\frac{M_0 \left(e^{\frac{\gamma}{\varepsilon} t} - e^{\frac{\gamma}{\varepsilon}(t - a_j - c_j u(t))}\right)}{\gamma} \sum_{i=1}^{N} \kappa_i,$$
$$u\,(t)\, e^{\frac{\gamma}{\varepsilon}(a_j + c_j u(t))} - u\,(t - a_j - c_j u\,(t)) \geqslant -\frac{M_0 \left(e^{\frac{\gamma}{\varepsilon}(a_j + c_j u(t))} - 1\right)}{\gamma} \sum_{i=1}^{N} \kappa_i. \tag{2.3.7}$$

Similar to the 1-delay case, the solution to the N-delay problem may either monotonically go to zero or oscillate. Suppose the latter case and suppose that the solution reaches a minimum at $u\left(t\right)=v$. Then at this point $\dot{u}\left(t\right)=0$. From the bounds on the solution,

$$u\left(t-a_1-c_1u(t)\right)\geqslant-\frac{\gamma}{\kappa_1}v-\frac{M_0}{\kappa_1}\sum_{i=2}^{N}\kappa_i.$$

Substitute this into (2.3.7) with $j=1$

$$\left(e^{\frac{\gamma}{\varepsilon}(a_1+c_1v)}+\frac{\gamma}{\kappa_1}\right)v\geqslant-\frac{M_0\left(e^{\frac{\gamma}{\varepsilon}(a_1+c_1v)}-1\right)}{\gamma}\sum_{i=1}^{N}\kappa_i-\frac{M}{\kappa_1}\sum_{i=2}^{N}\kappa_i,$$

$$\left(1+\frac{\gamma}{\kappa_1}e^{-\frac{\gamma}{\varepsilon}(a_1+c_1v)}\right)v+\frac{M_0}{\gamma}\left(\left(1-e^{-\frac{\gamma}{\varepsilon}(a_1+c_1v)}\right)\sum_{i=1}^{N}\kappa_i+\frac{\gamma}{\kappa_1}e^{-\frac{\gamma}{\varepsilon}(a_1+c_1v)}\sum_{i=2}^{N}\kappa_i\right)\geqslant 0.$$

This suggest defining a new function $H_N\left(v,w\right)$

$$H_N\left(v,w\right)=\left(1+\frac{\gamma}{\kappa_1}e^{-\frac{\gamma}{\varepsilon}(a_1+c_1v)}\right)v+\frac{w}{\gamma}\left[\left(1-e^{-\frac{\gamma}{\varepsilon}(a+cv)}\right)\sum_{i=1}^{N}\kappa_i+\frac{\gamma}{\kappa_1}e^{-\frac{\gamma}{\varepsilon}(a+cv)}\sum_{i=2}^{N}\kappa_i\right].$$

If $N=1$ this is $H_1\left(v,w\right)$ as defined in (2.3.2). Going back to (2.3), we can write this as $H_N\left(v,M_0\right)\geqslant 0$. Let $L_1<0$ solve $H_N\left(L_1,M\right)=0$. Then $v\geqslant L_1$ and L_1 is a new negative lower bound on the solution. After a time τ_0 this will be the bound on the relevant history function and using the same steps as in the derivation of (2.3), we derive

$$\left(1+\frac{\gamma}{\kappa_1}e^{-\frac{\gamma}{\varepsilon}(a_1+c_1v)}\right)v+\frac{L_1}{\gamma}\left(\left(1-e^{-\frac{\gamma}{\varepsilon}(a_1+c_1v)}\right)\sum_{i=1}^{N}\kappa_i+\frac{\gamma}{\kappa_1}e^{-\frac{\gamma}{\varepsilon}(a_1+c_1v)}\sum_{i=2}^{N}\kappa_i\right)\leqslant 0.$$

This can be written as $H_N\left(v,L_1\right)\leqslant 0$. Let M_1 be a positive solution to $H_N\left(x,L_1\right)=0$, then $v\leqslant M_1$ and M_1 is a new positive upper bound on the solution. As in the $N=1$ case, if we continue on in this manner, after every interval of length $2\tau_0$ we alternate between obtaining a new lower bound on the solution L_k and a new upper bound M_k that can be computed using the following relationships

$$H_N\left(L_{k+1},M_k\right)=0,\qquad H_N\left(M_{k+1},L_{k+1}\right)=0. \tag{2.3.8}$$

In the limit $k \to \infty$, the interval $[L_\infty, M_\infty]$ which bounds on the solution as $t \to \infty$ can be calculated using the following equations

$$H_N\left(L_\infty, M_\infty\right) = 0, \qquad H_N\left(M_\infty, L_\infty\right) = 0. \tag{2.3.9}$$

Figure 2–5 shows an example of these bounds for the $N = 2$ case as a function of κ_1 for fixed ε, a_1, a_2, c_1, c_2, γ and κ_2.

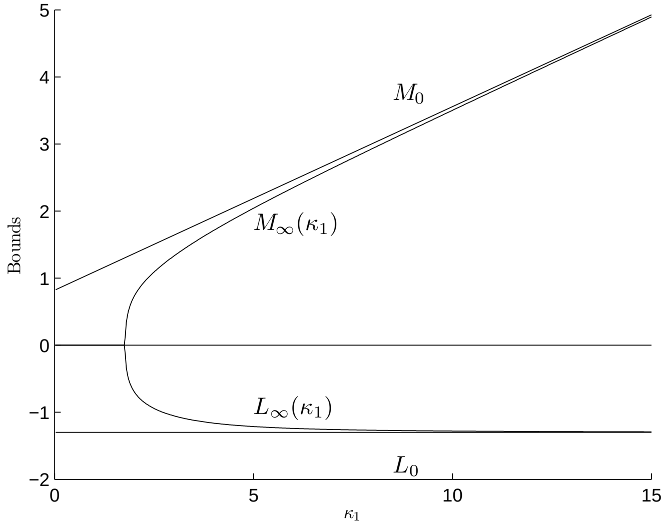


Figure 2–5: Bounds on the solution of (2.3.6) as a function of κ, for $N = 2$, $\varepsilon = c_1 = c_2 = 1$, $a_1 = 1.3$, $a_2 = 6$, $\gamma = 4.75$ and $\kappa_2 = 3$. The bounds L_∞ and M_∞ are given by (2.3.9).

CHAPTER 3
Stability of DDEs

In this chapter we consider the stability of the model DDE with one state dependent delay (1.1.1). We begin with the constant delay case ($c = 0$) and derive its characteristic equation. For fixed ε and a, the analytic stability region $\Sigma_\star$ of the zero solution to (1.1.1) in the (μ, σ) plane is derived from its characteristic equation by requiring all the roots to have a negative real part [15]. The mathematical description of $\Sigma_\star$ is well-known and can be found in the literature [1, 8, 10, 29]. Results by Györi and Hartung [25] show that $\Sigma_\star$ is also the stability region of the state dependent case ($c \neq 0$). In Section 3.2 we use a Gronwall argument to directly prove the asymptotic stability of the trivial solution for the state dependent DDE in part of $\Sigma_\star$. In Section 3.3 we consider the stability of RFDEs of the form (1.1.3) because there are well-established general results on stability of these equations [6, 29]. In particular we look at the generalisation of Lyapunov functions for ODEs to Lyapunov functionals for RFDEs by Krasovskii [34], and the switch back to Lyapunov functions for RFDEs using Lyapunov-Razumikhin theorems [51]. These ideas are then generalised to state dependent DDEs by applying them to our model DDE. The result is a proof of Lyapunov stability in a larger subset of $\Sigma_\star$ than that found using the Gronwall argument. These direct techniques to prove the stability of DDEs are extended in the next chapter into methods to prove the stability of numerical methods applied to DDEs.

3.1 The known stability region of the model problem

Here we restate the model DDE (1.1.1)

$$\begin{aligned} \varepsilon \dot{u}(t) &= \mu u(t) + \sigma u(t - a - cu(t)), \quad && t \geqslant 0, \\ u(t) &= \varphi(t), && t \leqslant 0. \end{aligned} \tag{3.1.1}$$

All parameters are in $\mathbb{R}$ and $\varepsilon, a > 0$. First set $c = 0$ to consider the constant delay case. Recall from Chapter 2 that it is always possible to rescale the equation so that $\varepsilon = a = 1$. Thus, μ and σ are effectively the only two free parameters in this equation. We would like to find the values of (μ, σ) for which the trivial solution of (3.1.1) is stable.

Let $u(t) = e^{\lambda t}$ to derive the characteristic equation of (3.1.1),

$$f(\lambda) = \varepsilon\lambda - \mu - \sigma e^{-\lambda a} = 0. \tag{3.1.2}$$

The equation $f(\lambda) = 0$ has infinitely many solutions in $\mathbb{C}$. Setting $\lambda = x + iy$ in (3.1.2) and taking the real and imaginary parts of this equation yields

$$-\mu + \varepsilon x - \sigma e^{-ax}\cos(ay) = 0, \quad \varepsilon y + \sigma e^{-ax}\sin(ay) = 0 \tag{3.1.3}$$

This equation is further manipulated to yield the equation of a curve (3.1.4) on which all the roots must lie. Since for $\lambda = x + iy$

$$\sigma^2 e^{-2ax}\cos(ay) = (-\mu + \varepsilon x)^2, \quad \sigma^2 e^{-2ax}\sin(ay) = \varepsilon^2 y^2,$$

then

$$(\varepsilon x - \mu)^2 + \varepsilon^2 y^2 = \sigma^2 e^{-2ax}. \tag{3.1.4}$$

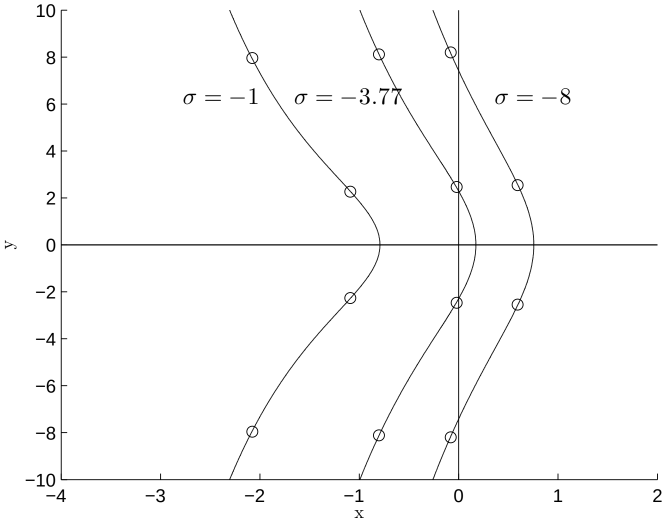


Figure 3–1: Illustration of the eigenvalues of (3.1.2) in the complex plane. The eigenvalues are indicated by open circles and for a given σ value they lie on the curve described by (3.1.4). The other parameter values are $\varepsilon = 1$, $\mu = -3$ and $a = 1$.

Figure 3–1 shows a complex conjugate pair of eigenvalues crossing the imaginary axis. Setting $x = 0$ in (3.1.3) gives the curve for which these crossings occur.

The stability region of (3.1.1) is well-known and is discussed in the books by Bellen and Zennaro [8], Bellman and Cooke [10] and Hale and Verduyn Lunel [29].

Definition 3.1.1 (Analytic Stability Region of (3.1.1) with $c = 0$)**.** The region bounded between the following curves

$$\ell_\star = \left\{(v, -v), v \in \left(-\infty, \tfrac{\varepsilon}{a}\right]\right\}$$

$$g_\star = \left\{(\mu(y), \sigma(y)), y \in \left(0, \tfrac{\pi}{a}\right)\right\}$$

where the functions $\mu(y)$ and $\sigma(y)$ are given by (3.1.5)

$$\mu(y) = \varepsilon y \cot(ay), \quad \sigma(y) = -\varepsilon y \csc(ay) \tag{3.1.5}$$

is the analytic stability region of the DDE (3.1.1). The stability region, to be denoted by $\Sigma_\star$, can be divided into three subregions: the cone $\overset{\Delta}{\Sigma}$, the wedge $\overset{w}{\Sigma}$ and the cusp $\overset{c}{\Sigma}$. These divisions are shown in Figure 3–2.

$$\Sigma_\star = \overset{\Delta}{\Sigma} \cup \overset{w}{\Sigma} \cup \overset{c}{\Sigma}, \qquad \partial\Sigma_\star = \ell_\star \cup g_\star$$

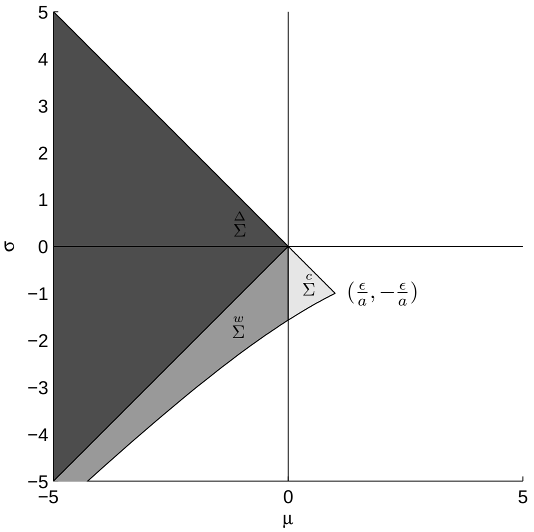


Figure 3–2: The analytic stability region $\Sigma_\star$ in the (μ, σ) plane for $\varepsilon = a = 1$.

Figure 3–3 shows how $\Sigma_\star$ depends on a. As $a \to 0$ (3.1.1) approaches the ODE $\dot{u}(t) = (\mu + \sigma)u(t)$ which is stable for $\mu + \sigma < 0$. Figure 3–3 shows that indeed $\Sigma_\star$ fills the region $\mu + \sigma < 0$ as $a \to 0$.

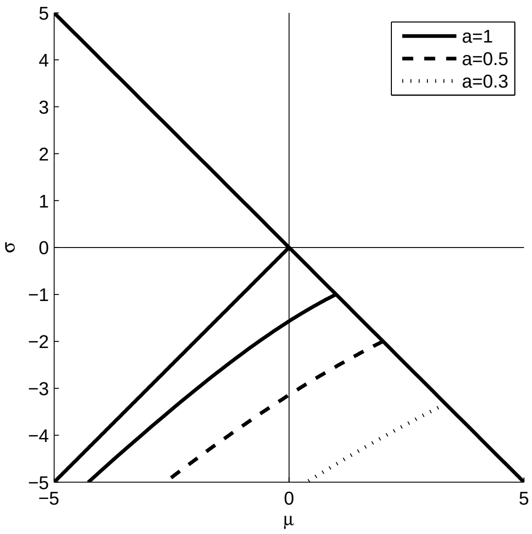


Figure 3–3: The dependence of $\Sigma_\star$ to the delay term a.

Now look for the set of (μ, σ) values for which there are real eigenvalues. In this case we set $\lambda = x \in \mathbb{R}$ and obtain

$$f(x) = \varepsilon x - \mu - \sigma e^{-ax},$$

$$f'(x) = \varepsilon + a\sigma e^{-ax}, \qquad f''(x) = -a^2 \sigma e^{-ax}.$$

There are two cases to be considered: $\sigma \geqslant 0$ and $\sigma < 0$. In the first case we have $f'(x) > 0$ and $\lim_{x \to -\infty} f(x) = -\infty$ and $\lim_{x \to +\infty} f(x) = +\infty$ so there is always one real root. In the second case we observe that f has to be concave up. We also observe that the function attains a minimum value at $x_1 = \frac{1}{a} \ln\left(-\frac{a\sigma}{\varepsilon}\right)$. The number of real eigenvalues in this case are thus given by whether or not $f(x_1)$ lies above or below zero.

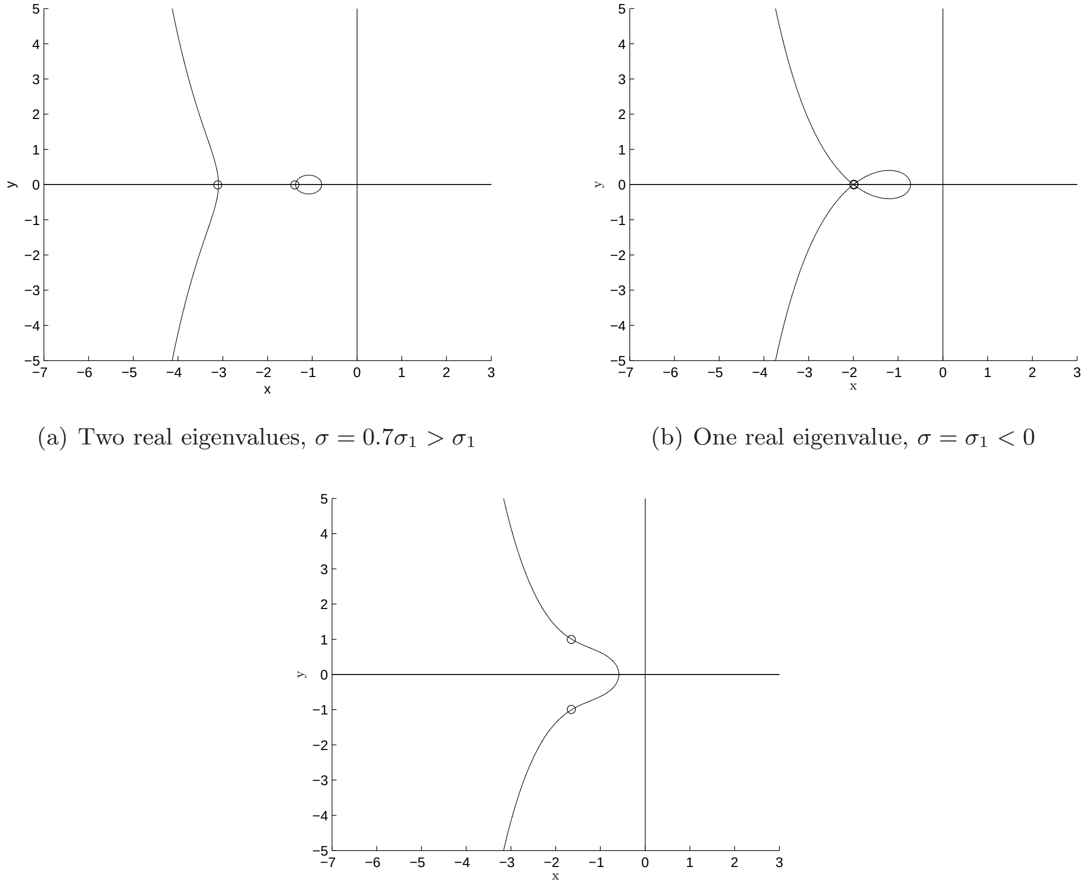
$$\begin{aligned} f(x_1) &= \frac{\varepsilon}{a} \ln\left(-\frac{a\sigma}{\varepsilon}\right) - \mu - \sigma e^{-\ln\left(-\frac{a\sigma}{\varepsilon}\right)} \\ &= \frac{\varepsilon}{a}\left(\ln\left(-\frac{a\sigma}{\varepsilon}\right) + 1\right) - \mu \end{aligned}$$

The zero of this function is $\sigma = -\frac{\varepsilon}{a} e^{\frac{a\mu}{\varepsilon} - 1}$. Define the curve

$$\sigma_1(\mu) = -\frac{\varepsilon}{a} e^{\frac{a\mu}{\varepsilon} - 1}. \tag{3.1.6}$$

Together with our result for $\sigma \geqslant 0$, the number of real roots are given by

$$\text{no. of real roots } = \begin{cases} 0, & \text{if } \sigma < \sigma_1(\mu), \\ 1, & \text{if } \sigma < 0 \text{ and } \sigma = \sigma_1(\mu) < 0, \text{ or } \sigma \geqslant 0, \\ 2, & \text{if } \sigma_1(\mu) < \sigma < 0. \end{cases}$$



(a) Two real eigenvalues, $\sigma = 0.7\sigma_1 > \sigma_1$

(b) One real eigenvalue, $\sigma = \sigma_1 < 0$

(c) No real eigenvalues, $\sigma = 1.7\sigma_1 < \sigma_1$

Figure 3–4: Illustration of the eigenvalues (solutions to (3.1.2)) in the complex plane. The eigenvalues are indicated by open circles. For each σ value they lie on the curve described by (3.1.4). Note that the curve is composed of two separate parts for the case $\sigma < \sigma_1$. The other parameter values are $\varepsilon = 1, \mu = -1$ and $a = 1$.

Now consider when the purely real eigenvalues change signs. Note that when $\sigma = 0$ then the sole eigenvalue is $\lambda = \frac{\mu}{\varepsilon}$ so it has the same sign as μ. Also when $\sigma < 0$, on the curve

$\sigma = \sigma_1(\mu)$ given by (3.1.6) there is only one real eigenvalue x_1,

$$x_1 = \frac{1}{a}\ln\left(-\tfrac{a\sigma_1}{\varepsilon}\right) = \frac{1}{a}\ln\left(e^{\frac{a\mu}{\varepsilon}-1}\right) = \frac{\mu}{\varepsilon} - \frac{1}{a}$$

So the one real eigenvalue on the curve defined by $\sigma = \sigma_1(\mu)$ changes sign from negative to positive when $\mu = \frac{\varepsilon}{a}$. Finally, we have a zero eigenvalue whenever $f(0) = 0$ which occurs when $0 - \mu - \sigma e^{-0} = 0 \Rightarrow \mu + \sigma = 0$.

In Figure 3–5 we show the boundaries at which some the eigenvalues of (3.1.2) change their nature. The blue curve is described by (3.1.5). On the blue curve there is a pair of complex conjugate eigenvalues on the imaginary axis and all other eigenvalues have negative real parts. The green and red curves represent $\sigma = \sigma_1(\mu)$. In the lower half-plane ($\sigma < 0$) and above this curve there are two real eigenvalues and below it there are none. On the curve itself there is only one real eigenvalue and it is negative on the green part ($\mu < \frac{\varepsilon}{a}$) and positive on the red part ($\mu > \frac{\varepsilon}{a}$). In the upper half-plane ($\sigma > 0$) there is only one real eigenvalue. The magenta line represents $\mu + \sigma = 0$ where there is a zero eigenvalue.

Further discussions are available in other references [10, 28, 29] which describe the properties of eigenvalues more comprehensively in each of the divisions shown in Figure 3–5. Here we just state that the stability region is found by splitting the (μ, σ) plane using the magenta and blue curves and taking the region on the left. When $\sigma \geqslant \sigma_1(\mu)$ the zero solution becomes unstable when it crosses the magenta line because an eigenvalue crosses zero and changes sign from negative to positive. When $\sigma < 0$ the zero solution becomes unstable when it crosses the blue curve because a pair of complex conjugate eigenvalues crossed the imaginary axis. These curves intersect at the point $\left(-\frac{\varepsilon}{a}, \frac{\varepsilon}{a}\right)$.

Now we consider the stability region of the state dependent case. Set $c \neq 0$. Following the results of Györi and Hartung [25], we can show that this actually stable in the same region as in the constant delay case.

Theorem 3.1.2 (Györi and Hartung [25])**.** *Consider the delay systems*

$$\dot{x}(t) = \sum_{i=1}^{m} A_i(t)x\left(t - \tau_i(t, x_t)\right) \tag{3.1.7}$$

$$\dot{y}(t) = \sum_{i=1}^{m} A_i(t)y\left(t - \tau_i(t, \mathbf{0})\right) \tag{3.1.8}$$

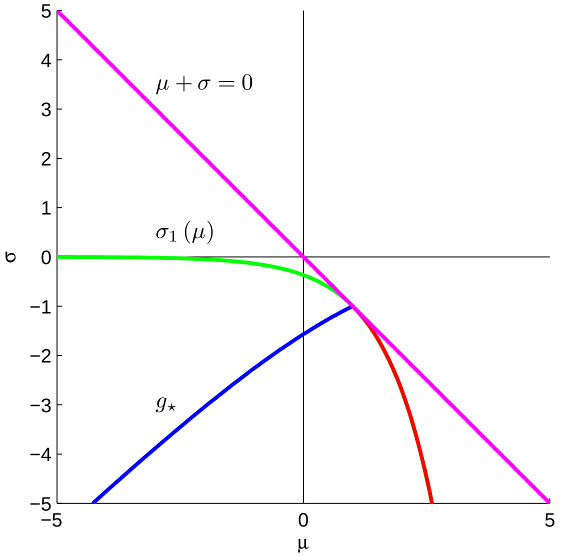


Figure 3–5: Divisions of the (μ,σ) plane based on what we know about the eigenvalues. The blue curve is described by (3.1.5). The green and red curves together represent $\sigma=\sigma_1(\mu)$. The magenta line represents $\mu+\sigma=0$ where there is a zero eigenvalue.

where $x_t \in C = C\left([-r,0],\mathbb{R}^d\right)$*. Suppose that the following are true:*

1. $A_i:[0,\infty)\to\mathbb{R}^{d\times d}$ *is continuous and* $|A_i(t)|\leqslant b_i$*,* $t\in[0,+\infty)$ *for* $i=1,...,m$*;*
2. $\varphi\in C$*;*
3. *the delay functions* $\tau_i:[0,\infty)\times C\to[0,r]$ *are continuous for* $i=1,...,m$*;*
4. *there exist a constant* $0<\gamma\leqslant\infty$ *and continuous functions* $\omega_i:[0,\gamma)\to[0,\infty)$*, such that*

$$|\tau_i(t,\psi)-\tau_i(t,\mathbf{0})|\leqslant\omega_i\left(\|\psi\|\right),\quad t\geqslant0,\quad\|\psi\|<\gamma,\quad i=1,...,m$$

   where $\omega_i(0)=0$ *(*$i=1,...,m$*).*
5. *the sets* $\{s\in[0,r]:s-\tau_i(s+t_0,\mathbf{0})=0\}$ *have Lebesgue measure 0 for* $i=1,...,m$ *and* $\tau_0\geqslant0$*.*

Conditions (1)-(3) guarantee existence but not uniqueness of the solutions (Driver [14]). Conditions (1)-(5) guarantee that the trivial solution of (3.1.7) *is exponentially stable if and only if the trivial solution of* (3.1.8) *is exponentially stable.*

This result can be applied to the auxiliary system (2.1.3) defined in Chapter 2 for the proof of existence and uniqueness of solutions. In this system we change the deviated argument in (3.1.1) to become $\alpha\left(t,u\left(t\right)\right)=\min\left\{t,t-a-cu\left(t\right)\right\}$. This auxiliary state dependent problem would be (3.1.7) and the constant delay problem would be (3.1.8). It is easy to show that (1)-(5) are satisfied in this case. Thus the auxiliary system also has $\Sigma_\star$ as its stability region. By choosing the bounds on initial function to be small enough so that no advances are allowed, the auxiliary system becomes equivalent to (3.1.1) so $\Sigma_\star$ is also the stability region of the zero solution to (3.1.1).

Finally, consider the following general state dependent delay equation,

$$\begin{aligned} &\varepsilon \dot{u}(t) = \mu u(t) + \sigma u(t - a - c(u(t))), \quad && t \geqslant 0, \\ &u(t) = \varphi(t), && t < 0. \end{aligned} \tag{3.1.9}$$

where $c(u)$ is a continuous function with $c(0) = 0$. Then using the result of Györi and Hartung [25] and the same steps as before, it is possible to show that $\Sigma_\star$ is the stability region of (3.1.9) as well.

3.2 Asymptotic stability of the model problem using a Gronwall argument

In this section we find sufficient conditions for the solution of (3.1.1) to remain inside an interval $[-\delta, \delta]$ for small enough δ and converge to zero. It is easy to see that the solutions are asymptotically stable for $(\mu, \sigma) \in \overset{\Delta}{\Sigma} = \{|\sigma| < -\mu\}$. Suppose a solution attains an extremum at time t. Then $\dot{u}(t) = 0$ so $u(t) = -\frac{\sigma}{\mu} u(t - a - cu(t))$. This shows a contraction in the size of the solution because $\left|\frac{\sigma}{\mu}\right| < 1$ and it becomes obvious that the solutions must go to zero. This asymptotic stability of the trivial solution for the constant delay case in $\overset{\Delta}{\Sigma}$ is proven in Section 3.3 using Lyapunov methods. Here we move on to proving asymptotic stability of the state dependent DDE in some subset of $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$, the delay dependent portion of the analytic stability region. In these regions we have $\sigma \leqslant \mu$, $\sigma < -\mu$, and automatically $\sigma < 0$.

We begin with a lemma on the behaviour of bounded solutions of (3.1.1). This is similar to Lemma 2.1.5 but it allows for positive values of μ.

Lemma 3.2.1. *Let $\varepsilon, a, c > 0$, $\mu + \sigma < 0$ and let $u(t)$ be a solution to* (3.1.1) *such that $u(t) \in [-\delta, \delta], \delta \in \left(0, \frac{a}{c}\right)$ for all t. Then $u(t)$ must be behave in one of the following manners:*

(A) There exists $\bar{T}$ such that $u(t) \downarrow 0$ for $t > \bar{T}$

(B) There exists $\bar{T}$ such that $u(t) \uparrow 0$ for $t > \bar{T}$

(C) For every $T > 0$ there exists $T_1, T_2 > T$ such that the solution attains a local maximum at T_1 and a local minimum at T_2.

Proof. Suppose there exists a time $\bar{T}$ such that $\dot{u}\left(t\right) \leqslant 0$ for all $t > \bar{T}$. Since the solutions are assumed to be bounded then there is $\bar{u} \in [-\delta, \delta]$, $u(t) \downarrow \bar{u}$ when $t \geqslant \bar{T}$. Then $\dot{u}\left(t\right) \to 0$. Since $t - a - c\delta \leqslant \alpha\left(t, u(t)\right) \leqslant t$ then $u\left(\alpha\left(t, u(t)\right)\right) \downarrow \bar{u}$ as well.

$$0 = \lim_{t\to\infty} \varepsilon\dot{u}(t) = \lim_{t\to\infty} \left[\mu u(t) + \sigma u\left(t - a - cu(t)\right)\right] = \left(\mu + \sigma\right)\bar{u}.$$

Since $\mu + \sigma < 0$ then $\bar{u} = 0$. Similarly, if there exists T such that $\dot{u}\left(t\right) \geqslant 0$ for all $t > T$ then we must have $u\left(t\right) \uparrow 0$ for $t > \bar{T}$. If there is no $\bar{T}$ such that the derivative does not change sign past $t > \bar{T}$ then it must behave as described in (C). □

Lemma 3.2.2. *Let* $\varepsilon, a, c > 0$, $\sigma \leqslant \mu$ *and* $\sigma < -\mu$. *Let* $-\frac{a}{c} < L < 0 < M$, $t > 0$ *and let the solution to* (2.1.1) *exist and satisfy* $u\left(s\right) \in [L, M]$ *for* $s \in [t - a - cM, t]$. *If* $\mu \neq 0$ *then the following inequalities hold*

$$u\left(t\right)e^{-\frac{\mu}{\varepsilon}t} - u\left(t - a - cu\left(t\right)\right)e^{-\frac{\mu}{\varepsilon}\left(t-a-cu(t)\right)} \geqslant -\frac{\sigma M}{\mu}\left(e^{-\frac{\mu}{\varepsilon}t} - e^{-\frac{\mu}{\varepsilon}\left(t-a-cu(t)\right)}\right), \qquad (3.2.1)$$

$$u\left(t\right)e^{-\frac{\mu}{\varepsilon}t} - u\left(t - a - cu\left(t\right)\right)e^{-\frac{\mu}{\varepsilon}\left(t-a-cu(t)\right)} \leqslant -\frac{\sigma L}{\mu}\left(e^{-\frac{\mu}{\varepsilon}t} - e^{-\frac{\mu}{\varepsilon}\left(t-a-cu(t)\right)}\right). \qquad (3.2.2)$$

If $\mu = 0$ *then the following inequalities hold,*

$$u\left(t\right) - u\left(t - a - cu\left(t\right)\right) \geqslant \frac{\sigma M}{\varepsilon}\left(a + cu(t)\right), \qquad (3.2.3)$$

$$u\left(t\right) - u\left(t - a - cu\left(t\right)\right) \leqslant \frac{\sigma L}{\varepsilon}\left(a + cu(t)\right). \qquad (3.2.4)$$

Proof. Since $u\left(s\right) \leqslant M$ then

$$\varepsilon\dot{u}\left(s\right) - \mu u\left(s\right) = \sigma u\left(s - a - cu\left(s\right)\right) \geqslant \sigma M.$$

for all $s \in [t - a - cM, t]$. Equation (3.2.1) is derived using the same Gronwall arguments that were used to derive both (2.1.6) and (2.3.3) in Chapter 2. The proof of (3.2.2), (3.2.3) and (3.2.4) are similar. Observe that taking the limit as $\mu \to 0$ in (3.2.1) and (3.2.2) yield (3.2.3) and (3.2.4) respectively. □

Lemma 3.2.3. *Let* $\varepsilon, a, c > 0$, $\sigma \leqslant \mu$ *and* $\sigma < -\mu$. *Let* $-\frac{a}{c} < L < 0 < M$, $t > 0$ *and let the solution to* (2.1.1) *exist and satisfy* $u\left(s\right) \in [L, M]$ *for* $s \in [t - a - cM, t]$. *Suppose that* $u\left(t\right) = v$

is a local extremum of the solution. If $\mu \neq 0$ *and* $1+\frac{\mu}{\sigma}e^{\frac{\mu}{\varepsilon}(a+cv)} > 0$ *then*

$$-\frac{\sigma}{\mu}\left[\frac{1-e^{\frac{\mu}{\varepsilon}(a+cv)}}{1+\frac{\mu}{\sigma}e^{\frac{\mu}{\varepsilon}(a+cv)}}\right]M \leqslant v \leqslant -\frac{\sigma}{\mu}\left[\frac{1-e^{\frac{\mu}{\varepsilon}(a+cv)}}{1+\frac{\mu}{\sigma}e^{\frac{\mu}{\varepsilon}(a+cv)}}\right]L. \tag{3.2.5}$$

If $\mu = 0$ *then*

$$\frac{\sigma M}{\varepsilon}\left(a+cv\right) \leqslant v \leqslant \frac{\sigma L}{\varepsilon}\left(a+cv\right) \tag{3.2.6}$$

Proof. Let $u\left(t\right) = v$ be a relative extremum of the solution for $t > 0$. Then $\dot{u}(t) = 0$ so we must have $u\left(t-a-cu\left(t\right)\right) = -\frac{\mu}{\sigma}v$. Using this and Lemma 3.2.2, we derive (3.2.5) and (3.2.6). Observe that taking the limit as $\mu \to 0$ in (3.2.5) yields (3.2.6). □

Definition 3.2.4. Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu$ and $\sigma < -\mu$. Define

$$r\left(v\right) = \begin{cases} \frac{\sigma}{\mu}\left[\frac{1-e^{\frac{\mu}{\varepsilon}(a+cv)}}{1+\frac{\mu}{\sigma}e^{\frac{\mu}{\varepsilon}(a+cv)}}\right], & \text{if } \mu \neq 0, \\ -\frac{\sigma}{\varepsilon}\left(a+cv\right), & \text{if } \mu = 0. \end{cases}$$

For fixed v, the expression $r\left(v\right)$ is continuous in μ, including at $\mu = 0$.

Lemma 3.2.5. *Let* $\varepsilon, a, c > 0$, $\sigma \leqslant \mu$ *and* $\sigma < -\mu$. *Let the model parameters satisfy* $r\left(0\right) \in \left(0,1\right)$. *Then there exists a sufficiently small* $\delta \in \left(0, \frac{a}{c}\right)$ *such that* $r\left(v\right) \in \left(0,1\right)$ *for all* $v \in \left[-\delta, \delta\right]$. *If the history function* $\varphi\left(t\right)$ *is continuous with* $\varphi\left(t\right) \in \left[\delta, \delta\right]$ *for* $t \in \left[-a-c\delta, 0\right]$ *then the solution* $u(t)$ *to* (3.1.1) *satisfies* $u\left(t\right) \in \left[-\delta, \delta\right]$ *for all* $t \geqslant 0$.

Proof. Let $r\left(0\right) \in \left(0,1\right)$. Since $r\left(v\right)$ is a continuous function of v then it is always possible to find a small enough $\delta \in \left(0, \frac{a}{c}\right)$ such that $r\left(v\right) \in \left(0,1\right)$ for all $v \in \left[-\delta, \delta\right]$.

Consider the sign of $r\left(v\right)$. This is always positive if $\mu < 0$. If $\mu > 0$ and $1+\frac{\mu}{\sigma}e^{\frac{\mu}{\varepsilon}(a+cv)} < 0$ then $r\left(v\right) < 0$. Thus, our choice of δ always excludes the case $1+\frac{\mu}{\sigma}e^{\frac{\mu}{\varepsilon}(a+cv)} < 0$. If $\mu = 0$ then we always have $r\left(0\right) \geqslant 0$ since $v \geqslant -\frac{a}{c}$.

First let $\varphi\left(t\right) \in \left[-\delta, \delta\right]$ for all $t \leqslant 0$. Suppose it is possible for $u\left(t\right)$ to leave the interval $\left[-\delta, \delta\right]$. Suppose that when this first happens the solution crosses its upper bound. Since $u\left(t\right)$ is continuous and differentiable for all $t > 0$, there exists $T_1 > 0$ such that $u\left(T_1\right) = \delta$ and $\dot{u}\left(T_1\right) \geqslant 0$.

$$0 \leqslant \varepsilon\dot{u}\left(T_1\right) = \mu u\left(T_1\right) + \sigma u\left(T_1 - a - cu\left(T_1\right)\right) = \mu\delta + \sigma u\left(T_1 - a - c\delta\right)$$

$$u\left(T_1 - a - c\delta\right) \leqslant -\frac{\mu}{\sigma}\delta \tag{3.2.7}$$

Since for all $t \leqslant T_1$ $u(t) \in [-\delta, \delta]$ then from (3.2.1) in Lemma 3.2.2 we obtain

$$u(t) - u(t - a - cu(t))\, e^{\frac{\mu}{\varepsilon}(a+cu(t))} \leqslant \frac{\sigma\delta}{\mu}\left(1 - e^{\frac{\mu}{\varepsilon}(a+cu(t))}\right).$$

At $t = T_1$ we have $u(T_1) = \delta$ and applying (3.2.7) we obtain

$$\left(1 + \frac{\mu}{\sigma} e^{\frac{\mu}{\varepsilon}(a+c\delta)}\right)\delta \leqslant \delta\frac{\sigma}{\mu}\left(1 - e^{\frac{\mu}{\varepsilon}(a+c\delta)}\right), \tag{3.2.8}$$

and hence

$$\frac{\sigma}{\mu}\left[\frac{1 - e^{\frac{\mu}{\varepsilon}(a+c\delta)}}{1 + \frac{\mu}{\sigma} e^{\frac{\mu}{\varepsilon}(a+c\delta)}}\right] \geqslant 1.$$

This implies $r(\delta) > 1$ which is a contradiction.

Similarly we obtain a contradiction to the solution leaving through the lower bound. Thus it is not possible for $u(t)$ to leave $[-\delta, \delta]$. Because of this we only require the history function $\varphi(t) \in [-\delta, \delta]$ for $t \in [-a - c\delta, 0]$. □

Theorem 3.2.6. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu$ and $\sigma < -\mu$. Let the model parameters satisfy $r(0) \in (0, 1)$. Then there exists a sufficiently small $\delta \in \left(0, \frac{a}{c}\right)$ such that if the history function $\varphi(t)$ is continuous with $\varphi(t) \in [-\delta, \delta]$ for all $t \in [-a - c\delta, 0]$ then the solution to* (3.1.1) *satisfies $u(t) \in [-\delta, \delta]$ for all $t \geqslant 0$ and $\lim_{t\to\infty} u(t) = 0$.*

Remark 3.2.7. In the limiting case $c = 0$ then $r(v) = r(0)$ for all $v \in [-\delta, \delta]$ which allows for $\delta \to \infty$. This reflects the global stability of zero solution when $r(0) \in (0, 1)$ for the constant delay case.

Proof. Choose δ as in Lemma 3.2.5 and define

$$r = \max_{v\in[-\delta,\delta]} r(v). \tag{3.2.9}$$

Then $r \in (0, 1)$ and by Lemma 3.2.5, $u(t) \in [-\delta, \delta]$ and $\alpha(t, u(t)) \in [t - a - c\delta, t]$ for $t \geqslant 0$. We call the solution on this interval the relevant history at time t.

Since the solutions are bounded then by Lemma 3.2.1, the solution may (A) eventually go to zero monotonically from above, or (B) eventually go to zero monotonically from below, or (C) oscillate. To complete the proof of this theorem we only have to prove that in the oscillating case we still get $u(t) \to 0$.

Suppose the solution attains a minimum at $u(S_1) = L_1$. Then by Lemma 3.2.3,

$$L_1 \geqslant -r(L_1)\delta \geqslant -r\delta > -\delta.$$

So $-r\delta$ is a new lower bound on the solution. For $t \geqslant S_1 + a + c\delta$, the relevant history is on the time interval $[0, a + c\delta]$ so $-r\delta$ is a lower bound on the relevant history. Suppose now that a local maximum occurs at $u(R_1) = M_1$. Then by Lemma 3.2.3,

$$M_1 \leqslant -r(M_1)(-r\delta) \leqslant r^2\delta$$

For $t \geqslant R_1 + a + c\delta$ the relevant history is bounded above by $r^2\delta$. Starting with these definitions of S_1 and R_1, for $n \geqslant 2$ define S_n to be the location of the first local minimum past $R_{n-1}+a+c\delta$, and R_n to be the location of the first local maximum past $S_n + a + c\delta$. Since we are assuming that the solutions behaves as described in (C) in Lemma 3.2.1, S_n and R_n exists for all $n \geqslant 0$. By iteratively applying Lemma 3.2.3, we get that for all $t \geqslant S_1$,

$$\begin{aligned} &u(t) \in \left[-r^n\delta, r^{n-1}\delta\right], \quad &&\text{if } t \in [S_n, R_n], \\ &u(t) \in [-r^n\delta, r^n\delta], &&\text{if } t \in [R_n, S_{n+1}]. \end{aligned}$$

Since $r \in (0, 1)$ then $u(t) \to 0$. □

The contraction by r argument is illustrated in Figure 3–6. The region in the (μ, σ) plane where $r(0) \in (0, 1)$ and we get asymptotic stability from Theorem 3.2.6 is shown in Figure 3–7.

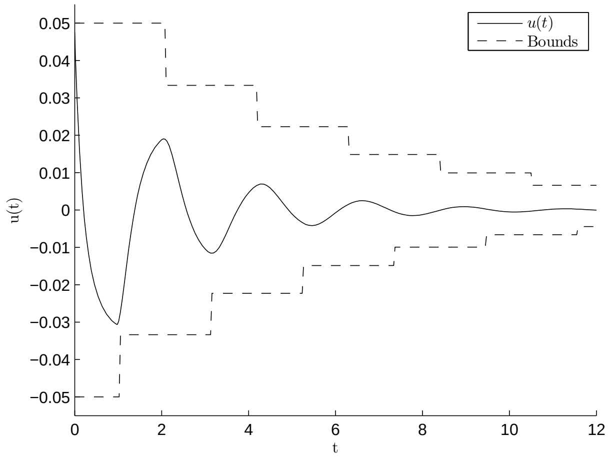


Figure 3–6: Illustration of the contraction of the bounds on the solution by $r \in (0, 1)$.

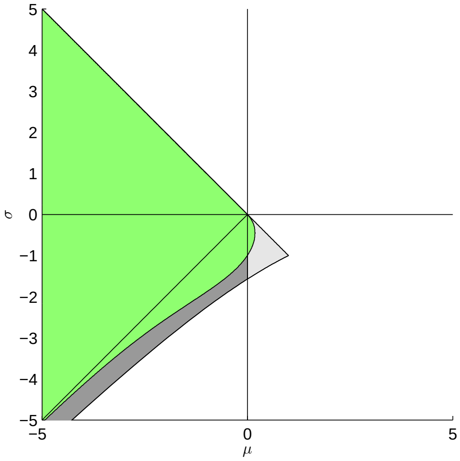


Figure 3–7: The set $\overset{\Delta}{\Sigma} \cup \{r(0) \in (0,1)\}$ is shaded green and plotted with $\varepsilon = a = c = 1$. The zero solution of the model DDE (3.1.1) is asymptotically stable in this region.

3.3 The Lyapunov-Krasovskii and Lyapunov-Razumikhin theorems for RFDEs

In this section we review the standard approaches to prove the stability of steady-state solutions to RFDEs. These results are then extended in the next section to prove similar results for our model state dependent DDE. Consider the RFDE

$$\begin{aligned} \dot{u}(t) &= F\left(t, u_t\right), \\ u_{t_0} &= \phi, \end{aligned} \tag{3.3.1}$$

where $F : \mathbb{R} \times C \to \mathbb{R}^n$ is completely continuous with $F(t, \mathbf{0}) = \mathbf{0}$. The value of the solution to (3.3.1) at time t is written as $u\left(t_0, \phi\right)(t)$ to show the dependence on the initial time and function ϕ, or simply as $u\left(t\right)$. The following definitions of different types of RFDE stability come from the text by Hale and Verduyn Lunel [29].

Definition 3.3.1 (Types of RFDE stability)**.** The zero solution to (3.3.1) is said to be *Lyapunov stable* if $\forall t_0 \in \mathbb{R}, \delta_1 > 0$, there is a $\delta_2 = \delta_2\left(t_0, \delta_1\right)$ such that if $\phi \in \mathcal{B}\left(0, \delta_2\right)$ then $u_t\left(t_0, \phi\right) \in \mathcal{B}\left(0, \delta_1\right)$ for all $t \geqslant t_0$. The zero solution is said to be *uniformly Lyapunov stable* if the δ_2 in the stability definition does not depend on t_0. We note that Hale and Verduyn Lunel simply call these two concepts stable and uniformly stable.

The zero solution to (3.3.1) is said to be *asymptotically stable* if it is Lyapunov stable and there exists $b_0 = b_0(t_0) > 0$ such that if $\phi \in \mathcal{B}\left(0, b_0\right)$ then $\lim\limits_{t\to\infty} u\left(t_0, \phi\right)(t) = 0$. The zero solution is *uniformly asymptotically stable* if it is uniformly Lyapunov stable and there exists $b_0 > 0$ such that for every $\eta > 0$, there is a $t_1(\eta)$ such that $\phi \in \mathcal{B}(0, b_0)$ then $u_t(t_0, \phi) \in \mathcal{B}(0, \eta)$ for $t \geqslant t_0 + t_1(\eta)$.

One of the standard methods to prove the stability of a stationary solution of an RFDE is using Lyapunov functionals (functions of the form $V(t, u_t)$). A Lyapunov functional $V : \mathbb{R} \times C \to \mathbb{R}$ is the extension of Lyapunov functions used for ODEs. Define

$$\dot{V}(t, \phi) = \limsup_{h \to 0^+} \frac{1}{h} \left[V(t + h, u_{t+h}(\phi)) - V(t, \phi) \right].$$

Theorem 3.3.2 (Stability using Lyapunov functionals (Hale and Verduyn Lunel [29]))**.** *Let $\omega_1, \omega_2, \omega_3 : \mathbb{R}^+ \to \mathbb{R}^+$ be continuous, nondecreasing functions. Let $\omega_1(s)$ and $\omega_2(s)$ be positive definite functions on $\mathbb{R}^+$. If there exists a continuous function $V : \mathbb{R} \times C \to \mathbb{R}$ such that*

$$\omega_1(|\phi(0)|) \leqslant V(t, \phi) \leqslant \omega_2(|\phi|),$$

$$\dot{V}(t, \phi) \leqslant -\omega_3(|\phi(0)|),$$

then the zero solution to (3.3.1) *is uniformly Lyapunov stable. If $\omega_1(s) \to \infty$ as $s \to \infty$ then the solutions to* (3.3.1) *are uniformly bounded (for every $\alpha > 0$ there is a constant $\beta = \beta(\alpha)$ such that if $|\phi| < \alpha$ then $|u_t(t_0, \phi)| < \beta$). If $\omega_3(s)$ is positive-definite then the zero solution is uniformly asymptotically stable.* □

For example, apply this to the constant delay DDE (3.1.1) with $c = 0$,

$$\varepsilon \dot{u}(t) = \mu u(t) + \sigma u(t - a).$$

Let the Lyapunov functional be $V(\phi) = \frac{1}{2}\phi(0)^2 + K \int_{-a}^{0} \phi^2(\theta)\, d\theta$ where $K > 0$. Then for any $\phi \in C$,

$$\frac{1}{2}\phi(0)^2 \leqslant V(\phi) \leqslant \left(\frac{1}{2} + Ka\right)|\phi|^2.$$

Applied to the DDE this functional can be written as $V(u_t) = \frac{1}{2}u(t)^2 + K \int_{s=t-a}^{t} u^2(s)\, ds$, so the derivative with respect to t is given by

$$\begin{aligned} \dot{V}(t, u_t) &= u(t)\, \dot{u}(t) + K\left(u^2(t) - u^2(t - a)\right), \\ &= \left(\frac{\mu}{\varepsilon} + K\right) u^2(t) + \frac{\sigma}{\varepsilon} u(t) u(t - a) - K u^2(t - a), \end{aligned}$$

$$\begin{aligned}\dot{V}\left(t,\phi\right) &= \left(\frac{\mu}{\varepsilon}+K\right)\phi^2\left(0\right)+\frac{\sigma}{\varepsilon}\phi(0)\phi\left(-a\right)-K\phi^2\left(-a\right),\\ &= -\begin{bmatrix}\phi(0) & \phi(-a)\end{bmatrix}\begin{bmatrix}-\frac{\mu}{\varepsilon}-K & -\frac{\sigma}{2\varepsilon}\\ -\frac{\sigma}{2\varepsilon} & K\end{bmatrix}\begin{bmatrix}\phi(0)\\ \phi(-a)\end{bmatrix}.\end{aligned}$$

This expression is always negative if the matrix is positive definite. From Sylvester's criterion, this is true if and only if both of the following conditions hold

$$-\frac{\mu}{\varepsilon}-K>0,\quad \left(-\frac{\mu}{\varepsilon}-K\right)K-\frac{\sigma^2}{4\varepsilon^2}>0.$$

Let $\varepsilon>0$ and $\mu<0$ be fixed. Then this condition is equivalent to

$$|\sigma|<2\varepsilon\sqrt{-K\left(\frac{\mu}{\varepsilon}+K\right)}.$$

The choice of $K\in\left[0,-\frac{\mu}{\varepsilon}\right]$ that yields the largest possible region for σ is $K=-\frac{\mu}{2\varepsilon}$. This yields Lyapunov stability of the zero solution if $|\sigma|\leqslant-\mu$ and asymptotic stability if $|\sigma|<-\mu$. Thus the zero solution to (3.1.1) with $c=0$ and $(\mu,\sigma)\in\overset{\Delta}{\Sigma}$ is Lyapunov stable and we have asymptotic stability in the interior of $\overset{\Delta}{\Sigma}$.

Razumikhin [51] showed that it is possible to use Lyapunov functions (functions of the form $V\left(u\left(t\right)\right)$) instead of the more complicated Lyapunov functionals. The stability theorems based on such functions are often called Lyapunov-Razumikhin theorems, or theorems of the Razumikhin-type. For simplicity we consider these theorems for autonomous RFDEs

$$\begin{aligned}\dot{u}(t) &= F\left(u_t\right),\\ u_{t_0} &= \phi.\end{aligned}\tag{3.3.2}$$

Without loss of generality set $t_0=0$. The value of the solution to (3.3.2) at time t is written as $u\left(t_0,\phi\right)(t)$ to show the dependence on the initial function ϕ, or simply as $u\left(t\right)$. Here we follow the proof of a Lyapunov-Razumikhin theorem for RFDEs by Barnea [6]. We define another type of stability used by Barnea which will be useful in the derivation.

Definition 3.3.3 (δ_0-stability)**.** Let $\delta_0\geqslant0$. The zero solution to (3.3.2) is said to be δ_0-stable if for every $\delta_1\geqslant\delta_0$ there exists a $\delta_2=\delta_2\left(\delta_1\right)>0$ such that if $\phi\in\mathcal{B}\left(\mathbf{0},\delta_2\right)$ then $u_t\left(t_0,\phi\right)\in\mathcal{B}\left(\mathbf{0},\delta_1\right)$ for all $t\geqslant0$. It follows that if $\delta_0=0$ then the zero solution is Lyapunov stable.

Definition 3.3.4. Let $V : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function satisfying

$$\omega_1 (|u|) \leqslant V(u) \leqslant \omega_2 (|u|), \quad \forall u \in \mathbb{R}^d \tag{3.3.3}$$

where ω_1, ω_2 are increasing functions in $C(\mathbb{R}^+, \mathbb{R}^+)$ such that $\omega_1(0) = \omega_2(0) = 0$ and $\omega_1(s)$, $\omega_2(s) \to \infty$ as $s \to \infty$. For $\phi \in C$, an integer $k \geqslant 0$, define

$$\bar{V} (\phi) = \sup_{s \in [-r, kr]} V (u (\phi) (s)), \tag{3.3.4}$$

$$\bar{V}' (\phi) = \lim_{h \to 0^+} \frac{1}{h} \Big(V (u_h (\phi)) - V (\phi) \Big). \tag{3.3.5}$$

Lemma 3.3.5. *For $\phi \in C$, if $\sup_{s \in [-r, kr]} |u (\phi) (s)| = \delta < \infty$ then $\omega_1 (\delta) \leqslant \bar{V} (\phi) \leqslant \omega_2 (\delta)$.*

Proof. This easily follows from the definition of $\bar{V}$, (3.3.3) and the fact that ω_1 and ω_2 are continuous, increasing functions. □

Lemma 3.3.6. *Recall the norms $\|\cdot\|$ and $|\cdot|$ defined in page 3. Let F satisfy the following Lipschitz property*

$$|F (\phi_1) - F (\phi_2)| < L \|\phi_1 - \phi_2\| \tag{3.3.6}$$

for all ϕ_1 and $\phi_2 \in C$. Then for all $t \geqslant 0$ the solution of (3.3.2) *satisfies*

$$|u (\phi_1) (t) - u (\phi_2) (t)| \leqslant \|\phi_1 - \phi_2\| e^{Lt}. \tag{3.3.7}$$

Proof. This proof is sketched in Halanay [27] who credits Krasovskii as his source. I reproduce the proof here in more detail. The inequality is obviously satisfied at $t = 0$. Assume there exists a time $t_1 \geqslant 0$ such that $|u (\phi_1) (t_1) - u (\phi_2) (t_1)| = \|\phi_1 - \phi_2\| e^{Lt_1}$ and that for an interval past this time the inequality becomes false. Then there exists $h > 0$ such that for all $t \in (t_1, t_1 + h)$,

$$|u (\phi_1) (t) - u (\phi_2) (t)| > \|\phi_1 - \phi_2\| e^{Lt}.$$

Then,

$$\frac{1}{t - t_1} \left[|u (\phi_1) (t) - u (\phi_2) (t)| - |u (\phi_1) (t_1) - u (\phi_2) (t_1)| \right] > \frac{e^{Lt} - e^{Lt_1}}{t - t_1} \|\phi_1 - \phi_2\|.$$

Hence,

$$\begin{aligned}
&\limsup_{t\to t_1^+}\frac{1}{t-t_1}\left[\left|u\left(\phi_1\right)(t)-u\left(\phi_2\right)(t)\right|-\left|u\left(\phi_1\right)(t_1)-u\left(\phi_2\right)(t_1)\right|\right],\\
&\qquad\geqslant \limsup_{t\to t_1^+}\frac{e^{Lt}-e^{Lt_1}}{t-t_1}\left\|\phi_1-\phi_2\right\|,\\
&\qquad= Le^{Lt_1}\left\|\phi_1-\phi_2\right\| = L\left|u\left(\phi_1\right)(t_1)-u\left(\phi_2\right)(t_1)\right|. \qquad (3.3.8)
\end{aligned}$$

However, using the reverse triangle inequality,

$$\begin{aligned}
&\limsup_{t\to t_1^+}\frac{1}{t-t_1}\left[\left|u\left(\phi_1\right)(t)-u\left(\phi_2\right)(t)\right|-\left|u\left(\phi_1\right)(t_1)-u\left(\phi_2\right)(t_1)\right|\right]\\
&\qquad\leqslant \limsup_{t\to t_1^+}\frac{1}{t-t_1}\Big|u\left(\phi_1\right)(t)-u\left(\phi_1\right)(t_1)-u\left(\phi_2\right)(t)+u\left(\phi_2\right)(t_1)\Big|,\\
&\qquad= \left|\lim_{t\to t_1^+}\frac{u\left(\phi_1\right)(t)-u\left(\phi_1\right)(t_1)}{t-t_1}-\lim_{t\to t_1^+}\frac{u\left(\phi_2\right)(t)-u\left(\phi_2\right)(t_1)}{t-t_1}\right|,
\end{aligned}$$

This is true because the limits exist. Thus,

$$\begin{aligned}
&\limsup_{t\to t_1^+}\frac{1}{t-t_1}\left[\left|u\left(\phi_1\right)(t)-u\left(\phi_2\right)(t)\right|-\left|u\left(\phi_1\right)(t_1)-u\left(\phi_2\right)(t_1)\right|\right]\\
&\qquad= \left|\frac{d}{dt}u\left(\phi_1\right)(t_1)-\frac{d}{dt}u\left(\phi_2\right)(t_1)\right| = \left|F\left(u_{t_1}(\phi_1)\right)-F\left(u_{t_1}(\phi_2)\right)\right|,\\
&\qquad< L\sup_{\theta\in[-r,0]}\left|u\left(t_1+\theta\right)\left(\phi_1\right)-u\left(t_1+\theta\right)\left(\phi_2\right)\right| = L\left|u\left(\phi_1\right)(t_1)-u\left(\phi_2\right)(t_1)\right|. \qquad (3.3.9)
\end{aligned}$$

The second to the last step comes from (3.3.6) and the last step is because (3.3.7) holds before time t_1. Since (3.3.9) contradicts (3.3.8), this proves the lemma. □

Lemma 3.3.7. *Let F satisfy (3.3.6) for some $L > 0$ and let $\tilde{\delta} > 0$. Suppose that $\bar{V}'(\phi) \leqslant 0$ for every $\phi \in C$ such that $\sup_{s\in[-r,kr]}|u(\phi)(s)| \geqslant \tilde{\delta}$. Then there is a $\delta_0 > 0$ such that for every $\delta_1 \geqslant \delta_0$ there is a $\delta_2 > 0$ such that $\phi \in \mathcal{B}(\mathbf{0}, \delta_2)$ implies*

I. The solution $u(\phi)(t)$ of (3.3.2) exists for all $t \geqslant 0$

II. $u_t(\phi) \in \mathcal{B}(\mathbf{0}, \delta_1)$ (which means that the zero solution of (3.3.2) is δ_0-stable).

If $\tilde{\delta} = 0$ then $\delta_0 = 0$ and (3.3.2) is Lyapunov stable.

Proof. Because of the properties of ω_1 and ω_2 we can always find a $\delta_0 > \tilde{\delta}$ such that $\omega_1(\delta_0) = \omega_2(\tilde{\delta})$. Pick $\phi \in C$ such that $\sup_{s\in[-r,kr]}|u(\phi)(s)| \leqslant \tilde{\delta}$. Such a ϕ exists because of Lemma 3.3.6.

We will first show that the solution exists and $|u(\phi)(t)| \leqslant \delta_0$ for all $t \geqslant 0$. Assume that this is false. Then there is a t' such that $u(\phi)(t)$ exists for $t \in [-r, t']$ and $|u(\phi)(t')| > \delta_0$. Then $t' > kr$ because $\delta_0 \geqslant \tilde{\delta}$. Using the left inequality in Lemma 3.3.5 and the fact that ω_1 is increasing,

$$\omega_1(\delta_0) \leqslant \omega_1\left(\sup_{s\in[-r,kr]} \left|u(\phi)\left(s + t' - kr\right)\right|\right) \leqslant \bar{V}\left(u_{t'-kr}(\phi)\right)$$

Now set $t^* = \max\{t : t \in [0, t' - kr], |u(\phi)(t)| \leqslant \tilde{\delta}\}$. This exists and $t^* < t'$ by continuity of u. Using the left inequality in Lemma 3.3.5,

$$\bar{V}\left(u_{t^*-kr}\right) \leqslant \omega_2\bigl(\tilde{\delta}\bigr) = \omega_1(\delta_0).$$

Then $\bar{V}\left(u_{t^*-kr}\right) \leqslant \bar{V}\left(u_{t'-kr}\right)$. By our assumption we must have $\bar{V}'(u_t) \leqslant 0$ for $t \in [t^* - kr, t - kr]$ which is a contradiction. Thus, $|u(\phi)(t)| \leqslant \delta_0$ for all $t \geqslant 0$. From Theorem 1.1.2, such boundedness implies that the solution can be continued for all $t \geqslant 0$.

To prove II, we use Lemma 3.3.6 and $F(\mathbf{0}) = \mathbf{0}$. This yields

$$|u(\phi)(t)| \leqslant \|\phi\| e^{Lkr}$$

for $t \in [-r, kr]$. Let $\delta_2 = \tilde{\delta}e^{-Lkr}$. If $\|\phi\| < \delta_2$ then $\sup_{s\in[-r,kr]} |u(\phi)(s)| \leqslant \tilde{\delta}$ and as proven in I, $|u(\phi)(t)| \leqslant \delta_0$ for all $t \geqslant 0$. Finally, if $\tilde{\delta} = 0$ then $\delta_0 = 0$ (because $\omega_1(0) = \omega_2(0) = 0$) and we get stability. □

Theorem 3.3.8. *Let F satisfy* (3.3.6) *for some $L > 0$ and let $k \geqslant 0$. Define V and $\bar{V}$ as in Definition 3.3.4. Let $M \geqslant 0$ and define the following set*

$$\Phi(M) = \left\{\phi \in C : \bar{V}(\phi) = V(u(\phi)(kr)) \geqslant M, \left.\frac{d}{ds}V(\phi)(s)\right|_{s=kr} > 0\right\}$$

If $\Phi(M)$ is empty for some $M > 0$ then there is a $\delta_0 > 0$ such that the zero solution to (3.3.2) *is δ_0-stable. If $M = 0$ then the zero solution is Lyapunov stable.*

Proof. Let $M \geqslant 0$ and let $\tilde{\delta} = \omega_1^{-1}(M) \geqslant 0$. Take any $\phi \in C$ and if $\sup_{s\in[-r,kr]} |u(\phi)(s)| \geqslant \tilde{\delta}$ then by Lemma 3.3.5,

$$M = \omega_1\bigl(\tilde{\delta}\bigr) \leqslant \bar{V}(\phi)$$

Suppose $\Phi(M)$ is empty. If we can show that $\sup_{s\in[-r,kr]} |u(\phi)(s)| \geqslant \tilde{\delta}$ implies $\bar{V}'(\phi) \leqslant 0$ then the proof is complete using Lemma 3.3.7. So assume $\sup_{s\in[-r,kr]} |u(\phi)(s)| \geqslant \tilde{\delta}$. Three cases are considered:

(i) The maximum value of V occurs in the interior of the interval $(-r, kr)$. If this holds then we can find $\theta_0 \in (-r, kr)$ such that $V(u(\theta_0)) = \bar{V}(\phi)$. Then for sufficiently small h, $\bar{V}(u_h) = \bar{V}(\phi)$ and this means $\bar{V}' = 0$.

(ii) The maximum value of V occurs at the endpoint $-r$. If this case holds then for sufficiently small h we have $\bar{V}(u_h) < \bar{V}(\phi)$ and this means $\bar{V}' < 0$.

(iii) The maximum value of V occurs at the endpoint kr. Then $\bar{V}(\phi) = V(u(\phi)(kr))$. Since $\Phi(M)$ is empty then $\bar{V}'(\phi) = \frac{d}{ds} V'(u(\phi)(s))\big|_{s=kr} \leqslant 0$.

Thus $\bar{V}'(\phi) \leqslant 0$ for all cases and the equation must be δ_0-stable from Lemma 3.3.7. Finally, note that if $M = 0$ then $\tilde{\delta} = 0$ and so also $\delta_0 = 0$. This proves the part about stability. □

Other versions of this theorem are also discussed in Hale and Verduyn Lunel [29], Ivanov, Liz and Trofimchuk [32], Krisztin [35] and many other papers on the topic.

Let us now return to our previous example, (3.1.1) with $c = 0$. Using a Lyapunov function $V(u) = \frac{1}{2}u^2$ instead of a functional,

$$V(u(t)) = \frac{1}{2}u(t)^2,$$

$$\dot{V}(u(t)) = u(t)\,\dot{u}(t) = u(t)\left(\frac{\mu}{\varepsilon}u(t) + \frac{\sigma}{\varepsilon}u(t-a)\right) = \frac{\mu}{\varepsilon}u^2(t) + \frac{\sigma}{\varepsilon}u(t)u(t-a).$$

Let $\varepsilon > 0$ and $\mu < 0$ be fixed. Suppose $|u(t+\theta)| \leqslant |u(t)|$ for $\theta \in [-a, 0]$. Then

$$\dot{V}(u(t)) = \frac{\mu}{\varepsilon}u^2(t) + \frac{\sigma}{\varepsilon}u(t)u(t-a) \leqslant \frac{\mu}{\varepsilon}u^2(t) + \left|\frac{\sigma}{\varepsilon}\right| u^2(t).$$

Thus $\dot{V}(u(t)) \leqslant 0$ if the parameters satisfy $|\sigma| \leqslant -\mu$. From Theorem 3.3.8, this proves that if $(\mu, \sigma) \in \overset{\Delta}{\Sigma}$ then the zero solution of (3.1.1) is Lyapunov stable for the constant delay case. A stronger version of Theorem 3.3.8 given in [29] shows asymptotic stability when $|\sigma| < -\mu$ but we move on now to applying this theorem to the case $\mu = 0$.

3.3.1 Stability of $\dot{u}(t) = \sigma u(t-a)$ using $k = 2$, Barnea [6]

Set $\mu = c = 0$ in (3.1.1).

$$\begin{aligned} \varepsilon\dot{u}(t) &= \sigma u\left(t-a\right), \quad && t \geqslant 0, \\ u\left(t\right) &= \phi\left(t\right), && t \leqslant 0. \end{aligned} \tag{3.3.10}$$

In this equation the bound on the delay is $r = a$. Using Theorem 3.3.8 with $k = 2$ it is possible to show that if $-\frac{3\varepsilon}{2a} < \sigma \leqslant 0$ the zero solution of (3.3.10) is stable. The following proof is based on the proof in Barnea [6]. Let $k = 2$ and $V = \frac{u^2}{2}$. We need to find the values of σ for which $\Phi\left(0\right)$ is empty. Define

$$B\left(\delta\right) = \left\{\phi \in C : \sup_{s\in[-r,kr]} \left|u\left(\phi\right)\left(s\right)\right| \leqslant \delta\right\},$$

$$\psi\left(\delta\right) = \left\{\phi : \phi \in \mathcal{B}\left(\delta\right), \left|u\left(\phi\right)\left(kr\right)\right| = \delta, \left.\frac{d}{ds}V\left(u\left(\phi\right)\left(s\right)\right)\right|_{s=kr} > 0\right\}.$$

Then $\Phi\left(0\right) = \cup_{\delta>0}\psi\left(\delta\right)$. Thus $\Phi\left(0\right)$ is empty if $\psi\left(\delta\right)$ is empty for all $\delta > 0$. But since this equation is linear then the values of σ for which $\psi\left(\delta\right)$ is empty are independent of δ. So to find a parameter region where the zero solution is stable we only need to find the region where $\psi\left(\delta\right)$ is empty for any $\delta > 0$. It is convenient to choose $\delta = 1$. However we will keep the δ term arbitrary so that we can compare the results here with those in the next section where we apply the method to state dependent delays. Since letting $v\left(t\right) = -u\left(t\right)$ yields the same equation $\varepsilon\dot{v}\left(t\right) = \sigma v\left(t-a\right)$, and V is symmetric then we can write $\psi\left(\delta\right)$ as

$$\psi\left(\delta\right) = \left\{\phi : \phi \in \mathcal{B}\left(\delta\right), u\left(\phi\right)\left(kr\right) = \delta, \left.\frac{d}{ds}V\left(u\left(\phi\right)\left(s\right)\right)\right|_{s=kr} > 0\right\}.$$

Let $-\frac{3\varepsilon}{2a} < \sigma \leqslant 0$ and $\delta > 0$. Assume that $\psi\left(\delta\right)$ is not empty. Then there must exist a $\phi \in C$ such that for some $\delta > 0$, $u(s) = u\left(\phi\right)\left(s\right)$ satisfies

(A) $|u(s)| \leqslant \delta, \quad \forall s \in [-a, 2a]$,

(B) $u(2a) = \delta$,

(C) $\dot{V} = \dot{u}\left(2a\right)u\left(2a\right) = \frac{\sigma}{\varepsilon}u(2a)u(a) > 0$.

By integrating (3.3.10) from $t = a$ to $2a$ we obtain

$$u\left(2a\right) = u\left(a\right) + \frac{\sigma}{\varepsilon}\int_0^a u(s)ds. \tag{3.3.11}$$

To prove that $\psi(\delta)$ is empty we will obtain a contradiction to (B) by showing that $u(2a) < \delta$. This is done by applying restrictions (A) and (C) on $u(s)$, $s \in [0,a]$. From (A) and (3.3.10), two of the restrictions are $|u(s)| \leqslant \delta$ and $|\dot{u}(s)| \leqslant \frac{|\sigma|}{\varepsilon}\delta$ for $s \in [0,a]$. From (C), we have $u(a) < 0$. Then

$$u(2a) \leqslant \sup_{\phi} \left[u(\phi)(2a) : \phi \in C, \sup_{s\in[0,a]} |u(\phi)(s)| \leqslant \delta, u(\phi)(a) < 0, \sup_{s\in[0,a]} |\dot{u}(\phi)(s)| \leqslant \frac{|\sigma|}{\varepsilon}\delta \right],$$

where $u(\phi)(2a)$ is from (3.3.11). Let $\eta = u_a(\phi)$ then

$$u(2a) \leqslant \sup_{\eta} \left[u(\eta)(a) : \eta \in C, \|\eta\| \leqslant \delta, \eta(0) < 0, \sup_{s\in[-a,0]} \left|\eta'(s)\right| \leqslant \frac{|\sigma|}{\varepsilon}\delta \right].$$

Let $P(\hat{u})$ be the value of the right hand side when we fix $\eta(0) = \hat{u}$. From the constraints, $\hat{u} \in [-\delta, 0]$. Then

$$P(\hat{u}) \equiv \sup_{\eta} \left[u(\eta)(a) : \eta \in C, \|\eta\| \leqslant \delta, \eta(0) = \hat{u}, \sup_{s\in[-a,0]} \left|\eta'(s)\right| \leqslant \frac{|\sigma|}{\varepsilon}\delta \right].$$

Since $\sigma < 0$, it is easy to see from (3.3.11) that the η that maximizes $P(\hat{u})$ is a function that stays at its most negative possible value of $-\delta$ for as long as possible and then increases linearly to $\hat{u}$. This function is

$$\tilde{\eta}(\theta) = \begin{cases} -\delta, & \theta \in \left[-a, \frac{\delta+\hat{u}}{\sigma\delta}\varepsilon\right], \\ \hat{u} - \frac{\sigma}{\varepsilon}\delta\theta, & \theta \in \left[\frac{\delta+\hat{u}}{\sigma\delta}\varepsilon, 0\right]. \end{cases} \tag{3.3.12}$$

In Barnea [6] this function is immediately integrated piecewise and yields the following expression

$$P(\hat{u}) = \hat{u} + \frac{\sigma}{\varepsilon}\int_{-a}^{0} \tilde{\eta}(\theta)d\theta = \hat{u} - \frac{\sigma}{\varepsilon}\delta a - \frac{1}{2\delta}(\delta + \hat{u})^2. \tag{3.3.13}$$

He found that the maximum of $P(\hat{u})$ occurs at $\hat{u} = 0$ which means

$$u(2a) \leqslant P(0) = -\delta\left(\frac{\sigma}{\varepsilon}a + \frac{1}{2}\right).$$

So if $-\frac{3\varepsilon}{2a} < \sigma \leqslant 0$ then $P(0) < \delta$. This is true for all $\delta > 0$ so $\Phi(0)$ is empty and the $u = 0$ solution of (3.3.10) is Lyapunov stable using Theorem 3.3.8.

Remark 3.3.9. The region found in [6] is correct but the derivation is not completely correct. The integration (3.3.13) was performed assuming that the point $\frac{\delta+\hat{u}}{\sigma\delta}\varepsilon \geqslant -a$ for all $\hat{u} \in [-\delta, 0]$. This assumption does not always hold. We have $\frac{\delta+\hat{u}}{\sigma\delta}\varepsilon < -a$ when $\hat{u} > -\delta - \frac{\sigma}{\varepsilon}a\delta$. So we have to take this case into account when

$$-\delta - \frac{\sigma}{\varepsilon}a\delta < 0 \quad \Rightarrow \quad \sigma > -\frac{\varepsilon}{a}$$

Instead of (3.3.12) we should use

$$\tilde{\eta}\left(\theta\right) = \left(\text{the restriction to } [-a,0] \text{ of } \bar{\eta}(\theta) = \begin{cases} -\delta, & \theta \in \left(-\infty, \frac{\delta+\hat{u}}{\sigma\delta}\varepsilon\right] \\ \hat{u} - \frac{\sigma}{\varepsilon}\delta\theta, & \theta \in \left[\frac{\delta+\hat{u}}{\sigma\delta}\varepsilon, 0\right] \end{cases}\right). \tag{3.3.14}$$

Let $\sigma > -\frac{\varepsilon}{a}$ and $\hat{u} > -\delta - \frac{\sigma}{\varepsilon}a\delta$. The integration of $P(\hat{u})$ yields

$$P\left(\hat{u}\right) = \hat{u}\left(1 + \frac{\sigma}{\varepsilon}a\right) + \frac{\sigma^2a^2}{2\varepsilon^2}\delta \leqslant \frac{\sigma^2a^2}{2\varepsilon^2}\delta \leqslant \frac{\delta}{2}, \tag{3.3.15}$$

where the last two inequalities both stem from $-\frac{\varepsilon}{a} \leqslant \sigma \leqslant 0$. Since $P\left(\hat{u}\right) < \delta$ then this case which might have been overlooked in [6] does not affect the stability region that was found.

The next step in this method would be to work with $k > 2$. Using the same ideas, $k = 3$ gives us an additional requirement on the η function: $|\eta''\left(\theta\right)| \leqslant \frac{\sigma^2}{\varepsilon^2}\delta$. Because of this, the new $\tilde{\eta}$ must be split into three parts: a horizontal part, a quadratic part with leading term $\frac{\sigma^2}{2\varepsilon^2}$ and then a linear with slope $\left|\frac{\sigma}{\varepsilon}\right|$. Performing the integration considering all possible cases confirms that the larger region $-\frac{37\varepsilon}{24a} < \sigma \leqslant 0$ is in the stability region. This integration is shown in Section 3.4.2. Using the characteristic equation of the DDE we know that the entire stability region of this DDE is actually $-\frac{\pi\varepsilon}{2a} < \sigma \leqslant 0$. Krisztin [35] is able to find this entire stability region for the constant delay equation using similar Razumikhin techniques with $k \to \infty$.

3.4 Stability of the model problem using a Razumikhin-style theorem

In this section we extend the ideas to the state dependent problem (3.1.1) with $\mu \neq 0$ and $c \neq 0$. If this equation is written as an autonomous RFDE (3.3.2), then $F\left(\phi\right) = \frac{\mu}{\varepsilon}\phi(0) + \frac{\sigma}{\varepsilon}\phi\left(-a - c\phi(0)\right)$. Aside from having no *a priori* bound on the delay term, it is easy to show that this F is not Lipschitz continuous in the space of continuous functions. Thus, Theorem 3.3.8 cannot be applied to our model DDE. Instead, a more direct proof of stability will be used for this problem. This approach will extend more naturally to a proof of stability of numerical methods applied to (3.1.1).

As in Section 3.3, the solution of (3.1.1) at time t will be denoted by $u\left(\varphi\right)\left(t\right)$ when we would like to be specific about the initial history function, and simply $u\left(t\right)$ otherwise. The focus of this section will again be in $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$ so set $\sigma \leqslant \mu$ and $\sigma < -\mu$ (again $\sigma < 0$). From

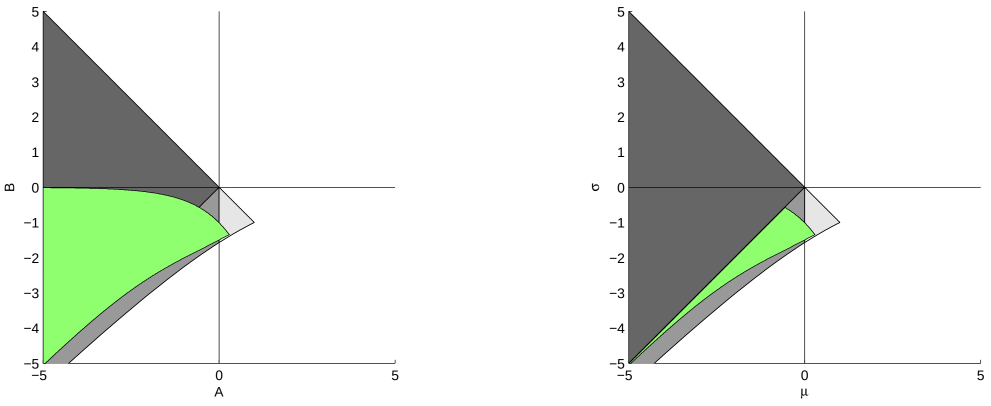
Lemma 2.1.2, if $u(0) > -\frac{a}{c}$ then $\tau(t, u(t)) = a + cu(t) \leqslant t$ so the delay cannot become an advance.

3.4.1 Results using $k = 2$

Barnea [6] claims that using the same techniques as in Section 3.3.1 for the case $c = 0$, the region $\mathcal{X}_2 = \{(\mu, \sigma) : 0 \leqslant s^* \leqslant a, P < 1\}$ (shown in Figure 3–8) where

$$s^* = -\frac{\varepsilon e^{\frac{\mu}{\varepsilon}a}}{\sigma}, \quad P = \frac{\sigma(\mu+\sigma)}{\mu^2}\left[e^{\frac{\mu}{\varepsilon}s^*} - \frac{\sigma}{\mu+\sigma}\right]$$

can be shown to be part of the stability region of (3.1.1). The shape of the region $\mathcal{X}_2$ appears a little strange and we note that it does not include the interval $-\frac{3\varepsilon}{2a} < \sigma \leqslant 0$ on the σ-axis which was proven to be stable in Section 3.3.1. Barnea did not graph $\mathcal{X}_2$ and did not present a derivation of how he obtained this expression, but he noted that setting $P = 1$ and letting $\mu \to 0$ yields that the point $\sigma = -\frac{3\varepsilon}{2a}$ is a boundary of $\mathcal{X}_2$ on the σ-axis. We observe that setting $s^* = 0$ and $\mu \to 0$ yields $\sigma = -\frac{\varepsilon}{a}$ is the other boundary on the σ-axis. Recall from the remark in page 54 that Barnea might have overlooked a case when $-\frac{\varepsilon}{a} < \sigma \leqslant 0$, precisely the region in the σ-axis that is now missing in Figure 3–8. It is likely be that when he extended his results to $\mu \neq 0$ he also omitted this case.



(a) $\mathcal{X}_2$

(b) $\mathcal{X}_2 - \overset{\Delta}{\Sigma}$

Figure 3–8: The set $\mathcal{X}_2 = \{(\mu, \sigma) : 0 \leqslant s^* \leqslant a, P < 1\}$ are shaded green. This is part of the stability region of (3.1.1) with $\varepsilon = a = 1$ and $c = 0$ according to Barnea [6] using the Razumkhin technique with $k = 2$.

In this section we derive new points in the stability region of (3.1.1) for arbitrary fixed ε, a and c with ε and $a > 0$. The new points for the case $\varepsilon = 1$, $a = 1$ and $c = 0$ are shown in

Figure 3–10. This region is noticeably different from the region in Figure 3–8(b). In particular, our new region contains the entire interval $-\frac{3\varepsilon}{2a} < \sigma \leqslant 0$ on the σ-axis and the points in $\mathcal{X}_2$ except for a small piece in the region $\mu > 0$. In Section 3.4.3 we improve upon the results of this section and derive a new region shown in Figure 3–13(d) which also contains the entire interval $-\frac{3\varepsilon}{2a} < \sigma \leqslant 0$ on the σ-axis and all the points in $\mathcal{X}_2 - \overset{\Delta}{\Sigma}$.

Using the same Razumikhin techniques as in the previous section we will show that for a given $\delta > 0$, in a subset of $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$ if the history function $\varphi(t)$ is small enough then the solution to (3.1.1) satisfies $u(t) \in [-\delta, \delta]$ for all $t \geqslant 0$. For the linear case ($c = 0$) this region is independent of δ. For the state dependent case, these regions change with δ as shown in Figure 3–9. As δ goes to zero these regions converge to the region for the constant delay case (Figure 3–10). We begin by showing that given δ, it is possible to find a bound for the history function so that $u(t) \in [-\delta, \delta]$ for any finite amount of time.

Lemma 3.4.1. *Let $L > \frac{|\mu|+|\sigma|}{\varepsilon}$, $T > 0$, $\delta \in \left(0, \left|\frac{a}{c}\right| e^{-LT}\right)$ and $|\varphi(s)| \leqslant \delta$ for $s \leqslant 0$. Then the solution of* (3.1.1) *satisfies $|u(\varphi)(t)| \leqslant \delta e^{Lt}$ for $t \in [0, T]$.*

Proof. Suppose not. Then there exists $t_1 \in [0, T]$ and $h > 0$ such that $|u(t_1)| = \delta e^{Lt_1}$ and $|u(t)| > \delta e^{Lt}$ for $t \in (t_1, t_1 + h)$. Suppose first that $u(t_1) = \delta e^{Lt_1}$ and $u(t) > \delta e^{Lt}$ for $t \in (t_1, t_1 + h)$. Then since the solutions to (3.1.1) are C^1 for $t > 0$ then

$$\lim_{t \to t_1^+} \frac{u(t) - u(t_1)}{t - t_1} \geqslant \lim_{t \to t_1^+} \left(\frac{e^{Lt} - e^{Lt_1}}{t - t_1} \right) \delta = L\delta e^{Lt_1}. \tag{3.4.1}$$

Since $t_1 - a - cu(t_1) \leqslant t_1$ then $|u(t_1 - a - cu(t_1))| \leqslant \delta e^{Lt_1}$. Thus,

$$\left| \lim_{t \to t_1^+} \frac{u(t) - u(t_1)}{t - t_1} \right| = \left| \frac{\mu}{\varepsilon} u(t_1) + \frac{\sigma}{\varepsilon} u(t_1 - a - cu(t_1)) \right| \leqslant \frac{|\mu| + |\sigma|}{\varepsilon} \delta e^{Lt_1} < L\delta e^{Lt_1}.$$

This contradicts (3.4.1). A similar contradiction can be obtained if we let $u(t_1) = -\delta e^{Lt_1}$ and $u(t) < -\delta e^{Lt}$ for $t \in (t_1, t_1 + h)$. Therefore there is no such $t_1 \in [0, T]$ and $h > 0$ such that $|u(t_1)| = \delta e^{Lt_1}$ and $|u(t)| > \delta e^{Lt}$ for $t \in (t_1, t_1 + h)$. □

Let $k = 2$ and $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$. Let $\tau_1 = a + |c|\,\delta$ (the upper bound of the delay term for $u(t) \in [-\delta, \delta]$). Then if $|\varphi(t)| \leqslant \delta_2 = \delta e^{-2L\tau_1}$ for some $L > \frac{|\mu|+|\sigma|}{\varepsilon}$ then from Lemma 3.4.1, $\sup_{s \in [-\tau_1, 2\tau_1]} |u(s)| \leqslant \delta$. Now take any $t \geqslant 2\tau_1$. If we can show that $\dot{u}(t)\,u(t) < 0$ if $|u(t)| = \delta$ then this proves that the solution must remain bounded inside $[-\delta, \delta]$.

First assume that $u(t) = \delta$, $\dot{u}(t) \geqslant 0$ for some $t \geqslant 2\tau_1$ (the case $u(t) = -\delta$, $\dot{u}(t) \leqslant 0$ will be considered later).

$$\frac{d}{dt}\left(e^{-\frac{\mu t}{\varepsilon}}u(t)\right) = e^{-\frac{\mu t}{\varepsilon}}\dot{u}(t) - \frac{\mu}{\varepsilon}e^{-\frac{\mu t}{\varepsilon}}u(t) = \frac{\sigma}{\varepsilon}e^{-\frac{\mu t}{\varepsilon}}u(t-a-cu(t))$$

using (3.1.1). Now integrating

$$u(t) = u(t_0)\,e^{\frac{\mu(t-t_0)}{\varepsilon}} + \frac{\sigma}{\varepsilon}e^{\frac{\mu t}{\varepsilon}}\int\limits_{t_0}^{t} e^{-\frac{\mu s}{\varepsilon}}u(s-a-cu(s))\,ds \tag{3.4.2}$$

Let $\tau_2 = \tau(t, u(t)) = \tau(t, \delta) = a + c\delta$. In general, $\tau_1 \geqslant \tau_2 > 0$ with $\tau_1 = \tau_2$ for $c > 0$. Since $t \geqslant 2\tau_1$ then $t - \tau_2 \geqslant \tau_1$. Suppose the equation has already been integrated from 0 to $t - \tau_2$. Set $t_0 = \alpha(t, u(t)) = t - \tau_2$ and use $\theta = s - t$

$$\begin{aligned} u(t) &= u(t-\tau_2)\,e^{\frac{\mu\tau_2}{\varepsilon}} + \frac{\sigma}{\varepsilon}\int\limits_{-\tau_2}^{0} e^{-\frac{\mu\theta}{\varepsilon}}u(t+\theta-a-cu(t+\theta))\,d\theta, \\ &= \eta(0)\,e^{\frac{\mu\tau_2}{\varepsilon}} + \frac{\sigma}{\varepsilon}\int\limits_{-\tau_2}^{0} e^{-\frac{\mu\theta}{\varepsilon}}\eta(\theta)\,d\theta, \end{aligned} \tag{3.4.3}$$

where $\eta(\theta) = u(t+\theta-a-cu(t+\theta))$ for $\theta \in [-\tau_2, 0]$. Since $t \geqslant 2\tau_1$ then $t + \theta \in [\tau_1, 2\tau_1]$ and $u(t+\theta) \in [-\delta, \delta]$ for all $\theta \in [-\tau_1, 0]$. Then also,

$$t + \theta - a - cu(t+\theta) \in [2\tau_1 - \tau_2 - a - |c|\,\delta, t - a + |c|\,\delta] \subseteq [0, t]\,.$$

So the η function must have properties stemming from the properties of $u(s)$ for $s \in [0, t]$. From the assumption that $\dot{u}(t) \geqslant 0$, one of these properties is $\mu u(t) + \sigma u(t-\tau_2) \geqslant 0$ hence,

$$u(t-\tau_2) \leqslant -\frac{\mu}{\sigma}\delta. \tag{3.4.4}$$

So we require $u(t-\tau_2) \in \left[-\delta, -\frac{\mu}{\sigma}\delta\right] \subseteq [-\delta, \delta]$. From the bounds on the solution we also obtain a bound on the derivative of $u(t)$

$$|\dot{u}(s)| \leqslant \frac{|\mu| + |\sigma|}{\varepsilon}\delta, \quad \text{for } s \in [0, t]. \tag{3.4.5}$$

From (3.4.4)-(3.4.5), we obtain

$$\eta(0) = u(t-\tau_2) \leqslant -\frac{\mu}{\sigma}\delta, \tag{3.4.6}$$

and $\eta'(\theta) = \dot{u}(t + \theta - a - cu(t + \theta))(1 - c\dot{u}(t + \theta))$ which implies

$$\left|\eta'(\theta)\right| \leqslant \frac{|\mu| + |\sigma|}{\varepsilon}\delta\left(1 + \frac{|\mu| + |\sigma|}{\varepsilon}\delta|c|\right) = D\delta, \tag{3.4.7}$$

where $D = \frac{|\mu|+|\sigma|}{\varepsilon}\left(1 + \frac{|\mu|+|\sigma|}{\varepsilon}\delta|c|\right)$. Following Section 3.3.1, given $\hat{u} = \eta(0)$, the η function that would yield the largest possible $u(t)$ in (3.4.3) is one that stays as negative as possible given the restrictions on η. Using such an η in (3.4.3) we get

$$u(t) \leqslant \sup_{\hat{u}\in\left[-\delta, -\frac{\mu}{\sigma}\delta\right]}\left(\hat{u}e^{\frac{\mu\tau_2}{\varepsilon}} + \frac{\sigma}{\varepsilon}\int\limits_{-\tau_2}^{0} e^{-\frac{\mu\theta}{\varepsilon}}\eta(\theta)\,d\theta\right).$$

Definition 3.4.2. Let $\varepsilon, a > 0$, $\sigma \leqslant \mu$ and $\sigma < -\mu$. For any $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$ and $\hat{u} \in \left[-\delta, -\frac{\mu}{\sigma}\delta\right]$, define $\eta_{(2)} \in C((-\infty, 0], [-\delta, \hat{u}])$ to be

$$\eta_{(2)}(\theta) = \begin{cases} -\delta, & \theta \in \left[-\infty, -\frac{\delta+\hat{u}}{D\delta}\right], \\ \hat{u} + D\delta\theta, & \theta \in \left[-\frac{\delta+\hat{u}}{D\delta}, 0\right]. \end{cases} \tag{3.4.8}$$

where $D = \frac{|\mu|+|\sigma|}{\varepsilon}\left(1 + \frac{|\mu|+|\sigma|}{\varepsilon}\delta|c|\right)$. Define $\mathcal{I}(\hat{u}, \delta, c, 2)$ to be

$$\mathcal{I}(\hat{u}, \delta, c, 2) = \hat{u}e^{\frac{\mu\tau_2}{\varepsilon}} + \frac{\sigma}{\varepsilon}\int\limits_{-\tau_2}^{0} e^{-\frac{\mu\theta}{\varepsilon}}\eta_{(2)}(\theta)\,d\theta,$$

The 2 in the argument of this function is from $k = 2$ (we consider general $k \geqslant 2$ in the next section). This function $\mathcal{I}(\hat{u}, \delta, c, 2)$ depends on c through τ_2 and D. If we use $|c|$ instead of c in the expression, the only change is τ_2 becomes τ_1;

$$\mathcal{I}(\hat{u}, \delta, |c|, 2) = \hat{u}e^{\frac{\mu\tau_1}{\varepsilon}} + \frac{\sigma}{\varepsilon}\int\limits_{-\tau_1}^{0} e^{-\frac{\mu\theta}{\varepsilon}}\eta_{(2)}(\theta)\,d\theta.$$

Let the function $P(\delta, c, 2)$ be defined as

$$P(\delta, c, 2) = \sup_{\hat{u}\in\left[-\delta, -\frac{\mu}{\sigma}\delta\right]}\mathcal{I}(\hat{u}, \delta, |c|, 2).$$

In the succeeding lemmas we fix $\varepsilon, a > 0$ and use the set notation $\{\cdot\}$ to denote regions in the (μ, σ) plane. For example, $\{P(\delta, c, 2) < \delta\}$ is the set of all (μ, σ) values for which the relation $P(\delta, c, 2) < \delta$ holds.

Lemma 3.4.3. *Let $\varepsilon, a > 0$, $\sigma \leqslant \mu$ and $\sigma < -\mu$. Let $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$ be fixed. Consider the partial derivative of $\mathcal{I}(\hat{u}, \delta, c, 2)$ with respect to τ_2,*

$$\frac{\partial}{\partial \tau_2}\mathcal{I}(\tau_2) \equiv \frac{\partial}{\partial \tau_2}\left(\hat{u}e^{\frac{\mu\tau_2}{\varepsilon}} + \frac{\sigma}{\varepsilon}\int\limits_{-\tau_2}^{0} e^{-\frac{\mu\theta}{\varepsilon}}\eta_{(2)}(\theta)\, d\theta\right) = \frac{e^{\frac{\mu\tau_2}{\varepsilon}}}{\varepsilon}\left[\mu\hat{u} + \sigma\eta_{(2)}(-\tau_2)\right]. \tag{3.4.9}$$

If $\frac{\partial}{\partial\tau_2}\mathcal{I}(\tau_2) \leqslant 0$ then $\mu > 0$ and $\mathcal{I}(\hat{u}, \delta, c, 2) < \delta$.

Proof. If either $\mu \leqslant 0$ holds or $\mu > 0$ and $\eta_{(2)}(-\tau_2) \leqslant 0$ hold then it is easy to show that $\mu\hat{u} + \sigma\eta_{(2)}(-\tau_2) > 0$. So consider the case when $\mu > 0$ and $\eta_{(2)}(-\tau_2) > 0$. Then $\frac{\delta+\hat{u}}{D\delta} > \tau_2$ and integrating yields

$$\mathcal{I}(\hat{u}, \delta, c, 2) = \hat{u}\left[e^{\frac{\mu}{\varepsilon}\tau_2} + \frac{\sigma}{\mu}\left(e^{\frac{\mu}{\varepsilon}\tau_2} - 1\right)\right] + \frac{\sigma}{\mu}D\delta\left[\frac{\varepsilon}{\mu}\left(e^{\frac{\mu}{\varepsilon}\tau_2} - 1\right) - \tau_2 e^{\frac{\mu}{\varepsilon}\tau_2}\right]$$

A very similar derivation of this expression is shown in (3.4.14). Let $\frac{\partial}{\partial\tau_2}\mathcal{I}(\tau_2) \leqslant 0$ then $\mu\hat{u} + \sigma\eta_{(2)}(-\tau_2) = (\mu + \sigma)\hat{u} + \sigma\tau_2 D\delta \leqslant 0$. Using this we derive

$$\begin{aligned}\mathcal{I}(\hat{u}, \delta, c, 2) &\leqslant \hat{u}\left[e^{\frac{\mu}{\varepsilon}\tau_2} + \frac{\sigma}{\mu}\left(e^{\frac{\mu}{\varepsilon}\tau_2} - 1\right)\right] + \delta\frac{\sigma\varepsilon D}{\mu^2}\left(e^{\frac{\mu}{\varepsilon}\tau_2} - 1\right) - (\mu + \sigma)\hat{u}\frac{e^{\frac{\mu\tau_2}{\varepsilon}}}{\mu} \\ &= -\hat{u}\frac{\sigma}{\mu} + \delta\frac{\sigma\varepsilon D}{\mu^2}\left(e^{\frac{\mu}{\varepsilon}\tau_2} - 1\right) \leqslant \delta + \delta\frac{\sigma\varepsilon D}{\mu^2}\left(e^{\frac{\mu}{\varepsilon}\tau_2} - 1\right) < \delta\end{aligned}$$

where the last two inequalities are because $\hat{u} \leqslant -\frac{\mu}{\sigma}\delta$ and $\sigma < 0$. □

Lemma 3.4.4. *Let $\varepsilon, a > 0$, $\sigma \leqslant \mu$ and $\sigma < -\mu$. Let $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$. Then*

$$\left\{P(\delta, c, 2) = \sup_{\hat{u} \in \left[-\delta, -\frac{\mu}{\sigma}\delta\right]} \mathcal{I}(\hat{u}, \delta, |c|, 2) < \delta\right\} \subseteq \left\{\sup_{\hat{u} \in \left[-\delta, -\frac{\mu}{\sigma}\delta\right]} \mathcal{I}(\hat{u}, \delta, -|c|, 2) < \delta\right\}.$$

Proof. Let $(\mu, \sigma) \in \{P(\delta, c, 2) < \delta\}$ and $\hat{u} \in \left[-\delta, -\frac{\mu}{\sigma}\delta\right] \subseteq [-\delta, \delta]$. Recall that changing the sign of c in $\mathcal{I}(\hat{u}, \delta, c, 2)$ only changes the value of $\tau_2 = a + c\delta$.

(i) If $\mu\hat{u} + \sigma\eta_{(2)}(-(a - |c|\delta)) \leqslant 0$ then by Lemma 3.4.3, $\mathcal{I}(\hat{u}, \delta, |c|, 2) < \delta$.

(ii) If $\mu\hat{u} + \sigma\eta_{(2)}(-(a - |c|\delta)) > 0$ and $\mu\hat{u} + \sigma\eta_{(2)}(-(a + |c|\delta)) \geqslant 0$ then $\mu\hat{u} + \sigma\eta_{(2)}(-\tau)) > 0$ for all $\tau \in (a - |c|\delta, a + |c|\delta)$. By Lemma 3.4.3, $\mathcal{I}(\hat{u}, \delta, -|c|, 2) \leqslant \mathcal{I}(\hat{u}, \delta, |c|, 2) < \delta$.

(iii) If $\mu\hat{u} + \sigma\eta_{(2)}(-(a - |c|\delta)) > 0$ and $\mu\hat{u} + \sigma\eta_{(2)}(-(a + |c|\delta)) < 0$ then there exists $x \in (a - |c|\delta, a + |c|\delta)$ such that $\mu\hat{u} + \sigma\eta_{(2)}(-x)) = 0$ and $\mu\hat{u} + \sigma\eta_{(2)}(-\tau)) > 0$ for $\tau \in$

$(a-|c|\delta, x)$. Then $\frac{\partial}{\partial\tau_2}\mathcal{I}\left(x\right)=0$ and hence by Lemma 3.4.3,

$$\mathcal{I}\left(\hat{u},\delta,-|c|,2\right)<\hat{u}e^{\frac{\mu x}{\varepsilon}}+\frac{\sigma}{\varepsilon}\int\limits_{-x}^{0}e^{-\frac{\mu\theta}{\varepsilon}}\eta_{(2)}\left(\theta\right)d\theta<\delta.$$

Cases (i), (ii) and (iii) all yield $\mathcal{I}\left(\hat{u},\delta,-|c|,2\right)<\delta$. The result easily follows. □

Lemma 3.4.5. *Let $\varepsilon, a>0$, $\sigma\leqslant\mu$ and $\sigma<-\mu$. If $0<\delta_*\leqslant\delta_{**}<\left|\frac{a}{c}\right|$ then*

$$\left\{P\left(\delta_{**},c,2\right)<\delta_{**}\right\}\subseteq\left\{P\left(\delta_{*},c,2\right)<\delta_{*}\right\}.$$

Proof. Increasing δ increases $\tau_1=a+|c|\delta$ and $D=\frac{|\mu|+|\sigma|}{\varepsilon}\left(1+\frac{|\mu|+|\sigma|}{\varepsilon}|c|\delta\right)$, the only nonlinearities in δ in the expression for $\mathcal{I}\left(\hat{u},\delta,|c|,2\right)$.

$$\frac{\partial}{\partial\delta}\left(\frac{\mathcal{I}\left(s\delta,\delta,c,2\right)}{\delta}\right)=\frac{e^{\frac{\mu\tau_1}{\varepsilon}}}{\varepsilon\delta}\left[\mu s\delta+\sigma\eta_{(2)}\left(-\tau_1\right)\right]|c|+\frac{\sigma}{\varepsilon}\int\limits_{-\min\{\tau_1,\frac{1+s}{D}\}}^{0}e^{-\frac{\mu\theta}{\varepsilon}}D\delta\theta d\theta. \qquad (3.4.10)$$

Notice that the second term is always positive. Let $\tau_1^*=a+|c|\delta_*$, $\tau_1^{**}=a+|c|\delta_{**}$ and $(\mu,\sigma)\in\left\{P\left(\delta_{**},c,2\right)<\delta_{**}\right\}$. Let $s\in\left[-1,-\frac{\mu}{\sigma}\right]$ and use the notation $\eta(\delta)\left(\theta\right)$ equal to the expression defined by (3.4.8). Consider the following cases:

(i) If $\mu s\delta_*+\sigma\eta(\delta_*)\left(-\tau_1^*\right)\leqslant 0$ then by Lemma 3.4.3, $\mathcal{I}\left(s\delta_*,\delta_*,|c|,2\right)<\delta_*$.

(ii) If $\mu s\delta_*+\sigma\eta(\delta_*)\left(-\tau_1^*\right)>0$ and $\mu s\delta_{**}+\sigma\eta(\delta_{**})\left(-\tau_1^{**}\right)\geqslant 0$ then $\frac{\partial}{\partial\delta}\frac{\mathcal{I}(s\delta,\delta,c,2)}{\delta}\geqslant 0$ for $\delta\in[\delta_*,\delta_{**}]$. Thus,

$$\frac{\mathcal{I}\left(s\delta_*,\delta_*,c,2\right)}{\delta_*}\leqslant\frac{\mathcal{I}\left(s\delta_{**},\delta_{**},c,2\right)}{\delta_{**}}\leqslant\frac{P\left(\delta_{**},c,2\right)}{\delta_{**}}<1,$$

(iii) If $\mu s\delta_*+\sigma\eta\left(-\tau_1^*\right)>0$ and $\mu s\delta_{**}+\sigma\eta\left(-\tau_1^{**}\right)<0$ then there exists $x\in(\delta_*,\delta_{**}]$ such that $\mu sx+\sigma\eta\left(-(a+|c|x)\right)=0$ and $\mu s\delta+\sigma\eta\left(-(a+|c|\delta)\right)>0$ for $\delta\in(\delta_*,x)$. By (3.4.10) and Lemma 3.4.3,

$$\frac{\mathcal{I}\left(s\delta_*,\delta_*,c,2\right)}{\delta_*}\leqslant\frac{\mathcal{I}\left(sx,x,c,2\right)}{x}<1.$$

Cases (i), (ii) and (iii) all yield $\mathcal{I}\left(s\delta_*,\delta_*,|c|,2\right)<\delta_*$. The result easily follows. □

Lemma 3.4.6. *Let $\varepsilon, a>0$, $\sigma\leqslant\mu$ and $\sigma<-\mu$. Let $\delta\in\left(0,\left|\frac{a}{c}\right|\right)$ and $(\mu,\sigma)\in\left\{P\left(\delta,c,2\right)<\delta\right\}$. Then there exists a $\delta_2\in(0,\delta]$ such that if $\varphi\left(t\right)\in[-\delta_2,\delta_2]$ for $t\in[-a-|c|\delta,0]$ then the solution of (3.1.1) satisfies $u\left(t\right)\in[-\delta,\delta]$ for all $t\geqslant 0$.*

Proof. Let $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$ and $(\mu, \sigma) \in \{P(\delta, c, 2) < \delta\}$. Let $L > \frac{|\mu|+|\sigma|}{\varepsilon}$, $\tau_1 = a + |c|\delta$, $\delta_2 = \delta e^{-2L\tau_1}$ and $|\varphi(t)| \in [-\delta_2, \delta_2]$. Then from Lemma 3.4.1, $\sup_{s\in[-\tau_1, 2\tau_1]} |u(s)| \leqslant \delta$. Now suppose that the solution exits the interval $[-\delta, \delta]$ for the first time at some $t > 2\tau_1$ through the upper bound. Then $u(t) = \delta$ and $\dot{u}(t) > 0$. We have derived that given these assumptions $u(t) \leqslant \sup_{\hat{u}\in\left[-\delta, -\frac{\mu}{\sigma}\delta\right]} \mathcal{I}(\hat{u}, \delta, c, 2) \leqslant P(\delta, c, 2)$. This is true for any sign of c by Lemma 3.4.4. But by our choice of (μ, σ), $P(\delta, c, 2) < \delta$ which contradicts our assumption. Thus, the solution cannot exit the interval $[-\delta, \delta]$ for the first time through the upper bound.

Now we need to show that in the same region we cannot have the solution exit the interval $[-\delta, \delta]$ for the first time through the lower bound either. Let $v(t) = -u(t)$ in (3.1.1). Then

$$\begin{aligned} \varepsilon \dot{v}(t) &= \mu v(t) + \sigma v(t - a + cv(t)), \quad && t \geqslant 0, \\ v(t) &= -\varphi(t), && t \leqslant 0. \end{aligned} \tag{3.4.11}$$

The problem is now to show that a solution of (3.4.11) cannot leave the interval $[-\delta, \delta]$ for the first time at some $t > 2\tau_1$ through the upper bound. But this is the same DDE as (3.1.1) except with c replaced by $-c$. By our discussion above and Lemma 3.4.4, if $(\mu, \sigma) \in \{P(\delta, c, 2) < \delta\}$ and φ is small enough then the solution to (3.4.11) cannot escape $[-\delta, \delta]$ through the upper bound. Hence the solution to (3.1.1) cannot escape $[-\delta, \delta]$ through the lower bound. □

Theorem 3.4.7. *Let $\varepsilon, a > 0$ and $(\mu, \sigma) \in \{P(1, 0, 2) < 1\}$. Then for every $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$ there exists a $\delta_2 > 0$ such that if $\varphi(t) \in [-\delta_2, \delta_2]$ for all $t \in [-a - |c|\delta, 0]$ then the solution to* (3.1.1) *satisfies $u(t) \in [-\delta, \delta]$ for all $t \geqslant 0$. This means that for $(\mu, \sigma) \in \{P(1, 0, 2) < 1\}$, the zero solution to* (3.1.1) *is Lyapunov stable.*

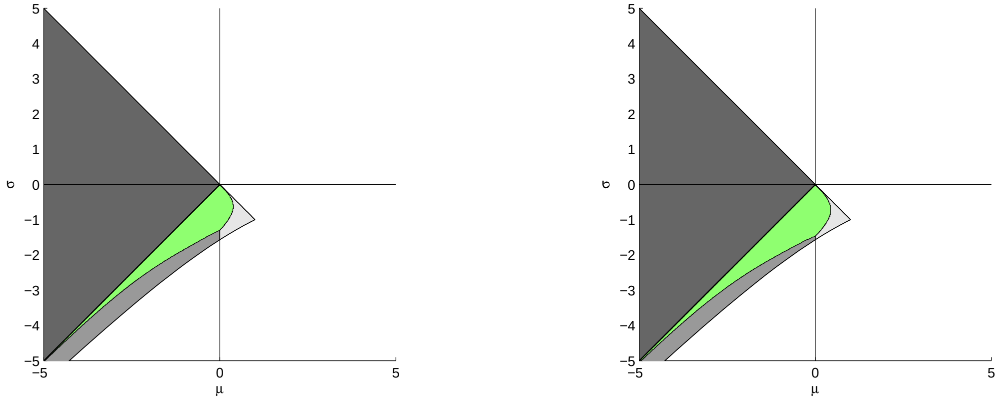
Proof. For this proof define

$$J = \bigcup_{\delta\in\left(0, \left|\frac{a}{c}\right|\right)} \{P(\delta, c, 2) < \delta\}.$$

Let $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$, $L > \frac{|\mu|+|\sigma|}{\varepsilon}$, $\tau_1 = a + |c|\delta$ and $(\mu, \sigma) \in J$. Then for some maximal $\delta_1 \in \left(0, \left|\frac{a}{c}\right|\right)$, $(\mu, \sigma) \in \{P(\delta_1, c, 2) < \delta_1\}$. If $\delta_1 < \delta$ set $\delta_2 = \delta_1 e^{-2L\tau_1}$. By Lemma 3.4.6, if $\varphi(t) \in [-\delta_2, \delta_2]$ for $t \in [-\tau_1, 0]$ then $u(t) \in [-\delta_1, \delta_1] \subseteq [-\delta, \delta]$ for all $t \geqslant 0$. If $\delta < \delta_1$ set $\delta_2 = \delta e^{-2L\tau_1}$. By Lemma 3.4.5, $(\mu, \sigma) \in \{P(\delta_1, c, 2) < \delta_1\} \subseteq \{P(\delta, c, 2) < \delta\}$. By Lemma 3.4.6, if $\varphi(t) \in [-\delta_2, \delta_2]$ for $t \in [-\tau_1, 0]$ then $u(t) \in [-\delta, \delta]$ for all $t \geqslant 0$. This proves that for $(\mu, \sigma) \in J$, the zero solution to (3.1.1) is Lyapunov stable.

Now we show that $J = \{P(1,0,2) < 1\}$. For all c, when $\delta \to 0$ then $\tau_1 \to a$, $D \to \frac{|\mu|+|\sigma|}{\varepsilon}$ (its value for $c = 0$) and so $\mathcal{I}(\hat{u}, \delta, |c|, 2) \to \mathcal{I}(\hat{u}, \delta, 0, 2)$. Thus, $P(\delta, c, 2) \to P(\delta, 0, 2)$. When $c = 0$ the nonlinear terms in δ (which appear in the D term) disappear and $\frac{P(\delta,0,2)}{\delta} = P(1,0,2)$. Thus, $\frac{P(\delta,c,2)}{\delta} \to P(1,0,2)$ as $\delta \to 0$. Because of this and Lemma 3.4.5, $J = \{P(1,0,2) < 1\}$. $\square$

The sets $\{P(\delta, c, 2) < \delta\}$ for different values of δ are shown in Figure 3–9. As $\delta \to 0$, these sets can be seen to be converging to the set $\{P(1,0,2) < 1\}$ shown in Figure 3–10.



(a) The region $\{P(\delta, 1, 2) < \delta\}$ for $\delta = 0.1$

(b) The region $\{P(\delta, 1, 2) < \delta\}$ for $\delta = 0.01$

Figure 3–9: The sets $\{P(\delta, 1, 2) < \delta\}$ for different values of δ with $\varepsilon = a = c = 1$. As $\delta \to 0$ the region converges to $\{P(1,0,2) < 1\}$ which is shown in Figure 3–10.

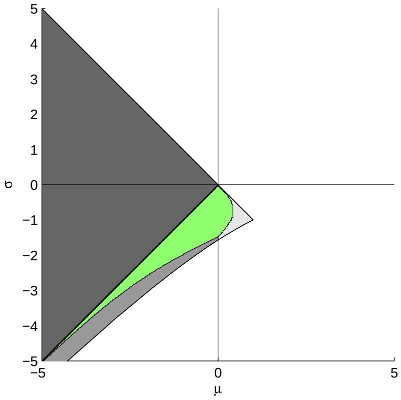


Figure 3–10: The set $\{P(1,0,2) < 1\}$ is shaded green and plotted with $\varepsilon = a = 1$. The function P is defined in Definition 3.4.2. This region is globally stable for the constant delay case and locally stable for the state dependent case.

Simplifying the integration term

Since this proof will be extended later on to prove the stability of numerical methods, it is necessary to perform the integration $\mathcal{I}\left(\hat{u}, \delta, |c|, 2\right)$ and find the $\hat{u}$ values for which the integral is a maximum. As pointed out in the remark in page 54, we first need to divide the region into cases according to which of $-\frac{\delta+\hat{u}}{D\delta}$ and $-\tau_1$ is larger.

CASE:$-\frac{\delta+\hat{u}}{D\delta} > -\tau_1$

In this case the integration has to be broken down into two parts so we call the integration for this case $\mathcal{I}_2$. When this case occurs $\hat{u}$ has another upper bound since

$$\frac{\delta+\hat{u}}{D\delta} < \tau_1 \quad \Rightarrow \quad \hat{u} < \left(\tau_1 D - 1\right)\delta.$$

Then $\hat{u} \in \left[-\delta, \min\left\{\left(\tau_1 D - 1\right)\delta, -\frac{\mu}{\sigma}\delta\right\}\right]$. First let $\mu \neq 0$.

$$\begin{aligned}
\mathcal{I}_2\left(\hat{u}, \delta\right) &= \hat{u}e^{\frac{\mu}{\varepsilon}\tau_1} + \frac{\sigma}{\varepsilon}\int_{-\tau_1}^{-\frac{\delta+\hat{u}}{D\delta}} e^{-\frac{\mu\theta}{\varepsilon}}\left(-\delta\right)d\theta + \frac{\sigma}{\varepsilon}\int_{-\frac{\delta+\hat{u}}{D\delta}}^{0} e^{-\frac{\mu\theta}{\varepsilon}}\left(\hat{u} + D\delta\theta\right)d\theta \\
&= \hat{u}e^{\frac{\mu}{\varepsilon}\tau_1} + \frac{\sigma}{\mu}\delta\left(e^{\frac{\mu}{\varepsilon}\frac{\delta+\hat{u}}{D\delta}} - e^{\frac{\mu}{\varepsilon}\tau_1}\right) - \frac{\sigma}{\mu}\hat{u}\left(1 - e^{\frac{\mu}{\varepsilon}\frac{\delta+\hat{u}}{D\delta}}\right) + \frac{\sigma}{\mu}D\delta\left(\frac{\varepsilon}{\mu}\left(e^{\frac{\mu}{\varepsilon}\frac{\delta+\hat{u}}{D\delta}} - 1\right) - \frac{\delta+\hat{u}}{D\delta}e^{\frac{\mu}{\varepsilon}\frac{\delta+\hat{u}}{D\delta}}\right) \\
&= \hat{u}\left(e^{\frac{\mu}{\varepsilon}\tau_1} - \frac{\sigma}{\mu}\right) + \frac{\sigma}{\mu}\delta\left[\frac{\varepsilon D}{\mu}\left(e^{\frac{\mu}{\varepsilon}\frac{\delta+\hat{u}}{D\delta}} - 1\right) - e^{\frac{\mu}{\varepsilon}\tau_1}\right] \qquad (3.4.12)
\end{aligned}$$

For the case $\mu = 0$ we can go back to the expression (3.3.13) in the previous section with a replaced by τ_1 and σ replaced by $\frac{\sigma}{\varepsilon}$.

$$\mathcal{I}_2\left(\hat{u}, \delta\right) = \hat{u} - \frac{\sigma\tau_1}{\varepsilon}\delta - \frac{1}{2\delta}\left(\delta + \hat{u}\right)^2 = -\frac{\sigma\tau_1}{\varepsilon}\delta - \frac{\delta}{2} - \frac{\hat{u}^2}{2\delta} \qquad (3.4.13)$$

It is easy to show that (3.4.12) approaches (3.4.13) as $\mu \to 0$ so we will always assume that $\mu \neq 0$, use (3.4.12) and just take the limit when we need to consider $\mu = 0$.

CASE: $-\frac{\delta+\hat{u}}{D\delta} \leqslant -\tau_1$

In this case the $\tilde{\eta}$ function does not have the flat part so the integration is only one-part. For this case to occur $\hat{u}$ must be in the interval $\left[\left(\tau_1 D - 1\right)\delta, -\frac{\mu}{\sigma}\delta\right]$. This is only possible in the region where $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$. Again, first let $\mu \neq 0$.

$$\begin{aligned}\mathcal{I}_1\left(\hat{u},\delta\right) &= \hat{u}e^{\frac{\mu}{\varepsilon}\tau_1} + \frac{\sigma}{\varepsilon}\int\limits_{-\tau_1}^{0} e^{-\frac{\mu\theta}{\varepsilon}}\left(\hat{u}+D\delta\theta\right)d\theta \\ &= \hat{u}e^{\frac{\mu}{\varepsilon}\tau_1} - \frac{\sigma}{\mu}\hat{u}\left(1-e^{\frac{\mu}{\varepsilon}\tau_1}\right) + \frac{\sigma}{\mu}D\delta\left[\frac{\varepsilon}{\mu}\left(e^{\frac{\mu}{\varepsilon}\tau_1}-1\right)-\tau_1 e^{\frac{\mu}{\varepsilon}\tau_1}\right] \\ &= \hat{u}\left[e^{\frac{\mu}{\varepsilon}\tau_1} + \frac{\sigma}{\mu}\left(e^{\frac{\mu}{\varepsilon}\tau_1}-1\right)\right] + \frac{\sigma}{\mu}D\delta\left[\frac{\varepsilon}{\mu}\left(e^{\frac{\mu}{\varepsilon}\tau_1}-1\right)-\tau_1 e^{\frac{\mu}{\varepsilon}\tau_1}\right] \qquad (3.4.14)\end{aligned}$$

When $\mu = 0$ we can go back to the expression in (3.3.15) in the previous section with a replaced by τ_1 and σ replaced by $\frac{\sigma}{\varepsilon}$.

$$\mathcal{I}_1\left(\hat{u},\delta\right) = \hat{u} + \frac{\sigma\tau_1}{\varepsilon}\hat{u} + \frac{\sigma^2\tau_1^2}{2\varepsilon^2}\delta \qquad (3.4.15)$$

The expression in (3.4.14) approaches (3.4.15) as $\mu \to 0$. Like in the previous case, we will always use (3.4.14) and just take limits when we need $\mu = 0$.

In Theorem 3.4.10 we prove that $P\left(\delta, c\right) = \mathcal{I}\left(-\frac{\mu}{\sigma}\delta, \delta, c, 2\right)$. The proof of this theorem requires the items proven in Lemmas 3.4.9 and 3.4.8.

Lemma 3.4.8. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu$ and $\sigma < -\mu$. Let $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$ and $\{\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}\}$. If $\mu \geqslant 0$ then $\sigma \geqslant -\frac{\varepsilon}{\tau_1}$. If $\mu < 0$ then $\mu \in \left[\left(-3+2\sqrt{2}\right)\frac{\varepsilon}{\tau_1}, 0\right]$ and*

$$\sigma \geqslant -\frac{\varepsilon}{\tau_1}\left[\frac{1}{2}\left(1+\frac{\mu\tau_1}{\varepsilon}\right)+\frac{1}{2}\sqrt{1+6\frac{\mu\tau_1}{\varepsilon}+\left(\frac{\mu\tau_1}{\varepsilon}\right)^2}\right] \geqslant -\frac{\varepsilon}{\tau_1}.$$

Proof. Let $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$. Then

$$\begin{aligned}\tau_1\frac{\operatorname{sign}\left(\mu\right)\mu-\sigma}{\varepsilon}-1 = \tau_1\frac{|\mu|+|\sigma|}{\varepsilon}-1 \leqslant \tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}, \\ \Rightarrow \quad \frac{\tau_1}{\varepsilon}\sigma^2 + \left(1 \text{ - } \operatorname{sign}\left(\mu\right)\frac{\mu\tau_1}{\varepsilon}\right)\sigma - \mu \leqslant 0. \qquad (3.4.16)\end{aligned}$$

The boundary of the region where this inequality holds is

$$\sigma = -\frac{\varepsilon}{\tau_1}\left[\frac{1}{2}\left(1 \text{ - } \operatorname{sign}\left(\mu\right)\frac{\mu\tau_1}{\varepsilon}\right) \pm \frac{1}{2}\sqrt{\left(1 \text{ - } \operatorname{sign}\left(\mu\right)\frac{\mu\tau_1}{\varepsilon}\right)^2 + 4\frac{\mu\tau_1}{\varepsilon}}\right]. \qquad (3.4.17)$$

If $\mu \geqslant 0$ then this simplifies to $\sigma = -\frac{\varepsilon}{\tau_1}$. Since $\mu = \sigma = 0$ satisfies (3.4.16) then this inequality holds for $\sigma \geqslant -\frac{\varepsilon}{\tau_1}$.

If $\mu < 0$ then (3.4.17) simplifies to

$$\sigma = -\frac{\varepsilon}{\tau_1}\left[\frac{1}{2}\left(1+\frac{\mu\tau_1}{\varepsilon}\right) \pm \frac{1}{2}\sqrt{1+6\frac{\mu\tau_1}{\varepsilon}+\left(\frac{\mu\tau_1}{\varepsilon}\right)^2}\right]. \tag{3.4.18}$$

Requiring $1+6\frac{\mu\tau_1}{\varepsilon}+\left(\frac{\mu\tau_1}{\varepsilon}\right)^2 \geqslant 0$ yields $\frac{\mu\tau_1}{\varepsilon} \in \left[-3+2\sqrt{2},0\right]$. The lower bound on σ can be found by taking the lower boundary in (3.4.18) which attains its minimum at $\mu = 0$. This yields $\sigma \geqslant -\frac{\varepsilon}{\tau_1}\left[\frac{1}{2}\left(1+\frac{\mu\tau_1}{\varepsilon}\right)+\frac{1}{2}\sqrt{1+6\frac{\mu\tau_1}{\varepsilon}+\left(\frac{\mu\tau_1}{\varepsilon}\right)^2}\right] \geqslant -\frac{\varepsilon}{\tau_1}$. □

Lemma 3.4.9. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu$ and $\sigma < -\mu$. Let $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$. Define*

$$\hat{u}_* = \left[\frac{\varepsilon}{\mu}D\ln\left(1-\frac{\mu}{\sigma}e^{\frac{\mu}{\varepsilon}\tau_1}\right)-1\right]\delta.$$

The following statements are true:

(A) If $\tau_1 D - 1 > -\frac{\mu}{\sigma}$ then the maximum of $\mathcal{I}_2(\hat{u},\delta)$ over $\hat{u} \in \left[-\delta, -\frac{\mu}{\sigma}\delta\right]$ occurs at either $-\frac{\mu}{\sigma}\delta$ or at $\hat{u}_*$ if $\hat{u}_* < -\frac{\mu}{\sigma}\delta$.

(B) If $\tau_1 D - 1 > -\frac{\mu}{\sigma}$ and $u_* < -\frac{\mu}{\sigma}\delta$ then $P(\delta, c, 2) \geqslant \delta$.

(C) If $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$ then $\sup_{\hat{u}\in[-\delta,(\tau_1 D-1)\delta]} \mathcal{I}_2(\hat{u},\delta) = \mathcal{I}_2((\tau_1 D-1)\delta,\delta)$

(D) If $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$ then $\sup_{\hat{u}\in\left[(\tau_1 D-1)\delta,-\frac{\mu}{\sigma}\delta\right]} \mathcal{I}_1(\hat{u},\delta) = \mathcal{I}_1\left(-\frac{\mu}{\sigma}\delta,\delta\right)$

(E) If $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$ then $\mathcal{I}_2((\tau_1 D-1)\delta,\delta) \leqslant \mathcal{I}_1\left(-\frac{\mu}{\sigma}\delta,\delta\right)$.

Proof of (A). To find the maximum of $\mathcal{I}_2$ with respect to $\hat{u}$, consider the derivative

$$\frac{\partial \mathcal{I}_2(\hat{u},\delta)}{\partial \hat{u}} = e^{\frac{\mu}{\varepsilon}\tau_1} + \frac{\sigma}{\mu}\left(e^{\frac{\mu}{\varepsilon}\frac{\delta+\hat{u}}{D\delta}}-1\right). \tag{3.4.19}$$

At $\hat{u} = -\delta$ this is positive. To find $\hat{u}_*$ where $\mathcal{I}_2(\hat{u}_*,\delta)$ is maximum, set the derivative equal to zero in (3.4.19),

$$\hat{u}_* = \left[\frac{\varepsilon}{\mu}D\ln\left(1-\frac{\mu}{\sigma}e^{\frac{\mu}{\varepsilon}\tau_1}\right)-1\right]\delta. \tag{3.4.20}$$

Since $1-\frac{\mu}{\sigma}e^{\frac{\mu\tau_1}{\varepsilon}} \in [0,1]$ if $\mu < 0$ and $1-\frac{\mu}{\sigma}e^{\frac{\mu\tau_1}{\varepsilon}} > 1$ if $\mu > 0$ then $\hat{u}_* > -\delta$ in both cases. □

Proof of (B). We first show that if $\tau_1 D - 1 > -\frac{\mu}{\sigma}$ and $\hat{u}_* < -\frac{\mu}{\sigma}\delta$ then we cannot have $\mu > 0$. Let $\hat{u}_* < -\frac{\mu}{\sigma}\delta$ and $\mu > 0$. Then $\frac{\partial \mathcal{I}_2\left(-\frac{\mu}{\sigma}\delta,\delta\right)}{\partial \hat{u}} < 0$. Consider the term $\frac{\varepsilon D}{\mu}$,

$$\frac{\varepsilon D}{\mu} = \frac{|\mu|+|\sigma|}{\mu}\left(1+\frac{|\mu|+|\sigma|}{\varepsilon}|c|\,\delta\right) \geqslant \frac{|\mu|+|\sigma|}{\mu} = \left(1-\frac{\sigma}{\mu}\right). \tag{3.4.21}$$

Now consider the exponent of the second term in (3.4.19) with $\hat{u} = -\frac{\mu}{\sigma}\delta$,

$$\frac{\mu}{\varepsilon}\frac{1-\frac{\mu}{\sigma}}{D} = \left(1-\frac{\mu}{\sigma}\right)\frac{\mu}{\varepsilon D} \leqslant \frac{1-\frac{\mu}{\sigma}}{1-\frac{\sigma}{\mu}} = -\frac{\mu}{\sigma}.$$

Thus,

$$e^{\frac{\mu\tau_1}{\varepsilon}} + \frac{\sigma}{\mu}\left(e^{-\frac{\mu}{\sigma}}-1\right) \leqslant \frac{\partial \mathcal{I}_2\left(-\frac{\mu}{\sigma}\delta,\delta\right)}{\partial\hat{u}} < 0.$$

Isolating τ_1 in this expression yields $\tau_1 < \frac{\varepsilon}{\mu}\ln\left[\frac{\sigma}{\mu}\left(1-e^{-\frac{\mu}{\sigma}}\right)\right]$. Let $x = \frac{\mu}{\sigma}$. Then $x \in (-1,0)$ and $\frac{1-e^{-x}}{x} > 1$. Also, $\left(1-\frac{1}{x}\right)\ln\left[\frac{1-e^{-x}}{x}\right] - 1 \leqslant -x$. Using these inequalities and (3.4.21) yields

$$\tau_1 D - 1 \leqslant \frac{\varepsilon D}{\mu}\ln\left[\frac{\sigma}{\mu}\left(1-e^{-\frac{\mu}{\sigma}}\right)\right] - 1 \leqslant \left(1-\frac{\sigma}{\mu}\right)\ln\left[\frac{\sigma}{\mu}\left(1-e^{-\frac{\mu}{\sigma}}\right)\right] - 1 \leqslant -\frac{\mu}{\sigma}.$$

It follows from this argument that if we require both $\tau_1 D - 1 > -\frac{\mu}{\sigma}$ and $\hat{u}_* < -\frac{\mu}{\sigma}\delta$ then we cannot have $\mu > 0$.

Now let $\mu < 0$ and $\hat{u}_* < -\frac{\mu}{\sigma}\delta$. Then by setting $\frac{\partial\mathcal{I}_2(\hat{u}_*,\delta)}{\partial\hat{u}} = 0$ in (3.4.19), $e^{\frac{\mu}{\varepsilon}\frac{\delta+\hat{u}_*}{D\delta}} - 1 = -\frac{\mu}{\sigma}e^{\frac{\mu}{\varepsilon}\tau_1}$. Also, $e^{\frac{\mu}{\varepsilon}\tau_1} - \frac{\sigma}{\mu} < 0$ because $\mu < 0$ and $\sigma \leqslant \mu < 0$. Thus,

$$\begin{aligned}\mathcal{I}_2\left(\hat{u}_*,\delta\right) &= \hat{u}_*\left(e^{\frac{\mu}{\varepsilon}\tau_1} - \frac{\sigma}{\mu}\right) + \frac{\sigma}{\mu}\delta\left[\frac{\varepsilon D}{\mu}\left(e^{\frac{\mu}{\varepsilon}\frac{\delta+\hat{u}_*}{D\delta}} - 1\right) - e^{\frac{\mu}{\varepsilon}\tau_1}\right],\\ &\geqslant -\frac{\mu}{\sigma}\delta\left(e^{\frac{\mu}{\varepsilon}\tau_1} - \frac{\sigma}{\mu}\right) + \frac{\sigma}{\mu}\delta\left[-\frac{\varepsilon D}{\sigma}e^{\frac{\mu}{\varepsilon}\tau_1} - e^{\frac{\mu}{\varepsilon}\tau_1}\right] = \delta - \left(\frac{\varepsilon D}{\mu} + \frac{\mu}{\sigma} + \frac{\sigma}{\mu}\right)\delta e^{\frac{\mu}{\varepsilon}\tau_1}.\end{aligned}$$

To complete the proof we need to show that $\left(\frac{\varepsilon D}{\mu} + \frac{\mu}{\sigma} + \frac{\sigma}{\mu}\right)\delta e^{\frac{\mu}{\varepsilon}\tau_1} < 0$. Since $D \geqslant \frac{|\mu|+|\sigma|}{\varepsilon} \geqslant \frac{\left|\frac{\mu}{\sigma}\right||\mu|+|\sigma|}{\varepsilon}$ then

$$D \geqslant \frac{\left|\frac{\mu}{\sigma}\right||\mu|+|\sigma|}{\varepsilon} = -\frac{1}{\varepsilon}\left(\frac{\mu^2}{\sigma} + \sigma\right) \Rightarrow \frac{\varepsilon D}{\mu} + \frac{\mu}{\sigma} + \frac{\sigma}{\mu} \leqslant 0.$$

This proves $\mathcal{I}_2\left(\hat{u}_*,\delta\right) > \delta$. □

Proof of (C). Let $\frac{\partial\mathcal{I}_2((\tau_1 D-1)\delta,\delta)}{\partial\hat{u}} < 0$. Then,

$$e^{\frac{\mu\tau_1}{\varepsilon}} + \frac{\sigma}{\mu}\left(e^{\frac{\mu\tau_1}{\varepsilon}} - 1\right) < 0. \tag{3.4.22}$$

This can be rewritten as

$$\sigma < \frac{\mu e^{\frac{\mu\tau_1}{\varepsilon}}}{1-e^{\frac{\mu\tau_1}{\varepsilon}}} = -\frac{\varepsilon}{\tau_1}\left(\frac{-\frac{\mu\tau_1}{\varepsilon}e^{\frac{\mu\tau_1}{\varepsilon}}}{1-e^{\frac{\mu\tau_1}{\varepsilon}}}\right). \tag{3.4.23}$$

We show that the expression in the right-hand-side is continuous and decreases as μ increases. Let $x = \frac{\mu\tau_1}{\varepsilon}$, then

$$\lim_{\mu\to 0} -\frac{\varepsilon}{\tau_1}\left(\frac{-\frac{\mu\tau_1}{\varepsilon}e^{\frac{\mu\tau_1}{\varepsilon}}}{1-e^{\frac{\mu\tau_1}{\varepsilon}}}\right) = \lim_{x\to 0} -\frac{\varepsilon}{\tau_1}\left(\frac{-xe^x}{1-e^x}\right) = -\frac{\varepsilon}{\tau_1}.$$

$$\frac{d}{d\mu}\left[-\frac{\varepsilon}{\tau_1}\left(\frac{-\frac{\mu\tau_1}{\varepsilon}e^{\frac{\mu\tau_1}{\varepsilon}}}{1-e^{\frac{\mu\tau_1}{\varepsilon}}}\right)\right] = \frac{d}{dx}\frac{xe^x}{1-e^x} = \frac{e^x\left(1+x-e^x\right)}{\left(1-e^x\right)^2} \leqslant 0.$$

The last inequality is because $1+x \leqslant e^x$ for all $x \in \mathbb{R}$. So in the region $\mu > 0$, a necessary condition for (3.4.22) to hold is $\sigma < -\frac{\varepsilon}{\tau_1}$. From Lemma 3.4.8, if $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$ and $\mu > 0$ then $\sigma \geqslant -\frac{\varepsilon}{\tau_1}$. Thus $\frac{\partial \mathcal{I}_2((\tau_1 D-1)\delta,\delta)}{\partial \hat{u}} \geqslant 0$ if $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$ and $\mu > 0$.

Now let $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$ and $\mu < 0$. From Lemma 3.4.8, $\frac{\mu\tau_1}{\varepsilon} \in \left[-3+2\sqrt{2}, 0\right]$ and $\sigma \geqslant -\frac{\varepsilon}{\tau_1}\left[\frac{1}{2}\left(1+\frac{\mu\tau_1}{\varepsilon}\right) - \frac{1}{2}\sqrt{1+6\frac{\mu\tau_1}{\varepsilon}+\left(\frac{\mu\tau_1}{\varepsilon}\right)^2}\right]$. Since $\frac{-xe^x}{1-e^x} \geqslant \frac{1}{2}\left(1+x\right)+\frac{1}{2}\sqrt{1+6x+x^2}$ for $x \in \left[-3+2\sqrt{2}, 0\right]$, then $\sigma \geqslant -\frac{\varepsilon}{\tau_1}\left(\frac{-\frac{\mu\tau_1}{\varepsilon}e^{\frac{\mu\tau_1}{\varepsilon}}}{1-e^{\frac{\mu\tau_1}{\varepsilon}}}\right)$. This contradicts with (3.4.23). Thus $\frac{\partial \mathcal{I}_2((\tau_1 D-1)\delta,\delta)}{\partial \hat{u}} \geqslant 0$ if $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$ and $\mu < 0$. Thus, we have proven that

$$\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma} \quad \Rightarrow \quad \frac{\partial \mathcal{I}_2\left(\left(\tau_1 D - 1\right)\delta, \delta\right)}{\partial \hat{u}} = e^{\frac{\mu\tau_1}{\varepsilon}} + \frac{\sigma}{\mu}\left(e^{\frac{\mu\tau_1}{\varepsilon}} - 1\right) \geqslant 0. \qquad (3.4.24)$$

This is illustrated in Figure 3–11. To complete the proof of (C), observe from (3.4.19) that the derivative $\frac{\partial \mathcal{I}_2(\hat{u},\delta)}{\partial \hat{u}}$ decreases as $\hat{u}$ increases. Then by (3.4.24), $\sup_{\hat{u}\in[-\delta,(\tau_1 D-1)\delta]} \mathcal{I}_2\left(\hat{u},\delta\right) = \mathcal{I}_2\left(\left(\tau_1 D - 1\right)\delta, \delta\right)$. □

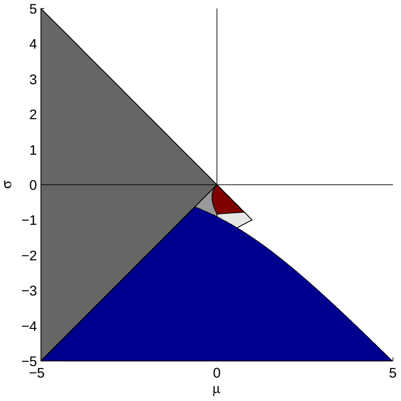


Figure 3–11: $\{\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}\}$ is shown in brown and $\{e^{\frac{\mu\tau_1}{\varepsilon}} + \frac{\sigma}{\mu}\left(e^{\frac{\mu\tau_1}{\varepsilon}} - 1\right) < 0\}$ is shown in blue ($\varepsilon = a = c = 1$ and $\delta = 0.1$). The two regions do not intersect.

Proof of (D). Let $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$. For all $\hat{u} \in \left[(\tau_1 D - 1)\delta, -\frac{\mu}{\sigma}\delta\right]$,

$$\frac{\partial \mathcal{I}_1(\hat{u}, \delta)}{\partial \hat{u}} = e^{\frac{\mu\tau_1}{\varepsilon}} + \frac{\sigma}{\mu}\left(e^{\frac{\mu\tau_1}{\varepsilon}} - 1\right) = \frac{\partial \mathcal{I}_2((\tau_1 D - 1)\delta, \delta)}{\partial \hat{u}}. \tag{3.4.25}$$

From (3.4.24), $\frac{\partial}{\partial \hat{u}}\mathcal{I}_1(\hat{u}, \delta) > 0$ for all $\hat{u} \in [-\delta, (\tau_1 D - 1)\delta]$. Thus, $\sup_{\hat{u}\in\left[(\tau_1 D - 1)\delta, -\frac{\mu}{\sigma}\delta\right]} \mathcal{I}_1(\hat{u}, \delta) = \mathcal{I}_1\left(-\frac{\mu}{\sigma}\delta, \delta\right)$. □

Proof of (E). Let $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$. The expression for $\mathcal{I}_2((\tau_1 D - 1)\delta, \delta)$ simplifies to

$$\mathcal{I}_2((\tau_1 D - 1)\delta, \delta) = (\tau_1 D - 1)\delta\left(e^{\frac{\mu}{\varepsilon}\tau_1} - \frac{\sigma}{\mu}\right) + \frac{\sigma}{\mu}\delta\left[\frac{\varepsilon D}{\mu}\left(e^{\frac{\mu}{\varepsilon}\tau_1} - 1\right) - e^{\frac{\mu}{\varepsilon}\tau_1}\right].$$

Compare $\mathcal{I}_1\left(-\frac{\mu}{\sigma}\delta, \delta\right)$ and $\mathcal{I}_2((\tau_1 D - 1)\delta, \delta)$ when $\tau_1 D - 1 < -\frac{\mu}{\sigma}$,

$$\begin{aligned}\mathcal{I}_1\left(-\frac{\mu}{\sigma}\delta, \delta\right) &- \mathcal{I}_2((\tau_1 D - 1)\delta, \delta)\\ &= \left(-\frac{\mu}{\sigma} - (\tau_1 D - 1)\right)\delta\left(e^{\frac{\mu}{\varepsilon}\tau_1} - \frac{\sigma}{\mu}\right) + \frac{\sigma}{\mu}e^{\frac{\mu}{\varepsilon}\tau_1}\left(-\frac{\mu}{\sigma}\delta\right) + \frac{\sigma}{\mu}\delta(-\tau_1 D + 1),\\ &= \left(-\frac{\mu}{\sigma} - (\tau_1 D - 1)\right)\delta\left[e^{\frac{\mu}{\varepsilon}\tau_1} + \frac{\sigma}{\mu}\left(e^{\frac{\mu}{\varepsilon}\tau_1} - 1\right)\right].\end{aligned}$$

This is non-negative because of (3.4.24). Thus, $\mathcal{I}_1\left(-\frac{\mu}{\sigma}\delta, \delta\right) \geqslant \mathcal{I}_2((\tau_1 D - 1)\delta, \delta)$. □

Theorem 3.4.10. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu$ and $\sigma < -\mu$. Let $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$. If $P(\delta, c, 2) < \delta$ then*

$$P(\delta, c, 2) = \begin{cases} \mathcal{I}_1\left(-\frac{\mu}{\sigma}\delta, \delta\right), & \textit{if } \tau_1 D - 1 \leqslant -\frac{\mu}{\sigma},\\ \mathcal{I}_2\left(-\frac{\mu}{\sigma}\delta, \delta\right), & \textit{if } \tau_1 D - 1 > -\frac{\mu}{\sigma}.\end{cases}$$

Proof. Note that this expression does not hold outside of $\{P(\delta, c, 2) < \delta\}$. In order to prove this theorem, we need the items (A)-(E) in Lemma 3.4.9.

Let $\tau_1 D - 1 > -\frac{\mu}{\sigma}$. Then we only have the two-part integration so $\mathcal{I}(\hat{u}, \delta, c, 2) = \mathcal{I}_2(\hat{u}, \delta)$. From (A) and (B), $P(\delta, c, 2) = \mathcal{I}_2\left(-\frac{\mu}{\sigma}, \delta\right)$ if $P(\delta, c, 2) < \delta$.

Let $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$. From (C) and (D), $\mathcal{I}(\hat{u}, \delta, c, 2) = \max\{\mathcal{I}_2((\tau_1 D - 1)\delta, \delta), \mathcal{I}_1\left(-\frac{\mu}{\sigma}\delta, \delta\right)\}$. From (E), $P(\delta, c, 2) = \mathcal{I}_1\left(-\frac{\mu}{\sigma}\delta, \delta\right)$. □

3.4.2 Results using arbitrary $k \geqslant 2$

In this section the results of Section 3.4.1 are generalised to arbitrary $k \geqslant 2$. The results of this section are summarized in Lemma 3.4.12 and Theorem 3.4.13 which parallel Lemma 3.4.6 and Theorem 3.4.7 in Section 3.4.1. Since we are still looking for new points in the stability region in $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$, we again set $\sigma \leqslant \mu$ and $\sigma < -\mu$.

Let $k \geqslant 2$, $L > \frac{|\mu|+|\sigma|}{\varepsilon}$ and $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$. Let $\tau_1 = a + |c|\delta$ and $\tau_2 = a + c\delta$. If $|\varphi(t)| \leqslant \delta_2 = \delta e^{-Lk\tau_1}$ then by Lemma 3.4.1, the solution to (3.1.1) satisfies $\sup_{s\in[-\tau_1,k\tau_1]} |u(s)| \leqslant \delta$. Now take any $t \geqslant k\tau_1$. As in Section 3.4.1, set $D = \frac{|\mu|+|\sigma|}{\varepsilon}\left(1 + \frac{|\mu|+|\sigma|}{\varepsilon}|c|\delta\right)$ and use (3.4.2) in page 58,

$$u(t) = \eta(0) e^{\frac{\mu\tau_2}{\varepsilon}} + \frac{\sigma}{\varepsilon}\int_{-\tau_2}^{0} e^{-\frac{\mu\theta}{\varepsilon}}\eta(\theta)\,d\theta. \tag{3.4.26}$$

We again have $\theta = s - t$ and $\eta(\theta) = u(t + \theta - a - cu(t+\theta))$ for $\theta \in [-\tau_2, 0]$. Since $t > k\tau_1$ then $t + \theta \geqslant (k-1)\tau_1$ and $u(t+\theta) \in [-\delta, \delta]$ for all θ. Also, $t + \theta - a - cu(t+\theta) \in [k\tau_1 - \tau_2 - a - |c|\delta, t - a + |c|\delta] \subseteq [(k-2)\tau_1, t]$. So we can require the η function to have the same properties as $u(s)$ when $s \in [(k-2)\tau_1, t]$. For the case $k = 2$, our only restrictions on $\eta(\theta)$ were $\eta(0) = \hat{u} \in \left[-\delta, -\frac{\mu}{\sigma}\delta\right]$ and $|\eta'(\theta)| \leqslant D\delta$. To extend our results to $k = 3$ we add a bound on the second derivative. For $t \in [\tau_1, 3\tau_1]$, $t - a - cu(t) \in [\tau_1, 2\tau_1]$ so $|\dot{u}(t)|$ and $|\dot{u}(t-a-cu(t))|$ are both bounded by $\frac{|\mu|+|\sigma|}{\varepsilon}\delta$ using the bounds on the solution. For $t \in [(k-2)\tau_1, t]$,

$$\begin{aligned}|\ddot{u}(t)| &= \frac{1}{\varepsilon}|\mu\dot{u}(t) + \sigma\dot{u}(t-a-cu(t))(1-c\dot{u}(t))|,\\ &\leqslant \frac{|\mu|}{\varepsilon}\frac{|\mu|+|\sigma|}{\varepsilon}\delta + \frac{|\sigma|}{\varepsilon}\frac{|\mu|+|\sigma|}{\varepsilon}\delta\left(1 + \frac{|\mu|+|\sigma|}{\varepsilon}|c|\delta\right) \leqslant \frac{|\mu|+|\sigma|}{\varepsilon}D\delta.\end{aligned}$$

Then the bound on the second derivative of η is

$$\begin{aligned}\left|\eta''(\theta)\right| &= \left|\frac{d^2}{d\theta^2}u(t+\theta-a-cu(t+\theta))\right|\\ &= |\ddot{u}(t+\theta-a-cu(t+\theta))(1-c\dot{u}(t+\theta)) + \dot{u}(t+\theta-a-cu(t+\theta))(-c\ddot{u}(t+\theta))|\\ &\leqslant \left|\frac{|\mu|+|\sigma|}{\varepsilon}D\delta\left(1+\frac{|\mu|+|\sigma|}{\varepsilon}|c|\delta\right) + \frac{|\mu|+|\sigma|}{\varepsilon}\delta|c|\frac{|\mu|+|\sigma|}{\varepsilon}D\delta\right|\\ &= D^2\delta + \left(\frac{|\mu|+|\sigma|}{\varepsilon}\right)^2|c|D\delta^2\end{aligned}$$

Define $D_{(0)} = 1$, $D_{(1)} = D$ and $D_{(2)} = D^2 + \left(\frac{|\mu|+|\sigma|}{\varepsilon}\right)^2|c|D\delta$ so that for $k = 3$ we have the following restrictions:

$$|\eta(\theta)| \leqslant D_{(0)}\delta, \quad \left|\eta'(\theta)\right| \leqslant D_{(1)}\delta, \quad \left|\eta''(\theta)\right| \leqslant D_{(2)}\delta.$$

For higher values of $k \geqslant 2$ we similarly derive bounds on $\left|\eta^{(i)}(\theta)\right| \leqslant D_{(i)}\delta$ for $i = 1, ..., k-1$. When $c = 0$ these simplify to $D_{(i)} = \left(\frac{|\mu|+|\sigma|}{\varepsilon}\right)^i$. If we let $\eta_{(k)}(\theta)$ be the function that satisfies these restrictions and stays as negative as possible then from (3.4.26) we must have

$$u(t) \leqslant \sup_{\hat{u}\in\left[-\delta,-\frac{\mu}{\sigma}\delta\right]} \left(\hat{u}e^{\frac{\mu\tau_2}{\varepsilon}} + \frac{\sigma}{\varepsilon}\int\limits_{-\tau_2}^{0} e^{-\frac{\mu\theta}{\varepsilon}}\eta_{(k)}(\theta)\,d\theta \right).$$

Definition 3.4.11. Let $k \in \mathbb{Z}$, $k \geqslant 2$, $\sigma \leqslant \mu$ and $\sigma < -\mu$. For any $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$ and $\hat{u} \in \left[-\delta, -\frac{\mu}{\sigma}\delta\right]$ define $\eta_{(k)}(\theta) \in C\left([-\tau_1, 0], \left[-\delta, -\frac{\mu}{\sigma}\delta\right]\right)$ to be the function that stays as negative as possible while satisfying the following conditions:

$$\begin{aligned} &\eta_{(k)}(0) = \hat{u}, \\ &\left|\eta_{(k)}^{(i)}(\theta)\right| \leqslant D_{(i)}\delta \quad \text{for } i = 0, ..., k-1. \end{aligned} \tag{3.4.27}$$

Also, define $\mathcal{I}(\hat{u}, \delta, |c|, k)$ and $P(\delta, c, k)$

$$\mathcal{I}(\hat{u}, \delta, |c|, k) = \hat{u}e^{\frac{\mu\tau_1}{\varepsilon}} + \frac{\sigma}{\varepsilon}\int\limits_{-\tau_1}^{0} e^{-\frac{\mu\theta}{\varepsilon}}\eta_{(k)}(\theta)\,d\theta,$$

$$P(\delta, c, k) = \sup_{\hat{u}\in\left[-\delta,-\frac{\mu}{\sigma}\delta\right]} \mathcal{I}(\hat{u}, \delta, |c|, k).$$

Sample $\eta_{(k)}(\theta)$ plots for $k = 2$ and 3 are shown in Figure 3.4.2. The $\eta_{(2)}$ function is given in Definition 3.4.2. For $k = 3$ it can be found by first defining

$$\bar{\eta}_{(3)}(\theta) = \begin{cases} -\delta, & \theta \leqslant 0, \\ -\delta + \frac{\delta}{2}D_{(2)}\theta^2, & 0 \leqslant \theta \leqslant \frac{D_{(1)}}{D_{(2)}}, \\ -\delta - \frac{\delta D_{(1)}^2}{2D_{(2)}} + \delta D_{(1)}\theta, & \frac{D_{(1)}}{D_{(2)}} \leqslant \theta. \end{cases} \tag{3.4.28}$$

This function needs to be shifted to the left in order to get the required value of $\hat{u}$ at $\theta = 0$. The shift depends on the value of $\hat{u}$.

$$\theta_{\text{shift}} = \begin{cases} \sqrt{\frac{2(\hat{u}+\delta)}{D_{(2)}\delta}}, & -\delta \leqslant \hat{u} \leqslant -\delta + \frac{\delta D_{(1)}^2}{2D_{(2)}}, \\ \frac{\hat{u}+\delta+\frac{\delta D_{(1)}^2}{2D_{(2)}}}{D_{(1)}\delta}, & -\delta + \frac{\delta D_{(1)}^2}{2D_{(2)}} < \hat{u}. \end{cases} \tag{3.4.29}$$

Now we define the η function for $k = 3$ as

$$\eta_{(3)}(\theta) = \bar{\eta}_{(3)}(\theta + \theta_{\text{shift}}).$$

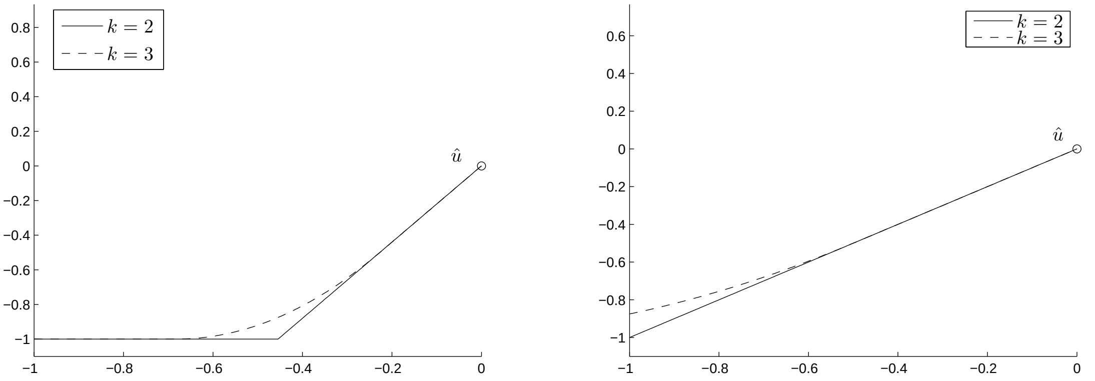
Using this we numerically evaluate $\{P(1,0,3) < 1\}$ and plot it in Figure 3–13(c). Observe that $\{P(1,0,2) < 1\} \subseteq \{P(1,0,3) < 1\} \subseteq \Sigma_\star$. Section 3.5 provides measurements on the size the regions that we have derived.

Lemma 3.4.12. *Let $\varepsilon, a > 0$, $\sigma \leqslant \mu$ and $\sigma < -\mu$. Let $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$, $k \in \mathbb{Z}$, $k \geqslant 2$ and $(\mu,\sigma) \in \{P(\delta,c,k) < \delta\}$. Then there exists a $\delta_2 \in (0,\delta)$ such that if $|\varphi(t)| \in [-\delta_2, \delta_2]$ for $t \in [-a-|c|\delta, 0]$ then the solution to* (3.1.1) *satisfies $u(t) \in [-\delta,\delta]$ for all $t \geqslant 0$.*

Proof. Similar to the proof of Lemma 3.4.6. □

Theorem 3.4.13. *Let $\varepsilon, a > 0$, $\sigma \leqslant \mu$ and $\sigma < -\mu$. Let $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$, $k \in \mathbb{Z}$, $k \geqslant 2$ and $(\mu,\sigma) \in \{P(1,0,k) < \delta\}$. Then for every $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$ there exists $\delta_2 \in (0,\delta)$ such that if $|\varphi(t)| \in [-\delta_2, \delta_2]$ for $t \in [-a-|c|\delta, 0]$ then $u(t) \in [-\delta,\delta]$ for all $t \geqslant 0$.*

Proof. Similar to the proof of Theorem 3.4.7. □



(a) $\mu = 1$, $\sigma = -1.2$, $\hat{u} = 0$

(b) $\mu = 0.4$, $\sigma = -0.6$, $\hat{u} = 0$

Figure 3–12: Sample $\tilde{\eta}(\theta)$ functions for $\varepsilon = a = 1$, $c = 0$ and $\delta = 1$.

Stability in the σ-axis

The segment of $\{P(1,0,3) < 1\}$ on the σ-axis is $\left(-\frac{37\varepsilon}{24a}, 0\right)$, an improvement over the segment $\left(-\frac{3\varepsilon}{2a}, 0\right)$ found using $k = 2$ in Section 3.3.1. We prove this now by evaluating $P(1,0,3)$ for the case $\varepsilon = 1$ and $\mu = c = 0$. In this case $D = -\sigma$ and $\mathcal{I}(\hat{u},\delta,0,3)$ simplifies to

$$\mathcal{I}(\hat{u},\delta,0,3) = \hat{u} + \sigma \int\limits_{-a+\theta_{\text{shift}}}^{\theta_{\text{shift}}} \eta_0(\theta)\, d\theta,$$

where η_0 and θ_{shift} are as defined in (3.4.28) and (3.4.29). Since the integration depends on the positions of $-a+\theta_{\text{shift}}$ and θ_{shift}, there are six cases to consider.

1. $-a+\theta_{\text{shift}} \leqslant 0$ and $\theta_{\text{shift}} \leqslant 0$ (only possible when $\hat{u} = -\delta$)

$$\mathcal{I}\left(\hat{u},\delta,0,3\right) = \hat{u} + \sigma \int\limits_{-a+\theta_{\text{shift}}}^{\theta_{\text{shift}}} -\delta d\theta = -(1+\sigma a)\delta$$

Then $\mathcal{I}\left(\hat{u},\delta,0,3\right) < \delta$ if $\sigma > -\frac{2}{a}$.

2. $-a+\theta_{\text{shift}} \leqslant 0$ and $0 < \theta_{\text{shift}} \leqslant \frac{1}{|\sigma|}$ (only possible if $\hat{u} \leqslant -\frac{\delta}{2}$). In this case, $\theta_{\text{shift}} = \frac{1}{|\sigma|}\sqrt{\frac{2(\hat{u}+\delta)}{\delta}}$. So we are considering $\sqrt{\frac{2(\hat{u}+\delta)}{\delta}} \leqslant |\sigma|a$. The integration yields

$$\begin{aligned}\mathcal{I}\left(\hat{u},\delta,0,3\right) &= \hat{u} + \sigma \int\limits_{-a+\theta_{\text{shift}}}^{0} -\delta d\theta + \sigma \int\limits_{0}^{\theta_{\text{shift}}} \left(-\delta + \frac{\delta}{2}\left|\sigma\right|^2\theta^2\right) d\theta, \\ &= \hat{u} - \sigma\delta a + \frac{\delta}{6}\sigma\left|\sigma\right|^2\theta_{\text{shift}}^3 = \hat{u} - \sigma\delta a - \frac{\delta}{6}\left(\frac{2\left(\hat{u}+\delta\right)}{\delta}\right)^{\frac{3}{2}}.\end{aligned}$$

In the range $\hat{u} \in \left[-\delta, -\frac{\delta}{2}\right]$ the right-hand-side is always increasing so the maximum is obtained at $\hat{u} = -\frac{\delta}{2}$. Then $\mathcal{I}\left(\hat{u},\delta,0,3\right) \leqslant -\frac{\delta}{2} - \sigma\delta a - \frac{\delta}{6} = -\sigma\delta a - \frac{2\delta}{3}$ and this is less than δ if $\sigma > -\frac{5}{3a}$.

3. $0 \leqslant -a+\theta_{\text{shift}} \leqslant \frac{1}{|\sigma|}$ and $0 < \theta_{\text{shift}} \leqslant \frac{1}{|\sigma|}$ (only possible if $\hat{u} \leqslant -\frac{\delta}{2}$). In this case we have $\theta_{\text{shift}} = \frac{1}{|\sigma|}\sqrt{\frac{2(\hat{u}+\delta)}{\delta}}$ and $-a + \frac{1}{|\sigma|}\sqrt{\frac{2(\hat{u}+\delta)}{\delta}} \geqslant 0$. This last inequality implies $|\sigma|a \leqslant \sqrt{\frac{2(\hat{u}+\delta)}{\delta}} \leqslant 1$ which implies $-\frac{1}{a} \leqslant \sigma$. But we already know from [6] that if $\sigma \in \left(-\frac{3}{2a}, 0\right)$ then $P\left(\delta,0,2\right) < \delta$. Thus in this case $P\left(\delta,0,3\right) \leqslant P\left(\delta,0,2\right) < \delta$.

4. $-a+\theta_{\text{shift}} \leqslant 0$ and $\frac{1}{|\sigma|} < \theta_{\text{shift}}$. This case requires $\hat{u} > -\frac{\delta}{2}$ and $\theta_{\text{shift}} = \frac{1}{|\sigma|}\frac{\hat{u}+\frac{3}{2}\delta}{\delta}$.

$$\begin{aligned}\mathcal{I}\left(\hat{u},\delta,0,3\right) &= \hat{u} + \sigma \int\limits_{-a+\theta_{\text{shift}}}^{0} -\delta d\theta + \sigma \int\limits_{0}^{\frac{1}{|\sigma|}} \left(-\delta + \frac{\delta}{2}\left|\sigma\right|^2\theta^2\right) d\theta + \sigma \int\limits_{\frac{1}{|\sigma|}}^{\theta_{\text{shift}}} \left(-\frac{3\delta}{2} + \delta\left|\sigma\right|\theta\right) d\theta \\ &= \hat{u} - \sigma a\delta - \sigma\theta_{\text{shift}}\frac{\delta}{2} - \frac{\delta}{6} + \sigma\left|\sigma\right|\theta_{\text{shift}}^2\frac{\delta}{2} = -\sigma a\delta - \frac{\hat{u}^2}{2\delta} - \frac{13\delta}{24} \\ &\leqslant -\sigma a\delta - \frac{13\delta}{24}\end{aligned}$$

The last line is because $\hat{u} \in [-\delta, 0]$. Thus, $\mathcal{I}\left(\hat{u},\delta,0,3\right) < \delta$ if $\sigma > -\frac{37}{24a}$. It turns out that this is the case that determines the stability interval.

5. $0 \leqslant -a+\theta_{\text{shift}} \leqslant \frac{1}{|\sigma|}$ and $\frac{1}{|\sigma|} < \theta_{\text{shift}}$. Here once again we have $\hat{u} > -\frac{\delta}{2}$ and $\theta_{\text{shift}} = \frac{1}{|\sigma|}\frac{\hat{u}+\frac{3}{2}\delta}{\delta}$. Also, for this case to occur we must have $0 \leqslant -a+\theta_{\text{shift}}$ which means

$$a \leqslant \theta_{\text{shift}} = \frac{1}{|\sigma|}\frac{\hat{u}+\frac{3}{2}\delta}{\delta} \quad \Rightarrow \quad \sigma > -\frac{3}{2a}.$$

But we already know from [6] that if $\sigma \in \left(-\frac{3}{2a}, 0\right)$ then $P(\delta, 0, 2) < \delta$. Thus in this case $P(\delta, 0, 3) < P(\delta, 0, 2) < \delta$.

6. $\frac{1}{|\sigma|} \leqslant -a+\theta_{\text{shift}}$ and $\frac{1}{|\sigma|} < \theta_{\text{shift}}$. For this case, $\hat{u} > -\frac{\delta}{2}$ and $|\sigma| a\delta < \hat{u} + \frac{\delta}{2}$. So we require $\hat{u} > \left(|\sigma| a - \frac{1}{2}\right)\delta$.

$$\begin{aligned}\mathcal{I}(\hat{u}, \delta, 0, 3) &= \hat{u} + \sigma \int\limits_{-a+\theta_{\text{shift}}}^{\theta_{\text{shift}}} \left(-\frac{3\delta}{2} + \delta|\sigma|\theta\right) d\theta = \hat{u} - \sigma a\frac{3\delta}{2} + \frac{\sigma|\sigma|}{2} a(-a+2\theta_{\text{shift}})\delta \\ &= (1+\sigma a)\hat{u} - \frac{\sigma|\sigma|}{2}a^2\delta\end{aligned}$$

We have to consider two more cases:

- If $\sigma \geqslant -\frac{1}{a}$ then the maximum occurs at $\hat{u} = 0$ and

  $$\mathcal{I}(\hat{u}, \delta, 0, 3) \leqslant -\frac{\sigma|\sigma|}{2}a^2\delta \leqslant \frac{\sigma^2 a^2\delta}{2} \leqslant \frac{1}{2}\delta.$$

  So this case always satisfies the criterion for being part of the stability region.

- If $\sigma < -\frac{1}{a}$ then the maximum occurs at $\hat{u} = \left(|\sigma| a - \frac{1}{2}\right)\delta$. Thus,

  $$\mathcal{I}(\hat{u}, \delta, 0, 3) = (1+\sigma a)\left(|\sigma| a - \frac{1}{2}\right)\delta - \frac{\sigma|\sigma|}{2}a^2\delta.$$

  For stability, we require $(1+\sigma a)\left(|\sigma| a - \frac{1}{2}\right)\delta - \frac{\sigma|\sigma|}{2}a^2\delta < \delta$ which leads to $\sigma^2a^2 + 3\sigma a + 3 > 0$. But this is true for any σa so this case does not provide restrictions.

The strictest restriction for $\{P(\delta, 0, 3) < \delta\}$ comes from Case 4 and that is $\sigma > -\frac{37}{24a}$. Thus we have shown Lyapunov stability of the zero solution to (3.1.1) if $\mu = 0$, $\varepsilon = 1$ and $\sigma \in \left(-\frac{37}{24a}, 0\right)$. Easily we can derive zero stability for $\sigma = 0$ ($\dot{u} = 0$) so we have zero stability for $\sigma \in \left(-\frac{37}{24a}, 0\right]$. For arbitrary $\varepsilon > 0$, this extends to Lyapunov stability for $\mu = 0$ and $\sigma \in \left(-\frac{37\varepsilon}{24a}, 0\right]$.

3.4.3 Results using $k = 2$ and $\eta^*_{(2)}(\theta)$

In this section we change the requirements on η again. Going back to the $k = 2$ case, instead of $|\eta'(\theta)| \leqslant D\delta$ we use $\eta'(\theta) \leqslant \frac{\bar{\mu}}{\varepsilon}\eta(\theta) + \frac{|\bar{\sigma}|}{\varepsilon}\delta$ where $\bar{\mu} = \mu\left(1 + \frac{|\mu|+|\sigma|}{\varepsilon}|c|\delta\right)$ and $\bar{\sigma} =$

$\sigma\left(1+\frac{|\mu|+|\sigma|}{\varepsilon}|c|\delta\right)$. This bound comes from the DDE (3.1.1) and $|u(t)|$ for $t\in[-\tau_1,2\tau_1]$. Given $\hat{u}\in\left[-\delta,-\frac{\mu}{\sigma}\delta\right]$, we want to find η which satisfies the bound on the derivative and stays as negative as possible. We find this by first defining $\bar{\eta}^*_{(2)}$,

$$\bar{\eta}^*_{(2)}(\theta)=\begin{cases}-\delta, & \theta\leqslant 0,\\ -\delta e^{\frac{\bar{\mu}}{\varepsilon}\theta}+\frac{|\sigma|\delta}{\mu}\left(e^{\frac{\bar{\mu}}{\varepsilon}\theta}-1\right), & \theta>0,\end{cases} \tag{3.4.30}$$

and then finding θ_{shift} such that $\bar{\eta}^*_{(2)}(\theta_{\text{shift}})=\hat{u}$. Solving this equation yields

$$\theta_{\text{shift}}=\frac{\varepsilon}{\bar{\mu}}\ln\left(\frac{\frac{|\sigma|}{\mu}+\frac{\hat{u}}{\delta}}{\frac{|\sigma|}{\mu}-1}\right). \tag{3.4.31}$$

The expression inside the logarithm is positive since $\sigma\leqslant\mu$, $\sigma<-\mu$ and $\hat{u}\in\left[-\delta,-\frac{\mu}{\sigma}\delta\right]$.

Definition 3.4.14. Let $\varepsilon,a>0$, $\sigma\leqslant\mu$ and $\sigma<-\mu$. For any $\delta\in\left(0,\left|\frac{a}{c}\right|\right)$ and $\hat{u}\in\left[-\delta,-\frac{\mu}{\sigma}\delta\right]$, define $\tilde{\eta}\in C\left((-\infty,0],[-\delta,\hat{u}]\right)$ to be $\eta^*_{(2)}(\theta)=\bar{\eta}^*_{(2)}(\theta+\theta_{\text{shift}})$ where $\bar{\eta}^*_{(2)}$ and θ_{shift} are given by (3.4.30) and (3.4.31). Also define

$$\mathcal{I}^*(\hat{u},\delta,|c|,2)=\hat{u}e^{\frac{\mu\tau_2}{\varepsilon}}+\frac{\sigma}{\varepsilon}\int_{-\tau_2}^{0}e^{-\frac{\mu\theta}{\varepsilon}}\eta^*_{(2)}(\theta)\,d\theta,$$

$$P^*(\delta,c,2)=\sup_{\hat{u}\in\left[-\delta,-\frac{\mu}{\sigma}\delta\right]}\mathcal{I}(\hat{u},\delta,|c|,2).$$

Theorem 3.4.15. *Let $\varepsilon,a>0$ and $(\mu,\sigma)\in\{P^*(1,0,2)<1\}$. Then for every $\delta\in\left(0,\left|\frac{a}{c}\right|\right)$ there exists a $\delta_2>0$ such that if $|\varphi(t)|\in[-\delta_2,\delta_2]$ for all $t\in[-a-c\delta,0]$ then the solution to* (3.1.1) *satisfies $u(t)\in[-\delta,\delta]$ for all $t\geqslant 0$. This means that for $(\mu,\sigma)\in\{P^*(1,0,2)<1\}$, the zero solution to* (3.1.1) *is Lyapunov stable.*

Proof. Similar to the proof of Theorem 3.4.7. □

The new points in the stability region found by using $\eta^*_{(2)}$ are shown in Figure 3–13(d). Note that the integration and maximum were evaluated numerically. Notice how this improves the stability region in $\overset{c}{\Sigma}$. We observe that this new region contains the region found by Barnea in [6] which is shown in Figure 3–8(b). But this region also contains the entire interval $-\frac{3\varepsilon}{2a}<\sigma\leqslant 0$ on the σ-axis.

3.5 Measurements of the regions

In this section we present measurement of the stability regions derived in Sections 3.2, 3.4.1, 3.4.2 and 3.4.3. For these tests we set $\varepsilon = a = 1$ and $c = 0$.

In Table 3–1, the difference between the numerical measurement of the region and the results found in Sections 3.4.1, 3.4.2 and 3.4.3 for $\mu = 0$ are very small. The tables show that by improving the $\eta(\theta)$ function in going from $\{P(1,0,2) < 1\}$ to $\{P(1,0,3) < 1\}$, we derive larger stability regions. However, the improvement is not very significant in $\overset{w}{\Sigma}$ with $\mu = -5$ and the difference gets smaller as μ gets more negative.

Table 3–1: Numerical measurements of the boundary of the derived regions of stability: Values of σ for fixed μ, $\varepsilon = a = 1$.

	σ at $\mu = -5$	σ at $\mu = -2$	σ at $\mu = 0$
$\{r(0) \in (0,1)\}$	-5.067385507762116	-2.557804062076531	-1
$\{P(1,0,2) < 1\}$	-5.067608565809856	-2.587564203387213	-1.499999765595120 $= -\frac{3}{2}$
$\{P(1,0,3) < 1\}$	-5.067926805638362	-2.612788724539893	-1.541665406581630 $= -\frac{37}{24}$
$\{P^*(1,0,2) < 1\}$	-5.067608565809855	-2.589826047227501	-1.499999765595120 $= -\frac{3}{2}$
$\Sigma_\star$	-5.660558641232643	-3.039605122412370	-1.570796326794897 $= -\frac{\pi}{2}$

Table 3–2: Numerical measurements of the boundary of the derived regions of stability: Values of μ for fixed σ, $\varepsilon = a = 1$.

	μ at $\sigma = -5$	μ at $\sigma = -2$
$\{r(0) \in (0,1)\}$	-4.928663352979243	-1.164014632535759
$\{P(1,0,2) < 1\}$	-4.928416273329012	-1.004116462539647
$\{P(1,0,3) < 1\}$	-4.928035196754609	-0.936121469769667
$\{P^*(1,0,2) < 1\}$	-4.928416273329013	-0.974926877322070
$\Sigma_\star$	-4.273422355721326	-0.638045048285238

Table 3–3: Numerical measurements of the boundary of the derived regions of stability: The value of μ at the rightmost boundary point for $\varepsilon = a = 1$.

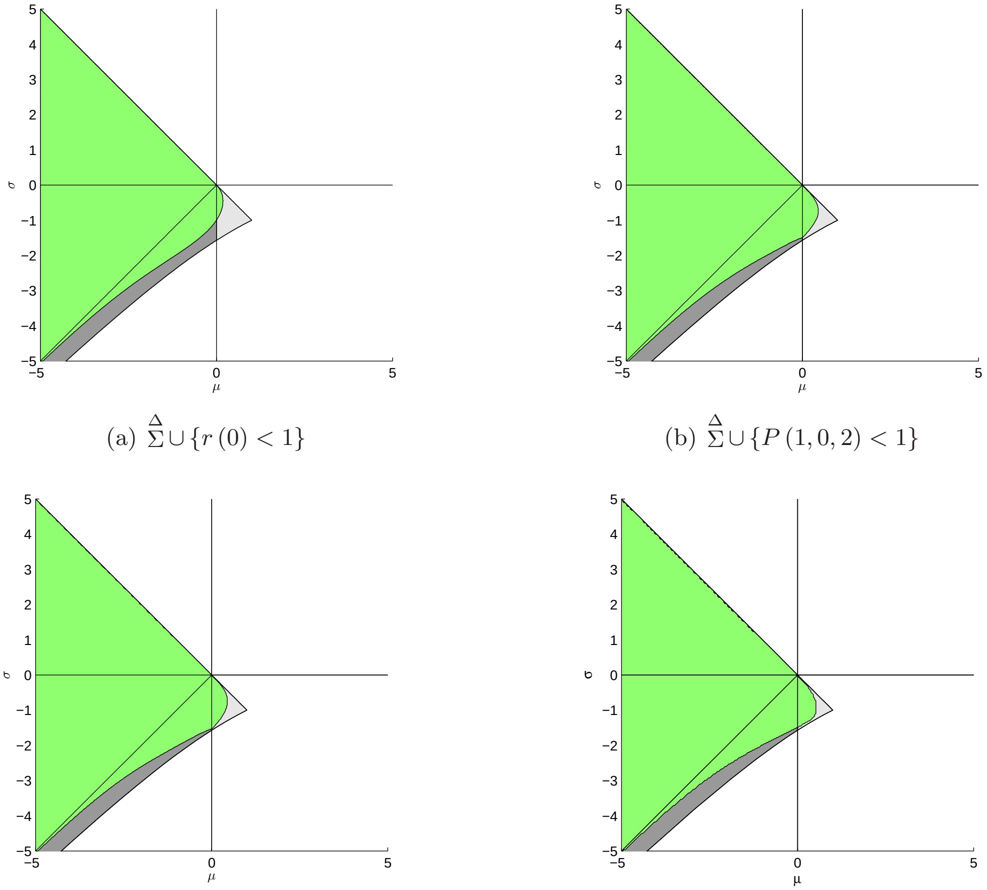
	Maximum value of μ
$\{r(0) \in (0,1)\}$	0.188226406459598
$\{P(1,0,2) < 1\}$	0.456988952862656
$\{P(1,0,3) < 1\}$	0.457021925023451
$\{P^*(1,0,2) < 1\}$	0.550544246705956
$\Sigma_\star$	1

The measured values of σ at $\mu = -5$ for $\{P(1,0,2) < 1\}$ and $\{P^*(1,0,2) < 1\}$ are almost the same in Table 3–1 for $\mu = -5$, and in Table 3–2 for $\sigma = -5$. This suggests that

$\{P^*(1,0,2) < 1\}$ is not a significant improvement over $\{P(1,0,2) < 1\}$ in $\overset{w}{\Sigma}$. We suspect that the reason a large part of $\overset{w}{\Sigma}$ cannot be included in the stability regions derived using Razumikhin-type arguments is because the solutions to (3.1.1) when $(\mu,\sigma) \in \overset{w}{\Sigma}$ display decaying oscillations with several cycles occurring over an interval of length a. It is likely that it will be necessary to use these oscillations to prove that solutions eventually decay to zero. Perhaps if we find a way to improve $\eta(\theta)$ to take this into account we could get more of $\overset{w}{\Sigma}$ in the stability region.

In Table 3–3 the μ values of the rightmost boundary point of the regions are measured. For the case $k = 2$ it is possible to calculate this point by finding the maximum μ value as a function of σ in the expression $\mathcal{I}\left(-\frac{\mu}{\sigma}, 1, 0, 2\right) = 1$. By solving for this we derive the maximum μ to be approximately $0.456971657679506\frac{\varepsilon}{a}$. This agrees well with the numerical measurement of the boundary point for $\{P(1,0,2) < 1\}$. The results of Table 3–3 were found using the golden section search algorithm discussed in [50].

These regions are all shown in Figure 3–13. In $\overset{\Delta}{\Sigma}$ the zero solution to (3.1.1) is asymptotically stable. The zero solution was also proven to be asymptotically stable in $\{r(0) \in (0,1)\}$. In the sets $\{P(1,0,k) < 1\}$ for $k \geqslant 2$ the zero solution was proven to be Lyapunov stable. All these regions are strict subsets of the known analytic stability region $\Sigma_\star$. It is encouraging that we were able to prove stability in a significant region of the delay dependent stability region $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$ using Gronwall and Lyapunov-Razumikhin techniques. These methods will be extended in the next chapter to find stability regions for backward Euler and other Θ methods. The regions that will be derived for backward Euler using the $k = 2$ Lyapunov-Razumikhin method depend on the stepsize used but always contain $\{P(1,0,2) < 1\}$ thus providing a region in which backward Euler is stable for any stepsize $h \in (0, a]$.



(a) $\overset{\triangle}{\Sigma}\cup\{r\,(0)<1\}$

(b) $\overset{\triangle}{\Sigma}\cup\{P\,(1,0,2)<1\}$

(c) $\overset{\triangle}{\Sigma}\cup\{P\,(1,0,3)<1\}$

(d) $\{P^{*}\,(1,0,2)<1\}$

Figure 3–13: Illustration of all the regions discussed in this chapter.

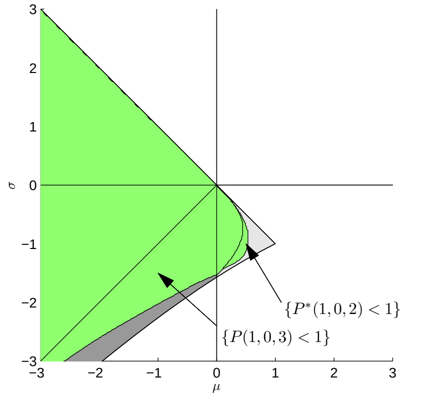


Figure 3–14: $\overset{\triangle}{\Sigma}\cup\{P\,(1,0,3)<1\}\cup\{P^{*}\,(1,0,2)<1\}$

CHAPTER 4
Stability of numerical methods for DDEs

In the first chapter we listed some of the issues involved in extending numerical methods for ODEs to solve DDEs. In this chapter we focus on the issue of stability failure. To clarify what we mean by this let us begin with an example from Bellen and Zennaro [8]. Consider numerical methods that use a constant stepsize of h. Recall that the A-stability region of an RK method is the set of complex numbers $h\lambda$ such that the numerical solution applied to the test problem $\dot{u}(t) = \lambda u$ converges to zero. A numerical method is A-stable if its A-stability region includes the set $\{\Re(h\lambda) < 0\}$. This means that when $\Re(\lambda) < 0$, an A-stable numerical method will reproduce the decay of the solutions to zero for any stepsize. On the other hand, methods with bounded stability domains such as explicit RK methods usually require very small stepsizes when $|\lambda|$ is large in order to exhibit the same behaviour.

Consider two well-known RK methods:

$$\text{midpoint rule: } \begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}, \qquad \text{trapezoidal rule: } \begin{array}{c|cc} 0 & 0 & \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Both methods are A-stable and of order 2 as ODE methods. Following the standard notation introduced in Chapter 1, RK methods can be transformed into continuous RK methods by replacing the weights b_i with polynomial functions. For the midpoint rule we set $b_1(\theta) = \theta$ and for the trapezoidal rule set $b_1(\theta) = \frac{1}{2}\theta$ and $b_2(\theta) = \frac{1}{2}\theta$. For both methods, the continuous extension we have just defined is simply linear interpolation between adjacent nodal values. Since we are only considering stability issues and long-term behaviour of numerical solutions in this chapter, we can ignore the tracking of discontinuity points and order conditions. Consider the problem

$$\begin{cases} \dot{u}(t) = -50u(t) + 40u(t-1), & t \geqslant 0 \\ u(t) = K \text{ (constant)}, & t \leqslant 0. \end{cases} \tag{4.0.1}$$

This is our model DDE (1.1.1) with $\varepsilon = a = 1$, $c = 0$, $\mu = -50$ and $\sigma = 40$. Since $|\sigma| < -\mu$ then $(\mu, \sigma) \in \overset{\Delta}{\Sigma}$. The delay in this problem is constant $\tau = a = 1$ so for all initial functions, the solutions of (4.0.1) converge to zero. Sample plots of the solutions are shown in Figure 4–1.

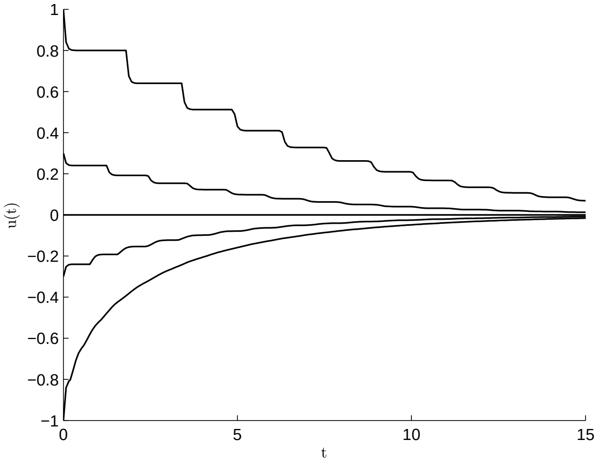
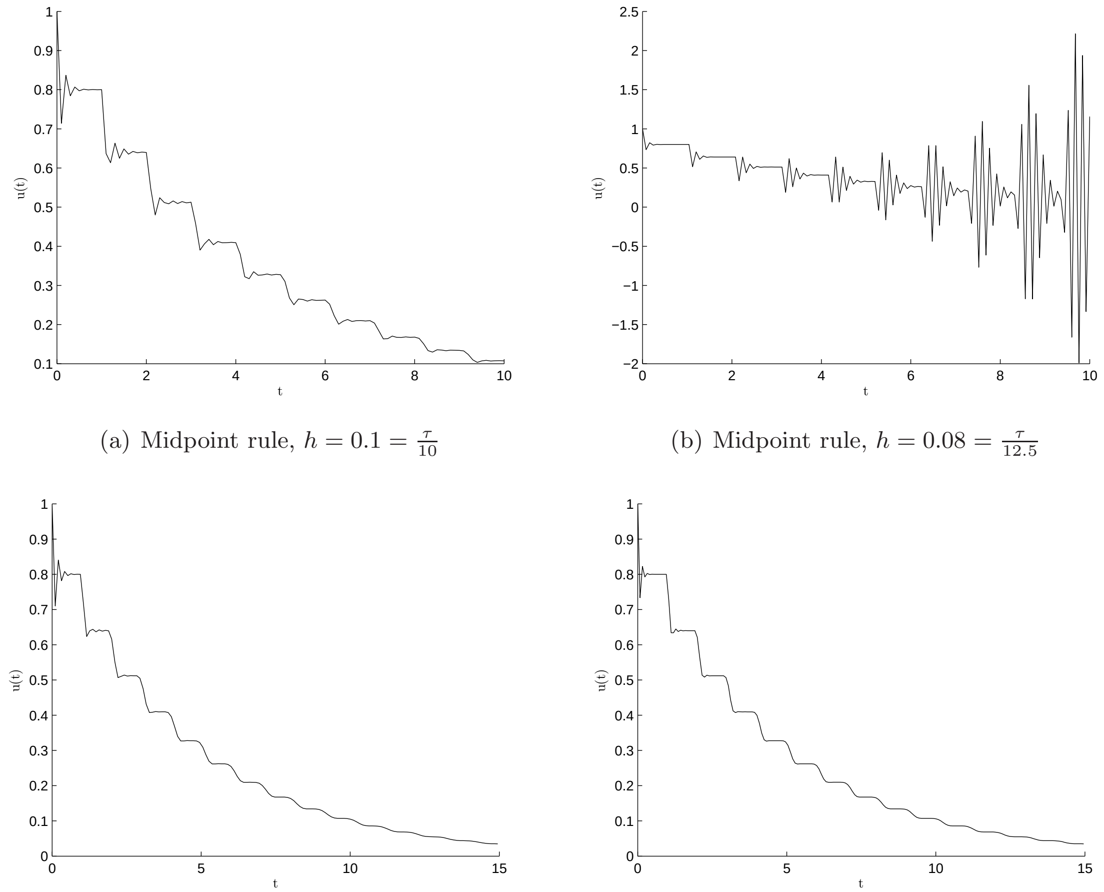


Figure 4–1: Sample time plots of the solutions to (4.0.1) for different constant initial functions. These simulations were performed using an SDIRK scheme discussed in Chapter 5.

Sample results of the numerical integration using midpoint and trapezoidal rules are shown in Figure 4–2. Observe that the numerical solution using the midpoint rule converges to zero in the case where the stepsize $h = 0.1 = \frac{\tau}{10}$, a submultiple of the delay. However, midpoint rule displays numerical instability when $h = 0.08 = \frac{\tau}{12.5}$, a non-submultiple of the delay. This is an example of what Bellen and Zennaro [8] call stability failure when the method is extended to solve DDEs. In contrast, trapezoidal rule provides stable solutions for both $h = 0.1$ and $h = 0.08$. This shows that of the two second order, A-stable ODE methods that are actually identical methods when applied to linear ODEs, the trapezoidal rule is more robust when extended to linear DDEs.

Another example in Bellen and Zennaro [8] shows that even the trapezoidal rule is inadequate for solving $\dot{u}\left(t\right) = \lambda\left(t\right) y\left(t\right) - \frac{4}{5}\lambda\left(t\right) u\left(t-1\right)$ where $\lambda\left(t\right) = -50\sin^2\left(\frac{2\pi}{3}\left(t-\frac{1}{4}\right)\right)$. The solutions of these equations are known to converge to zero. Although we will not discuss this example, it is interesting to note that the trapezoidal rule displays numerical instability in solving this equation even with the use of submultiple stepsizes. In later sections we consider backward Euler which satisfies a stronger form of stability called L-stability. Without going into details here, an RK method with matrix A, weights b and abscissae c is L-stable if its stability function $R\left(z\right) : \mathbb{C} \to \mathbb{R}$ given by $R\left(z\right) = \frac{\det\left(I-zA-zeb^T\right)}{\det\left(I-zA\right)}$ satisfies $R\left(z\right) \to 0$ as $|z| \to \infty$. In Chapter 5 we consider other methods that are L-stable.



(a) Midpoint rule, $h = 0.1 = \frac{\tau}{10}$

(b) Midpoint rule, $h = 0.08 = \frac{\tau}{12.5}$

(c) Trapezoidal rule, $h = 0.1 = \frac{\tau}{10}$

(d) Trapezoidal rule, $h = 0.08 = \frac{\tau}{12.5}$

Figure 4–2: Numerical solutions of (4.0.1) with $K = 1$ using midpoint and trapezoidal rules. This shows the stability failure of midpoint rule when using non-submultiple stepsize.

4.1 Definitions of stability for DDE methods

To extend the notions of stability to numerical methods for DDEs we use our model DDE with one state dependent delay as a test equation

$$\begin{aligned} &\varepsilon \dot{u}\,(t) = \mu u\,(t) + \sigma u\,(t - a - cu\,(t))\,, && t \geqslant 0, \\ &u\,(t) = \varphi\,(t)\,, && t \leqslant 0. \end{aligned} \tag{4.1.1}$$

Recall from Section 2.1 that for any nonzero ε and a, it is always possible to rescale the equation and set $\varepsilon = a = 1$. For any nonzero c, it is always possible to rescale and set $c = 1$. As in previous chapters we keep ε, a and c fixed and consider only nonnegative values, with $\varepsilon, a > 0$ and $c \geqslant 0$. Then the free parameters in the equation are $(\mu, \sigma) \in \mathbb{R}^2$. The history function

$\varphi(t)$ is a real-valued continuous function. General properties of this equation are discussed in Chapter 2. The analytic stability region $\Sigma_\star = \overset{\Delta}{\Sigma} \cup \overset{w}{\Sigma} \cup \overset{c}{\Sigma}$ of the constant delay case is derived in Chapter 3 (see Definition 3.1.1), and found to also be the stability region for the state dependent case [25].

Previous authors who have looked into the numerical stability of DDE methods only considered the stability of methods applied to the constant delay case [1, 4, 12, 20, 21, 37, 43, 57, 59, 61]. Many authors have also only considered the stability of numerical methods in the cone $\overset{\Delta}{\Sigma}$, the delay independent portion of $\Sigma_\star$. Barwell [7] first introduced the notion of P-stability in $\overset{\Delta}{\Sigma}$ but in his definition $(\mu, \sigma) \in \mathbb{C}^2$. Since complex coefficients will not be considered here, we instead define P(0)-stability.

Definition 4.1.1 (Adapted from Bellen and Zennaro [8])**.** A numerical method is said to be *P(0)-stable* if the numerical solution $\{u_n\}_{n\geqslant 0}$ derived from applying the numerical method to (4.1.1) with constant delay $\tau = a$ converges to zero for all $(\mu, \sigma) \in \overset{\Delta}{\Sigma}$, all initial functions and for constant stepsize $h = \frac{\tau}{m}$, $m \in \mathbb{N}$. Removing the constraint of h being a submultiple of τ and allowing for all $h \in (0, \tau)$ leads to the stronger concept of *GP(0)-stability.*

Zennaro [61] showed that any A-stable method is also P-stable (and therefore also P(0)-stable) for DDEs. This does not extend to GP(0)-stability as we can see from Figure 4.0.1. Although stability in $\overset{\Delta}{\Sigma}$ is important to consider in testing for the robustness of a numerical method, stability in $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$ should also be considered, and is perhaps more interesting. Recall that $\overset{\Delta}{\Sigma}$ is the delay independent portion of Σ_* so this region stays the same regardless of the value of a. On the other hand, $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$ is the delay dependent portion of Σ_* and its range depends on the values of ε and a.

Definition 4.1.2 (Adapted from Bellen and Zennaro [8])**.** A numerical method is said to be *D(0)-stable* if the numerical solution $\{u_n\}_{n\geqslant 0}$ derived from applying the numerical method to (4.1.1) with constant delay $\tau = a$ converges to zero for all $(\mu, \sigma) \in \Sigma_\star$, all initial functions and for constant stepsize $h = \frac{\tau}{m}$, $m \in \mathbb{N}$. Removing the constraint of h being a submultiple of τ and allowing for all $h \in (0, \tau)$ leads to the stronger concept of *GD(0)-stability.*

Methods that are known to be D(0)-stable are Θ methods for $\Theta \in \left[\frac{1}{2}, 1\right]$ [21], Radau IIA for $s = 2, 3$ [20], and Gauss RK methods with $s \geqslant 1$ [22]. Lobatto IIIc is not D(0) stable [20]. More results on the stability of DDE methods are available in Bellen and Zennaro [8]. There are currently no methods that have been proven to be GD(0)-stable.

In this thesis I consider the stability of numerical methods applied to state dependent problems. The natural test problem for this is (4.1.1) with $c \geqslant 0$ arbitrary. In this case it does not make sense to require a submultiple stepsize anymore so we consider all stepsizes $h > 0$. Following the stability proofs for the DDE in Section 3.2 using a Gronwall argument and in Section 3.4.1 using a Razumikhin-like proof, I prove the stability of backward Euler in $\overset{\Delta}{\Sigma}$ and a significant portion of $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$. The region derived using a Razumikhin-like proof is larger than that using the Gronwall argument, but in the latter region I prove the convergence of the backward Euler solutions to zero while I only prove a discrete version of Lyapunov stability in the former. For the $c = 0$ case, stability means global stability, but otherwise we always mean local stability. Both proofs are extended in Sections 4.9 and 4.10 to derive stepsize-dependent stability regions of general Θ methods. In the extension of the Razumikhin-style proof for Θ methods we derive analytic expressions for the stability regions which are then evaluated numerically.

4.2 Description of backward Euler

The backward Euler (BE) method is given by the following Butcher tableau:

$$\text{backward Euler:} \quad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

Backward Euler is an A-stable and L-stable ODE method. To solve DDEs, the method may be equipped with linear interpolation. As discussed in Section 1.2, this is done by changing from a constant $b_1 = 1$ to a function $b_1(\beta) = \beta$. Notice that we have switched our interpolation argument from θ to β. For the rest of the chapter we will always use β as the interpolation argument in order to avoid confusion later on when we look at the general Θ methods.

Consider the following DDE with one general delay,

$$\begin{aligned} \varepsilon \dot{u}(t) &= f\left(t, u(t), \alpha\left(t, u(t)\right)\right), && t \geqslant t_0, \\ u(t) &= \varphi(t), && t < t_0, \end{aligned} \tag{4.2.1}$$

where we assume $\alpha(t, u(t)) \leqslant t$ at all times. Backward Euler with constant stepsize h and linear interpolation applied to (4.2.1) can be written as

$$u_{n+1} = u_n + hf\left(t_{n+1}, y_{n+1}, \tilde{Y}_{n+1}^{(1)}\right) \tag{4.2.2}$$

where $\tilde{Y}_{n+1}^{(1)}$ is given by the value of the linear interpolation of the numerical solution at time $\tilde{t}_{n+1}^{(1)} = \alpha\left(t_{n+1}, u_{n+1}\right)$. In general, linear interpolation may be written as:

$$\eta\left(t_n + \beta h\right) = u_n + \beta h f\left(t_{n+1}, u_{n+1}, \tilde{Y}_{n+1}^{(1)}\right)$$

One may also think of the $\eta\left(t\right)$ in the following step-by-step manner: Suppose the numerical solution has been solved up to time t_n.

- If $t \leqslant t_0$ then $\eta\left(t\right) = \varphi\left(t\right)$.
- If $t \in [0, t_n]$ then solve the following system with $\beta \in [0, 1]$.

$$\begin{cases} m = \left\lceil \frac{t_n - t}{h} \right\rceil \\ t = \left(1-\beta\right) t_{n-m} + \beta t_{n-m+1} \\ \eta\left(t\right) = \left(1-\beta\right) u_{n-m} + \beta u_{n-m+1} \end{cases}$$

- If $t > t_n$ then solve the following system with $\beta > 0$.

$$\begin{cases} t = \left(1-\beta\right) t_n + \beta t_{n+1} \\ \eta\left(t\right) = \left(1-\beta\right) u_n + \beta u_{n+1} \end{cases}$$

This is how we will treat linear interpolation in this chapter.

Backward Euler applied to (4.1.1) yields

$$u_{n+1} = u_n + \frac{h}{\varepsilon}\left(\mu u_{n+1} + \sigma Y_{n+1}^{(1)}\right). \tag{4.2.3}$$

This equation has to be solved for u_{n+1} at every step of backward Euler. Suppose the BE solution to (4.1.1) is known up to time t_n and we would like to find the update at time t_{n+1}. Define the following function

$$g_{n+1}\left(v\right) = v - u_n - \frac{h}{\varepsilon}\left(\mu v + \sigma \tilde{Y}_{n+1}\left(v\right)\right), \tag{4.2.4}$$

where $\tilde{Y}_{n+1}\left(v\right) = \eta_v\left(t_{n+1} - a - cv\right)$ and

$$\eta_v\left(t\right) = \begin{cases} \eta\left(t\right), & \text{if } t \leqslant t_n, \\ \left(1-\beta\right) u_n + \beta v, \quad \beta = \frac{t - t_n}{h}, & \text{if } t > t_n. \end{cases}$$

The BE update u_{n+1} is any solution to $g_{n+1}\left(v\right) = 0$. As discussed in Section 1.2, if the stepsize h is small enough then this root exists and is unique even for the overlapping case. However, we are considering backward Euler because its has nice stability properties so we use it with larger stepsizes.

Lemma 4.2.1. *Let $\varepsilon, a > 0$ and $\mu + \sigma < 0$. Let the stepsize h be such that $\varepsilon - h\mu > 0$ (no restriction if $\mu < 0$) and let $L < 0 < M$. If $c > 0$ suppose that the history function $\varphi(t)$ is continuous and $\varphi(t) \in \left(-\frac{a}{c}, M\right)$ for $t \leqslant 0$. Then a BE solution to* (4.1.1) *$\{u_n\}_{n\geqslant 0}$ exists such that $u_n > -\frac{a}{c}$ for $n \geqslant 0$. If $c < 0$ suppose that $\varphi(t)$ is continuous and $\varphi(t) \in \left(L, -\frac{a}{c}\right)$ for $t \leqslant 0$. Then a BE solution to* (4.1.1) *$\{u_n\}_{n\geqslant 0}$ exists such that $u_n < -\frac{a}{c}$ for $n \geqslant 0$.*

Proof. Suppose first that $c > 0$. The proof will be by induction. For the base case $n = 0$, obviously $u_0 = \varphi(0)$ exists and from the bounds on φ, $u_0 > -\frac{a}{c}$. Now let $n \geqslant 0$ and suppose that u_n exists and satisfies $u_n > -\frac{a}{c}$. Rewrite (4.2.3) as

$$g_{n+1}(v) = \left(1 - \frac{h\mu}{\varepsilon}\right) v - u_n - \frac{h\sigma}{\varepsilon}\tilde{Y}_{n+1}(v). \tag{4.2.5}$$

Let $v = -\frac{a}{c}$. Then $t_{n+1} - a - cv = t_{n+1}$ so $\tilde{Y}_{n+1}(v) = -\frac{a}{c}$. Then,

$$g_{n+1}\left(-\frac{a}{c}\right) = \left(-1 + \frac{h(\mu+\sigma)}{\varepsilon}\right)\frac{a}{c} - u_n < 0.$$

Now consider letting $v \to \infty$. Then the deviated time $t_{n+1} - a - cv \to -\infty$ so $\tilde{Y}_{n+1}(v)$ takes its values from the history function which is bounded inside $\left(-\frac{a}{c}, M\right)$. From (4.2.5), and if $\mu > 0$ the stepsize restriction, $g_{n+1}(v) \to \infty$ as $v \to \infty$. Since $\tilde{Y}_{n+1}(v)$ is continuous then $g_{n+1}(v)$ is continuous and has a root in $\left(-\frac{a}{c}, \infty\right)$. This proves that a solution to $g_{n+1}(v) = 0$ exists and satisfies $v > -\frac{a}{c}$. Set the BE update $u_{n+1} = v$. The proof for the $c < 0$ case is similar. □

Lemma 4.2.1 enables us to always choose our BE update such that the delay does not become an advance. For the rest of the chapter we always restrict our BE solution accordingly.

Lemma 4.2.1 does not say anything about the uniqueness of the BE solution. Indeed, for general stepsizes the function $g_{n+1}(v)$ may have multiple roots. In Chapter 5 we take some care in choosing which root gives the best qualitative approximation to the DDE solution.

Properties of the BE solution when $\mu < 0$ and $\sigma < 0$.

Here we look at some properties of the BE solution when $\mu < 0$ and $\sigma < 0$. This is motivated by the properties derived in Section 2.1 for negative μ and σ. Recall the definitions in (2.1.2),

$$L_0 = -\frac{a}{c}, \qquad M_0 = \frac{a\sigma}{c\mu}, \qquad \tau_0 = a + cM_0. \tag{4.2.6}$$

Lemma 4.2.2. *Let $\varepsilon, a, c > 0$, $\mu < 0$ and $\sigma < 0$. Let the history function $\varphi(t)$ be continuous and $\varphi(t) \in (L_0, M_0)$ for $t \in [-\tau_0, 0]$. Then for any stepsize $h > 0$, a BE solution to* (4.1.1) *$\{u_n\}_{n\geqslant 0}$ exists such that $u_n \in (L_0, M_0)$ for $n \geqslant 0$.*

Proof. The existence of the BE solution and the lower bound are already guaranteed by Lemma 4.2.1. Since $\mu < 0$ then there is no restriction on the stepsize. So it only remains to prove that $u_n < M_0$ for all $n \geqslant 0$. Again, the proof will be by induction. For the base $n = 0$, obviously $u_0 = \varphi(0) < M_0$. Now let $n \geqslant 0$ and assume that $u_n < M_0$.

$$g_{n+1}(M_0) = \frac{a\sigma}{c\mu} - u_n - \frac{h}{\varepsilon}\left(\frac{a\sigma}{c} + \sigma\tilde{Y}_{n+1}\right) > \frac{a\sigma}{c\mu} - u_n > 0$$

The first inequality comes from the lower bound L_0 and the second inequality comes from $u_n < M_0$. Recall from the proof of Lemma 4.2.1 that $g_{n+1}(L_0) < 0$. Since $g_{n+1}(v)$ is continuous, this proves that a solution to $g_{n+1}(v) = 0$ exists and satisfies $v \in (L_0, M_0)$. Set the BE update $u_{n+1} = v$. □

Lemma 4.2.3. *Let $\varepsilon, a, c > 0$, $\mu < 0$ and $\sigma < 0$. Let the history function $\varphi(t)$ be continuous and $\varphi(t) \in (L_0, M_0)$ for $t \in [-\tau_0, 0]$. Then for any stepsize $h > 0$, any BE solution to* (4.1.1) *$\{u_n\}_{n\geqslant 0}$ must satisfy $u_n \in (L_0, M_0)$ for $n \geqslant 0$.*

Proof. This proof is by strong induction. The base case is easy. Now suppose the BE solution up to n, $\{u_0, ..., u_n\}$, is bounded inside (L_0, M_0). Let $v \leqslant L_0$. Then $\alpha(t_{n+1}, v) = t_{n+1} - a - cv > t_{n+1}$. Set $\alpha(t_n, v) = t_n + \beta h$. Then $\beta > 1$. The spurious stage is $\tilde{Y}_{n+1}(v) = (1-\beta)u_n + \beta v$ and then

$$\begin{aligned} g_{n+1}(v) &= v - u_n - \frac{h}{\varepsilon}\left(\mu v + \sigma\tilde{Y}_{n+1}(v)\right), \\ &= v - u_n - \frac{h}{\varepsilon}\left(\mu v + \sigma\left((1-\beta)u_n + \beta v\right)\right), \\ &= \left(1 - \frac{h\mu}{\varepsilon} - \frac{h\sigma\beta}{\varepsilon}\right)v - \left(1 + \frac{h\sigma(1-\beta)}{\varepsilon}\right)u_n. \end{aligned}$$

Since $\beta > 1$ then $\sigma(1-\beta) > 0$ and we can apply the lower bound on u_n in the expression above. Also applying $v \leqslant L_0$ yields

$$g_{n+1}(v) < \frac{h(\mu+\sigma)}{\varepsilon}\frac{a}{c} < 0.$$

Thus there is no solution to $g_{n+1}(v) = 0$ such that $v \leqslant L_0$. Now let $v \geqslant M_0$. Consider two possible cases:

- If $h \leqslant a + cv$ then $\alpha(t_{n+1}, v) \leqslant t_n$ so $\tilde{Y}_{n+1}(v) \in (L_0, M_0)$. Then,

$$g_{n+1}(v) \geqslant \frac{a\sigma}{c\mu} - u_n - \frac{h}{\varepsilon}\left(\frac{a\sigma}{c} + \sigma\tilde{Y}_{n+1}\right) > \frac{a\sigma}{c\mu} - u_n > 0. \tag{4.2.7}$$

  The first inequality comes from $v \geqslant M_0$, the second and third inequalities are from the lower bound of L_0.

- If $h > a + cv$ then $\alpha(t_{n+1}, v) \in [t_n, t_{n+1}]$. Let $\beta = \frac{\alpha(t_{n+1},v)-t_n}{h}$, so $\beta \in [0,1]$. Then $\tilde{Y}_{n+1} = (1-\beta)u_n + \beta v \geqslant L_0$ and we get the same series of inequalities as in (4.2.7).

Thus, there is no solution $v > M_0$ to $g_{n+1}(v) = 0$. This proves that any BE solution must remain bounded inside (L_0, M_0). □

Lemma 4.2.3 shows that the BE solution to (4.1.1) reproduces the bounds we found in Section 2.1 for the case when $\mu < 0$ and $\sigma < 0$. Now to show that it also has the same three possible types of behaviour as found in Lemma 3.2.1.

Lemma 4.2.4. *Let $\varepsilon, a, c > 0$, $\mu < 0$ and $\sigma < 0$. Let the history function $\varphi(t)$ be continuous and $\varphi(t) \in (L_0, M_0)$ for $t \in [-\tau_0, 0]$. Then for any stepsize $h > 0$, any BE solution to (4.1.1) must be behave in one of the following manners:*

(A) There exists $N \in \mathbb{N}$ such that $u_n \downarrow 0$ for $n > N$.

(B) There exists $N \in \mathbb{N}$ such that $u_n \uparrow 0$ for $n > N$.

(C) For every $N > 0$ there exists $N_1, N_2 \in \mathbb{N}$ and $N_1, N_2 > N$ such that the solution attains a positive maximum at N_1 and a negative minimum at N_2.

Proof. From Lemma 4.2.3, for all $n > 0$, $u_n \in (L_0, M_0)$. Using these bounds we can use Lemma 4.4.1 which states that bounded solutions must behave as in (A), (B) or (C) for the more general case when $\mu + \sigma < 0$. □

4.3 Stability of BE in $\overset{\Delta}{\Sigma}$

Recall that the cone $\overset{\Delta}{\Sigma}$ is the delay independent portion of the analytic stability region of the model DDE (4.1.1). It is known that for the constant delay case, backward Euler is stable independent of stepsize in $\overset{\Delta}{\Sigma}$ (see Guglielmi [21] and Maset [43] for the complex coefficients case). In this section we prove that if $(\mu, \sigma) \in \overset{\Delta}{\Sigma}$, the BE solution to (4.1.1) converge to zero for the state dependent case for all stepsizes.

Recall (4.2.3), the BE equation applied to the model problem (4.1.1)

$$u_{n+1} = u_n + \frac{h}{\varepsilon}\left(\mu u_{n+1} + \sigma Y^{(1)}_{n+1}\right),$$

$$u_{n+1} = \frac{\varepsilon}{\varepsilon - h\mu} u_n + \frac{h\sigma}{\varepsilon - h\mu} Y^{(1)}_{n+1}. \tag{4.3.1}$$

Assume that the numerical solution up to u_n and the entire history function is bounded inside $[-\delta, \delta]$. Assume also that there is no overlapping. Then $\tilde{Y}^{(1)}_{n+1}$ is also bounded by δ and

$$|u_{n+1}| \leqslant \frac{|\varepsilon| + |h\sigma|}{|\varepsilon - h\mu|}\delta. \tag{4.3.2}$$

We will return to this equation later on. First consider how this changes if there is overlapping. Since we exclude any BE solutions below $-\frac{a}{c}$ then $t_{n+1} - a - cu_{n+1} \in [t_n, t_{n+1}]$ and we have $\tilde{Y}^{(1)}_{n+1} = (1-\beta)\, u_n + \beta u_{n+1}$ with $\beta \in [0,1]$. Going back to (4.3.1) and solving for u_{n+1} yields

$$u_{n+1} = \left(\frac{\varepsilon + h\sigma\,(1-\beta)}{\varepsilon - h\mu - h\sigma\beta}\right) u_n. \tag{4.3.3}$$

Let $\varepsilon - h\mu > 0$. If $\sigma \leqslant 0$ then since $\beta \in [0,1]$ we easily derive (4.3.2) again. If $\sigma > 0$ this not so easy. Let $\sigma > 0$ and consider $\frac{\varepsilon + h\sigma(1-\beta)}{\varepsilon - h\mu - h\sigma\beta}$ as a function of $\beta \in [0,1]$. If $\mu + \sigma < 0$ then this function is positive at $\beta = 1$ and its derivative is always negative. Thus, $\left|\frac{\varepsilon + h\sigma(1-\beta)}{\varepsilon - h\mu - h\sigma\beta}\right|$ is maximum at $\beta = 0$. Using this, we derive (4.3.2) again,

$$|u_{n+1}| \leqslant \frac{|\varepsilon| + |h\sigma|}{|\varepsilon - h\mu|}\delta. \tag{4.3.4}$$

We have now shown that this inequality is true for both the overlapping and non overlapping cases if $\varepsilon - h\mu > 0$ and $\mu + \sigma < 0$. Consider the case when $\frac{|\varepsilon| + |h\sigma|}{|\varepsilon - h\mu|} < 1$,

$$\frac{|\varepsilon| + |h\sigma|}{|\varepsilon - h\mu|} = \frac{\varepsilon + |h\sigma|}{\varepsilon - h\mu} < 1 \quad \Leftrightarrow \quad |\sigma| < -\mu \quad \Leftrightarrow \quad (\mu, \sigma) \in \overset{\Delta}{\Sigma}.$$

Theorem 4.3.1. *Let* $\varepsilon, a, c > 0$, $(\mu, \sigma) \in \overset{\Delta}{\Sigma}$ *and* $h > 0$. *For every* $\delta > 0$, *if the history function* $\varphi(t)$ *is continuous and* $\varphi(t) \in [-\delta, \delta]$ *for all* $t \in [-a - c\delta, 0]$ *then the BE solution to* (4.1.1) $\{u_n\}_{n \geqslant 0}$ *satisfies* $u_n \in [-\delta, \delta]$ *for all* $n \geqslant 0$ *and* $\lim_{n\to\infty} u_n = 0$.

Proof. Let $r = \frac{|\varepsilon| + |h\sigma|}{|\varepsilon - h\mu|}$. If $(\mu, \sigma) \in \overset{\Delta}{\Sigma}$ then $r \in (0,1)$. Suppose $\varphi(t) \in [-\delta, \delta]$ for all $t \leqslant 0$. It is easy to see from (4.3.4) that $u_n \in [-\delta, \delta]$ for all $n \geqslant 0$. Thus $\varphi(t)$ is only required to be known, continuous and $\varphi(t) \in [-\delta, \delta]$ for $t \in [-a - c\delta, 0]$. Also, $\alpha(t, u_{n+1}) \in [t - a - c\delta, t]$ for $t \geqslant 0$. We call the solution on this interval the relevant history at time t.

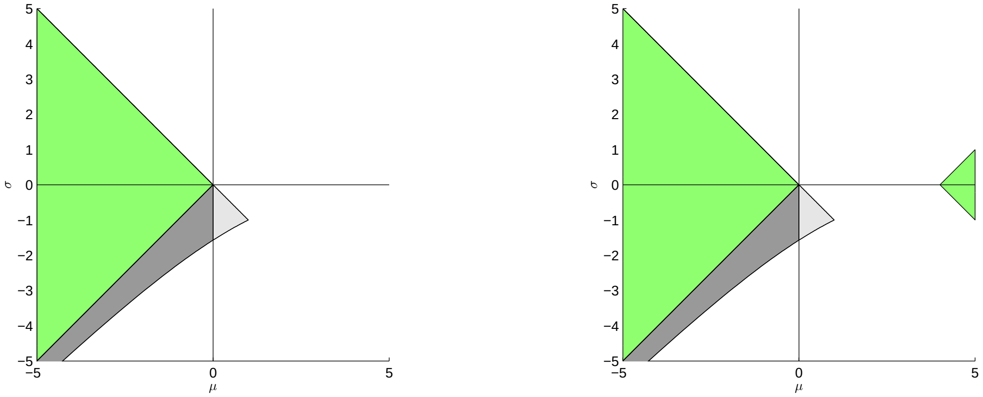
By (4.3.4), $|u_{n+1}| \leqslant r\delta$ for $n \geqslant 0$. Let $\bar{N} = \left\lceil \frac{a+c\delta}{h} \right\rceil$. For $n \geqslant \bar{N}$, the relevant history will be on the time interval $[0, t_{n+1}]$ so $r\delta$ is the bound on the relevant history. Using $r\delta$ instead of δ in (4.3.4), $|u_{n+1}| \leqslant r^2\delta$ for $n \geqslant \bar{N}$.

After $\bar{N}$ more steps, the relevant history of the BE solution is bounded above by $r^2\delta$. By iteratively applying (4.3.4) we get that for every $n \geqslant 0$ if we let $\bar{n} = \left\lfloor \frac{n}{\bar{N}} \right\rfloor$ then $u_n \in [-r^{\bar{n}}\delta, r^{\bar{n}}\delta]$. Since $r \in (0, 1)$ then $u_n \to 0$. □

Now consider the case when $\varepsilon - h\mu < 0$. This is only possible if $\mu > 0$.

$$\frac{|\varepsilon| + |h\sigma|}{|\varepsilon - h\mu|} = \frac{\varepsilon + |h\sigma|}{-(\varepsilon - h\mu)} < 1 \quad \Rightarrow \quad |\sigma| < h\mu - 2\varepsilon.$$

For $\mu > 0$ the points where we have a contraction must satisfy $|\sigma| < h\mu - 2\varepsilon$. On these points we always have $\mu + \sigma > 0$ so if we set $\beta = 1$ in (4.3.3) we get $\frac{\varepsilon + h\sigma(1-\beta)}{\varepsilon - h\mu - h\sigma\beta} = \frac{\varepsilon}{\varepsilon - h(\mu+\sigma)} > 1$ and we do not have a contraction of the solutions as described in Theorem 4.3.1. Thus the region we have derived for the case $\mu > 0$ only yields a contraction if there is no overlapping. This can be guaranteed if $h \in (0, a)$ and the bound δ is chosen to be small enough. Sample plots of the stability regions are shown in Figure 4–3.



(a) $h = 0.1$

(b) $h = 0.5$

Figure 4–3: The sets $\left\{ \frac{|\varepsilon| + |h\sigma|}{|\varepsilon - h\mu|} < 1 \right\}$ are shaded green and plotted with $\varepsilon = a = 1$. Each set restricted to $\mu < 0$ is $\overset{\Delta}{\Sigma}$ and we have shown that if (μ, σ) is in that set then the BE solution to (4.1.1) converges to zero for all stepsizes. The set restricted to $\mu > 0$ is only stable if there is no overlapping.

4.4 Stability of BE using a discrete Gronwall argument

Let us now consider the stability of the backward Euler method in $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$, the delay dependent portion of $\Sigma_{\star}$. In these regions, $\sigma \leqslant \mu$ and $\sigma < -\mu$. So automatically, $\sigma \leqslant 0$. This section is based on a discrete Gronwall argument which is a generalization of the proof of the asymptotic stability of the zero solution to (4.1.1) in Section 3.2 which used a continuous Gronwall argument .

Lemma 4.4.1. *Let $\varepsilon, a, c > 0$ and $\mu + \sigma < 0$. Let the history function $\varphi(t)$ and the BE solution to (4.1.1) be bounded by $[-\delta, \delta]$ for all $t \leqslant 0$ and $n \geqslant 0$. Then for any stepsize $h > 0$, the BE solution must behave in one of the following manners:*

(A) There exists $N \in \mathbb{N}$ such that $u_n \downarrow 0$ for $n > N$

(B) There exists $N \in \mathbb{N}$ such that $u_n \uparrow 0$ for $n > N$

(C) For every $N > 0$ there exists $N_1, N_2 \in \mathbb{N}$ and $N_1, N_2 > N$ such that the solution attains a positive maximum at N_1 and a negative minimum at N_2.

Proof. From the bounds on the solution we must have

$$t_{n+1} - a - c\delta \leqslant \alpha\left(t_{n+1}, u_{n+1}\right) \leqslant t_{n+1}. \tag{4.4.1}$$

Suppose there exists a point $N \in \mathbb{N}$ such that $u_{n+1} \leqslant u_n$ for all $n > N$. Then for some $\bar{u} \in [-\delta, \delta]$, $u_n \downarrow \bar{u}$. Then also, $u_{n+1} - u_n \to 0$. By (4.4.1), we also have $\tilde{Y}^{(1)}_{n+1} \to \bar{u}$.

$$0 = \lim_{n\to\infty} \varepsilon\left(u_{n+1} - u_n\right) = \lim_{n\to\infty} h\left(\mu u_{n+1} + \sigma \tilde{Y}^{(1)}_{n+1}\right) = h\left(\mu + \sigma\right)\bar{u}$$

Since $\mu + \sigma < 0$ then $\bar{u} = 0$. Thus in this case $u_n \downarrow 0$ for $n > N$.

Similarly, if there exists a point $N \in \mathbb{N}$ such that $u_{n+1} \geqslant u_n$ for all $n > N$ then we must have $u_{n+1} \uparrow 0$. If there is no N past which the BE solution becomes monotonic then since the solution is bounded its behaviour must be given by (C). □

At every time step the value of the spurious stage $\tilde{Y}^{(1)}_{n+1} = \eta\left(\alpha\left(t_{n+1}, u_{n+1}\right)\right)$ is found by setting

$$m = \left\lfloor \frac{a + c u_{n+1}}{h} \right\rfloor, \quad \beta = m + 1 - \frac{a + c u_{n+1}}{h}. \tag{4.4.2}$$

Then $\alpha\left(t_{n+1}, u_{n+1}\right) = t_{n+1} - a - c u_{n+1} = t_{n-m} + \beta h = \left(1 - \beta\right) t_{n-m} + \beta t_{n-m+1}$ so

$$Y^{(1)}_{n+1} = \left(1 - \beta\right) u_{n-m} + \beta u_{n-m+1}. \tag{4.4.3}$$

In the succeeding lemmas and theorems we always assume that $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$, $\sigma < -\mu$, and automatically $\sigma < 0$. These restrictions are always satisfied by points in $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$.

Lemma 4.4.2. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let the stepsize h be such that $\varepsilon - h\mu > 0$ (no restriction if $\mu < 0$, automatically satisfied for $\mu > 0$ when $h \in (0, a]$). Let $-\frac{a}{c} < L < 0 < M$, $n \in \mathbb{N}$, $n \geqslant 0$ and let the BE solution to (4.1.1) $\{u_i\}_{i=0}^{n}$ satisfy $u_i \in [L, M]$ for $i = \max\{0, n - \lceil \frac{a+cM}{h} \rceil\}, ..., n$. If $t_{n+1} - a - cM < 0$ then let $\varphi(s) \in [L, M]$ for $s \in [t_{n+1} - a - cM, 0]$. Let $\mu \neq 0$. If $u_{n+1} \leqslant u_n$ then*

$$u_{n+1} - \frac{\varepsilon}{\varepsilon - h\mu(1-\beta)} \left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^m \tilde{Y}_{n+1}^{(1)} \geqslant -\frac{\sigma M}{\mu} \left[1 - \frac{\varepsilon}{\varepsilon - h\mu(1-\beta)} \left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^m\right], \quad (4.4.4)$$

and if $u_{n+1} \geqslant u_n$ then

$$u_{n+1} - \frac{\varepsilon}{\varepsilon - h\mu(1-\beta)} \left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^m \tilde{Y}_{n+1}^{(1)} \leqslant -\frac{\sigma L}{\mu} \left[1 - \frac{\varepsilon}{\varepsilon - h\mu(1-\beta)} \left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^m\right]. \quad (4.4.5)$$

Let $\mu = 0$. If $u_{n+1} \leqslant u_n$ then

$$u_{n+1} - \tilde{Y}_{n+1}^{(1)} \geqslant (a + cu_{n+1}) \frac{\sigma M}{\varepsilon}, \quad (4.4.6)$$

and if $u_{n+1} \geqslant u_n$ then

$$u_{n+1} - \tilde{Y}_{n+1}^{(1)} \leqslant (a + cu_{n+1}) \frac{\sigma L}{\varepsilon}. \quad (4.4.7)$$

Proof. Start with the BE equation applied to the model problem (4.1.1).

$$u_{n+1} = u_n + \frac{h}{\varepsilon}\left[\mu u_{n+1} + \sigma \tilde{Y}_{n+1}^{(1)}\right],$$

Suppose $u_{n+1} \leqslant u_n$. Then even in the overlapping case $\tilde{Y}_{n+1}^{(1)} \leqslant M$.

$$u_{n+1} = \frac{\varepsilon}{\varepsilon - h\mu} u_n + \frac{h\sigma}{\varepsilon - h\mu} \tilde{Y}_{n+1}^{(1)} \geqslant \frac{\varepsilon}{\varepsilon - h\mu} u_n + \frac{h\sigma}{\varepsilon - h\mu} M \quad (4.4.8)$$

Consider first the case $\mu \neq 0$. Applying the discrete Gronwall lemma to the inequality above,

$$u_{n+1} \geqslant \frac{\frac{h\sigma M}{\varepsilon - h\mu}}{1 - \frac{\varepsilon}{\varepsilon - h\mu}} \left(1 - \left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{m+1}\right) + \left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{m+1} u_{n-m},$$

$$u_{n+1} \geqslant -\frac{\sigma M}{\mu} \left(1 - \left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{m+1}\right) + \left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{m+1} u_{n-m}. \quad (4.4.9)$$

Similarly, we can derive

$$u_{n+1} \geqslant -\frac{\sigma M}{\mu}\left(1-\left(\frac{\varepsilon}{\varepsilon-h\mu}\right)^m\right)+\left(\frac{\varepsilon}{\varepsilon-h\mu}\right)^m u_{n-m+1}. \tag{4.4.10}$$

Take $(1-\beta)\frac{\varepsilon-h\mu}{\varepsilon}\times$ (4.4.9)+$\beta\times$ (4.4.10). This yields

$$\left(\frac{\varepsilon-h\mu}{\varepsilon}(1-\beta)+\beta\right)u_{n+1} \geqslant -\frac{\sigma M}{\mu}\left(\frac{\varepsilon-h\mu}{\varepsilon}(1-\beta)+\beta-\left(\frac{\varepsilon}{\varepsilon-h\mu}\right)^m\right)$$
$$+\left(\frac{\varepsilon}{\varepsilon-h\mu}\right)^m\left((1-\beta)u_{n-m}+\beta u_{n-m+1}\right),$$

which simplifies to

$$\frac{\varepsilon-h\mu(1-\beta)}{\varepsilon}u_{n+1} \geqslant -\frac{\sigma M}{\mu}\left(\frac{\varepsilon-h\mu(1-\beta)}{\varepsilon}-\left(\frac{\varepsilon}{\varepsilon-h\mu}\right)^m\right)+\left(\frac{\varepsilon}{\varepsilon-h\mu}\right)^m\tilde{Y}^{(1)}_{n+1},$$

$$u_{n+1} \geqslant -\frac{\sigma M}{\mu}\left[1-\frac{\varepsilon}{\varepsilon-h\mu(1-\beta)}\left(\frac{\varepsilon}{\varepsilon-h\mu}\right)^m\right]+\frac{\varepsilon}{\varepsilon-h\mu(1-\beta)}\left(\frac{\varepsilon}{\varepsilon-h\mu}\right)^m\tilde{Y}^{(1)}_{n+1}.$$

Rearranging yields (4.4.4). The proof of (4.4.5) follows similarly. Now let $\mu=0$. Then (4.4.8) becomes simply $u_{n+1}\geqslant u_n+\frac{h\sigma M}{\varepsilon}$. Applying a discrete Gronwall inequality yields

$$u_{n+1} \geqslant u_{n-m}+(m+1)\frac{h\sigma M}{\varepsilon}, \quad u_{n+1}\geqslant u_{n-m+1}+m\frac{h\sigma M}{\varepsilon} \tag{4.4.11}$$

Taking $(1-\beta)\times$ the first equation + $\beta\times$ the second equation yields

$$u_{n+1} \geqslant (1-\beta)u_{n-m}+\beta u_{n-m+1}+(m+1-\beta)h\sigma M=\tilde{Y}^{(1)}_{n+1}+(a+cu_{n+1})\frac{\sigma M}{\varepsilon}.$$

By rearranging we get (4.4.6). The proof of (4.4.7) follows similarly. □

Definition 4.4.3. Recall $r(v)$ from Definition 3.2.4,

$$r(v)=\begin{cases}\frac{\sigma}{\mu}\left[\frac{1-e^{\frac{\mu}{\varepsilon}(a+cv)}}{1+\frac{\mu}{\sigma}e^{\frac{\mu}{\varepsilon}(a+cv)}}\right], & \text{if } \mu\neq 0,\\ -\frac{\sigma}{\varepsilon}(a+cv), & \text{if } \mu=0.\end{cases}$$

Also define the following $R_{\Theta=1,h}(v)$ function

$$R_{\Theta=1,h}(v)=\begin{cases}\frac{\sigma}{\mu}\left[\frac{1-A^*_v}{1+\frac{\mu}{\sigma}A^*_v}\right], & \text{if } \mu\neq 0,\\ -\frac{\sigma}{\varepsilon}(a+cv), & \text{if } \mu=0.\end{cases}$$

where for each v,

$$A_v^* = \frac{\varepsilon}{\varepsilon - h\mu\left(1-\beta_v\right)}\left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{m_v},$$

and m_v and β_v are defined by

$$m_v = \left\lfloor \frac{a+cv}{h} \right\rfloor, \quad \beta_v = m_v + 1 - \frac{a+cv}{h}.$$

For fixed v, one may show using L'Hopital's rule that the expression $R_{\Theta=1,h}\left(v\right)$ is continuous in μ, including at $\mu = 0$. So both these functions are continuous at $\mu = 0$. To show that $R_{\Theta=1,h}\left(v\right)$ is continuous in v, consider a point v_* at which $\frac{a+cv_*}{h}$ is an integer. In the limit $v \downarrow v_*$ then $m_v \to m_{v_*}$ and $\beta_v \to 1$. In the limit that $v \uparrow v_*$, $m_v \to m_{v_*} - 1$, and $\beta_v \to 0$. Thus,

$$\lim_{v\to v_*^+} \frac{\varepsilon}{\varepsilon - h\mu\left(1-\beta_v\right)}\left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{m_v} = \left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{m_{v_*}} = \lim_{v\to v_*^-} \frac{\varepsilon}{\varepsilon - h\mu\left(1-\beta_v\right)}\left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{m_v}.$$

So $R_{\Theta=1,h}\left(v\right)$ is a continuous function of v even though m_v and β_v are not continuous.

Lemma 4.4.4. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let the stepsize h be such that $\varepsilon - h\mu > 0$ (no restriction if $\mu < 0$, automatically satisfied for $\mu > 0$ when $h \in (0, a]$). Let $-\frac{a}{c} < L < 0 < M$, $n \in \mathbb{N}$, $n \geqslant 0$ and let the BE solution to* (4.1.1) *$\{u_i\}_{i=0}^{n}$ satisfy $u_i \in [L, M]$ for $i = \max\{0, n - \lceil \frac{a+cM}{h} \rceil\}, ..., n$. If $t_{n+1} - a - c\delta < 0$ then let $\varphi\left(s\right) \in [L, M]$ for $s \in [t_{n+1} - a - c\delta, 0]$. Also let $1 + \frac{\mu}{\sigma}\frac{\varepsilon}{\varepsilon - h\mu(1-\beta)}\left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{m} > 0$ where m and β are as defined in* (4.4.2)*. If $u_{n+1} \leqslant u_n$ then*

$$u_{n+1} \geqslant -R_{\Theta=1,h}\left(u_{n+1}\right) M. \tag{4.4.12}$$

If $u_{n+1} \geqslant u_n$ then

$$u_{n+1} \leqslant -R_{\Theta=1,h}\left(u_{n+1}\right) L. \tag{4.4.13}$$

Proof. Suppose $u_{n+1} \leqslant u_n$. Then,

$$u_{n+1} = u_n + \frac{h}{\varepsilon}\left(\mu u_{n+1} + \sigma \tilde{Y}_{n+1}^{(1)}\right) \leqslant u_n \quad \Rightarrow \quad \tilde{Y}_{n+1}^{(1)} \geqslant -\frac{\mu}{\sigma} u_{n+1}.$$

Since the relevant history at $n+1$ is bounded above by M then we may use (4.4.4) and (4.4.6) in Lemma 4.4.2. Using $\tilde{Y}_{n+1}^{(1)} \geqslant -\frac{\mu}{\sigma} u_{n+1}$ in those equations yield (4.4.12). Equation (4.4.13) can be derived similarly. □

Lemma 4.4.5. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let the stepsize h be such that $\varepsilon - h\mu > 0$ (no restriction if $\mu < 0$, automatically satisfied for $\mu > 0$ when $h \in (0, a]$). Let the model parameters and the stepsize h satisfy $R_{\Theta=1,h}(0) \in (0,1)$. Then there exists a sufficiently small $\delta_* \in \left(0, \left|\frac{a}{c}\right|\right)$ such that $R_{\Theta=1,h}(v) \in (0,1)$ for all $v \in [-\delta_*, \delta_*]$. Let $\delta = \left|\frac{\varepsilon - h\mu}{\varepsilon - h\sigma}\right| \delta_*$. If $\varphi(t)$ is continuous and $\varphi(t) \in [-\delta, \delta]$ for all $t \in [-a - c\delta, 0]$ then the BE solution to (4.1.1) $\{u_n\}_{n \geqslant 0}$ satisfies $u_n \in [-\delta, \delta]$ for all $n \geqslant 0$.*

Proof. Let $R_{\Theta=1,h}(0) \in (0,1)$. Since $R_{\Theta=1,h}(v)$ is a continuous function of v then it is always possible to find a small enough $\delta_* \in \left(0, \frac{a}{c}\right)$ such that $R_{\Theta=1,h}(v) \in (0,1)$ for all $v \in [-\delta_*, \delta_*]$.

Now consider the sign of $1 + \frac{\mu}{\sigma} \frac{\varepsilon}{\varepsilon - h\mu(1-\beta_v)} \left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{m_v}$. This is always positive if $\mu < 0$. If $\mu > 0$ and this term is negative then $R_{\Theta=1,h}(v) < 0$. So from our choice of δ_*, we must have $1 + \frac{\mu}{\sigma} \frac{\varepsilon}{\varepsilon - h\mu(1-\beta_v)} \left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{m_v} > 0$ for all $v \in [-\delta_*, \delta_*]$.

Suppose it is possible for u_n to leave the interval $[-\delta, \delta]$. Suppose that when this first happens the solution crosses its upper bound. But up to the n-th step the BE solution is still bounded inside $[-\delta, \delta]$. Recall from (4.3.4) that if the history function and the BE solution up to n are bounded inside $[-\delta, \delta]$ and $\mu + \sigma < 0$ then

$$|u_{n+1}| \leqslant \frac{|\varepsilon| + |h\sigma|}{|\varepsilon - h\mu|} \delta = \delta_*$$

This is true whether or not there is overlapping. Then we have the case $u_n \leqslant \delta \leqslant u_{n+1} \leqslant \delta_*$. Using $L = \delta$ in Lemma 4.4.4,

$$u_{n+1} \leqslant R_{\Theta=1,h}(u_{n+1}) \delta.$$

But $u_{n+1} \geqslant \delta$ so this means, $1 \leqslant R_{\Theta=1,h}(u_{n+1})$. But since $u_{n+1} \in [0, \delta_*]$ then we must have $R_{\Theta=1,h}(u_{n+1}) \in (0,1)$ by our choice of δ_*. This is a contradiction and so the BE solution cannot leave the interval $[-\delta, \delta]$ through the upper bound. Similarly, assuming that the BE solution leaves through the lower bound also leads to a contradiction. Thus, the BE solution cannot exit the interval $[-\delta, \delta]$. □

Theorem 4.4.6. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let the stepsize h be such that $\varepsilon - h\mu > 0$ (no restriction if $\mu < 0$, automatically satisfied for $\mu > 0$ when $h \in (0, a]$). Let the model parameters and the stepsize h satisfy $R_{\Theta=1,h}(0) \in (0,1)$. Then there exists $\delta \in \left(0, \frac{a}{c}\right)$ such that if the history function $\varphi(t)$ is continuous and $\varphi(t) \in [-\delta, \delta]$ for all $t \in [-a - c\delta, 0]$ then the BE solution to (4.1.1) $\{u_n\}_{n \geqslant 0}$ satisfies $u_n \in [-\delta, \delta]$ for all $n \geqslant 0$ and $\lim_{n \to \infty} u_n = 0$.*

Remark 4.4.7. Note that if we had the case $c \to 0$ then we can let $\delta \to \infty$ which means global stability of the numerical solution in the region where the parameter satisfy $R_{\Theta=1,h}(0) \in (0,1)$.

Proof. Let $R_{\Theta=1,h}(0) \in (0,1)$. Choose δ as in Lemma 4.4.5 so that by setting $\varphi(t) \in [-\delta,\delta]$ for all $t \in [-a-c\delta, 0]$ then $u_n \in [-\delta,\delta]$ for $n \geqslant 0$. This choice of δ also yields $R_{\Theta=1,h}(v) \in (0,1)$ for all $v \in [-\delta,\delta]$. Then if we define $r = \max\limits_{v\in[-\delta,\delta]} R_{\Theta=1,h}(v)$, then $r \in (0,1)$. Since $u_n \in [-\delta,\delta]$ for all $n \geqslant 0$, then $\alpha(t_{n+1}, u_{n+1}) \in [t_{n+1} - a - c\delta, t_{n+1}]$ for all n. We call the values of the BE solution (including the continuous extension) on this interval the relevant history at step n.

Since the solutions are bounded then by Lemma 4.4.1, the BE solution may (A) eventually go to zero monotonically from above, or (B) eventually go to zero monotonically from below, or (C) it may display oscillations. To complete the proof of this theorem we only have to prove that in the oscillating case we still get $u_n \to 0$.

Suppose that the BE solution attains a local minimum at $n = S_1$ and let $u_{S_1} = L_1$. Then by Lemma 4.4.4,

$$L_1 \geqslant -R_{\Theta=1,h}(L_1)\,\delta \geqslant -r\delta.$$

So $-r\delta$ is a new lower bound on the solution. Let $\bar{N} = \left\lceil \frac{a+c\delta}{h} \right\rceil$. For $n \geqslant S_1 + \bar{N}$, the relevant history will be on the time interval $\left[t_{S_1}, t_{S_1+\bar{N}}\right]$ so $-r\delta$ is a lower bound on the relevant history. Suppose now that a local maximum occurs at $n = R_1$ and with $u_{R_1} = M_1$. Then by Lemma 4.4.4,

$$M_1 \leqslant -R_{\Theta=1,h}(M_1)(-r\delta) \leqslant r^2\delta.$$

For $n \geqslant R_1 + \bar{N}$ steps, the relevant history of the BE solution is bounded above by $r^2\delta$. Starting with these definitions of S_1 and R_1, for $i \geqslant 2$ define S_i to be the location of the first local minimum past $R_{i-1} + \bar{N}$, and define R_i to be the location of the first local maximum past $S_i + \bar{N}$. Since we are assuming that the solutions behaves as described in (C) in Lemma 4.4.1, S_i and R_i exist for all $i \geqslant 0$. By iteratively applying Lemma 4.4.4, we get that for all $n \geqslant S_1$,

$$\begin{aligned} u_n &\in \left[-r^i\delta, r^{i-1}\delta\right], \quad \text{if } n = S_i, ..., R_i, \\ u_n &\in \left[-r^i\delta, r^i\delta\right], \qquad \text{if } n = R_i, ..., S_{i+1}. \end{aligned}$$

Since $r \in (0,1)$ then $u_n \to 0$. □

Theorem 4.4.6 gives stepsize-dependent analytic expressions for regions where the BE solution with small enough history functions converge to zero. These regions can be written as the sets $\{R_{\Theta=1,h}(0) \in (0,1)\}$. Sample plots of these sets are shown in Figure 4–4. From these figures we see that there is a region contained in all these sets independent of stepsize. This region is actually $(\mu, \sigma) : \{r(0) \in (0,1), \mu < 0\}$, part of the region in which we proved the asymptotic stability of the zero solution of (4.1.1) in Theorem 4.4.9.

Lemma 4.4.8. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < 0$ and $\sigma < -\mu$. Let the stepsize h be such that $\varepsilon - h\mu > 0$ (no restriction if $\mu < 0$, automatically satisfied for $\mu > 0$ when $h \in (0, a]$). Then for all $v \geqslant -\frac{a}{c}$ it holds that $R_{\Theta=1,h}(v) \leqslant r(v)$.*

Proof. Since $1 + x \leqslant e^x$ for all $x \in \mathbb{R}$,

$$\frac{\varepsilon}{\varepsilon - h\mu(1-\beta)} = \left(1 - \frac{h\mu(1-\beta)}{\varepsilon}\right)^{-1} \geqslant e^{\frac{h\mu(1-\beta)}{\varepsilon}},$$

$$\left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{m} = \left(1 - \frac{h\mu}{\varepsilon}\right)^{-m} \geqslant e^{\frac{h\mu m}{\varepsilon}}.$$

Thus,

$$\frac{\varepsilon}{\varepsilon - h\mu(1-\beta)}\left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{m} \geqslant e^{\frac{h\mu(m+1-\beta)}{\varepsilon}} = e^{\frac{\mu(a+cv)}{\varepsilon}},$$

Since $\mu < 0$,

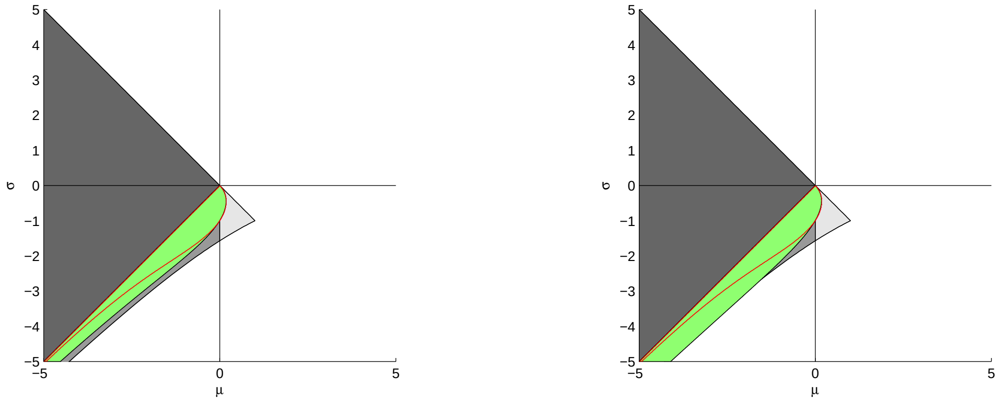
$$\frac{\sigma}{\mu}\left[\frac{1 - \frac{\varepsilon}{\varepsilon - h\mu(1-\beta)}\left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{m}}{1 + \frac{\mu}{\sigma}\frac{\varepsilon}{\varepsilon - h\mu(1-\beta)}\left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{m}}\right] \leqslant \frac{\sigma}{\mu}\left[\frac{1 - e^{\frac{\mu}{\varepsilon}(a+cv)}}{1 + \frac{\mu}{\sigma}e^{\frac{\mu}{\varepsilon}(a+cv)}}\right],$$

which completes the proof of the lemma. □

Theorem 4.4.9. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < 0$ and $\sigma < -\mu$. Let the model parameters satisfy $r(0) \in (0,1)$. Then there exists $\delta \in \left(0, \frac{a}{c}\right)$ such that for all stepsizes $h > 0$, if $\varphi(t) \in [-\delta, \delta]$ for $t \in [-a - c\delta, 0]$ then the BE solution to (4.1.1) $\{u_n\}_{n \geqslant 0}$ satisfies $u_n \in [-\delta, \delta]$ and $\lim_{n \to \infty} u_n = 0$.*

Proof. If $\mu < 0$ then $R_{\Theta=1,h}(v) > 0$ for all $v \geqslant -\frac{a}{c}$. By Lemma 4.4.8, $R_{\Theta=1,h}(0) \leqslant r(0) < 1$. Thus Theorem 4.4.6 applies for any stepsize $h > 0$. □

The stepsize-independent stability region $\{r(0) \in (0,1), \mu < 0\}$ given by Theorem 4.4.9 is confined to the left half-plane so we still do not have a stepsize independent stability region in $\overset{c}{\Sigma}$. In the next section we adapt the Razumikhin-style argument we had from Section 3.4.1 to prove that in a larger parameter region we have a discrete version of Lyapunov stability, and in Section 4.6 it is proven that $u_n \to 0$ in this case when $\mu < 0$ and $h > a$. Although the results



(a) $h = 0.3$ (b) $h = 0.5$

Figure 4–4: The set $\{R_{\Theta=1,h}(0) \in (0,1)\}$ is shaded green and plotted with $\varepsilon = a = c = 1$. This figure also shows the boundary of the set $\{r(0) \in (0,1)\}$ (in red) which is contained inside $\{R_{\Theta=1,h}(0) \in (0,1)\}$ when $\mu < 0$.

may be extended later on to prove convergence to zero for all $h > 0$, currently we have only proven convergence to zero for all stepsizes in this section, using the Gronwall argument.

4.5 Stability of BE using a Razumikhin-style proof

Recall that in Section 3.4.1 (Theorems 3.4.7 and 3.4.10) we used a Razumikhin-style proof to show that for every $\delta > 0$, if $(\mu, \sigma) \in \{P(\delta, c, 2) < \delta\}$ and the history function is small enough then the solution to the model problem (4.1.1) $u(t) \in [-\delta, \delta]$ for all $t \geqslant 0$. Also, if $(\mu, \sigma) \in \{P(1, 0, 2) < 1\}$ then the zero solution to (4.1.1) is Lyapunov stable. From Theorem 3.4.10, the function P can be written as

$$P(\delta, c, 2) = \begin{cases} \mathcal{I}_1\left(-\frac{\mu}{\sigma}\delta, \delta\right), & \text{if } \tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}, \\ \mathcal{I}_2\left(-\frac{\mu}{\sigma}\delta, \delta\right), & \text{if } \tau_1 D - 1 > -\frac{\mu}{\sigma}, \end{cases} \tag{4.5.1}$$

$$\mathcal{I}_1(\hat{u}, \delta) = \hat{u}\left[e^{\frac{\mu}{\varepsilon}\tau_1} + \frac{\sigma}{\mu}\left(e^{\frac{\mu}{\varepsilon}\tau_1} - 1\right)\right] + \frac{\sigma}{\mu}D\delta\left[\frac{\varepsilon}{\mu}\left(e^{\frac{\mu}{\varepsilon}\tau_1} - 1\right) - \tau_1 e^{\frac{\mu}{\varepsilon}\tau_1}\right],$$

$$\mathcal{I}_2(\hat{u}, \delta) = \hat{u}\left(e^{\frac{\mu}{\varepsilon}\tau_1} - \frac{\sigma}{\mu}\right) + \frac{\sigma}{\mu}\delta\left[\frac{\varepsilon D}{\mu}\left(e^{\frac{\mu}{\varepsilon}\frac{\delta+\hat{u}}{D\delta}} - 1\right) - e^{\frac{\mu}{\varepsilon}\tau_1}\right],$$

where $\tau_1 = a + |c|\delta$ and $D = \frac{|\mu|+|\sigma|}{\varepsilon}\left(1 + \frac{|\mu|+|\sigma|}{\varepsilon}\delta\,|c|\right)$.

In this section the Razumikhin-style proof of Lyapunov stability are adapted to prove that the BE solution to (4.1.1) satisfies similar properties. Let $c \in \mathbb{R}$, but set $\varepsilon, a > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Automatically, $\sigma < 0$. These restrictions are always satisfied by points in the delay dependent stability region $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$ of (4.1.1). Given a constant step-size h, backward Euler applied to (4.1.1) is given by

$$u_{n+1} = u_n + \frac{h}{\varepsilon}\left(\mu u_{n+1} + \sigma \tilde{Y}^{(1)}_{n+1}\right).$$

where $\tilde{Y}_{n+1}$ is the value of the linear interpolation η at $t_{n+1} - a - cu_{n+1}$. Rearranging yields

$$u_{n+1} = \frac{\varepsilon}{\varepsilon - h\mu} u_n + \frac{h\sigma}{\varepsilon - h\mu} \tilde{Y}^{(1)}_{n+1}. \tag{4.5.2}$$

Here we only consider the case when $\varepsilon - h\mu > 0$ to keep $\frac{\varepsilon}{\varepsilon - h\mu} > 0$. This is always true if $\mu \leqslant 0$ and it provides a step-size restriction otherwise. But the stability domain of the test equation requires $\mu < \frac{\varepsilon}{a}$ so in the case $\mu > 0$ choosing $h \in (0, a]$ is sufficient.

Lemma 4.5.1. *Let $\varepsilon, a > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$, $\sigma < -\mu$. Let the stepsize h be such that $\varepsilon - h\mu > 0$ (no restriction if $\mu < 0$, automatically satisfied for $\mu > 0$ when $h \in (0, a]$). Define $\|\varphi\| = \sup_{s \leqslant 0} |\varphi(s)|$. If $\|\varphi\| < \left|\frac{a}{c}\right|$ (no restriction if $c = 0$) then*

$$|u_n| \leqslant \left(\frac{\varepsilon - h\sigma}{\varepsilon - h\mu}\right)^n \|\varphi\|.$$

Proof. The proof is by strong induction. For the case $n = 0$ this is obviously true. Suppose for all $i = 1, ..., n$ we have

$$|u_i| \leqslant \left(\frac{\varepsilon - h\sigma}{\varepsilon - h\mu}\right)^i \|\varphi\|.$$

Before going on to the $n + 1$ case, consider the term $\frac{\varepsilon - h\sigma}{\varepsilon - h\mu}$. Since we are interested in $\sigma \leqslant \mu$ then $\varepsilon - h\sigma \geqslant \varepsilon - h\mu$. So $\frac{\varepsilon - h\sigma}{\varepsilon - h\mu} \geqslant 1$. So in fact, for all $i = 1, ..., n$

$$|u_i| \leqslant \left(\frac{\varepsilon - h\sigma}{\varepsilon - h\mu}\right)^n \|\varphi\|.$$

Now look at the $n + 1$ case with no overlapping ($t_{n+1} - a - cu_{n+1} \leqslant t_n$). From (4.5.2),

$$\begin{aligned} |u_{n+1}| &\leqslant \frac{\varepsilon}{\varepsilon - h\mu} |u_n| + \frac{h|\sigma|}{\varepsilon - h\mu} \left|\tilde{Y}^{(1)}_{n+1}\right|, \\ &\leqslant \frac{\varepsilon}{\varepsilon - h\mu} \left(\frac{\varepsilon - h\sigma}{\varepsilon - h\mu}\right)^n \|\varphi\| - \frac{h\sigma}{\varepsilon - h\mu} \left(\frac{\varepsilon - h\sigma}{\varepsilon - h\mu}\right)^n \|\varphi\|, \\ &\leqslant \left(\frac{\varepsilon - h\sigma}{\varepsilon - h\mu}\right)^{n+1} \|\varphi\|. \end{aligned}$$

The second to the last step is because the continuous extension is always a convex combination of stages and we are assuming that there is no overlapping. Now consider the $n+1$ case with overlapping. Recall from Lemma 4.2.1 that we may always choose our BE solution such that u_{n+1} is bounded away from $\frac{a}{c}$. Then $t_n < t_{n+1} - a - cu_{n+1} \leqslant t_{n+1}$. Write $t_{n+1} - a - cu_{n+1} = t_n + \beta h$ where $\beta \in [0,1]$. Then,

$$u_{n+1} = u_n + \frac{h}{\varepsilon}\left[\mu u_{n+1} + \sigma\left(\left(1-\beta\right)u_n + \beta u_{n+1}\right)\right]$$

$$\left(1 - \frac{h\mu}{\varepsilon} - \frac{h\sigma\beta}{\varepsilon}\right)u_{n+1} = \left(1 + \frac{h\sigma\left(1-\beta\right)}{\varepsilon}\right)u_n$$

$$u_{n+1} = \left(\frac{\varepsilon + h\sigma\left(1-\beta\right)}{\varepsilon - h\mu - h\sigma\beta}\right)u_n$$

$$|u_{n+1}| = \left|\frac{\varepsilon + h\sigma\left(1-\beta\right)}{\varepsilon - h\mu - h\sigma\beta}\right||u_n| \leqslant \left(\frac{\varepsilon - h\sigma}{\varepsilon - h\mu}\right)|u_n| \leqslant \left(\frac{\varepsilon - h\sigma}{\varepsilon - h\mu}\right)^{n+1}\|\varphi\|.$$

This completes the proof by strong induction. □

By Lemma 4.5.1, for every $\delta_1 > 0$ it is always possible to bound a finite segment of the BE solution by δ_1 by bounding the history function by an appropriate δ_2. If we can prove that past that finite segment it is not possible to have $u_{n+1} > \delta_1 \geqslant u_n$ or $u_{n+1} < -\delta_1 \leqslant u_n$ then the BE solution must always remain bounded inside δ_1. This is the discrete version of the Razumikhin-style proof of stability used in Section 3.4.1.

Let $\delta_1 \in \left(0, \left|\frac{a}{c}\right|\right)$, $\delta \in \left(\delta, \left|\frac{a}{c}\right|\right)$, $M = \left\lceil\frac{a+|c|\delta}{h}\right\rceil$ and $\delta_2 = \delta_1\left(\frac{\varepsilon-h\sigma}{\varepsilon-h\mu}\right)^{-2M}$. If $\varphi\left(t\right) \in [-\delta_2, \delta_2]$ for $t \leqslant 0$ then by Lemma 4.5.1, the segment of the BE solution to the model problem (4.1.1) $\{u_n\}_{n\geqslant 0}$ from $n = 0$ to $2M$ must satisfy $u_n \in [-\delta_1, \delta_1]$ for $n = 0, ..., 2M$. Because of this bound on the numerical solution then we actually only need $\varphi\left(t\right) \in [-\delta_2, \delta_2]$ for $t \in [-a - |c|\delta_1, 0]$.

We now look for parameter regions for which the BE solution cannot exit $[-\delta_1, \delta_1]$. Suppose that the BE solution $\{u_n\}_{n\geqslant 0}$ escapes the interval for the first time through the upper bound at the $(n+1)$-st step for some $n \geqslant 2M$. Then $u_{n+1} > \delta_1 \geqslant u_n$. In the parameter region where we can obtain a contradiction to this assumption then BE solution must remain inside $[-\delta, \delta]$. Suppose $u_{n+1} = \delta$. Since $u_{n+1} > u_n$ then we get the following condition on $\tilde{Y}^{(1)}_{n+1}$

$$\mu u_{n+1} + \sigma\tilde{Y}^{(1)}_{n+1} > 0 \quad \Rightarrow \quad \tilde{Y}^{(1)}_{n+1} < -\frac{\mu}{\sigma}u_{n+1} = -\frac{\mu}{\sigma}\delta. \tag{4.5.3}$$

As in the previous section, define

$$m = \left\lfloor \frac{a + cu_{n+1}}{h} \right\rfloor = \left\lfloor \frac{a + c\delta}{h} \right\rfloor, \qquad \beta = m + 1 - \frac{a + c\delta}{h} \tag{4.5.4}$$

The definitions of m and β were chosen so that $t_{n+1} - a - c\delta = (1 - \beta) t_{n-m} + \beta t_{n-m+1}$. Thus because we are using linear interpolation we also have,

$$\tilde{Y}_{n+1}^{(1)} = (1 - \beta) u_{n-m} + \beta u_{n-m+1} \tag{4.5.5}$$

Since $n \geqslant 2M$ this implies that for $i = n - m, .., n$,

$$t_{i+1} - a - cu_{n+1} \in [t_{n-m+1} - a - |c| \delta, t_{n+1} - a + |c| \delta] \subseteq [0, t_{n+1}].$$

This means that $\tilde{Y}_{i+1}^{(1)} = \eta (t_{i+1} - a - cu_{i+1})$ must have properties stemming from the properties of the continuous extension on $[0, t_{n+1}]$. We derive some of these properties first and see later on why these properties are important.

The change from $\tilde{Y}_i^{(1)}$ to $\tilde{Y}_{i+1}^{(1)}$ depends on the maximum change in the mesh values in a single step. Since δ_1 is a bound on the history function and $\{u_i\}_{i=0}^n$ then δ is also a bound on the history function and $\{u_i\}_{i=0}^n$. Since $u_{n+1} = \delta$ then it is also a bound on the continuous extension $\eta(t)$ for $t \in [0, t_{n+1}]$. As a result, for $i = 0, ..., n$,

$$|u_{i+1} - u_i| \leqslant \frac{|\mu u_{i+1}| + \left|\sigma \tilde{Y}_{i+1}^{(1)}\right|}{\varepsilon} h \leqslant \frac{|\mu| + |\sigma|}{\varepsilon} \delta h.$$

The quantity $\left|\tilde{Y}_{i+1}^{(1)} - \tilde{Y}_i^{(1)}\right|$ is bounded by the number of time steps between $t_{i+1} - a - cu_{i+1}$ and $t_i - a - cu_i$ (including fractions of a step since we are working with linear interpolation), multiplied by the maximum change in the mesh values in a single step.

$$\begin{aligned} \left|\tilde{Y}_{i+1}^{(1)} - \tilde{Y}_i^{(1)}\right| &\leqslant \left|\frac{(t_{i+1} - a - cu_{i+1}) - (t_i - a - cu_i)}{h}\right| \frac{|\mu| + |\sigma|}{\varepsilon} \delta h, \\ &\leqslant \left(1 + \frac{|u_{i+1} - u_i|}{h} |c|\right) \frac{|\mu| + |\sigma|}{\varepsilon} \delta h, \\ &\leqslant \left(1 + \frac{|\mu| + |\sigma|}{\varepsilon} |c| \delta\right) \frac{|\mu| + |\sigma|}{\varepsilon} \delta h, \\ &= D\delta h, \end{aligned} \tag{4.5.6}$$

where $D = \left(1 + \frac{|\mu|+|\sigma|}{\varepsilon} |c| \delta\right) \frac{|\mu|+|\sigma|}{\varepsilon}$ as in Section 3.4.1.

Lemma 4.5.2. *If $A > 0$ and $u_{n+1} = Au_n + v_n$ then*

$$u_{n+1} = A^{m+1} u_{n-m} + \sum_{i=n-m}^{n} A^{n-i} v_i$$

Proof. Fix n and use induction on m. □

Writing (4.5.2) in the form of Lemma 4.5.2 then $u_{n+1} = Au_n + v_n$ with

$$A = \frac{\varepsilon}{\varepsilon - h\mu}, \quad v_i = \frac{h\sigma}{\varepsilon - h\mu} \tilde{Y}^{(1)}_{n+1}. \tag{4.5.7}$$

Applying Lemma 4.5.2 yields $u_{n+1} = A^{m+1} u_{n-m} + \sum\limits_{i=n-m}^{n} A^{n-i} v_i$. We would like to maximize the right hand side of this equation in order to get a bound on the value of u_{n+1}. To do this we need to make the sequence $\{v_i\}$ as large as possible. This is done by using the most negative possible sequence of $\tilde{Y}_{i+1}$ given a fixed value of $\tilde{Y}_{n+1}$. Let this sequence be $\{w_i\}$. Using the bounds on the solution and the restrictions (4.5.3) and (4.5.6), define $\{w_i\}$ to be

$$w_i = \begin{cases} -\delta, & n - m \leqslant i \leqslant n - \ell, \\ \hat{u} - (n-i)\, D\delta h, & n - \ell + 1 \leqslant i \leqslant n, \end{cases} \tag{4.5.8}$$

where $\hat{u} = \tilde{Y}^{(1)}_{n+1} \in \left[-\delta, -\frac{\mu}{\sigma}\delta\right]$ and

$$\ell = \left\lceil \frac{\hat{u} + \delta}{D\delta h} \right\rceil, \qquad \chi = \ell - \frac{\hat{u} + \delta}{D\delta h}. \tag{4.5.9}$$

Set $\tilde{v}_i = \frac{h\sigma}{\varepsilon - h\mu} w_i$. Then using Lemma 4.5.2, we derive

$$u_{n+1} = A^{m+1} u_{n-m} + \sum_{i=n-m}^{n} A^{n-i} \tilde{v}_i \leqslant A^{m+1} u_{n-m} + \sum_{i=n-m}^{n} A^{n-i} \tilde{v}_i, \tag{4.5.10}$$

$$u_{n+1} = A^{m} u_{n-m+1} + \sum_{i=n-m+1}^{n} A^{n-i} \tilde{v}_i \leqslant A^{m} u_{n-m+1} + \sum_{i=n-m+1}^{n} A^{n-i} \tilde{v}_i. \tag{4.5.11}$$

We performed two summations so that we can combine the u_{n-m} and u_{n-m+1} terms using $\hat{u} = (1-\beta)\, u_{n-m} + \beta u_{n-m+1}$ from (4.5.5). Take $\frac{(1-\beta)}{A}\times$ (4.5.10) plus $\beta\times$ (4.5.11),

$$\left(\frac{1-\beta}{A} + \beta\right) u_{n+1} \leqslant \left[(1-\beta)\, u_{n-m} + \beta u_{n-m+1}\right] A^m$$

$$+ \frac{1-\beta}{A} \sum_{i=n-m}^{n} A^{n-i}\tilde{v}_i + \beta \sum_{i=n-m+1}^{n} A^{n-i}\tilde{v}_i,$$

$$\left(\frac{1-\beta}{A} + \beta\right) u_{n+1} = \hat{u}A^m + \frac{1-\beta}{A} \sum_{i=n-m}^{n} A^{n-i}\tilde{v}_i + \beta \sum_{i=n-m+1}^{n} A^{n-i}\tilde{v}_i. \tag{4.5.12}$$

From this equation we would like to derive discrete versions of the expressions $\mathcal{I}_1$ and $\mathcal{I}_2$. Recall from Section 3.4.1 that $\mathcal{I}_2$ is used when the $\eta_{(2)}$ function over the integration interval can be split between the flat part at $-\delta$ and a line with slope $D\delta$ (corresponding to the increasing part when $i = n-\ell+1, ..., n$ in the definition of $\{w_i\}$). The expression $\mathcal{I}_1$ is used when the function to be integrated consists of only the straight line with slope $D\delta$. We now define $\mathcal{S}_1$ and $\mathcal{S}_2$ where $\mathcal{S}_2$ is used when the $\{w_i\}$ sequence consists of a flat part and then an increasing part (occurs when $\ell \leqslant m$), and $\mathcal{S}_1$ is used when the $\{w_i\}$ sequence consists of only the increasing part (occurs when $\ell > m$). As in the derivation of $\mathcal{I}_1$ and $\mathcal{I}_2$ we first assume $\mu \neq 0$.

Deriving $\mathcal{S}_2$, the discrete version of $\mathcal{I}_2$.

Consider the case when $\ell \leqslant m$.

$$\left\lceil \frac{\delta + \hat{u}}{D\delta h} \right\rceil \leqslant m \quad \Rightarrow \quad \hat{u} \leqslant (mhD-1)\,\delta \leqslant (\tau_2 D - 1)\,\delta$$

In this case $\hat{u} \in \left[-\delta, \max\left\{-\frac{\mu}{\sigma}\delta, (mhD-1)\,\delta\right\}\right]$. Consider the last term in (4.5.12),

$$\sum_{i=n-m}^{n} A^{n-i}\tilde{v}_i = \frac{h\sigma}{\varepsilon - h\mu}\left[-\sum_{i=n-m}^{n-\ell} A^{n-i}\delta + \sum_{i=n-\ell+1}^{n} A^{n-i}\left(\hat{u} - (n-i)\,D\delta h\right)\right],$$

$$= \frac{h\sigma}{\varepsilon - h\mu}\left[\frac{A^{m+1} - A^{\ell}}{1-A}\delta + \sum_{i=n-\ell+1}^{n} A^{n-i}\left(\hat{u} - (n-i)\,D\delta h\right)\right]. \tag{4.5.13}$$

The second term in (4.5.13) can be written as

$$\sum_{i=n-\ell+1}^{n} A^{n-i}\left(\hat{u} - (n-i)\,D\delta h\right) = (\hat{u} - nD\delta h)\,\frac{1-A^{\ell}}{1-A} + D\delta h A^n \sum_{i=n-\ell+1}^{n} i\left(\frac{1}{A}\right)^i. \tag{4.5.14}$$

Using the formula $\sum\limits_{i=0}^{n} ix^i = \frac{x-(n+1)x^{n+1}+nx^{n+2}}{(x-1)^2}$, the summation term at the end of (4.5.14) simplifies to

$$\sum_{i=n-\ell+1}^{n} i\left(\frac{1}{A}\right)^i = \frac{\left(\frac{1}{A}-\frac{n+1}{A^{n+1}}+\frac{n}{A^{n+2}}\right)-\left(\frac{1}{A}-\frac{n-\ell+1}{A^{n-\ell+1}}+\frac{n-\ell}{A^{n-\ell+2}}\right)}{\left(\frac{1}{A}-1\right)^2},$$
$$= \frac{1}{(1-A)A^n}\left[n\left(1-A^\ell\right)+\ell A^{\ell+1}+\frac{A\left(A^\ell-1\right)}{1-A}\right].$$

Substitute this back to (4.5.14) and simplifying yields

$$\sum_{i=n-\ell+1}^{n} A^{n-i}\left(\hat{u}-(n-i)D\delta h\right) = \frac{1-A^\ell}{1-A}\hat{u}+D\delta h\frac{\ell A^\ell}{1-A}+D\delta h\frac{A\left(A^\ell-1\right)}{(1-A)^2}. \tag{4.5.15}$$

Now substitute this back to (4.5.13)

$$\sum_{i=n-m}^{n} A^{n-i}\tilde{v}_i = \frac{h\sigma}{\varepsilon-h\mu}\left[\frac{A^{m+1}-A^\ell}{1-A}\delta+\frac{1-A^\ell}{1-A}\hat{u}+D\delta h\frac{\ell A^\ell}{1-A}+D\delta h\frac{A\left(A^\ell-1\right)}{(1-A)^2}\right],$$
$$= \frac{h\sigma}{(\varepsilon-h\mu)(1-A)}\left[-(\hat{u}+\delta)A^\ell+\delta A^{m+1}+\hat{u}+D\delta h\ell A^\ell+D\delta h\frac{A\left(A^\ell-1\right)}{1-A}\right]. \tag{4.5.16}$$

Since $A=\frac{\varepsilon}{\varepsilon-h\mu}$ then $\frac{h\sigma}{(\varepsilon-h\mu)(1-A)}=-\frac{\sigma}{\mu}$ and $\frac{A}{1-A}=-\frac{\varepsilon}{h\mu}$. Using this expression in (4.5.16) simplifies to

$$\sum_{i=n-m}^{n} A^{n-i}\tilde{v}_i = \frac{\sigma}{\mu}\left[(\hat{u}+\delta)A^\ell-\delta A^{m+1}-\hat{u}-D\delta h\ell A^\ell+\frac{\varepsilon D\delta}{\mu}\left(A^\ell-1\right)\right].$$

From the definition of ℓ and χ in (4.5.9), $\hat{u}+\delta-D\delta h\ell=-D\delta h\chi$. Thus,

$$\sum_{i=n-m}^{n} A^{n-i}\tilde{v}_i = \frac{\sigma}{\mu}\left[-D\delta h\chi A^\ell-\delta A^{m+1}-\hat{u}+\frac{\varepsilon D\delta}{\mu}\left(A^\ell-1\right)\right],$$
$$= \frac{\sigma}{\mu}\left[\frac{\varepsilon D\delta}{\mu}\left(\left(1-\frac{h\mu\chi}{\varepsilon}\right)A^\ell-1\right)-\hat{u}-\delta A^{m+1}\right]. \tag{4.5.17}$$

Similarly, if we performed the summation from $n-m+1$ to n instead, we derive

$$\sum_{i=n-m+1}^{n} A^{n-i}\tilde{v}_i = \frac{\sigma}{\mu}\left[\frac{\varepsilon D\delta}{\mu}\left(\left(1-\frac{h\mu\chi}{\varepsilon}\right)A^\ell-1\right)-\hat{u}-\delta A^{m}\right]. \tag{4.5.18}$$

Note that this expression works even in the boundary case $\ell = m$. The summation term in (4.5.12) can now be written as

$$\frac{1-\beta}{A}\sum_{i=n-m}^{n} A^{n-i}\tilde{v}_i + \beta \sum_{i=n-m+1}^{n} A^{n-i}\tilde{v}_i$$
$$= \frac{\sigma}{\mu}\left[\left(\frac{1-\beta}{A}+\beta\right)\left[\frac{\varepsilon D\delta}{\mu}\left(\left(1-\frac{h\mu\chi}{\varepsilon}\right)A^{\ell}-1\right)-\hat{u}\right]-\delta A^{m}\right]$$

Using $\frac{1-\beta}{A}+\beta = (1-\beta)\frac{\varepsilon-h\mu}{\varepsilon}+\beta = \frac{\varepsilon-h\mu(1-\beta)}{\varepsilon}$ and solving for u_{n+1} in (4.5.12) we get

$$u_{n+1} \leqslant \hat{u}\left(A^{*}-\frac{\sigma}{\mu}\right)+\frac{\sigma}{\mu}\delta\left[\frac{\varepsilon D}{\mu}\left(\left(1-\frac{h\mu\chi}{\varepsilon}\right)A^{\ell}-1\right)-A^{*}\right]$$

where $A^{*} = \frac{\varepsilon}{\varepsilon-h\mu(1-\beta)}A^{m}$. This expression gives a discrete version of $\mathcal{I}_2$. We summarize what we have done so far in Lemma 4.5.4.

Definition 4.5.3. Let $\varepsilon, a > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let the stepsize h be such that $\varepsilon - h\mu > 0$. For $\delta_1 \in \left(0, \left|\frac{a}{c}\right|\right)$, $\delta \in \left(\delta_1, \left|\frac{a}{c}\right|\right)$ define $\tau_1 = a + |c|\delta$, $\tau_2 = a + c\delta$, $M = \left\lceil \frac{\tau_1}{h}\right\rceil$ and $\delta_2 = \delta_1\left(\frac{\varepsilon-h\sigma}{\varepsilon-h\mu}\right)^{-2M}$. As in (4.5.4), (4.5.9), (4.5.7), define

$$m = \left\lfloor \frac{a+cu_{n+1}}{h}\right\rfloor = \left\lfloor\frac{a+c\delta}{h}\right\rfloor, \qquad \beta = m+1-\frac{a+c\delta}{h}, \qquad \ell = \left\lceil\frac{\hat{u}+\delta}{D\delta h}\right\rceil, \qquad \chi = \ell - \frac{\hat{u}+\delta}{D\delta h},$$

$$A = \frac{\varepsilon}{\varepsilon-h\mu}, \qquad A^{*} = \frac{\varepsilon}{\varepsilon-h\mu\,(1-\beta)}A^{m}.$$

Lemma 4.5.4. *Let $\varepsilon, a > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let the stepsize h be such that $\varepsilon - h\mu > 0$. Let $\delta_1 \in \left(0, \left|\frac{a}{c}\right|\right)$, $\delta \in \left(\delta_1, \left|\frac{a}{c}\right|\right)$ and $\hat{u} \in \left[-\delta, \min\{-\frac{\mu}{\sigma}\delta, (mhD-1)\delta\}\right]$. Define τ_1, τ_2, M, δ_2, m, β, ℓ, χ, A and A^{*} as in Definition 4.5.3. If $\varphi(t) \in [-\delta_2, \delta_2]$ for $t \in [-a-|c|\delta, 0]$ then the BE solution $\{u_n\}_{n\geqslant 0}$ satisfies $u_n \in [-\delta_1, \delta_1]$ for $n = 0, ..., 2M$. Suppose that the BE solution escapes the interval for the first time through the upper bound at $u_{n+1} = \delta > \delta_1$, $n \geqslant 2M$ and $\hat{u} = \tilde{Y}^{(1)}_{n+1} = \eta\,(t_{n+1} - a - cu_{n+1})$. Define*

$$\mathcal{S}_2(\hat{u},\delta) = \hat{u}\left(A^{*}-\frac{\sigma}{\mu}\right)+\frac{\sigma}{\mu}\delta\left[\frac{\varepsilon D}{\mu}\left(\left(1-\frac{h\mu\chi}{\varepsilon}\right)A^{\ell}-1\right)-A^{*}\right]. \tag{4.5.19}$$

Then $u_{n+1} \leqslant \mathcal{S}_2(\hat{u},\delta)$. □

When $\mu = 0$ we define $\mathcal{S}_2(\hat{u},\delta)$ as the limit of the right hand side of (4.5.19) as $\mu \to 0$. For brevity we do not consider this case separately.

Deriving $\mathcal{S}_1$, the discrete version of $\mathcal{I}_1$.

Now consider the case when $\ell > m$.

$$\left\lceil \frac{\delta + \hat{u}}{D\delta h} \right\rceil > m \quad \Rightarrow \quad \hat{u} > (mhD - 1)\,\delta$$

Then $\hat{u} \in \left[(mhD-1)\,\delta, -\frac{\mu}{\sigma}\delta\right]$. Consider the last term in (4.5.12). The summation does not have to be split up in this case. Using the same steps as in the derivation of (4.5.15) we derive

$$\begin{aligned}\sum_{i=n-m}^{n} A^{n-i}\tilde{v}_i &= \frac{h\sigma}{\varepsilon - h\mu}\sum_{i=n-m}^{n} A^{n-i}\left(\hat{u} - (n-i)\,D\delta h\right),\\ &= \frac{h\sigma}{(\varepsilon - h\mu)(1-A)}\left[\left(1 - A^{m+1}\right)\hat{u} + D\delta h\,(m+1)\,A^{m+1} + D\delta h\frac{A\left(A^{m+1}-1\right)}{1-A}\right],\\ &= \frac{\sigma}{\mu}\left[\left(A^{m+1}-1\right)\hat{u} - D\delta h\,(m+1)\,A^{m+1} + \frac{\varepsilon D\delta}{\mu}\left(A^{m+1}-1\right)\right].\end{aligned}$$

Similarly, if the summation were performed from $n-m+1$ to n this yields

$$\sum_{i=n-m+1}^{n} A^{n-i}\tilde{v}_i = \frac{\sigma}{\mu}\left[\left(A^{m}-1\right)\hat{u} - D\delta hmA^{m} + \frac{\varepsilon D\delta}{\mu}\left(A^{m}-1\right)\right].$$

Combining these two terms as in (4.5.12),

$$\begin{aligned}&\frac{1-\beta}{A}\sum_{i=n-m}^{n} A^{n-i}\tilde{v}_i + \beta\sum_{i=n-m+1}^{n} A^{n-i}\tilde{v}_i,\\ &= \frac{\sigma}{\mu}\left[-\left(\frac{1-\beta}{A}+\beta\right)\left(\hat{u}+\frac{\varepsilon D\delta}{\mu}\right) + \hat{u}A^{m} - D\delta h\,(m+1-\beta)\,A^{m} + \frac{\varepsilon D\delta}{\mu}A^{m}\right],\\ &= \frac{\sigma}{\mu}\left[-\left(\frac{1-\beta}{A}+\beta\right)\left(\hat{u}+\frac{\varepsilon D\delta}{\mu}\right) + \left(\hat{u} - D\delta\tau_2 + \frac{\varepsilon D\delta}{\mu}\right)A^{m}\right].\end{aligned}$$

Solving for u_{n+1} in (4.5.12) and using $A^* = \frac{\varepsilon}{\varepsilon - h\mu(1-\beta)}A^m$ we get

$$\begin{aligned}u_{n+1} &\leqslant \hat{u}A^* + \frac{\sigma}{\mu}\left[-\hat{u} - \frac{\varepsilon D\delta}{\mu} + \left(\hat{u} - D\delta\tau_2 + \frac{\varepsilon D\delta}{\mu}\right)A^*\right]\\ &= \hat{u}\left[A^* + \frac{\sigma}{\mu}\left(A^*-1\right)\right] + \frac{\sigma}{\mu}D\delta\left[\frac{\varepsilon}{\mu}\left(A^*-1\right) - \tau_2 A^*\right]\end{aligned}$$

This expression gives a discrete version of $\mathcal{I}_1$.

Lemma 4.5.5. *Let $\varepsilon, a > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let the stepsize h be such that $\varepsilon - h\mu > 0$. Let $\delta_1 \in \left(0, \left|\frac{a}{c}\right|\right)$, $\delta \in \left(\delta_1, \left|\frac{a}{c}\right|\right)$ and $\hat{u} \in \left[(mhD-1)\delta, -\frac{\mu}{\sigma}\delta\right]$. Define τ_1, τ_2, M, δ_2, m, β, ℓ, χ, A and A^* as in Definition 4.5.3. If $\varphi(t) \in [-\delta_2, \delta_2]$ for $t \in [-a-|c|\delta, 0]$ then the*

BE solution $\{u_n\}_{n\geqslant 0}$ *satisfies* $u_n \in [-\delta_1, \delta_1]$ *for* $n = 0, ..., 2M$*. Suppose that the BE solution escapes the interval for the first time through the upper bound at* $u_{n+1} = \delta \geqslant \delta_1$*,* $n \geqslant 2M$ *and* $\hat{u} = \tilde{Y}^{(1)}_{n+1} = \eta\left(t_{n+1} - a - cu_{n+1}\right)$*. Define*

$$\mathcal{S}_1\left(\hat{u}, \delta\right) = \hat{u}\left[A^* + \frac{\sigma}{\mu}\left(A^* - 1\right)\right] + \frac{\sigma}{\mu}D\delta\left[\frac{\varepsilon}{\mu}\left(A^* - 1\right) - \tau_2 A^*\right] \tag{4.5.20}$$

Then $u_{n+1} \leqslant \mathcal{S}_1\left(\hat{u}, \delta\right)$*.* □

When $\mu = 0$ we define $\mathcal{S}_1\left(\hat{u}, \delta\right)$ as the limit of the right hand side of (4.5.20) as $\mu \to 0$. Again, we do not consider this case separately for brevity.

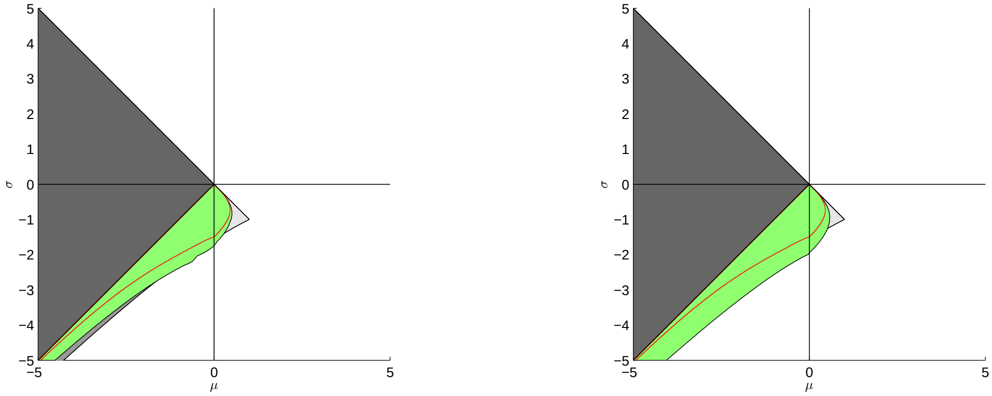
Definition 4.5.6. Let $\varepsilon, a > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let the stepsize h be such that $\varepsilon - h\mu > 0$. Let $\delta_1 \in \left(0, \left|\frac{a}{c}\right|\right)$, $\delta \in \left(\delta_1, \left|\frac{a}{c}\right|\right)$ and $\hat{u} \in \left[(mhD-1)\delta, -\frac{\mu}{\sigma}\delta\right]$. Define

$$\mathcal{S}\left(\Theta = 1, h\right)\left(\hat{u}, \delta, c, 2\right) = \begin{cases} \mathcal{S}_1\left(\hat{u}, \delta\right), & \ell > m \\ \mathcal{S}_2\left(\hat{u}, \delta\right), & \ell \leqslant m, \end{cases}$$

using (4.5.19), (4.5.20) and Definition 4.5.3. Also define the function

$$P_{\Theta=1,h}\left(\delta, c, 2\right) = \sup_{\hat{u} \in \left[-\delta, -\frac{\mu}{\sigma}\delta\right]} \mathcal{S}\left(\Theta = 1, h\right)\left(\hat{u}, \delta, |c|, 2\right).$$

Note that by using $|c|$ instead of c in $\mathcal{S}$ we only need to set τ_2 and m to be equal to τ_1 and M respectively.



(a) $h = 0.3, \delta = 0.1$

(b) $h = 0.5, \delta = 0.1$

Figure 4–5: The sets $\{P_{\Theta=1}\left(\delta, c, 2\right) < \delta\}$ are shaded green and plotted with $\varepsilon = a = c = 1$. This figure also shows the boundary of the set $\{P\left(1, 0, 2\right) < 1\}$ (in red) which is always contained inside $\{P_{\Theta=1}\left(\delta, c, 2\right) < \delta\}$ for small enough δ (Theorem 3.4.10).

Sample plots of the sets $\{P_{\Theta=1,h}(\delta,c,2)<\delta\}$ in the (μ,σ) plane are shown in Figure 4–5. We prove in Lemma 4.5.10 that if $P_{\Theta=1,h}(\delta,c,2)<\delta$ then the BE solution to (4.1.1) cannot exit the interval $[-\delta,\delta]$. Figure 4–5 also shows that the set $\{P(1,0,2)<1\}$ is contained inside $\{P_{\Theta=1,h}(\delta,c,2)<\delta\}$ for small δ. This is proven in Theorem 3.4.10 where we also show that a discrete version of Lyapunov stability holds for $(\mu,\sigma)\in\{P(1,0,2)<1\}$. To prove these results we first need to consider how the sets $\{P_{\Theta=1,h}(\delta,c,2)<\delta\}$ change when we change δ or c.

Lemma 4.5.7. *Let $\varepsilon,a>0$, $\sigma\leqslant\mu<\frac{\varepsilon}{a}$ and $\sigma<-\mu$. Let the stepsize h be such that $\varepsilon-h\mu>0$. Let $\delta\in\left(0,\left|\frac{a}{c}\right|\right)$, $m\in\mathbb{N}$ and $\hat{u}\in\left[-\delta,-\frac{\mu}{\sigma}\delta\right]$ be fixed. The partial derivative of $\mathcal{S}(\Theta=1,h)(\hat{u},\delta,c,2)$ with respect to τ_2 is defined for $\tau_2\in(mh,(m+1)h)$ and it is given by*

$$\frac{\partial}{\partial\tau_2}\mathcal{S}(\tau_2)\equiv\frac{\partial}{\partial\tau_2}\begin{cases}\mathcal{S}_1(\hat{u},\delta), & \ell>m\\ \mathcal{S}_2(\hat{u},\delta), & \ell\leqslant m\end{cases}=\frac{A^*}{\varepsilon-h\mu(1-\beta)}(\mu\hat{u}+\sigma w_{n-m}). \tag{4.5.21}$$

If $\frac{\partial}{\partial\tau_2}\mathcal{S}(\tau_2)\leqslant 0$ then $\mu>0$ and $\mathcal{S}(\Theta=1,h)(\hat{u},\delta,c,2)<\delta$.

Proof. If either $\mu\leqslant 0$ holds or $\mu>0$ and $w_{n-m}\leqslant 0$ holds then it is easy to show that $\mu\hat{u}+\sigma w_{n-m}>0$. So consider the case when $\mu>0$ and $w_{n-m}>0$. Then we must have $\ell\leqslant m$ and $\mathcal{S}(\Theta=1,h)(\hat{u},\delta,c,2)=\mathcal{S}_1(\hat{u},\delta)$. Let $\frac{\partial}{\partial\tau_2}\mathcal{S}(\tau_2)\leqslant 0$ then $\mu\hat{u}+\sigma w_{n-m}=(\mu+\sigma)\hat{u}+\sigma mhD\delta\leqslant 0$. Using this we derive

$$\begin{aligned}\mathcal{S}_1(\hat{u},\delta)&\leqslant\hat{u}\left[A^*+\frac{\sigma}{\mu}(A^*-1)\right]+\frac{\sigma}{\mu}D\delta\frac{\varepsilon}{\mu}(A^*-1)-(\mu+\sigma)\hat{u}\frac{A^*}{\mu}-\sigma D\delta(1-\beta)\\ &=-\hat{u}\frac{\sigma}{\mu}+\delta\frac{\sigma\varepsilon D}{\mu^2}(A^m-1)-(\mu+\sigma)\hat{u}\frac{e^{\frac{\mu\tau_2}{\varepsilon}}}{\mu}\leqslant\delta+\delta\frac{\sigma\varepsilon D}{\mu^2}\left(e^{\frac{\mu}{\varepsilon}\tau_2}-1\right)<\delta.\end{aligned}$$

where the last line is because $\hat{u}\leqslant-\frac{\mu}{\sigma}\delta$ and $\sigma<0$. □

Lemma 4.5.8. *Let $\varepsilon,a>0$, $\sigma\leqslant\mu<\frac{\varepsilon}{a}$ and $\sigma<-\mu$. Let the stepsize h be such that $\varepsilon-h\mu>0$ (no restriction if $\mu\leqslant 0$, automatically satisfied for $\mu>0$ when $h\in(0,a]$). Let $\delta\in\left(0,\left|\frac{a}{c}\right|\right)$. Then $\mathcal{S}(\Theta=1,h)(\hat{u},\delta,-|c|,2)\leqslant\mathcal{S}(\Theta=1,h)(\hat{u},\delta,|c|,2)$ and*

$$\left\{P_{\Theta=1,h}(\delta,c,2)=\sup_{\hat{u}\in\left[-\delta,-\frac{\mu}{\sigma}\delta\right]}\mathcal{S}(\Theta=1,h)(\hat{u},\delta,|c|,2)<\delta\right\}$$
$$\subseteq\left\{\sup_{\hat{u}\in\left[-\delta,-\frac{\mu}{\sigma}\delta\right]}\mathcal{S}(\Theta=1,h)(\hat{u},\delta,-|c|,2)<\delta\right\}.$$

Proof. Similar to the proof of Lemma 3.4.4. We use the continuity of $\mathcal{S}\left(\Theta=1,h\right)\left(\hat{u},\delta,c,2\right)$ with respect to τ_2 even though it is not differentiable at points where $\tau_2 = mh$, $m \in \mathbb{N}$. □

Lemma 4.5.9. *Let $\varepsilon, a > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let the stepsize h be such that $\varepsilon - h\mu > 0$ (no restriction if $\mu \leqslant 0$, automatically satisfied for $\mu > 0$ when $h \in (0,a]$). If $0 < \delta \leqslant \delta_* < \left|\frac{a}{c}\right|$ then $\{P_{\Theta=1,h}\left(\delta_*, c, 2\right) < \delta_*\} \subseteq \{P_{\Theta=1,h}\left(\delta, c, 2\right) < \delta\}$.*

Proof. Similar to the proof of Lemma 3.4.5. □

Lemmas 4.5.9 describes how the sets $\{P_{\Theta=1,h}\left(\delta, c, 2\right) < \delta\}$ change with δ and Lemma 4.5.8 describes how these sets change with the sign of c. These are used along with Lemmas 4.5.4 and 4.5.5 to prove that the BE solution cannot escape $[-\delta, \delta]$ if $P_{\Theta=1,h}\left(\delta, c, 2\right) < \delta$. This result is given in the following lemma.

Lemma 4.5.10. *Let $\varepsilon, a > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let the stepsize h be such that $\varepsilon - h\mu > 0$ (no restriction if $\mu \leqslant 0$, automatically satisfied for $\mu > 0$ when $h \in (0,a]$). Let $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$, $M = \left\lceil\frac{a+|c|\delta}{h}\right\rceil$, $\delta_2 = \delta\left(\frac{\varepsilon - h\sigma}{\varepsilon - h\mu}\right)^{-2M-1}$ and $(\mu, \sigma) \in \{P_{\Theta=1,h}\left(\delta, c, 2\right) < \delta\}$. Then if $\varphi(t) \in [-\delta_2, \delta_2]$ for $t \in [-a - |c|\delta, 0]$ then the BE solution to (4.1.1) $\{u_n\}_{n \geqslant 0}$ satisfies $u_n \in [-\delta, \delta]$ for $n \geqslant 0$.*

Proof. Let $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$, $\delta_1 = \left(\frac{\varepsilon - h\sigma}{\varepsilon - h\mu}\right)^{-1}\delta$ and $\delta_2 = \delta\left(\frac{\varepsilon - h\sigma}{\varepsilon - h\mu}\right)^{-2M-1}$. By Lemma 4.5.1, if $\varphi\left(t\right) \in [-\delta_2, \delta_2]$ for $t \in [-a - c\delta, 0]$ then $u_n \in [-\delta_1, \delta_1]$ for $n = 0, ..., 2M$. Assume the solution first exits this interval at the $(n+1)$st step through the upper bound with $u_{n+1} = \delta_3 > \delta_1$. Then by Lemmas 4.5.4 and 4.5.5, $u_{n+1} \leqslant P_{\Theta=1,h}\left(\delta_3, c, 2\right)$. If $(\mu, \sigma) \in \{P_{\Theta=1,h}\left(\delta_3, c, 2\right) < \delta_3\}$ then we have contradiction. Thus the BE solution cannot exit $[-\delta_1, \delta_1]$ for the first time through the upper bound.

Recall from (4.3.4) that if the history function and the BE solution up to n are bounded inside $[-\delta_1, \delta_1]$, $\varepsilon - h\mu > 0$ and $\mu + \sigma < 0$ then

$$|u_{n+1}| \leqslant \frac{|\varepsilon| + |h\sigma|}{|\varepsilon - h\mu|}\delta_1 = \delta.$$

Thus, $u_{n+1} = \delta_3 > \delta_1$ is only possible if $\delta_3 \in [\delta_1, \delta]$. By Lemma 4.5.9,

$$\bigcap_{\delta_3 \in [\delta_1, \delta]} \{P_{\Theta=1,h}\left(\delta_3, c, 2\right) < \delta_3\} = \{P_{\Theta=1,h}\left(\delta, c, 2\right) < \delta\}$$

Let $(\mu, \sigma) \in \{P_{\Theta=1,h}\left(\delta, c, 2\right) < \delta\}$ and $\varphi(t) \in [-\delta_2, \delta_2]$. Then by our discussion above, the BE solution cannot escape $[-\delta_1, \delta_1] \subseteq [-\delta, \delta]$ through the upper bound.

Now consider the case in which the BE solution leaves $[-\delta, \delta]$ through the lower bound by considering the system $v(t) = -u(t)$,

$$\begin{aligned} \varepsilon \dot{v}(t) &= \mu v(t) + \sigma v(t - a + cv(t)), \quad && t \geqslant 0, \\ v(t) &= -\varphi(t), && t \leqslant 0. \end{aligned} \tag{4.5.22}$$

This is effectively the same system (4.1.1) except with c replaced by $-c$. By our discussion above and Lemma 4.5.8, if $(\mu, \sigma) \in \{P_{\Theta=1,h}(\delta, c, 2) < \delta\}$ and φ is small enough then the BE solution to (4.5.22) cannot escape $[-\delta, \delta]$ through the upper bound. Thus the BE solution to (4.1.1) cannot escape $[-\delta, \delta]$ through the lower bound either. □

Lemma 4.5.10 proves that if $(\mu, \sigma) \in \{P_{\Theta=1}(\delta, c, 2) < \delta\}$, then it is possible to bound the history function such that $u_n \in [-\delta, \delta]$ for $n \geqslant 0$. To extend this idea to Lyapunov stability, we need a region where we can bound the BE solution like this for all $\delta > 0$. In Theorem 4.5.15 we show that for small enough δ, $\{P(1, 0, 2) < 1\} \subseteq \{P_{\Theta=1}(\delta, c, 2) < \delta\}$. Thus we get a discrete version of Lyapunov stability of the BE solution to (4.1.1) if $(\mu, \sigma) \in \{P(1, 0, 2) < 1\}$. The proof of Theorem 4.5.15 requires the items proven in Lemma 4.5.14. This lemma requires Lemmas 4.5.11, 4.5.12 and 4.5.13.

Lemma 4.5.11. *Let $\varepsilon, a > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let the stepsize h be such that $\varepsilon - h\mu > 0$. Let $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$ and use Definition 4.5.3. The following inequalities hold*

$$\frac{\varepsilon}{\varepsilon - h\mu(1-\beta)} A^m = A^* \geqslant e^{\frac{\mu\tau_1}{\varepsilon}}, \quad \left(1 - \frac{h\mu\chi}{\varepsilon}\right) A^\ell \geqslant e^{\frac{\mu}{\varepsilon}\frac{\delta+\hat{u}}{D\delta}}. \tag{4.5.23}$$

If $\mu > 0$ then the following inequalities also hold

$$\left(1 - \frac{h\mu}{\varepsilon}\right)^{-\frac{\tau_1}{h}} \geqslant \frac{\varepsilon}{\varepsilon - h\mu(1-\beta)} A^m = A^* \geqslant e^{\frac{\mu\tau_1}{\varepsilon}}, \tag{4.5.24}$$

$$\left(1 - \frac{h\mu\chi}{\varepsilon}\right) A^\ell \geqslant \left(1 - \frac{h\mu}{\varepsilon}\right)^{-\frac{\delta+\hat{u}}{D\delta h}} \geqslant e^{\frac{\mu}{\varepsilon}\frac{\delta+\hat{u}}{D\delta}}. \tag{4.5.25}$$

Proof. Since $1 + x \leqslant e^x$ for all $x \in \mathbb{R}$,

$$\frac{\varepsilon}{\varepsilon - h\mu(1-\beta)} = \left(1 - \frac{h\mu(1-\beta)}{\varepsilon}\right)^{-1} \geqslant e^{\frac{h\mu(1-\beta)}{\varepsilon}},$$

$$\left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^m = \left(1 - \frac{h\mu}{\varepsilon}\right)^{-m} \geqslant e^{\frac{h\mu m}{\varepsilon}}.$$

Together these two inequalities yield the first inequality in (4.5.23). For the second inequality let $s = \frac{h\mu}{\varepsilon}$. Since $\chi \in [0,1]$ then for all $s \in \mathbb{R}$,

$$\frac{1-\chi s}{1-s} \geqslant e^{(1-\chi)s},$$

with the equality only occurring at $s = 0$. Note that $\ell \geqslant 1$. Since $1 + x \leqslant e^x$ for all $x \in \mathbb{R}$,

$$\left(\frac{\varepsilon}{\varepsilon - h\mu}\right)^{\ell-1} = \left(1 - \frac{h\mu}{\varepsilon}\right)^{-\ell+1} \geqslant e^{\frac{h\mu(\ell-1)}{\varepsilon}}.$$

The two inequalities used together yield the second inequality in (4.5.23). Now let $\mu > 0$. Since $\varepsilon - h\mu > 0$ then $s = \frac{h\mu}{\varepsilon} \in [0,1)$. Let $\omega = 1 - \beta \in (0,1]$. Then,

$$(1-\omega s)^{-1} = 1 + \omega s + \omega^2 s^2 + \omega^3 s^3 + ...$$

$$(1-s)^{-\omega} = 1 + \omega s + \frac{\omega(\omega+1)}{2!}s^2 + \frac{\omega(\omega+1)(\omega+2)}{3!}s^3 + ...$$

From these series expansions, $(1-\omega s)^{-1} \leqslant (1-s)^{-\omega}$. Also, since $1 + x \leqslant e^x$ for $x \in \mathbb{R}$, $1 - \omega s \leqslant e^{-\omega s}$ and $1 - s \leqslant e^{-s}$. Thus for $m \geqslant 0$,

$$e^{(m+\omega)s} \leqslant (1-\omega s)^{-1}(1-s)^{-m} \leqslant (1-s)^{-(m+\omega)}.$$

By substituting in $\omega = 1 - \beta$, $s = \frac{h\mu}{\varepsilon}$ and $m + 1 - \beta = \frac{\tau_1}{h}$ this equation becomes (4.5.24). For the next inequality we consider the series expansion

$$(1-s)^{\omega} = 1 - \omega s - \frac{\omega(1-\omega)}{2!}s^2 - \frac{\omega(1-\omega)(2-\omega)}{3!}s^3 - ....$$

Then $1 - \omega s \geqslant (1-s)^{\omega}$. Thus it follows that

$$(1-\omega s)(1-s)^{-\ell} \geqslant (1-s)^{-(\ell-\omega)}. \qquad (4.5.26)$$

And again from $1 + x \leqslant e^x$ we have $1 - s \leqslant e^{-s}$ which immediately yields

$$(1-s)^{-(\ell-\omega)} \geqslant e^{(\ell-\omega)s} \qquad (4.5.27)$$

From (4.5.26) and (4.5.27), $(1-\omega s)(1-s)^{-\ell} \geqslant (1-s)^{-(\ell-\omega)} \geqslant e^{(\ell-\omega)s}$. Setting $\omega = \chi$, $s = \frac{h\mu}{\varepsilon}$ and $\ell - \chi = \frac{\delta+\hat{u}}{D\delta h}$ this inequality becomes (4.5.25). □

Lemma 4.5.12. *Let $\varepsilon, a > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$ and use Definition 4.5.3. Let $h \in (0, \tau_1)$. If $mhD - 1 \leqslant -\frac{\mu}{\sigma}$ then $\mu \geqslant -\frac{\varepsilon}{2\tau_1}$, $\sigma \geqslant -\frac{2\varepsilon}{\tau_1}$.*

Proof. Let $mhD - 1 \leqslant -\frac{\mu}{\sigma}$. Then by the same proof as in Lemma 3.4.8, we must have $\mu \geqslant \left(-3 + 2\sqrt{2}\right) \frac{\varepsilon}{mh}$ and $\sigma \geqslant -\frac{\varepsilon}{mh}$. For all $\tau_1 > h > 0$, $\frac{\tau_1}{mh} = \frac{\tau_1}{\left\lfloor \frac{\tau_1}{h} \right\rfloor} \in [1, 2]$. Then $\sigma \geqslant -\frac{2\varepsilon}{\tau_1}$ and $\mu \geqslant 2\left(-3 + 2\sqrt{2}\right) \frac{\varepsilon}{\tau_1} \approx -0.343 \frac{\varepsilon}{\tau_1} > -\frac{\varepsilon}{2\tau_1}$. □

Recall that in the sets $\{P(\delta, c, 2) < \delta\}$ we always have $\sigma \leqslant \mu$ and $\sigma < -\mu$. In Section 3.4.3, we also calculated that in the limit set $\{P(1, 0, 2) < 1\}$ we have $\mu \leqslant s\frac{\varepsilon}{a}$ where $s \approx 0.456971657679506$. Thus, for all $(\mu, \sigma) \in \{P(1, 0, 2) < \delta\}$, if we restrict $\delta \in \left(0, \left|\frac{a}{c}\right| \left(\frac{1}{2s} - 1\right)\right)$ then $\mu \leqslant s\frac{\varepsilon}{a} \leqslant \frac{\varepsilon}{2\tau_1}$. Choosing $\delta \in \left(0, \frac{a}{4|c|}\right)$ suffices. Since $\{P(\delta, c, 2) < \delta\} \subseteq \{P(1, 0, 2) < 1\}$ then $\mu \leqslant \frac{\varepsilon}{2\tau_1}$ as well in the set $\{P(\delta, c, 2) < \delta\}$. This is necessary for the proofs of Lemmas 4.5.13 and 4.5.14.

Lemma 4.5.13. *Let $\varepsilon, a > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let $h \in (0, a]$, $\delta \in \left(0, \frac{a}{4|c|}\right)$ so that $\tau_1 = a + |c|\delta \leqslant \frac{\varepsilon}{2\mu}$ for all $(\mu, \sigma) \in \{P(\delta, c, 2) < \delta\}$. Let $m = \left\lfloor \frac{\tau_1}{h} \right\rfloor$ and $\hat{u}_0 = \frac{\sigma}{\mu}\delta \frac{1 - \frac{\varepsilon}{\mu\tau_1}}{1 + \frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}}$. If $mhD - 1 \leqslant -\frac{\mu}{\sigma}$ and $\mu > 0$ then $(mhD - 1)\delta \geqslant \hat{u}_0$.*

Proof. Let $mhD - 1 \leqslant -\frac{\mu}{\sigma}$ and $\mu > 0$. Since $\frac{\tau_1}{mh} \in [1, 2]$ (shown in Lemma 4.5.12) then $mh \in \left[\frac{\tau_1}{2}, \tau_1\right]$ and

$$\hat{u}_0 - (mhD - 1)\delta \leqslant \frac{\sigma}{\mu}\delta \frac{1 - \frac{\varepsilon}{\mu\tau_1}}{1 + \frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}} - \left(\frac{\tau_1 D}{2} - 1\right)\delta = \delta \frac{1 + \frac{\sigma}{\mu}\left(1 - \frac{\varepsilon D}{2\mu}\right) - \frac{\tau_1 D}{2}}{1 + \frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}}. \tag{4.5.28}$$

Since $\sigma < -\mu$ and $\frac{\mu\tau_1}{\varepsilon} \leqslant \frac{1}{2}$ then the denominator is negative. Consider the numerator. The term $\frac{\varepsilon D}{2\mu} \geqslant \frac{\mu - \sigma}{2\mu} \geqslant 1$. Thus since $\frac{\sigma}{\mu} < -1$,

$$1 + \frac{\sigma}{\mu}\left(1 - \frac{\varepsilon D}{2\mu}\right) - \frac{\tau_1 D}{2} \geqslant \frac{\varepsilon D}{2\mu} - \frac{\tau_1 D}{2} = \frac{D}{2}\left(\frac{\varepsilon}{\mu} - \tau_1\right) > 0$$

Going back to (4.5.28), this yields the required result $(mhD - 1)\delta \geqslant \hat{u}_0$. □

Next we prove Lemma 4.5.14 which is used to prove Theorem 4.5.15.

Lemma 4.5.14. *Let $\varepsilon, a > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let the stepsize $h > 0$. If $\mu > 0$ restrict $h \in (0, a]$. Let $\delta \in \left(0, \frac{a}{4|c|}\right)$ so that $\tau_1 = a + |c|\delta$ satisfies $\frac{\mu\tau_1}{\varepsilon} \leqslant \frac{1}{2}$ for all $(\mu, \sigma) \in \{P(\delta, c, 2) < \delta\}$. Then for all $(\mu, \sigma) \in \{P(\delta, c, 2) < \delta\}$, the following items (A)-(G) are true:*

(A) If $\mu < 0$ then for all $\hat{u} \leqslant 0$, $\mathcal{S}_2(\hat{u}, \delta) \leqslant \mathcal{I}_2(\hat{u}, \delta)$.

(B) If $\mu > 0$ then for all $u \geqslant \frac{\sigma}{\mu}\delta\frac{1-\frac{\varepsilon}{\mu\tau_1}}{1+\frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}}$, $\mathcal{S}_2(\hat{u},\delta) \leqslant \mathcal{I}_2(\hat{u},\delta)$. Also,

$$\sup_{\hat{u}\in\left[-\delta,\min\{-\frac{\mu}{\sigma}\delta,(\tau_1 D-1)\delta\}\right]} \mathcal{S}_2(\hat{u},\delta) < \delta.$$

(C) If $\mu < 0$ then for all $\hat{u} \leqslant 0$, $\mathcal{S}_1(\hat{u},\delta) \leqslant \mathcal{I}_1(\hat{u},\delta)$.

(D) If $\mu > 0$ then for all $\hat{u} \geqslant -\delta$, $\mathcal{S}_1(\hat{u},\delta) \leqslant \mathcal{I}_1(\hat{u},\delta)$.

(E) If $\tau_1 D - 1 > -\frac{\mu}{\sigma} > \lfloor\frac{\tau_1}{h}\rfloor hD - 1$ and $h \in (0,\tau_1]$ then $\sup_{\hat{u}\in\left[\left(\lfloor\frac{\tau_1}{h}\rfloor hD-1\right)\delta,-\frac{\mu}{\sigma}\delta\right]} \mathcal{S}_1(\hat{u},\delta) \leqslant \mathcal{I}_2\left(-\frac{\mu}{\sigma}\delta,\delta\right)$. If $h > \tau_1$ and $\mu < 0$ then $\sup_{\hat{u}\in\left[\left(\lfloor\frac{\tau_1}{h}\rfloor hD-1\right)\delta,-\frac{\mu}{\sigma}\delta\right]} \mathcal{S}_1(\hat{u},\delta) < \delta$.

(F) If $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$ then $\sup_{\hat{u}\in[-\delta,(\tau_1 D-1)\delta]} \mathcal{S}_2(\hat{u},\delta) = \mathcal{S}_2((\tau_1 D-1)\delta,\delta)$.

(G) If $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$ then $\mathcal{S}_2((\tau_1 D-1)\delta,\delta) \leqslant \sup_{\hat{v}\in\left[(\tau_1 D-1)\delta,-\frac{\mu}{\sigma}\delta\right]} \mathcal{S}_1(\hat{v},\delta)$.

Proof of (A). Recall the definition of $\mathcal{S}_2(\hat{u},\delta)$ in (4.5.19) and compare this with $\mathcal{I}_2(\hat{u},\delta)$ in (3.4.14),

$$\mathcal{S}_2(\hat{u},\delta) - \mathcal{I}_2(\hat{u},\delta) = \left(\hat{u} - \frac{\sigma}{\mu}\delta\right)\left(A^* - e^{\frac{\mu\tau_1}{\varepsilon}}\right) + \frac{\sigma\varepsilon D}{\mu^2}\delta\left(\left(1 - \frac{h\mu\chi}{\varepsilon}\right)A^\ell - e^{\frac{\mu}{\varepsilon}\frac{\delta+\hat{u}}{D\delta}}\right). \quad (4.5.29)$$

When $\mu < 0$ then $\hat{u} \in \left[-\delta, -\frac{\mu}{\sigma}\delta\right] \subseteq [-\delta, 0]$ and so $\hat{u} - \frac{\sigma}{\mu}\delta < 0$. Also, $\frac{\sigma\varepsilon D}{\mu^2}\delta < 0$. Using Lemma 4.5.11, when $\mu \leqslant 0$ then $\mathcal{S}_2(\hat{u},\delta) \leqslant \mathcal{I}_2(\hat{u},\delta)$. □

Proof of (B). First we would like to show that $\mathcal{S}_2(\hat{u},\delta)$ is continuous with respect to $\hat{u}$. The only term in the expression that gives us trouble with this is $\left(1 - \frac{h\mu\chi}{\varepsilon}\right)A^\ell$ whenever ℓ changes values. Let $i \geqslant 0$ be an integer. Then

$$\lim_{\frac{\delta+\hat{u}}{D\delta h}\to i^+} \left(1 - \frac{h\mu\chi}{\varepsilon}\right)A^\ell = \lim_{\chi\to 1}\left(1 - \frac{h\mu\chi}{\varepsilon}\right)A^{i+1} = \left(1 - \frac{h\mu}{\varepsilon}\right)A^{i+1} = A^i,$$

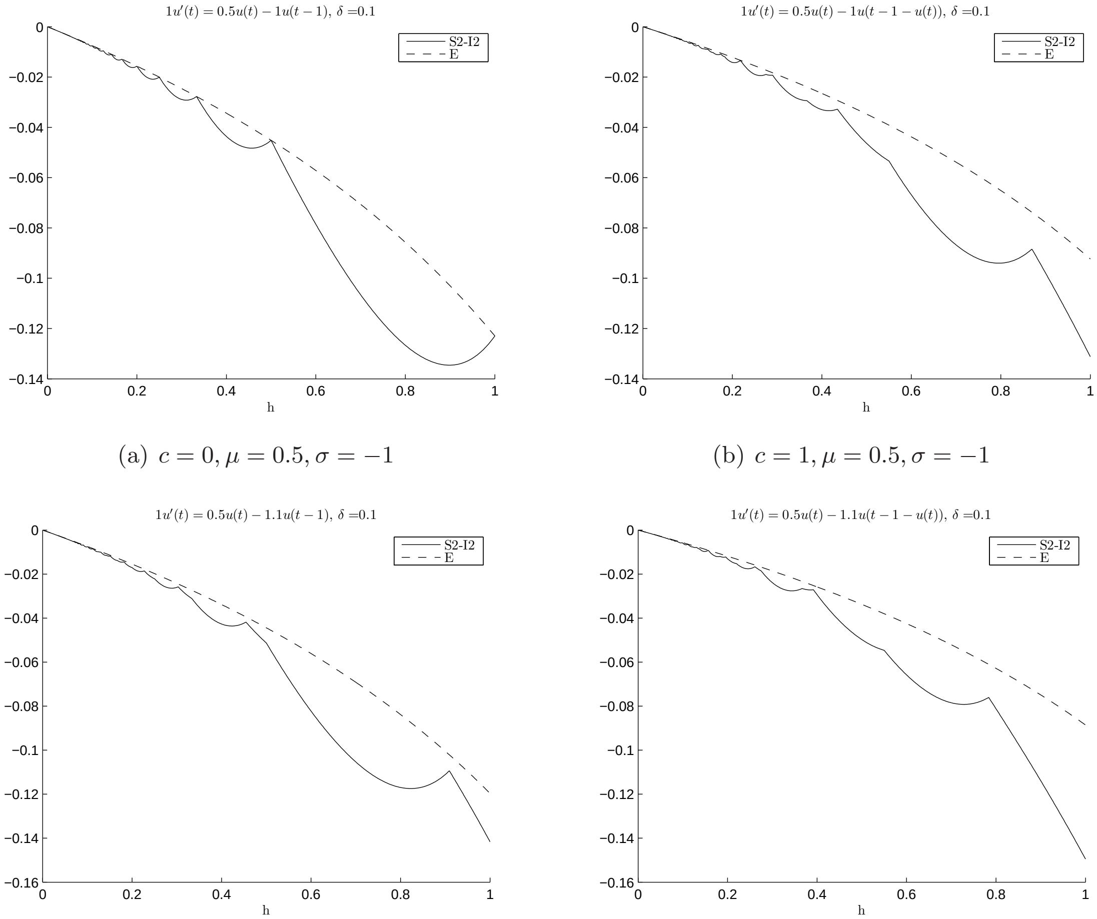
$$\lim_{\frac{\delta+\hat{u}}{D\delta h}\to i^-} \left(1 - \frac{h\mu\chi}{\varepsilon}\right)A^\ell = \lim_{\chi\to 0}\left(1 - \frac{h\mu\chi}{\varepsilon}\right)A^{i} = A^i.$$

Let $\mu > 0$ and $h \in (0,a]$. Then $\hat{u} - \frac{\sigma}{\mu}\delta > 0$ for $\hat{u} \in \left[-\delta, \min\left\{-\frac{\mu}{\sigma}\delta, (\tau_1 D - 1)\delta\right\}\right]$. From (4.5.29) we see that $\mathcal{S}_2(-\delta,\delta) \geqslant \mathcal{I}_2(-\delta,\delta)$ because the second term disappears. So we cannot show $\mathcal{S}_2(\hat{u},\delta) \geqslant \mathcal{I}_2(\hat{u},\delta)$ for all $\hat{u}$. From (4.5.29) and Lemma 4.5.11, we derive a function $E(h)$

defined as follows

$$
\begin{aligned}
&\mathcal{S}_2\left(\hat{u},\delta\right)-\mathcal{I}_2\left(\hat{u},\delta\right)\\
&\leqslant E\left(h\right)=\left(\hat{u}-\frac{\sigma}{\mu}\delta\right)\left[\left(1-\frac{h\mu}{\varepsilon}\right)^{-\frac{\tau_1}{h}}-e^{\frac{\mu\tau_1}{\varepsilon}}\right]+\frac{\sigma\varepsilon D}{\mu^2}\delta\left[\left(1-\frac{h\mu}{\varepsilon}\right)^{-\frac{\delta+\hat{u}}{D\delta h}}-e^{\frac{\mu}{\varepsilon}\frac{\delta+\hat{u}}{D\delta}}\right]. \quad (4.5.30)
\end{aligned}
$$

Figure 4–6 shows the comparison between $\mathcal{S}_2\left(\hat{u},\delta\right)-\mathcal{I}_2\left(\hat{u},\delta\right)$ and $E\left(h\right)$. The function $E\left(h\right)$ always goes to zero as $h\to 0$. So if we would like to keep $\mathcal{S}_2\left(\hat{u},\delta\right)-\mathcal{I}_2\left(\hat{u},\delta\right)$ negative for all h such that $\varepsilon-h\mu>0$, then it suffices to require $\frac{dE(h)}{dh}\leqslant 0$ for all $h\in(0,a]$.



(a) $c=0,\mu=0.5,\sigma=-1$

(b) $c=1,\mu=0.5,\sigma=-1$

(c) $c=0,\mu=0.5,\sigma=-1.1$

(d) $c=1,\mu=0.5,\sigma=-1.1$

Figure 4–6: Sample plots of $\mathcal{S}_2\left(-\frac{\mu}{\sigma}\delta,\delta\right)-\mathcal{I}_2\left(-\frac{\mu}{\sigma}\delta,\delta\right)$ and the estimate $E\left(h\right)$ given in (4.5.30) versus the step-size h for $\varepsilon=a=1$ and different values of σ and c.

$$\begin{aligned}\frac{dE(h)}{dh} &= \left(\hat{u}-\frac{\sigma}{\mu}\delta\right)\left(1-\frac{h\mu}{\varepsilon}\right)^{-\frac{\tau_1}{h}}\tau_1\left(\frac{\ln\left(1-\frac{h\mu}{\varepsilon}\right)}{h^2}+\frac{\mu}{\varepsilon h\left(1-\frac{h\mu}{\varepsilon}\right)}\right)\\ &\quad+\frac{\sigma\varepsilon D}{\mu^2}\delta\left(1-\frac{h\mu}{\varepsilon}\right)^{-\frac{\delta+\hat{u}}{D\delta h}}\frac{\delta+\hat{u}}{D\delta}\left(\frac{\ln\left(1-\frac{h\mu}{\varepsilon}\right)}{h^2}+\frac{\mu}{\varepsilon h\left(1-\frac{h\mu}{\varepsilon}\right)}\right)\end{aligned}$$

Series expansions show that the term $\frac{\ln\left(1-\frac{h\mu}{\varepsilon}\right)}{h^2}+\frac{\mu}{\varepsilon h\left(1-\frac{h\mu}{\varepsilon}\right)}$ is always positive. Thus,

$$\begin{aligned}&\frac{dE(h)}{dh}<0\\ \Leftrightarrow\quad &\left(1+\frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}\left(1-\frac{h\mu}{\varepsilon}\right)^{\frac{\tau_1-\frac{\delta+\hat{u}}{D\delta}}{h}}\right)\hat{u}-\frac{\sigma}{\mu}\delta\left(1-\frac{\varepsilon}{\mu\tau_1}\left(1-\frac{h\mu}{\varepsilon}\right)^{\frac{\tau_1-\frac{\delta+\hat{u}}{D\delta}}{h}}\right)\leqslant 0.\end{aligned}\tag{4.5.31}$$

Consider the sign of the term $1-\frac{\varepsilon}{\mu\tau_1}\left(1-\frac{h\mu}{\varepsilon}\right)^{\frac{\tau_1-\frac{\delta+\hat{u}}{D\delta}}{h}}$,

$$1-\frac{\varepsilon}{\mu\tau_1}\left(1-\frac{h\mu}{\varepsilon}\right)^{\frac{\tau_1-\frac{\delta+\hat{u}}{D\delta}}{h}}\leqslant 0\quad\Leftrightarrow\quad\frac{\delta+\hat{u}}{D\delta}\geqslant\tau_1-\frac{h\ln\left(\frac{\mu\tau_1}{\varepsilon}\right)}{\ln\left(1-\frac{h\mu}{\varepsilon}\right)}.\tag{4.5.32}$$

Recall we chose δ small enough such that $\frac{\mu\tau_1}{\varepsilon}\leqslant\frac{1}{2}$. Then $\ln\left(\frac{\mu\tau_1}{\varepsilon}\right)<0$. Since $h\in(0,a]$ then $x=\frac{\mu h}{\varepsilon}\leqslant\frac{\mu\tau_1}{\varepsilon}\leqslant\frac{1}{2}$. The function $\frac{x}{\ln(1-x)}$ is an increasing function of $x=\frac{\mu h}{\varepsilon}\in\left[0,\frac{\mu\tau_1}{\varepsilon}\right]$, thus

$$\tau_1\left(1-\frac{\ln\left(\frac{\mu\tau_1}{\varepsilon}\right)}{\ln\left(1-\frac{\mu\tau_1}{\varepsilon}\right)}\right)=\tau_1-\frac{\varepsilon\ln\left(\frac{\mu\tau_1}{\varepsilon}\right)}{\mu}\frac{\frac{\mu\tau_1}{\varepsilon}}{\ln\left(1-\frac{\mu\tau_1}{\varepsilon}\right)}\geqslant\tau_1-\frac{\varepsilon\ln\left(\frac{\mu\tau_1}{\varepsilon}\right)}{\mu}\frac{\frac{h\mu}{\varepsilon}}{\ln\left(1-\frac{h\mu}{\varepsilon}\right)}.$$

Since $1-\frac{\ln(x)}{\ln(1-x)}\leqslant 0$ for all $x=\frac{\mu\tau_1}{\varepsilon}\in\left(0,\frac{1}{2}\right]$. Then $\frac{\delta+\hat{u}}{D\delta}\geqslant 0\geqslant\tau_1\left(1-\frac{\ln\left(\frac{\mu\tau_1}{\varepsilon}\right)}{\ln\left(1-\frac{\mu\tau_1}{\varepsilon}\right)}\right)$. Then by (4.5.32) and $\sigma<-\mu<0$,

$$1+\frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}\left(1-\frac{h\mu}{\varepsilon}\right)^{\frac{\tau_1-\frac{\delta+\hat{u}}{D\delta}}{h}}\leqslant 1-\frac{\varepsilon}{\mu\tau_1}\left(1-\frac{h\mu}{\varepsilon}\right)^{\frac{\tau_1-\frac{\delta+\hat{u}}{D\delta}}{h}}\leqslant 0.$$

Going back to (4.5.31), $\frac{dE(h)}{dh} < 0$ if $\hat{u}$ satisfies

$$\hat{u} \geqslant \frac{\sigma}{\mu}\delta \frac{1 - \frac{\varepsilon}{\mu\tau_1}\left(1 - \frac{h\mu}{\varepsilon}\right)^{\frac{\tau_1 - \frac{\delta+\hat{u}}{D\delta}}{h}}}{1 + \frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}\left(1 - \frac{h\mu}{\varepsilon}\right)^{\frac{\tau_1 - \frac{\delta+\hat{u}}{D\delta}}{h}}}.$$

Consider the function $\frac{\sigma}{\mu}\frac{1 - \frac{\varepsilon}{\mu\tau_1}x}{1 + \frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}x}$ of $x = \left(1 - \frac{h\mu}{\varepsilon}\right)^{\frac{\tau_1 - \frac{\delta+\hat{u}}{D\delta}}{h}}$. The derivative of this for $\mu > 0$ is positive so the function is maximum at the maximum value of x. This occurs at $\hat{u} = (\tau_1 D - 1)\,\delta$ and $x = 1$. Thus,

$$\frac{\sigma}{\mu}\delta \frac{1 - \frac{\varepsilon}{\mu\tau_1}}{1 + \frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}} \geqslant \frac{\sigma}{\mu}\delta \frac{1 - \frac{\varepsilon}{\mu\tau_1}\left(1 - \frac{h\mu}{\varepsilon}\right)^{\frac{\tau_1 - \frac{\delta+\hat{u}}{D\delta}}{h}}}{1 + \frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}\left(1 - \frac{h\mu}{\varepsilon}\right)^{\frac{\tau_1 - \frac{\delta+\hat{u}}{D\delta}}{h}}}.$$

It suffices to require $\hat{u} \geqslant \hat{u}_0 = \frac{\sigma}{\mu}\delta\frac{1 - \frac{\varepsilon}{\mu\tau_1}}{1 + \frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}}$ to get $\frac{dE(h)}{dh} < 0$. This proves the first part of (B).

In order to show the second part of (B), consider the derivative of $\mathcal{S}_2(\hat{u}, \delta)$ with respect to $\hat{u}$. When $\frac{\delta+\hat{u}}{D\delta h}$ is not an integer, the derivative of $\mathcal{S}_2(\hat{u}, \delta)$ is defined and given by

$$\frac{d}{d\hat{u}}\mathcal{S}_2(\hat{u}, \delta) = A^* - \frac{\sigma}{\mu} + \frac{\sigma}{\mu}A^\ell. \tag{4.5.33}$$

This shows that $\mathcal{S}_2$ is piecewise linear. The function reaches a maximum at the start of the first interval where this derivative is negative. Let ℓ be fixed and assume that this derivative is negative. Then,

$$\frac{d}{d\hat{u}}\mathcal{S}_2(\hat{u}, \delta) = A^* - \frac{\sigma}{\mu} + \frac{\sigma}{\mu}A^\ell \leqslant 0 \quad \Rightarrow \quad \sigma \leqslant -\mu\frac{A^*}{A^\ell - 1}$$

Since $\ell \leqslant m$ then $-\mu\frac{A^*}{A^\ell - 1}$ is a decreasing function of μ. Thus the maximum occurs at $\mu = 0$. Using L'Hopital's rule, the value at this point is $-\frac{\varepsilon}{h\ell}$. Thus,

$$\frac{d}{d\hat{u}}\mathcal{S}_2(\hat{u}, \delta) \leqslant 0 \quad \Rightarrow \quad \sigma \leqslant -\frac{\varepsilon}{h\ell} \tag{4.5.34}$$

We will come back to this later. Let $\hat{u}_0 = \frac{\sigma}{\mu}\delta\frac{1 - \frac{\varepsilon}{\mu\tau_1}}{1 + \frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}}$ and

$$\ell_0 = \left\lceil \frac{\delta + \hat{u}_0}{D\delta h} \right\rceil = \left\lceil \frac{\delta + \frac{\sigma}{\mu}\delta\frac{1 - \frac{\varepsilon}{\mu\tau_1}}{1 + \frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}}}{D\delta h} \right\rceil = \left\lceil \frac{1 + \frac{\sigma}{\mu}}{\left(1 + \frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}\right)Dh} \right\rceil. \tag{4.5.35}$$

Suppose $\ell_0 > 1$. From our choice of δ, $\frac{\varepsilon}{\mu\tau_1} > \frac{1}{2}$. Since $\sigma < -\mu$ then $\frac{1+\frac{\sigma}{\mu}}{1+\frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}} < 1$. Thus for $\ell_0 > 1$ we must have $Dh < 1$.

$$Dh < 1 \quad \Rightarrow \quad \frac{\mu-\sigma}{\varepsilon}h < 1 \quad \Rightarrow \quad \mu - \frac{\varepsilon}{h} < \sigma.$$

In particular this means $\sigma > -\frac{\varepsilon}{h}$. From (4.5.34), we cannot have $\frac{d}{d\hat{u}}\mathcal{S}_2\left(\hat{u}_0, \delta\right) \leqslant 0$ if $\ell_0 > 1$. Thus in this case the maximum value of $\mathcal{S}_2\left(\hat{u}, \delta\right)$ must occur past $\hat{u}_0$. We have already proven that $\mathcal{S}_2\left(\hat{u}, \delta\right) \leqslant \mathcal{I}_2\left(\hat{u}, \delta\right)$ past $\hat{u}_0$. So if $\ell_0 > 1$ then for all $\hat{u}_1 \geqslant \hat{u}_0$,

$$\sup_{\hat{u}\in[-\delta,\hat{u}_1]} \mathcal{S}_2\left(\hat{u}, \delta\right) \leqslant \sup_{\hat{u}\in[-\delta,\hat{u}_1]} \mathcal{I}_2\left(\hat{u}, \delta\right).$$

Since $\mu > 0$ then $\hat{u}_0 < 0 < -\frac{\mu}{\sigma}\delta$. If $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$ then $mhD - 1 \leqslant -\frac{\mu}{\sigma}$ and from Lemma 4.5.13, $\hat{u}_0 \leqslant \left(mhD - 1\right)\delta \leqslant \left(\tau_1 D - 1\right)\delta$. Thus, $\min\{-\frac{\mu}{\sigma}\delta, \left(\tau_1 D - 1\right)\delta\} \geqslant \hat{u}_0$. This proves the second part of (B) for the case $\ell_0 > 1$.

Now consider the case $\ell_0 = 0$ or 1. In either case, the maximum of $\mathcal{S}_2\left(\hat{u}, \delta\right)$ must occur at $\hat{u} = -\delta$. If we can prove that $\mathcal{S}_2\left(\hat{u}, \delta\right) < \delta$ then we are done. Suppose not. Then,

$$\mathcal{S}_2\left(-\delta, \delta\right) = -\delta\left(A^* - \frac{\sigma}{\mu} + \frac{\sigma}{\mu}A^*\right) \geqslant \delta \quad \Rightarrow \quad \sigma \leqslant -\mu\frac{A^*+1}{A^*-1}.$$

The term $-\mu\frac{A^*+1}{A^*-1}$ is an increasing function of A^*. Using this and Lemma 4.5.11,

$$\sigma \leqslant -\mu\frac{A^*+1}{A^*-1} \leqslant -\mu\frac{\left(1-\frac{h\mu}{\varepsilon}\right)^{-\frac{\tau_1}{h}}+1}{\left(1-\frac{h\mu}{\varepsilon}\right)^{-\frac{\tau_1}{h}}-1} \leqslant -\mu\frac{\left(1-\frac{\mu\tau_1}{\varepsilon}\right)^{-1}+1}{\left(1-\frac{\mu\tau_1}{\varepsilon}\right)^{-1}-1} = \mu - \frac{2\varepsilon}{\tau_1}.$$

By our choice of δ, $\tau_1 \leqslant \frac{5}{4}a$. Thus, $\sigma \leqslant \mu - \frac{8\varepsilon}{5a}$. But if $\left(\mu, \sigma\right) \in \{P\left(\delta, c, 2\right) < \delta\} \subseteq \{P\left(1, 0, 2\right) < 1\}$ then $\sigma \geqslant \mu - \frac{3\varepsilon}{2a}$. This is a contradiction. Then we must have $S_2\left(-\delta, \delta\right) < \delta$. □

Proof of (C). Let $\mu < 0$. Then

$$\mathcal{S}_1\left(\hat{u}, \delta\right) - \mathcal{I}_1\left(\hat{u}, \delta\right) = \left[\hat{u}\left(1+\frac{\sigma}{\mu}\right) + \frac{\sigma D}{\mu}\delta\left(\frac{\varepsilon}{\mu} - \tau_1\right)\right]\left(A^* - e^{\frac{\mu\tau_1}{\varepsilon}}\right). \tag{4.5.36}$$

From Lemma 4.5.11, the second factor is positive. Since $\sigma \leqslant \mu < 0$ then the first factor, $\hat{u}\left(1+\frac{\sigma}{\mu}\right) + \frac{\sigma D}{\mu}\delta\left(\frac{\varepsilon}{\mu} - \tau_1\right)$ is negative for all $\hat{u} \leqslant 0$. This is enough to show $\mathcal{S}_1\left(\hat{u}, \delta\right) \leqslant \mathcal{I}_1\left(\hat{u}, \delta\right)$ for all $\hat{u} \in \left[-\delta, -\frac{\mu}{\sigma}\delta\right]$. □

Proof of (D). Consider (4.5.36) with $\mu \geqslant 0$. From Lemma 4.5.11, the second factor is positive. So we need to show that $\hat{u}\left(1+\frac{\sigma}{\mu}\right)+\frac{\sigma D}{\mu}\delta\left(\frac{\varepsilon}{\mu}-\tau_1\right) \leqslant 0$. Since $1+\frac{\sigma}{\mu} \leqslant 0$ then set $\hat{u}=-\delta$ to get the worst case.

$$-\delta\left(1+\frac{\sigma}{\mu}\right)+\frac{\sigma}{\mu}D\delta\left(\frac{\varepsilon}{\mu}-\tau_1\right) \leqslant 0 \quad \Leftrightarrow \quad 1-\frac{\mu\tau_1}{\varepsilon} \geqslant \mu\frac{\frac{\mu}{\sigma}+1}{\varepsilon D} \tag{4.5.37}$$

Since $D=\frac{|\mu|+|\sigma|}{\varepsilon}\left(1+\frac{|\mu|+|\sigma|}{\varepsilon}\delta|c|\right) \geqslant \frac{|\mu|+|\sigma|}{\varepsilon}$ then $\frac{\mu}{\sigma}\frac{\frac{\mu}{\sigma}+1}{\frac{\mu}{\sigma}-1} \geqslant \mu\frac{\frac{\mu}{\sigma}+1}{\varepsilon D}$. Maximizing the function $x\left(\frac{x+1}{x-1}\right)$ over $x=\frac{\mu}{\sigma} \in [-1,0]$ yields a maximum value of $3-\sqrt{2}$. Hence,

$$0.8284 \approx 3-2\sqrt{2} \geqslant \frac{\mu}{\sigma}\frac{\frac{\mu}{\sigma}+1}{\frac{\mu}{\sigma}-1} \geqslant \mu\frac{\frac{\mu}{\sigma}+1}{\varepsilon D}. \tag{4.5.38}$$

By our choice of δ, $\frac{\mu\tau_1}{\varepsilon} \leqslant \frac{1}{2}$ so $1-\frac{\mu\tau_1}{\varepsilon} > 3-2\sqrt{2}$. By this, (4.5.36), (4.5.37) and (4.5.38) we must have $\mathcal{S}_1(-\delta,\delta) \leqslant \mathcal{I}_1(-\delta,\delta)$. Then $\mathcal{S}_1(\hat{u},\delta) \leqslant \mathcal{I}_1(\hat{u},\delta)$ for all $\hat{u} \geqslant -\delta$. □

Proof of (E). Let $\tau_1 D-1 > -\frac{\mu}{\sigma} > mhD-1$ and $\hat{u} \in \left[(mhD-1)\delta, -\frac{\mu}{\sigma}\delta\right]$. In this case $\ell=m+1$ so the relevant summation is $\mathcal{S}_1(\hat{u},\delta)$. First consider $h \in (0,a]$. We begin by proving that in this case, $\mathcal{S}_1((\tau_1 D-1)\delta,\delta) \leqslant \mathcal{I}_1((\tau_1 D-1)\delta,\delta)$. For $\mu>0$ this is guaranteed by (D). For $\mu<0$ the result is guaranteed by (C) only if $(\tau_1 D-1)\delta<0$ which is not always true. Let $\mu<0$. From (4.5.36), the sign of $\mathcal{S}_1((\tau_1 D-1)\delta,\delta)-\mathcal{I}_1((\tau_1 D-1)\delta,\delta)$ is the same as the sign of $(\tau_1 D-1)\left(1+\frac{\sigma}{\mu}\right)+\frac{\sigma D}{\mu}\left(\frac{\varepsilon}{\mu}-\tau_1\right)$. Since $\sigma<-\mu$ then $1+\frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}<0$. Thus,

$$\begin{aligned}
&(\tau_1 D-1)\left(1+\frac{\sigma}{\mu}\right)+\frac{\sigma D}{\mu}\left(\frac{\varepsilon}{\mu}-\tau_1\right)=\left(1+\frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}\right)\tau_1 D-1-\frac{\sigma}{\mu},\\
&\leqslant\left(1+\frac{\sigma}{\mu}\frac{\varepsilon}{\mu\tau_1}\right)\tau_1\frac{|\mu|+|\sigma|}{\varepsilon}-1-\frac{\sigma}{\mu},\\
&=-\frac{(\mu+\sigma)\tau_1}{\varepsilon}-1-2\frac{\sigma}{\mu}-\frac{\sigma^2}{\mu^2} \leqslant -\frac{(\mu+\sigma)\tau_1}{\varepsilon}-4.
\end{aligned} \tag{4.5.39}$$

By Lemma 4.5.12, $\frac{\mu\tau_1}{\varepsilon} \geqslant -\frac{1}{2}$ and $\frac{\sigma\tau_1}{\varepsilon} \geqslant -2$. Then the term in (4.5.39) is negative so $\mathcal{S}_1((\tau_1 D-1)\delta,\delta) \leqslant \mathcal{I}_1((\tau_1 D-1)\delta,\delta)$.

From this result and from the continuity at the point where we switch between $\mathcal{I}_1$ and $\mathcal{I}_2$,

$$\mathcal{S}_1((\tau_1 D-1)\delta,\delta) \leqslant \mathcal{I}_1((\tau_1 D-1)\delta,\delta)=\mathcal{I}_2((\tau_1 D-1)\delta,\delta).$$

From (A), (B) and continuity at the point where we switch between $\mathcal{S}_1$ and $\mathcal{S}_2$,

$$\mathcal{S}_1\left(\left(\left\lfloor\frac{\tau_1}{h}\right\rfloor hD-1\right)\delta,\delta\right)=\mathcal{S}_2\left(\left(\left\lfloor\frac{\tau_1}{h}\right\rfloor hD-1\right)\delta,\delta\right)\leqslant\mathcal{I}_2\left(\left(\left\lfloor\frac{\tau_1}{h}\right\rfloor hD-1\right)\delta,\delta\right).$$

Since $\mathcal{S}_1(\hat{u},\delta)$ is monotonic and $\mathcal{I}_2(\hat{u},\delta)$ is concave downwards with respect to $\hat{u}$ (these properties of $\mathcal{I}_2$ are discussed in Theorem 3.4.10), then the two functions cannot cross between $\hat{u}=\left(\lfloor\frac{\tau_1}{h}\rfloor hD-1\right)\delta$ and $(\tau_1 D-1)\delta$. So $\mathcal{S}_1(\hat{u},\delta)\leqslant\mathcal{I}_2(\hat{u},\delta)$ in the entire interval if $h\in(0,\tau_1]$.

The case $h>\tau_1$ is only allowed when $\mu\leqslant 0$. Then Lemma 4.5.12 does not apply but in this case $m=0$ so easily $\mathcal{S}_1(\hat{u},\delta)=\hat{u}-\frac{\sigma}{\mu}\tau_1 D\delta<\delta$. □

Proof of (F). Recall the expression for $\frac{d}{d\hat{u}}\mathcal{S}_2(\hat{u},\delta)$ from (4.5.33). If $\mu\leqslant 0$ then $A<1$ and increasing $\hat{u}$ decreases the derivative. If $\mu>0$ then $A>1$ and increasing $\hat{u}$ also decreases the derivative. So to prove (F) we just have to show that $\frac{d}{d\hat{u}}\mathcal{S}_2((\tau_1 D-1)\delta,\delta)>0$. Let $\hat{u}=(\tau_1 D-1)\delta$. Then $\ell=\left\lceil\frac{\tau_1}{h}\right\rceil$. Then generally $\ell=m+1$.

$$\frac{d}{d\hat{u}}\mathcal{S}_2(\hat{u},\delta)\geqslant\frac{\varepsilon}{\varepsilon-h\mu(1-\beta)}A^m-\frac{\sigma}{\mu}+\frac{\sigma}{\mu}\frac{\varepsilon}{\varepsilon-h\mu(1-\beta)}A^m$$

So it suffices to prove $\frac{\varepsilon}{\varepsilon-h\mu(1-\beta)}A^m-\frac{\sigma}{\mu}+\frac{\sigma}{\mu}\frac{\varepsilon}{\varepsilon-h\mu(1-\beta)}A^m\geqslant 0$. In the case $\mu<0$,

$$\frac{\varepsilon}{\varepsilon-h\mu(1-\beta)}A^m-\frac{\sigma}{\mu}+\frac{\sigma}{\mu}\frac{\varepsilon}{\varepsilon-h\mu(1-\beta)}A^m>e^{\frac{\mu}{\varepsilon}\tau_1}-\frac{\sigma}{\mu}+\frac{\sigma}{\mu}e^{\frac{\mu}{\varepsilon}\tau_1}>0,$$

where the first inequality is from Lemma 4.5.11 and the second is from Lemma 3.4.9 (C). Now let $\mu>0$. Suppose $\frac{\varepsilon}{\varepsilon-h\mu(1-\beta)}A^m-\frac{\sigma}{\mu}+\frac{\sigma}{\mu}\frac{\varepsilon}{\varepsilon-h\mu(1-\beta)}A^m<0$. Solving for σ we get

$$\sigma<-\mu\left(\frac{\frac{\varepsilon}{\varepsilon-h\mu(1-\beta)}A^m}{\frac{\varepsilon}{\varepsilon-h\mu(1-\beta)}A^m-1}\right)$$

The right hand side can be shown to be a decreasing function in μ. In the limit $\mu\to 0$, the right hand side goes to $-\frac{\varepsilon}{\tau_1}$ using L'Hopital's rule. So for $\mu>0$,

$$\frac{d}{d\hat{u}}\mathcal{S}_2(\hat{u},\delta)<0\quad\Rightarrow\quad\sigma<-\frac{\varepsilon}{\tau_1}$$

But from Lemma 3.4.8, $\tau_1 D-1\leqslant-\frac{\mu}{\sigma}$ implies $\sigma\geqslant-\frac{\varepsilon}{\tau_1}$. Thus if $\tau_1 D-1\leqslant-\frac{\mu}{\sigma}$ we must have $\frac{d}{d\hat{u}}\mathcal{S}_2(\hat{u},\delta)\geqslant 0$. □

Proof of (G). Compare $\mathcal{S}_2\left(\left(\tau_1 D-1\right)\delta,\delta\right)$ and $\mathcal{S}_1\left(\hat{v},\delta\right)$ where $\hat{v}\in\left[\left(\tau_1 D-1\right)\delta,-\frac{\mu}{\sigma}\delta\right]$.

$$\begin{aligned}\mathcal{S}_2\left(\left(\tau_1 D-1\right)\delta,\delta\right)-\mathcal{S}_1\left(\hat{v},\delta\right)&<\left(\hat{u}-\hat{v}\right)\left(A^*-\frac{\sigma}{\mu}\right)+\left(\left(\tau_1 D-1\right)-\hat{v}\right)\frac{\sigma}{\mu}A^*\\&=\left(\tau_1 D-1-\hat{v}\right)\left(A^*-\frac{\sigma}{\mu}+\frac{\sigma}{\mu}A^*\right)\end{aligned}$$

In the proof of (F) we showed that the second factor is nonnegative if $\tau_1 D-1\leqslant-\frac{\mu}{\sigma}$. So $\mathcal{S}_2\left(\left(\tau_1 D-1\right)\delta,\delta\right)\leqslant S_1\left(\hat{v},\delta\right)$ for $v>\left(\tau_1 D-1\right)\delta$. □

This finally brings us Theorem 4.5.15, the main result of this section.

Theorem 4.5.15. *Let $\varepsilon,a>0$, $\sigma\leqslant\mu<\frac{\varepsilon}{a}$ and $\sigma<-\mu$. Let the stepsize $h>0$. If $\mu>0$ restrict $h\in(0,a]$. Let $(\mu,\sigma)\in\{P\left(1,0,2\right)<1\}$ where P is as defined in Definition 3.4.2. Then for any $\delta\in\left(0,\frac{a}{4|c|}\right)$ there exists $\delta_2>0$ such that if $\varphi(t)\in[-\delta_2,\delta_2]$ for $t\in[-a-|c|\delta,0]$ then the BE solution to* (4.1.1) *$\{u_n\}_{n\geqslant0}$ satisfies $u_n\in[-\delta,\delta]$ for $n\geqslant0$.*

Proof. The first part of this proof is similar to the proof of Theorem 3.4.7. Define

$$J=\bigcup_{\delta\in\left(0,\left|\frac{a}{c}\right|\right)}\left\{P_{\Theta=1,h}\left(\delta,c,2\right)<\delta\right\}.$$

Let $\delta\in\left(0,\left|\frac{a}{c}\right|\right)$ and $(\mu,\sigma)\in J$. Then for some maximal $\delta_1\in\left(0,\left|\frac{a}{c}\right|\right)$ we have $(\mu,\sigma)\in\{P_{\Theta=1,h}\left(\delta_1,c,2\right)<\delta_1\}$. If $\delta_1<\delta$ set $\delta_2=\delta_1\left(\frac{\varepsilon-h\sigma}{\varepsilon-h\mu}\right)^{-M-1}$. By Lemma 4.5.10, if $\varphi\left(t\right)\in[-\delta_2,\delta_2]$ for $t\in[-a-|c|\delta,0]$ then $u_n\in[-\delta_1,\delta_1]\subseteq[-\delta,\delta]$ for all $n\geqslant0$. If $\delta<\delta_1$ set $\delta_2=\delta\left(\frac{\varepsilon-h\sigma}{\varepsilon-h\mu}\right)^{-M-1}$. In this case, $(\mu,\sigma)\in\{P_{\Theta=1,h}\left(\delta_1,c,2\right)<\delta_1\}\subseteq\{P_{\Theta=1,h}\left(\delta,c,2\right)<\delta\}$ by Lemma 4.5.9. By Lemma 4.5.10, if $\varphi\left(t\right)\in[-\delta_2,\delta_2]$ for $t\in[-a-|c|\delta,0]$ then $u_n\in[-\delta,\delta]$ for all $n\geqslant0$. This proves Lyapunov stability.

We now need to prove that $\{P\left(1,0,2\right)<1\}\subseteq J$ for any $h\in(0,a]$ and $\{P\left(1,0,2\right)<1,\mu<0\}\subseteq J$ for any $h>0$. Recall that $\{P\left(1,0,2\right)<1\}=\cup_{\delta\in\left(0,\left|\frac{a}{c}\right|\right)}\{P\left(\delta,c,2\right)<\delta\}$. It suffices to show that for all small enough $\delta>0$, (i) for any $h\in(0,a]$, if $P\left(\delta,c,2\right)<\delta$ then $P_{\Theta=1,h}\left(\delta,c,2\right)<\delta$ and (ii) for $h>0$ and $\mu<0$, if $P\left(\delta,c,2\right)<\delta$ then $P_{\Theta=1,h}\left(\delta,c,2\right)<\delta$. We prove this using items (A)-(G) from Lemma 4.5.14 so small enough δ means $\delta\in\left(0,\frac{a}{4|c|}\right)$. Let $(\mu,\sigma)\in\{P\left(\delta,c,2\right)<\delta\}$. If $\mu<0$ then let $h>0$, otherwise let $h\in(0,a]$. Here are all the possible cases:

I. $\tau_1 D-1>\lfloor\frac{\tau_1}{h}\rfloor hD-1\geqslant-\frac{\mu}{\sigma}$. In this case we have $\ell\leqslant m$ for all $\hat{u}\in\left[-\delta,-\frac{\mu}{\sigma}\delta\right]$. From (A) and (B), $P_{\Theta=1,h}\left(\delta,c,2\right)=\sup_{\hat{u}\in\left[-\delta,-\frac{\mu}{\sigma}\delta\right]}\mathcal{S}_2\left(\hat{u},\delta\right)<\delta$.

II. $\tau_1 D - 1 > -\frac{\mu}{\sigma} > \lfloor \frac{\tau_1}{h} \rfloor hD - 1$. There are two cases:

i. $\ell \leqslant m$. From (A) and (B), $\sup_{\hat{u} \in \left[-\delta, \left(\lfloor \frac{\tau_1}{h} \rfloor hD - 1\right)\delta\right]} \mathcal{S}_2\left(\hat{u}, \delta\right) < \delta$.

ii. $\ell = m + 1$. From (E) $\sup_{\hat{u} \in \left[\left(\lfloor \frac{\tau_1}{h} \rfloor hD - 1\right)\delta, -\frac{\mu}{\sigma}\delta\right]} \mathcal{S}_1\left(\hat{u}, \delta\right) \leqslant \mathcal{I}_2\left(-\frac{\mu}{\sigma}\delta, \delta\right)$ for all $h \in (0, \tau_1]$. If $\mu < 0$ then also for $h > a$, $\sup_{\hat{u} \in \left[\left(\lfloor \frac{\tau_1}{h} \rfloor hD - 1\right)\delta, -\frac{\mu}{\sigma}\delta\right]} \mathcal{S}_1\left(\hat{u}, \delta\right) < \delta$.

Thus, since $\mathcal{I}_2\left(-\frac{\mu}{\sigma}\delta\right) = P\left(\delta, c, 2\right) < \delta$ then $P_{\Theta=1,h}\left(\delta, c, 2\right) < \delta$.

III. $\tau_1 D - 1 \leqslant -\frac{\mu}{\sigma}$. There are two possible cases:

i. $\ell \leqslant m$. From (F) and (G), $\sup_{\hat{u} \in [-\delta, (\tau_1 D - 1)\delta]} \mathcal{S}_2\left(\hat{u}, \delta\right) \leqslant \sup_{\hat{u} \in [(\tau_1 D - 1)\delta, \delta]} \mathcal{S}_1\left(\hat{u}, \delta\right)$. From (C) and (D), $\sup_{\hat{u} \in [(\tau_1 D - 1)\delta, \delta]} \mathcal{S}_1\left(\hat{u}, \delta\right) = \mathcal{I}_1\left(-\frac{\mu}{\sigma}\delta, \delta\right)$.

ii. $\ell \geqslant m$. From (C) and (D), $\sup_{\hat{u} \in \left[(\tau_1 D - 1)\delta, -\frac{\mu}{\sigma}\delta\right]} \mathcal{S}_2\left(\hat{u}, \delta\right) \leqslant \sup_{\hat{u} \in \left[(\tau_1 D - 1)\delta, -\frac{\mu}{\sigma}\delta\right]} \mathcal{I}_2\left(\hat{u}, \delta\right)$.

Thus, $P_{\Theta=1,h}\left(\delta, c, 2\right) \leqslant \mathcal{I}_1\left(-\frac{\mu}{\sigma}\delta\right) = P\left(\delta, c, 2\right) < \delta$.

Cases I-III show that $P_{\Theta=1,h}\left(\delta, c, 2\right) < \delta$ for all $(\mu, \sigma) \in \{P\left(\delta, c, 2\right) < \delta\}$.

Finally, for any $h > 0$ (not restricted to $h \in (0, a]$) the same proof shows that for all small enough δ, $\{P\left(\delta, c, 2\right) < \delta, \mu < 0\} \subseteq \{P_{\Theta=1,h}\left(\delta, c, 2\right) < \delta, \mu < 0\}$. □

Theorem 4.5.15 states that if $(\mu, \sigma) \in \{P\left(1, 0, 2\right) < 1\}$ then the BE solution to (4.1.1) is stable for any $h > 0$ if $\mu \geqslant 0$, and for $h \in (0, a]$ if $\mu > 0$. This is the first expression for a stability region that we have found for BE in $\overset{c}{\Sigma}$ that does not depend on the stepsize. However, the stepsize is still restricted to $h \in (0, a]$, a restriction also used by other authors when considering the constant delay case. In the next section we consider $h > a$.

4.6 Stability of BE with $h > a$

In this section we consider the case $h > a$. As usual, let $\varepsilon, a > 0$, $\sigma \leqslant \mu \leqslant \frac{\varepsilon}{a}$ and $\sigma < -\mu$. We also add the other restriction that $\sigma > \mu - \frac{2\varepsilon}{a}$. These restrictions are always satisfied by points in the analytic stability region $\Sigma_\star$ of (4.1.1). Given a constant step-size $h > a$, backward Euler applied to (4.1.1) is given by the zero of $g_{n+1}\left(v\right)$ defined in (4.2.4)

$$g_{n+1}\left(v\right) = v - u_n - \frac{h}{\varepsilon}\left(\mu v + \sigma \tilde{Y}_{n+1}\left(v\right)\right), \tag{4.6.1}$$

where $\tilde{Y}_{n+1}\left(v\right) = \eta_v\left(t_{n+1} - a - cv\right)$ and

$$\eta_v\left(t\right) = \begin{cases} \eta\left(t\right), & \text{if } t \leqslant t_n, \\ \left(1 - \beta\right) u_n + \beta v, \quad \beta = \frac{t - t_n}{h}, & \text{if } t > t_n. \end{cases}$$

For all $v \in \left(-\frac{a}{c}, \frac{h-a}{c}\right)$ we have the overlapping case $t_{n+1} - a - cv \geqslant t_n$. Then $\tilde{Y}^{(1)}_{n+1}\left(v\right) = \left(1 - \beta\right) u_n + \beta v$ where $\beta = \frac{t_{n+1} - a - cv - t_n}{h} = 1 - \frac{a + cv}{h}$. Using this (4.6.1) and grouping powers of

v together yields

$$g_{n+1}\left(v\right) = \frac{\sigma c}{\varepsilon}v^2 + \left(1 + \frac{\sigma\left(a - cu_n\right)}{\varepsilon} - \frac{h\left(\mu+\sigma\right)}{\varepsilon}\right)v - \left(1 + \frac{\sigma a}{\varepsilon}\right)u_n. \tag{4.6.2}$$

Lemma 4.6.1. *Let $\varepsilon, a, c > 0$, $(\mu, \sigma) \in \Sigma_\star$ and $h > a$. Let $|u_n| \in \left(0, \min\left\{\frac{1}{4}\left|\frac{\mu+\sigma}{\sigma}\right|\frac{h-a}{c}, \left|\frac{\varepsilon-h\mu}{\varepsilon+h\sigma}\right|\frac{h-a}{c}, \frac{h-a}{c}, \frac{a}{c}, \right\}\right)$. If $u_n \neq 0$ then there exists $v_* \in \left[-\frac{a}{c}, \frac{h-a}{c}\right]$ such that $g_{n+1}\left(v_*\right) = 0$, $|v_*| < |u_n|$. If $u_n = 0$ then $v_* = 0$ solves $g_{n+1}\left(v_*\right) = 0$. If $\varepsilon - h\mu > 0$ then there are no other solutions to $g_{n+1}\left(v_*\right) = 0$ in $\left[-\frac{a}{c}, \frac{h-a}{c}\right]$.*

Proof. Consider the values of $g_{n+1}\left(v\right)$ at $v = u_n$, 0 and $-u_n$. By our choice of u_n, $t_{n+1} - a - cv \geqslant t_n$ for all three values of v so we may use (4.6.2).

$$\begin{aligned}
g_{n+1}\left(u_n\right) &= -\frac{h\left(\mu+\sigma\right)}{\varepsilon}u_n \\
g_{n+1}\left(0\right) &= -\left(1 + \frac{\sigma a}{\varepsilon}\right)u_n \\
g_{n+1}\left(-u_n\right) &= \left(2\frac{\sigma c}{\varepsilon}u_n + \frac{h\left(\mu+\sigma\right)}{\varepsilon} - 2\left(1 + \frac{\sigma a}{\varepsilon}\right)\right)u_n
\end{aligned}$$

Obviously if $u_n = 0$ then $v_* = 0$ is a solution to $g_{n+1}\left(v\right) = 0$. So let $u_n \neq 0$. Consider the term $\frac{h(\mu+\sigma)}{\varepsilon} - 2\left(1 + \frac{\sigma a}{\varepsilon}\right)$ in the expression for $g_{n+1}\left(-u_n\right)$. Since $(\mu, \sigma) \in \Sigma_\star$ then $\sigma \geqslant \mu - \frac{2\varepsilon}{a}$ and

$$\begin{aligned}
\frac{h\left(\mu+\sigma\right)}{\varepsilon} - 2\left(1 + \frac{\sigma a}{\varepsilon}\right) &= -2 - \frac{\sigma a}{\varepsilon} + \frac{h\mu}{\varepsilon} + \frac{\sigma a}{\varepsilon}\left(\frac{h}{a} - 1\right), \\
&\leqslant -2 - \frac{a}{\varepsilon}\left(\mu - \frac{2\varepsilon}{a}\right) + \frac{h\mu}{\varepsilon} + \frac{\sigma a}{\varepsilon}\left(\frac{h}{a} - 1\right), \\
&= \frac{\mu+\sigma}{\varepsilon}\left(h - a\right).
\end{aligned}$$

Since $h > a$ and $\mu + \sigma < 0$ then $\frac{h(\mu+\sigma)}{\varepsilon} - 2\left(1 + \frac{\sigma a}{\varepsilon}\right) < 0$. By our choice of u_n,

$$2\frac{\sigma c}{\varepsilon}u_n + \frac{h\left(\mu+\sigma\right)}{\varepsilon} - 2\left(1 + \frac{\sigma a}{\varepsilon}\right) < 2\frac{|\sigma|\, c}{\varepsilon}\left|\frac{\mu+\sigma}{4\sigma}\right|\frac{h-a}{c} + \frac{\mu+\sigma}{\varepsilon}\left(h-a\right) < \frac{\mu+\sigma}{2\varepsilon}\left(h-a\right) < 0.$$

Thus, the sign of $g_{n+1}\left(-u_n\right)$ is the opposite sign of u_n. Also, since $\mu + \sigma < 0$, $g_{n+1}\left(u_n\right)$ has the same sign as u_n. Together these results imply that $g_{n+1}\left(-\left|u_n\right|\right) < 0$ and $g_{n+1}\left(\left|u_n\right|\right) > 0$. Then there must be a root v to $g_{n+1}\left(v\right)$ such that $v \in \left[-\left|u_n\right|, \left|u_n\right|\right] \subseteq \left[-\frac{a}{c}, \frac{h-a}{c}\right]$.

Now consider the value of the function at the two endpoints,

$$\begin{aligned} g_{n+1}\left(-\frac{a}{c}\right) &= -u_n + \frac{a}{c}\left(\frac{(\mu+\sigma)\,h}{\varepsilon} - 1\right), \\ g_{n+1}\left(\frac{h-a}{c}\right) &= -u_n\left(1+\frac{h\sigma}{\varepsilon}\right) + \frac{h-a}{c}\left(1-\frac{h\mu}{\varepsilon}\right). \end{aligned}$$

Since $\mu+\sigma<0$ and $|u_n|<\frac{a}{c}$ then easily $g_{n+1}\left(-\frac{a}{c}\right)<0$. Since we assume $|u_n|<\left|\frac{\varepsilon-h\mu}{\varepsilon+h\sigma}\right|\frac{h-a}{c}$ then $\operatorname{sign}\left(g_{n+1}\left(\frac{h-a}{c}\right)\right)=\operatorname{sign}\left(\varepsilon-h\mu\right)$.

Thus, we now have $g_{n+1}\left(-\frac{a}{c}\right)<0$, $g_{n+1}\left(-|u_n|\right)<0$, $g_{n+1}\left(|u_n|\right)>0$ and $\operatorname{sign}\left(g_{n+1}\left(\frac{h-a}{c}\right)\right)=\operatorname{sign}\left(\varepsilon-h\mu\right)$. If $\mu<0$ and $\sigma<0$ then $g_{n+1}(v)$ is a concave down quadratic function that is positive at $\frac{h-a}{c}$. Then this function has only one root in $\left[-\frac{a}{c},\frac{h-a}{c}\right]$. If $\mu<0$ and $\sigma>0$ then $g_{n+1}(v)$ is a concave up quadratic function that is negative at $-\frac{a}{c}$. Then this function has only one root in the interval $\left[-\frac{a}{c},\frac{h-a}{c}\right]$ as well. If $\mu>0$ then automatically $\sigma<0$ in order to satisfy $\mu+\sigma<0$. Then $g_{n+1}(v)$ is a concave down quadratic function with one root in $\left[-\frac{a}{c},\frac{h-a}{c}\right]$ if $\varepsilon-h\mu>0$ and two if $\varepsilon-h\mu<0$. The case $\varepsilon-h\mu=0$ has one root if $u_n>0$ and two roots if $u_n<0$. In the case where there are two roots, one root is between $|u_n|$ and $\frac{h-a}{c}$. □

Remark 4.6.2. In particular, if $1+\frac{\sigma a}{\varepsilon}=0$ then $g_{n+1}(0)=0$ so we can choose $v_*=0$. If $1+\frac{\sigma a}{\varepsilon}>0$ then $g_{n+1}(0)$ has the opposite sign as $g_{n+1}(u_n)$ so v_* is between 0 and u_n. If $1+\frac{\sigma a}{\varepsilon}<0$ then v_* is between 0 and $-u_n$.

Theorem 4.6.3. *Let* $\varepsilon,a,c>0$, $(\mu,\sigma)\in\Sigma_\star$ *and* $h>a$. *Let* $\delta\in\left(0,\min\left\{\frac{1}{4}\left|\frac{\mu+\sigma}{\sigma}\right|\frac{h-a}{c},\left|\frac{\varepsilon-h\mu}{\varepsilon+h\sigma}\right|\frac{h-a}{c},\frac{h-a}{c},\frac{a}{c},\right\}\right)$. *If* $\varphi(t)$ *is continuous and* $\varphi(t)\in[-\delta,\delta]$ *for all* $t\in[-a-c\delta,0]$ *then there exists a BE solution to* (4.1.1) $\{u_n\}_{n\geqslant 0}$ *such that the following are satisfied:*

(A) $u_n\in[-\delta,\delta]$ *for all* $n\geqslant 0$.

(B) For all $n\geqslant 0$, $|u_{n+1}|<|u_n|$ *if* $u_n\neq 0$ *and* $u_{n+1}=0$ *if* $u_n=0$.

(C) $\lim\limits_{n\to\infty}u_n=0$.

Proof. The proof of (A) and (B) is by strong induction which easily follows from Lemma 4.6.1. The convergence to zero (C) follows from the fact that $g_{n+1}(u_n)$ and $g_{n+1}(-u_n)$ are nonzero unless $u_n=0$. Note that there may be other BE solutions to (4.1.1). □

Theorem 4.6.4. *Let* $\varepsilon,a,c>0$, $\mu<0$, $(\mu,\sigma)\in\{P(1,0,2)<1\}$ *(P is as defined in Definition 3.4.2) and* $h>a$. *Then for any* $\delta\in\left(0,\min\left\{\frac{1}{4}\left|\frac{\mu+\sigma}{\sigma}\right|\frac{h-a}{c},\left|\frac{\varepsilon-h\mu}{\varepsilon+h\sigma}\right|\frac{h-a}{c},\frac{h-a}{c},\frac{a}{4c},\right\}\right)$ *there* *exists $\delta_2 > 0$ such that if $\varphi(t) \in [-\delta_2, \delta_2]$ for $t \in [-a - |c|\delta, 0]$ then any BE solution to* (4.1.1) $\{u_n\}_{n \geqslant 0}$ *satisfies $u_n \in [-\delta, \delta]$ for $n \geqslant 0$ and $\lim_{n\to\infty} u_n = 0$.*

Proof. By Theorem 4.5.15 there exists $\delta_2 > 0$ such that if $\varphi(t) \in [-\delta_2, \delta_2]$ for $t \in [-a - |c|\delta, 0]$ then any BE solution to (4.1.1) $\{u_n\}_{n \geqslant 0}$ satisfies $u_n \in [-\delta, \delta]$. By our choice of δ we are always in the overlapping regime and Lemma 4.6.1 applies. Since $\mu < 0$ then $\varepsilon - h\mu > 0$ and at every time step there is always only one root to $g_{n+1}(v)$ so the BE solution is unique by Lemma 4.6.1. By the same lemma, for all $n \geqslant 0$, $|u_{n+1}| < |u_n|$ if $u_n \neq 0$ and $u_{n+1} = 0$ if $u_n = 0$. Since $g_{n+1}(u_n)$ and $g_{n+1}(-u_n)$ are nonzero unless $u_n = 0$, $\lim_{n\to\infty} u_n = 0$. □

4.7 Description of Theta methods

The Θ methods are RK methods parameterized by $\Theta \in [0, 1]$. They are given by the following Butcher tableau,

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1-\Theta & \Theta \\ \hline & 1-\Theta & \Theta \end{array}.$$

The method with $\Theta = 0$ is forward Euler, $\Theta = \frac{1}{2}$ is the trapezoidal rule and $\Theta = 1$ is backward Euler. The trapezoidal rule is of order two while the other methods of order one. Recall from Theorem 1.2.2 that in order to preserve the order p of an ODE method when it is extended to solve DDEs we need to use a continuous extension that has at least order $p - 1$. Since linear interpolation is of order one, then if we use this as the continuous extension of a general Θ method we preserve the orders of the methods. In particular, we preserve the second order trapezoidal rule. Theorem 1.2.2 lists other requirements necessary to preserve the order of a method but in this chapter we focus on stability. Issues involved in the implementation of methods are discussed in Chapter 5.

Consider the following DDE with one general delay,

$$\begin{aligned} \dot{u}(t) &= f(t, u(t), \alpha(t, u(t))), & t &\geqslant t_0, \\ u(t) &= \varphi(t), & t &< t_0, \end{aligned} \tag{4.7.1}$$

where we assume $\alpha(t, u(t)) \leqslant t$ for all times. The Θ method with constant stepsize h and linear interpolation applied to (4.7.1) can be written as

$$u_{n+1} = u_n + (1-\Theta) hf\left(t_n, u_n, \tilde{Y}^{(1)}_{n+1}\right) + \Theta hf\left(t_{n+1}, u_{n+1}, \tilde{Y}^{(2)}_{n+1}\right) \tag{4.7.2}$$

where the spurious stages are given by values of the continuous extension $\tilde{Y}_{n+1}^{(1)} = \eta\left(\alpha\left(t_n, u_n\right)\right)$ and $\tilde{Y}_{n+1}^{(2)} \eta\left(\alpha\left(t_{n+1}, u_{n+1}\right)\right)$. The continuous extension we use here is linear interpolation,

$$\eta\left(t_n + \beta h\right) = u_n + \left(1 - \Theta\right) \beta h f\left(t_n, u_n, \tilde{Y}_{n+1}^{(1)}\right) + \Theta \beta h f\left(t_{n+1}, u_{n+1}, \tilde{Y}_{n+1}^{(2)}\right).$$

One may also think of linear interpolation in the step-by-step manner described on page 84.

4.8 Stability of Theta methods in $\overset{\Delta}{\Sigma}$

Here we derive the delay independent stability regions of the Θ methods using the same arguments as in Section 4.3. This method does not yield stability in all of the cone $\overset{\Delta}{\Sigma}$ for any method other than backward Euler ($\Theta = 1$). These results may not be sharp since it is known that the Θ methods are GP-stable if and only if $\Theta \in \left[\frac{1}{2}, 1\right]$ [37]. Although GP-stability deals with the constant delay case, numerical simulations show that trapezoidal rule appears to be stable also in all of $\overset{\Delta}{\Sigma}$ for all $h \in (0, a)$. Nevertheless, since there are currently no results on stability of methods for our model state dependent DDE, we continue here with the contraction argument as in Section 4.3.

The Θ method applied to (4.1.1) has the following form:

$$u_{n+1} = u_n + \frac{h}{\varepsilon}\left[\left(1 - \Theta\right)\left(\mu u_n + \sigma \tilde{Y}_{n+1}^{(1)}\right) + \Theta\left(\mu u_{n+1} + \sigma \tilde{Y}_{n+1}^{(2)}\right)\right]. \tag{4.8.1}$$

For convenience use $\tilde{Y}_{n+1} = \left(1 - \Theta\right) \tilde{Y}_{n+1}^{(1)} + \Theta \tilde{Y}_{n+1}^{(2)}$. We derive an expression for u_{n+1} ignoring the dependence of $Y_{n+1}^{(2)}$ on u_{n+1},

$$u_{n+1} = \left(\frac{\varepsilon + h\mu\left(1 - \Theta\right)}{\varepsilon - h\mu\Theta}\right) u_n + \frac{h\sigma}{\varepsilon - h\mu\Theta} \tilde{Y}_{n+1}. \tag{4.8.2}$$

Assume that the numerical solution up to u_n and the entire history function is bounded inside $[-\delta, \delta]$. Assume also that there is no overlapping. Then $\tilde{Y}_{n+1}$ is also bounded by δ and

$$\left|u_{n+1}\right| \leqslant \left[\frac{\left|\varepsilon + h\mu\left(1 - \Theta\right)\right| + \left|h\sigma\right|}{\left|\varepsilon - h\mu\Theta\right|}\right] \delta. \tag{4.8.3}$$

Consider the parameter region where

$$\frac{\left|\varepsilon + h\mu\left(1 - \Theta\right)\right| + \left|h\sigma\right|}{\left|\varepsilon - h\mu\Theta\right|} < 1.$$

For $\mu < 0$ then we can rewrite this as

$$|\sigma| < \begin{cases} -\mu, & \text{if } \mu \geqslant -\frac{\varepsilon}{h(1-\Theta)}, \\ \frac{2\varepsilon + h\mu(1-2\Theta)}{h}, & \text{if } \mu < -\frac{\varepsilon}{h(1-\Theta)} \ . \end{cases}$$

For $\Theta < \frac{1}{2}$ this ends at the μ axis when $2\varepsilon + h\mu\,(1-2\Theta) = 0$.

Now consider the case with overlapping ($t_n < t_{n+1} - a - cu_{n+1} \leqslant t_{n+1}$). We will prove that (4.8.3) still holds in this case if $\frac{|\varepsilon + h\mu(1-\Theta)| + |h\sigma|}{|\varepsilon - h\mu\Theta|} < 1$. Since we are assuming overlapping, $\tilde{Y}_{n+1} = (1-\Theta)\,\tilde{Y}^{(1)}_{n+1} + \Theta\,[(1-\beta)\,u_n + \beta u_{n+1}]$. Using this in (4.8.2) we derive

$$u_{n+1} = \left(\frac{\varepsilon + h\mu\,(1-\Theta)}{\varepsilon - h\mu\Theta}\right) u_n + \frac{h\sigma}{\varepsilon - h\mu\Theta}\left[(1-\Theta)\,\tilde{Y}^{(1)}_{n+1} + \Theta\,[(1-\beta)\,u_n + \beta u_{n+1}]\right].$$

Solving for u_{n+1} leads to

$$u_{n+1} = \left(\frac{\varepsilon + h\mu\,(1-\Theta) + h\sigma\Theta\,(1-\beta)}{\varepsilon - h\mu\Theta - h\sigma\Theta\beta}\right) u_n + \frac{h\sigma\,(1-\Theta)}{\varepsilon - h\mu\Theta - h\sigma\Theta\beta}\tilde{Y}^{(1)}_{n+1},$$

$$|u_{n+1}| \leqslant \left[\frac{|\varepsilon + h\mu\,(1-\Theta) + h\sigma\Theta\,(1-\beta)| + |h\sigma\,(1-\Theta)|}{|\varepsilon - h\mu\Theta - h\sigma\Theta\beta|}\right]\delta.$$

Let $\varepsilon - h\mu\Theta > 0$. If $\sigma \leqslant 0$ then since $\beta \in [0,1]$ this easily leads to (4.8.3) again. So let $\sigma > 0$ and consider $\frac{\varepsilon + h\mu(1-\Theta) + h\sigma\Theta(1-\beta)}{\varepsilon - h\mu\Theta - h\sigma\Theta\beta}$ as a function of $\beta \in [0,1]$. If $\mu + \sigma < 0$ then the derivative is always negative. If $\varepsilon + h\mu\,(1-\Theta) \geqslant 0$ then this function is nonnegative at $\beta = 1$ so its absolute value is maximized at $\beta = 0$. This easily leads to (4.8.3) again. If $\varepsilon + h\mu\,(1-\Theta) < 0$ then

$$\begin{aligned} &\frac{|\varepsilon + h\mu\,(1-\Theta) + h\sigma\Theta\,(1-\beta)| + |h\sigma\,(1-\Theta)|}{|\varepsilon - h\mu\Theta - h\sigma\Theta\beta|} \\ &\leqslant \max\left\{\frac{|\varepsilon + h\mu\,(1-\Theta) + h\sigma\Theta| + |h\sigma\,(1-\Theta)|}{|\varepsilon - h\mu\Theta|}, \frac{|\varepsilon + h\mu\,(1-\Theta)| + |h\sigma\,(1-\Theta)|}{|\varepsilon - h\mu\Theta - h\sigma\Theta|}\right\}, \\ &\leqslant \max\left\{\frac{|\varepsilon + h\mu\,(1-\Theta)| + |h\sigma|}{|\varepsilon - h\mu\Theta|}, \frac{|\varepsilon + h\mu\,(1-\Theta)| + |h\sigma\,(1-\Theta)|}{|\varepsilon - h\mu\Theta - h\sigma\Theta|}\right\}. \end{aligned} \tag{4.8.4}$$

Since we are considering the case, $\varepsilon - h\mu\Theta > 0$, $\mu + \sigma < 0$, $\sigma > 0$ and $\varepsilon + h\mu\,(1-\Theta) < 0$ then

$$\frac{|\varepsilon + h\mu\,(1-\Theta)| + |h\sigma\,(1-\Theta)|}{|\varepsilon - h\mu\Theta - h\sigma\Theta|} - \frac{|\varepsilon + h\mu\,(1-\Theta)| + |h\sigma|}{|\varepsilon - h\mu\Theta|} = -h\sigma\frac{\varepsilon\Theta - h\mu\Theta^2 + \varepsilon - h\sigma\Theta}{(\varepsilon - h\mu\Theta - h\sigma\Theta)\,(\varepsilon - h\mu\Theta.)} \tag{4.8.5}$$

Assuming that $\frac{|\varepsilon+h\mu(1-\Theta)|+|h\sigma|}{|\varepsilon-h\mu\Theta|} < 1$ then $\sigma < \frac{2\varepsilon+h\mu(1-2\Theta)}{h}$, and

$$\begin{aligned}\varepsilon\Theta - h\mu\Theta^2 + \varepsilon - h\sigma\Theta &> \varepsilon\Theta - h\mu\Theta^2 + \varepsilon - (2\varepsilon\Theta + h\mu\Theta\,(1-2\Theta)) \\ &= (1-\Theta)\,(\varepsilon - h\mu\Theta) > 0.\end{aligned}$$

By (4.8.5), $\frac{|\varepsilon+h\mu(1-\Theta)|+|h\sigma(1-\Theta)|}{|\varepsilon-h\mu\Theta-h\sigma\Theta|} < \frac{|\varepsilon+h\mu(1-\Theta)|+|h\sigma|}{|\varepsilon-h\mu\Theta|}$. From (4.8.4), we derive (4.8.3) again.

$$|u_{n+1}| \leqslant \left[\frac{|\varepsilon + h\mu\,(1-\Theta)| + |h\sigma|}{|\varepsilon - h\mu\Theta|}\right]\delta \tag{4.8.6}$$

We have now shown that this inequality is true for the non overlapping cases and for the overlapping cases where the stepsize and parameters satisfy $\varepsilon - h\mu\Theta > 0$, $\mu + \sigma < 0$ and one of the following conditions:

1. $\sigma \leqslant 0$,
2. $\sigma > 0$, $\varepsilon + h\mu\,(1-\Theta) \geqslant 0$,
3. $\sigma > 0$, $\varepsilon + h\mu\,(1-\Theta) < 0$, $\frac{|\varepsilon+h\mu(1-\Theta)|+|h\sigma|}{|\varepsilon-h\mu\Theta|} < 1$.

This shows that (4.8.6) holds for all points for which the parameters satisfy $\frac{|\varepsilon+h\mu(1-\Theta)|+|h\sigma|}{|\varepsilon-h\mu\Theta|} < 1$.

Theorem 4.8.1. *Let $\varepsilon, a, c > 0$, $\mu < 0$, $\frac{|\varepsilon+h\mu(1-\Theta)|+|h\sigma|}{|\varepsilon-h\mu\Theta|} < 1$ and $h > 0$. For every $\delta > 0$, if the history function $\varphi(t)$ is continuous and $\varphi(t) \in [-\delta, \delta]$ for all $t \in [-a - c\delta, 0]$ then the Θ method solution to* (4.1.1) *$\{u_n\}_{n\geqslant 0}$ satisfies $u_n \in [-\delta, \delta]$ for all $n \geqslant 0$ and $\lim_{n\to\infty} u_n = 0$.*

Proof. Similar to the proof of Theorem 4.3.1. □

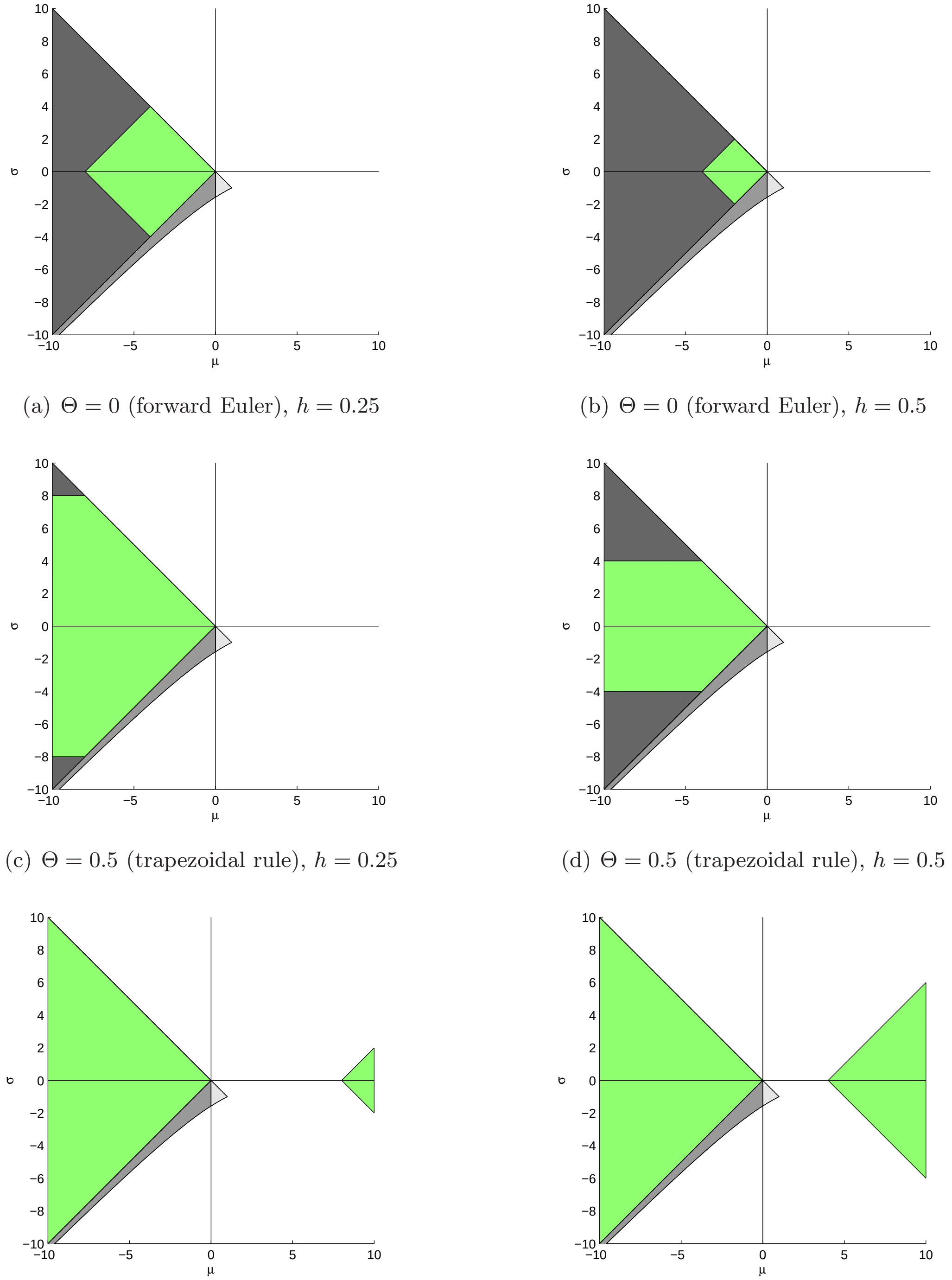
Sample plots of the sets $\left\{\frac{|\varepsilon+h\mu(1-\Theta)|+|h\sigma|}{|\varepsilon-h\mu\Theta|} < 1\right\}$ are shown in Figure 4–7. Note that our bounds give the correct known stability region of Θ methods for the case when $\sigma = 0$ and (4.1.1) is an ODE. If (μ, σ) is in one of those sets with $\mu > 0$, the solutions also converge to zero as long as there is no overlapping. This can be guaranteed if $h \in (0, a)$ and the bound δ is chosen to be small enough. The part of the set in the $\mu > 0$ half-plane can be written as

$$|\sigma| < \frac{-2\varepsilon + h\mu\,(2\Theta - 1)}{h}$$

If $\Theta \leqslant \frac{1}{2}$ then we do not have this region at all.

4.9 Stability of Theta methods using a discrete Gronwall argument

In this section we generalise the results in Section 4.4 to general Θ methods. We prove the asymptotic stability of Θ method solutions to (4.1.1) in a portion of $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$. Once again we set $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$, $\sigma < -\mu$, and automatically $\sigma < 0$.



(a) $\Theta = 0$ (forward Euler), $h = 0.25$

(b) $\Theta = 0$ (forward Euler), $h = 0.5$

(c) $\Theta = 0.5$ (trapezoidal rule), $h = 0.25$

(d) $\Theta = 0.5$ (trapezoidal rule), $h = 0.5$

(e) $\Theta = 1$ (backward Euler), $h = 0.25$

(f) $\Theta = 1$ (backward Euler), $h = 0.5$

Figure 4–7: The sets $\left\{\frac{|\varepsilon+h\mu(1-\Theta)|+|h\sigma|}{|\varepsilon-h\mu\Theta|} < 1\right\}$ for different values of Θ and h are shaded green and plotted with $\varepsilon = a = 1$. We have shown that if (μ, σ) is in this set and $\mu < 0$ then the Θ method solution to (4.1.1) goes zero for all stepsizes.

Lemma 4.9.1. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu$ and $\sigma < -\mu$. Let the history function $\varphi(t)$ and the Θ method solution $\{u_n\}$ to (4.1.1) be bounded by $[-\delta, \delta]$ for all $t \leqslant 0$ and $n \geqslant 0$. Then for any stepsize $h > 0$, the Θ method solution must behave in one of the following manners:*

(A) There exists $N \in \mathbb{N}$ such that $u_n \downarrow 0$ for $n > N$

(B) There exists $N \in \mathbb{N}$ such that $u_n \uparrow 0$ for $n > N$

(C) For every $N > 0$ there exists $N_1, N_2 \in \mathbb{N}$ and $N_1, N_2 > N$ such that the solution attains a positive maximum at N_1 and a negative minimum at N_2.

Proof. Similar to the proof of Lemma 4.4.1. □

At every time step, the value of the spurious stages can be found by first solving for m_1, m_2, β_1 and β_2,

$$\begin{array}{ll} m_1 = \left\lceil \frac{a+cu_n}{h} \right\rceil, & \beta_1 = m_1 - \frac{a+cu_n}{h}, \\ m_2 = \left\lfloor \frac{a+cu_{n+1}}{h} \right\rfloor, & \beta_2 = m_2 + 1 - \frac{a+cu_{n+1}}{h} \end{array}. \tag{4.9.1}$$

This is valid whether or not there has been overlapping. From these relations we derive

$$\alpha(t_n, u_n) = t_n - a - cu_n = t_{n-m_1} + \beta_1 h = (1-\beta_1)\, t_{n-m} + \beta_1 t_{n-m_1+1},$$

$$\alpha(t_{n+1}, u_{n+1}) = t_{n+1} - a - cu_{n+1} = t_{n-m_2} + \beta_2 h = (1-\beta_2)\, t_{n-m_2} + \beta_2 t_{n-m_2+1},$$

Thus using linear interpolation, the values of the spurious stages $\tilde{Y}^{(1)}_{n+1}$ and $\tilde{Y}^{(2)}_{n+1}$ are

$$Y^{(1)}_{n+1} = (1-\beta_1)\, u_{n-m_1} + \beta u_{n-m_1+1} \in \left[\frac{m_1 h - a}{c}, \frac{(m_1+1)\,h - a}{c}\right], \tag{4.9.2}$$

$$Y^{(2)}_{n+1} = (1-\beta_2)\, u_{n-m_2} + \beta u_{n-m_2+1} \in \left[\frac{m_2 h - a}{c}, \frac{(m_2+1)\,h - a}{c}\right]. \tag{4.9.3}$$

Also define

$$B = \frac{1}{(1-\Theta)\frac{\varepsilon - h\mu(\Theta-\beta_1)}{\varepsilon + h\mu(1-\Theta)}\left(\frac{\varepsilon - h\mu\Theta}{\varepsilon + h\mu(1-\Theta)}\right)^{m_1} + \Theta\frac{\varepsilon - h\mu(\Theta-\beta_2)}{\varepsilon + h\mu(1-\Theta)}\left(\frac{\varepsilon - h\mu\Theta}{\varepsilon + h\mu(1-\Theta)}\right)^{m_2}}, \tag{4.9.4}$$

$$\tilde{Y}_{n+1} = (1-\Theta)\,\tilde{Y}^{(1)}_{n+1} + \Theta\tilde{Y}^{(2)}_{n+1}. \tag{4.9.5}$$

Lemma 4.9.2. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let $\Theta \in [0,1]$ and let the stepsize $h > 0$ be such that $\frac{\varepsilon + h\mu(1-\Theta)}{\varepsilon - h\mu\Theta} > 0$. Let $-\frac{a}{c} < L < 0 < M$, $n \in \mathbb{N}$, $n \geqslant 0$ and let the Θ method solution to (4.1.1) $\{u_i\}_{i=0}^{n}$ satisfy $u_i \in [L, M]$ for $i = \max\{0, n - \left\lceil \frac{a+cM}{h} \right\rceil\}, ..., n$. If $t_n - a - c\delta < 0$ then* *let $\varphi(s) \in [L, M]$ for $s \in [t_n - a - c\delta, 0]$. Let $\mu \neq 0$. If $u_{n+1} \leqslant u_n$ then*

$$\frac{1}{B} u_{n+1} - \tilde{Y}_{n+1} \geqslant -\frac{\sigma M}{\mu} \left[\frac{1}{B} - 1 \right], \tag{4.9.6}$$

and if $u_{n+1} \geqslant u_n$ then

$$\frac{1}{B} u_{n+1} - \tilde{Y}_{n+1} \leqslant -\frac{\sigma L}{\mu} \left[\frac{1}{B} - 1 \right], \tag{4.9.7}$$

Let $\mu = 0$. If $u_{n+1} \leqslant u_n$ then

$$u_{n+1} - \tilde{Y}_{n+1} \geqslant \left[(1 - \Theta)(a + cu_n + h) + \Theta(a + cu_{n+1}) \right] \frac{\sigma M}{\varepsilon}, \tag{4.9.8}$$

and if $u_{n+1} \geqslant u_n$ then

$$u_{n+1} - \tilde{Y}_{n+1} \leqslant \left[(1 - \Theta)(a + cu_n + h) + \Theta(a + cu_{n+1}) \right] \frac{\sigma L}{\varepsilon}. \tag{4.9.9}$$

Proof. Start with the equation for the Θ method applied to (4.1.1) and solve for u_{n+1},

$$u_{n+1} = u_n + \frac{h}{\varepsilon} \left[(1 - \Theta) \left(\mu u_n + \sigma \tilde{Y}^{(1)}_{n+1} \right) + \Theta \left(\mu u_{n+1} + \sigma \tilde{Y}^{(2)}_{n+1} \right) \right]$$

$$u_{n+1} = \frac{\varepsilon + h\mu(1 - \Theta)}{\varepsilon - h\mu\Theta} u_n + \frac{h\sigma}{\varepsilon - h\mu\Theta} \tilde{Y}_{n+1} \geqslant \frac{\varepsilon + h\mu(1 - \Theta)}{\varepsilon - h\mu\Theta} u_n + \frac{h\sigma}{\varepsilon - h\mu\Theta} M \tag{4.9.10}$$

Let $\mu \neq 0$. Use $m = m_i$ and $\beta = \beta_i$ for $i = 1$. Using the discrete Gronwall lemma,

$$u_{n+1} \geqslant \frac{\frac{h\sigma M}{\varepsilon - h\mu\Theta}}{1 - \frac{\varepsilon + h\mu(1-\Theta)}{\varepsilon - h\mu\Theta}} \left(1 - \left(\frac{\varepsilon + h\mu(1 - \Theta)}{\varepsilon - h\mu\Theta} \right)^{m+1} \right) + \left(\frac{\varepsilon + h\mu(1 - \Theta)}{\varepsilon - h\mu\Theta} \right)^{m+1} u_{n-m}$$

This simplifies to

$$u_{n+1} \geqslant -\frac{\sigma M}{\mu} \left(1 - \left(\frac{\varepsilon + h\mu(1 - \Theta)}{\varepsilon - h\mu\Theta} \right)^{m+1} \right) + \left(\frac{\varepsilon + h\mu(1 - \Theta)}{\varepsilon - h\mu\Theta} \right)^{m+1} u_{n-m}. \tag{4.9.11}$$

Similarly, we also have

$$u_{n+1} \geqslant -\frac{\sigma M}{\mu} \left(1 - \left(\frac{\varepsilon + h\mu(1 - \Theta)}{\varepsilon - h\mu\Theta} \right)^{m} \right) + \left(\frac{\varepsilon + h\mu(1 - \Theta)}{\varepsilon - h\mu\Theta} \right)^{m} u_{n-m+1}. \tag{4.9.12}$$

Take $\frac{\varepsilon-h\mu\Theta}{\varepsilon+h\mu(1-\Theta)}(1-\beta)\times$ (4.9.11) $+\beta\times$ (4.9.12),

$$\begin{aligned}\left(\frac{\varepsilon-h\mu\Theta}{\varepsilon+h\mu\left(1-\Theta\right)}\left(1-\beta\right)+\beta\right)u_{n+1}&\\ \geqslant -\frac{\sigma M}{\mu}\left(\frac{\varepsilon-h\mu\Theta}{\varepsilon+h\mu\left(1-\Theta\right)}\left(1-\beta\right)+\beta-\left(\frac{\varepsilon+h\mu\left(1-\Theta\right)}{\varepsilon-h\mu\Theta}\right)^{m}\right)&\\ +\left(\frac{\varepsilon+h\mu\left(1-\Theta\right)}{\varepsilon-h\mu\Theta}\right)^{m}\left(\left(1-\beta\right)u_{n-m}+\beta u_{n-m+1}\right).&\end{aligned}$$

Using $\frac{\varepsilon-h\mu\Theta}{\varepsilon+h\mu(1-\Theta)}(1-\beta)+\beta=\frac{\varepsilon-h\mu(\Theta-\beta)}{\varepsilon+h\mu(1-\Theta)}$ and $(1-\beta)u_{n-m}+\beta u_{n-m+1}=Y_{n+1}^{(i)}$ this becomes

$$\begin{aligned}\frac{\varepsilon-h\mu\left(\Theta-\beta\right)}{\varepsilon+h\mu\left(1-\Theta\right)}u_{n+1}&\\ \geqslant -\frac{\sigma M}{\mu}\left(\frac{\varepsilon-h\mu\left(\Theta-\beta\right)}{\varepsilon+h\mu\left(1-\Theta\right)}-\left(\frac{\varepsilon+h\mu\left(1-\Theta\right)}{\varepsilon-h\mu\Theta}\right)^{m}\right)+\left(\frac{\varepsilon+h\mu\left(1-\Theta\right)}{\varepsilon-h\mu\Theta}\right)^{m}\tilde{Y}_{n+1}^{(i)}.&\end{aligned}$$

Rewrite as

$$\begin{aligned}\frac{\varepsilon-h\mu\left(\Theta-\beta_1\right)}{\varepsilon+h\mu\left(1-\Theta\right)}\left(\frac{\varepsilon-h\mu\Theta}{\varepsilon+h\mu\left(1-\Theta\right)}\right)^{m_1}u_{n+1}&\\ \geqslant -\frac{\sigma M}{\mu}\left(\frac{\varepsilon-h\mu\left(\Theta-\beta_1\right)}{\varepsilon+h\mu\left(1-\Theta\right)}\left(\frac{\varepsilon-h\mu\Theta}{\varepsilon+h\mu\left(1-\Theta\right)}\right)^{m_1}-1\right)+\tilde{Y}_{n+1}^{(1)}.&\quad(4.9.13)\end{aligned}$$

Using the same derivation as before, we derive (4.9.14) using $m=m_i$ and $\beta=\beta_i$ with $i=2$,

$$\begin{aligned}\frac{\varepsilon-h\mu\left(\Theta-\beta_2\right)}{\varepsilon+h\mu\left(1-\Theta\right)}\left(\frac{\varepsilon-h\mu\Theta}{\varepsilon+h\mu\left(1-\Theta\right)}\right)^{m_2}u_{n+1}&\\ \geqslant -\frac{\sigma M}{\mu}\left(\frac{\varepsilon-h\mu\left(\Theta-\beta_2\right)}{\varepsilon+h\mu\left(1-\Theta\right)}\left(\frac{\varepsilon-h\mu\Theta}{\varepsilon+h\mu\left(1-\Theta\right)}\right)^{m_2}-1\right)+\tilde{Y}_{n+1}^{(2)}&\quad(4.9.14)\end{aligned}$$

Taking $(1-\Theta)\times$ (4.9.13) + $\Theta\times$ (4.9.14) yields (4.9.6). The derivation of (4.9.7) is similar. Now let $\mu=0$. Then (4.9.10) becomes simply $u_{n+1}\geqslant u_n+\frac{h\sigma M}{\varepsilon}$. Let $m=m_i$ and $\beta=\beta_i$ with $i=1$. Applying a discrete Gronwall inequality yields

$$u_{n+1}\geqslant u_{n-m_i}+\left(m_i+1\right)\frac{h\sigma M}{\varepsilon},\quad u_{n+1}\geqslant u_{n-m_i+1}+m_i\frac{h\sigma M}{\varepsilon}$$

Taking $(1-\beta_i)\times$ the first equation + $\beta_i\times$ the second equation yields

$$u_{n+1}\geqslant\left(1-\beta_i\right)u_{n-m_i}+\beta_i u_{n-m_i+1}+\left(m_i+1-\beta_i\right)h\sigma M\geqslant\tilde{Y}_{n+1}^{(i)}+\left(m_i+1-\beta_i\right)h\sigma M.$$

Taking $(1-\Theta)\times$ this equation with $i=1$ + $\Theta\times$ this equation with $i=2$ and using (4.9.1) yields (4.9.8). The derivation of (4.9.9) is similar. $\square$

Definition 4.9.3. Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let $\Theta \in [0,1]$ and let the stepsize $h>0$ be such that $\frac{\varepsilon + h\mu(1-\Theta)}{\varepsilon - h\mu\Theta} > 0$. For $(v,w) \in \left(-\frac{a}{c}, \infty\right)^2$ define

$$r_{\Theta,h}(v,w) = \frac{\sigma}{\mu} \frac{1 - \left[(1-\Theta)\, e^{-\frac{\mu(a+cv+h)}{\varepsilon + h\mu(1-\Theta)}} + \Theta e^{-\frac{\mu(a+cw)}{\varepsilon+h\mu(1-\Theta)}}\right]^{-1}}{1 + \frac{\mu}{\sigma}\left[(1-\Theta)\, e^{-\frac{\mu(a+cv+h)}{\varepsilon + h\mu(1-\Theta)}} + \Theta e^{-\frac{\mu(a+cw)}{\varepsilon+h\mu(1-\Theta)}}\right]^{-1}}$$

Note that $r_{\Theta=1,h}(v,w) = r(w)$ from Definitions 3.2.4 and 4.4.3. Also define the following $R_{\Theta,h}(v,w)$ function

$$R_{\Theta,h}(v,w) = \begin{cases} \frac{\sigma}{\mu}\left[\frac{1-B(v,w)}{1+\frac{\mu}{\sigma}B(v,w)}\right], & \text{if } \mu \neq 0, \\ -\frac{\sigma}{\varepsilon}\left[(1-\Theta)(a+cv+h) + \Theta(a+cw)\right], & \text{if } \mu = 0. \end{cases}$$

where for each v and w,

$$B(v,w) = \frac{1}{(1-\Theta)\frac{\varepsilon - h\mu(\Theta-\beta_v)}{\varepsilon + h\mu(1-\Theta)}\left(\frac{\varepsilon - h\mu\Theta}{\varepsilon+h\mu(1-\Theta)}\right)^{m_v} + \Theta\frac{\varepsilon - h\mu(\Theta-\beta_w)}{\varepsilon+h\mu(1-\Theta)}\left(\frac{\varepsilon-h\mu\Theta}{\varepsilon+h\mu(1-\Theta)}\right)^{m_w}},$$

and m_v, β_v, m_w and β_w are defined by

$$\begin{aligned} m_v &= \left\lceil \tfrac{a+cv}{h} \right\rceil, \quad & \beta_1 &= m_1 - \tfrac{a+cv}{h}, \\ m_w &= \left\lfloor \tfrac{a+cw}{h} \right\rfloor, \quad & \beta_2 &= m_2 + 1 - \tfrac{a+cw}{h} \end{aligned}$$

For fixed v and w, one may show using L'Hopital's rule that $\lim\limits_{\mu\to 0} r_{\Theta,h}(v,w) = r_{\Theta,h}(v,w)$ and $\lim\limits_{\mu\to 0} R_{\Theta,h}(v,w) = R_{\Theta,h}(v,w)$. So these functions are continuous at $\mu=0$. Also, $R_{\Theta,h}(v,w)$ is a continuous function of v and w.

Lemma 4.9.4. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let $\Theta \in [0,1]$ and let the stepsize $h > 0$ be such that $\frac{\varepsilon+h\mu(1-\Theta)}{\varepsilon-h\mu\Theta} > 0$. Let $-\frac{a}{c} < L < 0 < M$, $n \in \mathbb{N}$, $n \geqslant 0$ and let the Θ method solution to (4.1.1) $\{u_i\}_{i=0}^{n}$ satisfy $u_i \in [L,M]$ for $i = \max\{0, n - \lceil \frac{a+cM}{h} \rceil\}, ..., n$. If $t_n - a - c\delta < 0$ then let $\varphi(s) \in [L,M]$ for $s \in [t_n - a - c\delta, 0]$. If $u_{n+1} \leqslant u_n$ then*

$$u_{n+1} \geqslant -R_{\Theta,h}(u_n, u_{n+1})\, M. \tag{4.9.15}$$

If $u_{n+1} \geqslant u_n$ then

$$u_{n+1} \geqslant -R_{\Theta,h}\left(u_n, u_{n+1}\right) L. \tag{4.9.16}$$

Proof. If $u_{n+1} \leqslant u_n$ then,

$$(1-\Theta)\left(\mu u_n + \sigma \tilde{Y}_{n+1}^{(1)}\right) + \Theta\left(\mu u_{n+1} + \sigma \tilde{Y}_{n+1}^{(2)}\right) \leqslant 0.$$

Using (4.9.5),

$$\tilde{Y}_{n+1} \geqslant -\frac{\mu}{\sigma}\left((1-\Theta) u_n + \Theta u_{n+1}\right)$$

Solving for u_n in (4.8.1) yields,

$$u_n = \frac{\varepsilon - h\mu\Theta}{\varepsilon + h\mu(1-\Theta)} u_{n+1} - \frac{h\sigma}{\varepsilon + h\mu(1-\Theta)} \tilde{Y}_{n+1}.$$

Apply this to the previous inequality and solve for $\tilde{Y}_{n+1}$. This simplifies to $\tilde{Y}_{n+1} \geqslant -\frac{\mu}{\sigma} u_{n+1}$. Substitute this into (4.9.6) and solving for u_{n+1} yields (4.9.15). The derivation of (4.9.16) is similar. □

Lemma 4.9.5. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let $\Theta \in [0,1]$ and let the stepsize $h > 0$ be such that $\frac{\varepsilon + h\mu(1-\Theta)}{\varepsilon - h\mu\Theta} > 0$. Let the model parameters and the stepsize h satisfy $R_{\Theta,h}(0,0) \in (0,1)$. Then there exists a sufficiently small $\delta_* \in \left(0, \left|\frac{a}{c}\right|\right)$ such that $R_{\Theta=1,h}(v,w) \in (0,1)$ for all $v, w \in [-\delta_*, \delta_*]$. Let $\delta = \left|\frac{\varepsilon - h\mu\Theta}{\varepsilon + h\mu(1-\Theta) - h\sigma}\right| \delta_*$. If $\varphi(t)$ is continuous and $\varphi(t) \in [-\delta, \delta]$ for all $t \in [-a - c\delta, 0]$ then the Θ method solution to (4.1.1) $\{u_n\}_{n \geqslant 0}$ satisfies $u_n \in [-\delta, \delta]$ for all $n \geqslant 0$.*

Proof. Similar to the proof of Lemma 4.4.5. □

Theorem 4.9.6. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let $\Theta \in [0,1]$ and let the stepsize $h > 0$ be such that $\frac{\varepsilon + h\mu(1-\Theta)}{\varepsilon - h\mu\Theta} > 0$. Let the model parameters and the stepsize h satisfy $R_{\Theta,h}(0,0) \in (0,1)$. Then there exists $\delta \in \left(0, \frac{a}{c}\right)$ such that if the history function $\varphi(t)$ is continuous and $\varphi(t) \in [-\delta, \delta]$ for all $t \in [-a - c\delta, 0]$ then the Θ method solution to (4.1.1) $\{u_n\}_{n \geqslant 0}$ satisfies $u_n \in [-\delta, \delta]$ for all $n \geqslant 0$ and $\lim_{n \to \infty} u_n = 0$.*

Proof. Similar to the proof of Theorem 4.4.6. □

Lemma 4.9.7. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let $\Theta \in [0,1]$ and let the stepsize $h > 0$ be such that $\frac{\varepsilon + h\mu(1-\Theta)}{\varepsilon - h\mu\Theta} > 0$. Then*

$$R_{\Theta,h}\left(v,w\right) \leqslant r_{\Theta,h}\left(v,w\right) \tag{4.9.17}$$

Proof. Since $1 + x \leqslant e^x$ for all $x \in \mathbb{R}$ then for any $\beta_z \in [0,1]$ and $m_z \in \mathbb{N}$,

$$\frac{\varepsilon - h\mu\left(\Theta - \beta_z\right)}{\varepsilon + h\mu\left(1-\Theta\right)} = 1 - \frac{h\mu\left(1-\beta_z\right)}{\varepsilon + h\mu\left(1-\Theta\right)} \leqslant e^{-\frac{\mu h(1-\beta_z)}{\varepsilon + h\mu(1-\Theta)}},$$

$$\left(\frac{\varepsilon - h\mu\Theta}{\varepsilon + h\mu\left(1-\Theta\right)}\right)^{m_z} = \left(1 - \frac{h\mu}{\varepsilon + h\mu\left(1-\Theta\right)}\right)^{m_z} \leqslant e^{-\frac{\mu h m_z}{\varepsilon + h\mu(1-\Theta)}}.$$

Thus,

$$\frac{\varepsilon - h\mu\left(\Theta - \beta_z\right)}{\varepsilon + h\mu\left(1-\Theta\right)}\left(\frac{\varepsilon - h\mu\Theta}{\varepsilon + h\mu\left(1-\Theta\right)}\right)^{m_z} \leqslant e^{-\frac{\mu h\left(m_z + 1 - \beta_i\right)}{\varepsilon + h\mu(1-\Theta)}}.$$

From these relations we derive that for $v, w > -\frac{a}{c}$,

$$B(v,w) \geqslant \left[\left(1-\Theta\right)e^{-\frac{\mu h(m_v + 1 - \beta_v)}{\varepsilon + h\mu(1-\Theta)}} + \Theta e^{-\frac{\mu h(m_w + 1 - \beta_w)}{\varepsilon + h\mu(1-\Theta)}}\right]^{-1}.$$

Since $\mu < 0$, then

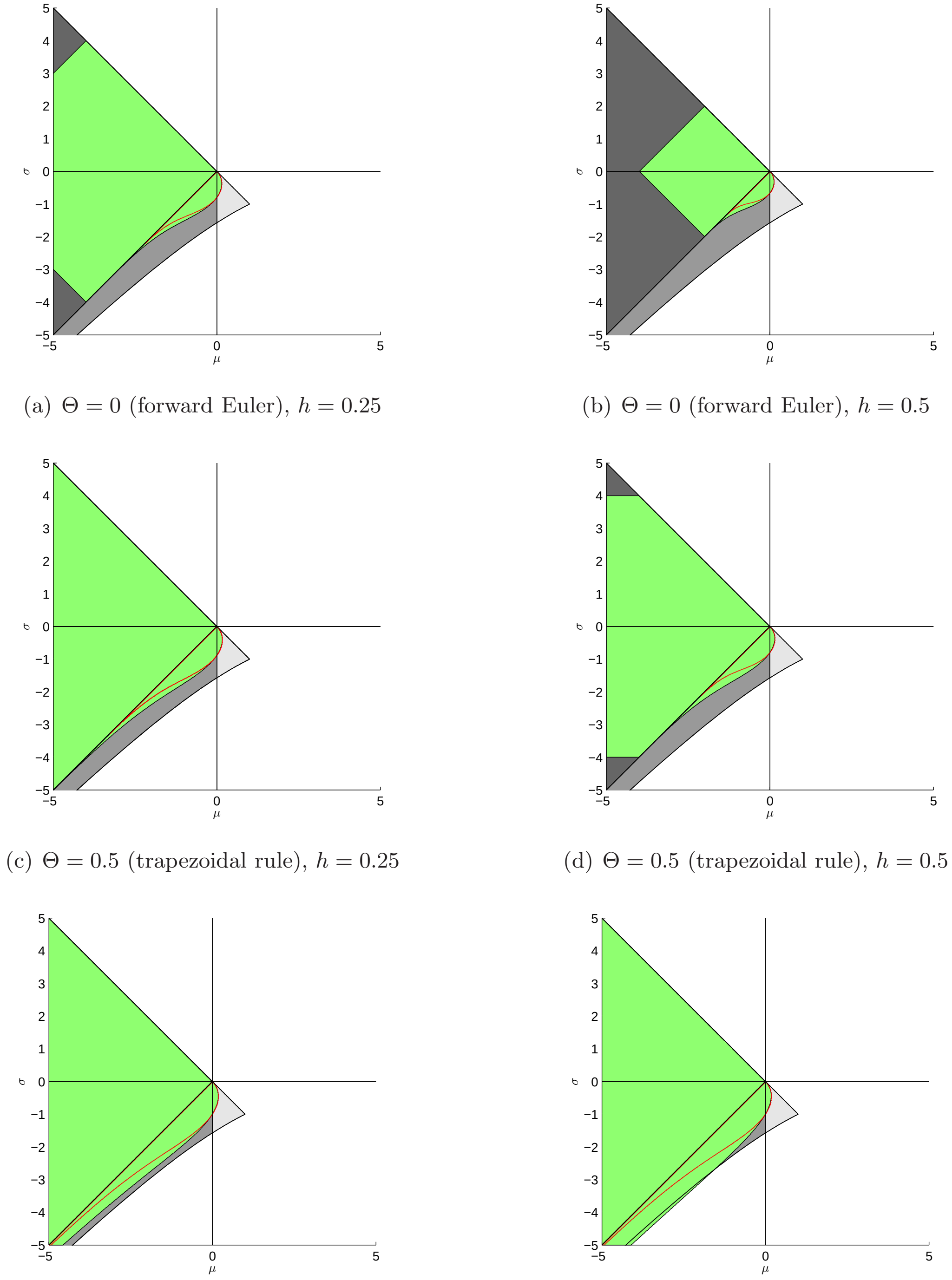
$$R_{\Theta,h}\left(v,w\right) = \frac{\sigma}{\mu}\left[\frac{1 - B\left(v,w\right)}{1 + \frac{\mu}{\sigma}B\left(v,w\right)}\right] \leqslant \frac{\sigma}{\mu}\frac{1 - \left[\left(1-\Theta\right)e^{-\frac{\mu h(m_1 + 1 - \beta_1)}{\varepsilon + h\mu(1-\Theta)}} + \Theta e^{-\frac{\mu h(m_1 + 1 - \beta_1)}{\varepsilon + h\mu(1-\Theta)}}\right]^{-1}}{1 + \frac{\mu}{\sigma}\left[\left(1-\Theta\right)e^{-\frac{\mu h(m_1 + 1 - \beta_1)}{\varepsilon + h\mu(1-\Theta)}} + \Theta e^{-\frac{\mu h(m_1 + 1 - \beta_1)}{\varepsilon + h\mu(1-\Theta)}}\right]^{-1}}.$$

Using $\left(m_v + 1 - \beta_v\right)h = a + cv + h$ and $\left(m_w + 1 - \beta_w\right)h = a + cw$ we derive (4.9.17). □

Theorem 4.9.8. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < 0$ and $\sigma < -\mu$. Let $\Theta \in [0,1]$ and let the stepsize $h > 0$ be such that $\frac{\varepsilon + h\mu(1-\Theta)}{\varepsilon - h\mu\Theta} > 0$. Let the model parameters satisfy $r_{\Theta,h}\left(0,0\right) \in (0,1)$. Then there exists $\delta \in \left(0, \frac{a}{c}\right)$ such that for all stepsizes $h > 0$, if $\varphi\left(t\right) \in [-\delta, \delta]$ for $t \in [-a - c\delta, 0]$ then the Θ method solution to (4.1.1) $\{u_n\}_{n \geqslant 0}$ satisfies $u_n \in [-\delta, \delta]$ and $\lim_{n\to\infty} u_n = 0$.*

Proof. If $\mu < 0$ then $R_{\Theta,h}\left(v,w\right) > 0$ for all $v \geqslant -\frac{a}{c}$. By Lemma 4.9.7, $R_{\Theta,h}\left(0,0\right) \leqslant r_{\Theta,h}\left(0,0\right) < 1$. Thus Theorem 4.9.6 applies for any stepsize $h > 0$. □

See Figure 4–8 for plots of the regions $\{R_{\Theta,h}\left(0,0\right) \in (0,1)\}$ and $\{r_{\Theta,h}\left(0,0\right) \in (0,1)\}$.



(a) $\Theta = 0$ (forward Euler), $h = 0.25$

(b) $\Theta = 0$ (forward Euler), $h = 0.5$

(c) $\Theta = 0.5$ (trapezoidal rule), $h = 0.25$

(d) $\Theta = 0.5$ (trapezoidal rule), $h = 0.5$

(e) $\Theta = 1$ (backward Euler), $h = 0.25$

(f) $\Theta = 1$ (backward Euler), $h = 0.5$

Figure 4–8: The set $\{R_{\Theta,h}(0,0) \in (0,1)\}$ is shaded green and plotted with $\varepsilon = a = c = 1$. This figure also shows the boundary of the set $\{r_{\Theta,h}(0,0) \in (0,1)\}$ (in red) which is contained inside $\{R_{\Theta,h}(0) \in (0,1)\}$ when $\mu < 0$.

4.10 Stability of Theta methods using a Razumikhin-style proof

In this section we extend the results of Section 4.5 on backward Euler to general Θ methods. In this section we let $\tilde{Y}_{n+1} = (1-\Theta)\,\tilde{Y}^{(2)}_{n+1} + \Theta\tilde{Y}^{(2)}_{n+1}$. Recall from (4.8.2) the general Θ method applied to the model problem (4.1.1)

$$u_{n+1} = \left(\frac{\varepsilon + h\mu\,(1-\Theta)}{\varepsilon - h\mu\Theta}\right) u_n + \frac{h\sigma}{\varepsilon - h\mu\Theta}\tilde{Y}_{n+1}. \tag{4.10.1}$$

We only consider the case when $\varepsilon - h\mu\Theta > 0$ and $\varepsilon + h\mu\,(1-\Theta) > 0$. This implies a stepsize restriction of $h < \frac{\varepsilon}{\mu\Theta}$ when $\mu > 0$ and $\Theta \neq 0$, and $h < -\frac{\varepsilon}{\mu(1-\Theta)}$ when $\mu < 0$ and $\Theta \neq 1$.

Lemma 4.10.1. *Let $\varepsilon, a > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$, $\sigma < -\mu$. Let $\Theta \in [0,1]$ and let the parameters satisfy $\frac{\varepsilon + h\mu(1-\Theta)}{\varepsilon - h\mu\Theta} > 0$. Define $\|\varphi\| = \sup_{s\leqslant 0}|\varphi(s)|$. If $\|\varphi\| < \left|\frac{a}{c}\right|$ (no restriction if $c=0$) then*

$$|u_n| \leqslant \left(\frac{\varepsilon + h\mu\,(1-\Theta) - h\sigma}{\varepsilon - h\mu\Theta}\right)^n \|\varphi\|$$

for all n such that $\left(\frac{\varepsilon + h\mu(1-\Theta) - h\sigma}{\varepsilon - h\mu\Theta}\right)^n \|\varphi\| < \left|\frac{a}{c}\right|$ (no restriction if $c = 0$).

Proof. This proof is by strong induction. For the case $n = 0$ this is obviously true. Suppose this is true up to n. Then for all $i = 1, ..., n$ we have

$$|u_i| \leqslant \left(\frac{\varepsilon + h\mu\,(1-\Theta) - h\sigma}{\varepsilon - h\mu\Theta}\right)^i \|\varphi\|$$

Before we look at the $n+1$ case, consider the term $\frac{\varepsilon - h\mu(1-\Theta) - h\sigma}{\varepsilon - h\mu\Theta}$. Since $\sigma < \mu$ then we must have $\frac{\varepsilon + h\mu(1-\Theta) - h\sigma}{\varepsilon - h\mu\Theta} > 1$. This means that for all $i = 1, ..., n$ we have

$$|u_i| \leqslant \left(\frac{\varepsilon + h\mu\,(1-\Theta) - h\sigma}{\varepsilon - h\mu\Theta}\right)^n \|\varphi\|$$

Now consider the $n+1$ case with no overlapping ($t_{n+1} - a - cu_{n+1} \leqslant t_n$). Using (4.9.5),

$$\begin{aligned}
|u_{n+1}| &\leqslant \frac{\varepsilon + h\mu\,(1-\Theta)}{\varepsilon - h\mu\Theta}\,|u_n| + \frac{h\,|\sigma|}{\varepsilon - h\mu\Theta}\left|\tilde{Y}_{n+1}\right| \\
&\leqslant \frac{\varepsilon + h\mu\,(1-\Theta)}{\varepsilon - h\mu\Theta}\left(\frac{\varepsilon + h\mu\,(1-\Theta) - h\sigma}{\varepsilon - h\mu\Theta}\right)^n \|\varphi\| - \frac{h\sigma}{\varepsilon - h\mu\Theta}\left(\frac{\varepsilon + h\mu\,(1-\Theta) - h\sigma}{\varepsilon - h\mu\Theta}\right)^n \|\varphi\| \\
&= \left(\frac{\varepsilon + h\mu\,(1-\Theta) - h\sigma}{\varepsilon - h\mu\Theta}\right)^{n+1} \|\varphi\|
\end{aligned}$$

Now consider the $n+1$ case with overlapping $(t_n < t_{n+1} - a - cu_{n+1} \leqslant t_{n+1})$. Then we have $\tilde{Y}^{(2)}_{n+1} = (1-\beta)\,u_n + \beta u_{n+1}$ and solving for u_{n+1} in (4.10.1) yields

$$u_{n+1} = \left(\frac{\varepsilon + h\mu\,(1-\Theta) + h\sigma\Theta\,(1-\beta)}{\varepsilon - h\mu\Theta - h\sigma\Theta\beta}\right) u_n + \frac{h\sigma\,(1-\Theta)}{\varepsilon - h\mu\Theta - h\sigma\Theta\beta}\tilde{Y}^{(1)}_{n+1}$$

$$\begin{aligned}|u_{n+1}| &\leqslant \left|\frac{\varepsilon + h\mu\,(1-\Theta) + h\sigma\Theta\,(1-\beta)}{\varepsilon - h\mu\Theta - h\sigma\Theta\beta}\right| |u_n| + \frac{h\,|\sigma|\,(1-\Theta)}{\varepsilon - h\mu\Theta - h\sigma\Theta\beta}\left|\tilde{Y}^{(1)}_{n+1}\right| \\ &\leqslant \frac{|\varepsilon + h\mu\,(1-\Theta) + h\sigma\Theta\,(1-\beta)| - h\sigma\,(1-\Theta)}{\varepsilon - h\mu\Theta - h\sigma\Theta\beta}\left(\frac{\varepsilon + h\mu\,(1-\Theta) - h\sigma}{\varepsilon - h\mu\Theta}\right)^n \|\varphi\|\end{aligned}$$

Due to our restriction that the solution remain bounded away from $\frac{a}{c}$, $\beta \in [0,1]$. Since $\varepsilon + h\mu\,(1-\Theta) > 0$ then $|\varepsilon + h\mu\,(1-\Theta) + h\sigma\Theta\,(1-\beta)| \leqslant \varepsilon + h\mu\,(1-\Theta) - h\sigma\Theta$, and

$$\begin{aligned}u_{n+1} &\leqslant \frac{\varepsilon + h\mu\,(1-\Theta) - h\sigma}{\varepsilon - h\mu\Theta - h\sigma\Theta\beta}\left(\frac{\varepsilon + h\mu\,(1-\Theta) - h\sigma}{\varepsilon - h\mu\Theta}\right)^n \|\varphi\| \\ &\leqslant \left(\frac{\varepsilon + h\mu\,(1-\Theta) - h\sigma}{\varepsilon - h\mu\Theta}\right)^{n+1} \|\varphi\|\end{aligned}$$

This completes the proof by strong induction. □

Recall the method of proof in Section 4.5. By Lemma 4.10.1, for every $\delta_1 > 0$ it is always possible to bound a finite segment of the Θ method solution by δ_1 by bounding the history function by an appropriate δ_2. Let $\delta_1 \in \left(0, \left|\frac{a}{c}\right|\right)$, $\delta \in \left(\delta_1, \left|\frac{a}{c}\right|\right)$, $M = \left\lceil \frac{a+|c|\delta}{h} \right\rceil$ and $\delta_2 = \delta_1 \left(\frac{\varepsilon + h\mu(1-\Theta) - h\sigma}{\varepsilon - h\mu\Theta}\right)^{-2M}$. If $\varphi\,(t) \in [-\delta_2, \delta_2]$ for $t \leqslant 0$ then from Lemma 4.10.1, the segment of the Θ method solution to the model problem (4.1.1) $\{u_n\}_{n=0}^{2M}$ must satisfy $u_n \in [-\delta_1, \delta_1]$ for $n = 0, ..., 2M$. Because of this bound on the numerical solution then we actually only need $\varphi\,(t) \in [-\delta_2, \delta_2]$ for $t \in [-a - |c|\delta, 0]$.

As in Section 4.5, we look for parameter regions for which the Θ method solution cannot exit $[-\delta_1, \delta_1]$. We do this by first supposing that the Θ method solution $\{u_n\}_{n\geqslant 0}$ escapes the interval for the first time through the upper bound at the $(n+1)$-st step for some $n > 2M$. Then $u_{n+1} > \delta_1 \geqslant u_n$. In the parameter regions where we can obtain a contradiction to this assumption the Θ method solution must remain inside $[-\delta, \delta]$. Let $u_{n+1} = \delta$. Since $u_{n+1} > u_n$ then $(1-\Theta)\left(\mu u_n + \sigma\tilde{Y}^{(1)}_{n+1}\right) + \Theta\left(\mu u_{n+1} + \sigma\tilde{Y}^{(2)}_{n+1}\right) > 0$ and

$$\tilde{Y}_{n+1} = (1-\Theta)\,\tilde{Y}^{(2)}_{n+1} + \Theta\tilde{Y}^{(2)}_{n+1} < -\frac{\mu}{\sigma}\left[(1-\Theta)\,u_n + \Theta u_{n+1}\right].$$

But from (4.10.1), $u_n = \left(\frac{\varepsilon-h\mu\Theta}{\varepsilon+h\mu(1-\Theta)}\right)u_{n+1} - \frac{h\sigma}{\varepsilon+h\mu(1-\Theta)}\tilde{Y}_{n+1}$ so

$$\tilde{Y}_{n+1} < -\frac{\mu}{\sigma}\left[(1-\Theta)\left(\left(\frac{\varepsilon-h\mu\Theta}{\varepsilon+h\mu(1-\Theta)}\right)u_{n+1} - \frac{h\sigma}{\varepsilon+h\mu(1-\Theta)}\tilde{Y}_{n+1}\right)+\Theta u_{n+1}\right],$$

and solving for $\tilde{Y}_{n+1}$ gives

$$\tilde{Y}_{n+1} < -\frac{\mu}{\sigma}u_{n+1} = -\frac{\mu}{\sigma}\delta. \tag{4.10.2}$$

Define m_1, m_2, β_1 and β_2 as in (4.9.1)

$$\begin{array}{ll} m_1 = \left\lceil \frac{a+cu_n}{h}\right\rceil, & \beta_1 = m_1 - \frac{a+cu_n}{h}, \\ m_2 = \left\lfloor \frac{a+cu_{n+1}}{h}\right\rfloor, & \beta_2 = m_2 + 1 - \frac{a+cu_{n+1}}{h} \end{array}. \tag{4.10.3}$$

Let $m = \max\{m_1, m_2\}$. Since $n > 2M$ then for $i = n-m, .., n$ we have

$$t_i - a - cu_i \in [t_{n-m} - a - |c|\,\delta, t_n - a + |c|\,\delta] \subseteq [0, t_n]$$

$$t_{i+1} - a - cu_{i+1} \in [t_{n-m+1} - a - |c|\,\delta, t_{n+1} - a + |c|\,\delta] \subseteq [0, t_{n+1}]$$

This means that $\tilde{Y}^{(1)}_{i+1} = \eta(t_i - a - cu_i)$ and $\tilde{Y}^{(2)}_{i+1} = \eta(t_{i+1} - a - cu_{i+1})$ must have properties stemming from the properties of the continuous extension on $[0, t_{n+1}]$. Let us derive some of these properties first and then we will see later on why these properties are important.

The change from $\tilde{Y}^{(j)}_i$ to $\tilde{Y}^{(j)}_{i+1}$ for $j = 1, 2$ depends on the maximum change in the mesh values in a single step. So we need to determine the maximum change in mesh values in a single step. Since δ_1 is a bound on the absolute value of the past mesh values and $u_{n+1} = \delta > \delta_1$ then δ is a bound on the absolute value of the past mesh values, the present one at t_{n+1} and the continuous extension $\eta(t)$ for $t \in [0, t_{n+1}]$. As a result, for $i = 0, ..., n$ we have

$$\begin{aligned} |u_{i+1} - u_i| &= \frac{h}{\varepsilon}\left|(1-\Theta)\left(\mu u_i + \sigma\tilde{Y}^{(1)}_{i+1}\right) + \Theta\left(\mu u_{i+1} + \sigma\tilde{Y}^{(2)}_{i+1}\right)\right| \\ &\leqslant \frac{h}{\varepsilon}\left[|\mu|\,|(1-\Theta)\,u_i + \Theta u_i| + |\sigma|\left|\tilde{Y}_{i+1}\right|\right] \leqslant \frac{|\mu|+|\sigma|}{\varepsilon}\delta h \end{aligned}$$

The quantity $\left|\tilde{Y}^{(1)}_{i+1} - \tilde{Y}^{(1)}_i\right|$ is bounded by the maximum number of time steps (including fractions of a step since we are working with linear interpolation) between $t_i - a - cu_i$ and $t_{i-1} - a - cu_{i-1}$, multiplied by the maximum change in the mesh values in a single step. Similarly, the quantity $\left|\tilde{Y}^{(2)}_{i+1} - \tilde{Y}^{(2)}_i\right|$ is bounded by the maximum number of time steps between $t_{i+1} - a - cu_{i+1}$ and $t_i - a - cu_i$, multiplied by the maximum change in the mesh values in a single step. Thus,

$$\begin{aligned}
\left|\tilde{Y}_{i+1} - \tilde{Y}_i\right| &\leqslant \Bigg[(1-\Theta)\left|\frac{(t_i - a - cu_i) - (t_{i-1} - a - cu_{i-1})}{h}\right| \\
&\qquad + \Theta\left|\frac{(t_{i+1} - a - cu_{i+1}) - (t_i - a - cu_i)}{h}\right|\Bigg]\frac{|\mu| + |\sigma|}{\varepsilon}\delta h, \\
&\leqslant \left(1 + \frac{|\mu| + |\sigma|}{\varepsilon}\,|c|\,\delta\right)\frac{|\mu| + |\sigma|}{\varepsilon}\delta h, \\
&= D\delta h,
\end{aligned} \tag{4.10.4, 4.10.5}$$

where as before we define $D = \left(1 + \frac{|\mu|+|\sigma|}{\varepsilon}\,|c|\,\delta\right)\frac{|\mu|+|\sigma|}{\varepsilon}$. Now recall Lemma 4.5.2. If we write (4.10.1) in the form $u_{n+1} = Au_n + v_n$ then

$$A = \frac{\varepsilon + h\mu\,(1-\Theta)}{\varepsilon - h\mu\Theta}, \qquad v_i = \frac{h\sigma}{\varepsilon - h\mu\Theta}\tilde{Y}_{i+1}$$

Applying Lemma 4.5.2 yields $u_{n+1} = A^{m+1}u_{n-m} + \sum\limits_{i=n-m}^{n} A^{n-i}v_i$. We would like to get the right hand side of this equation as large as possible in order to get a bound on the value of u_{n+1}. This is done by using the most negative possible sequence of $\tilde{Y}_{i+1}$. Let that sequence be $\{w_i\}$. Using the bound on the solution and the conditions (4.10.2)-(4.10.5), we define $\{w_i\}$ as

$$w_i = \begin{cases} -\delta, & i \leqslant n - \ell \\ \hat{u} - (n-i)\,D\delta h, & n - \ell + 1 \leqslant i \leqslant n \end{cases}$$

where $\hat{u} = \tilde{Y}_{n+1}^{(1)} \in \left[-\delta, -\frac{\mu}{\sigma}\delta\right]$ and

$$\ell = \left\lceil \frac{\hat{u} + \delta}{D\delta h} \right\rceil, \qquad \chi = \ell - \frac{\hat{u} + \delta}{D\delta h}. \tag{4.10.6}$$

Set $\tilde{v}_i = \frac{h\sigma}{\varepsilon - h\mu}w_i$. For $j = 1$ or 2, using Lemma 4.5.2, we derive

$$u_{n+1} = A^{m+1}u_{n-m_j} + \sum_{i=n-m_j}^{n} A^{n-i}\tilde{v}_i \leqslant A^{m_j+1}u_{n-m} + \sum_{i=n-m_j}^{n} A^{n-i}\tilde{v}_i, \tag{4.10.7}$$

$$u_{n+1} = A^{m}u_{n-m_j+1} + \sum_{i=n-m_j+1}^{n} A^{n-i}\tilde{v}_i \leqslant A^{m_j}u_{n-m+1} + \sum_{i=n-m_j+1}^{n} A^{n-i}\tilde{v}_i. \tag{4.10.8}$$

Suppose $\ell \leqslant m_j$. As in Section 4.5, take $\frac{(1-\beta)}{A}\times$ (4.10.7) plus $\beta\times$ (4.10.8),

$$\begin{aligned}
&\left(\frac{1-\beta_j}{A}+\beta_j\right)u_{n+1} &(4.10.9)\\
&\qquad\leqslant \left[(1-\beta_j)\,u_{n-m_j}+\beta_j u_{n-m_j+1}\right]A^{m_j}+\frac{1-\beta_j}{A}\sum_{i=n-m_j}^{n}A^{n-i}\tilde{v}_i+\beta_j\sum_{i=n-m_j+1}^{n}A^{n-i}\tilde{v}_i\\
&\qquad=\tilde{Y}_{n+1}^{(j)}A^{m_j}+\frac{1-\beta_j}{A}\sum_{i=n-m_j}^{n}A^{n-i}\tilde{v}_i+\beta_j\sum_{i=n-m_j+1}^{n}A^{n-i}\tilde{v}_i. &(4.10.10)
\end{aligned}$$

We skip the steps in performing the summation because they are very similar to the steps in Section 4.5. The result is

$$\begin{aligned}
&\frac{1-\beta_j}{A}\sum_{i=n-m_j}^{n}A^{n-i}\tilde{v}_i+\beta_j\sum_{i=n-m_j+1}^{n}A^{n-i}\tilde{v}_i\\
&=\frac{\sigma}{\mu}\left[\left(\frac{1-\beta_j}{A}+\beta_j\right)\left[\frac{(\varepsilon+h\mu\,(1-\Theta))\,D\delta}{\mu}\left(\left(1-\frac{h\mu\chi}{\varepsilon+h\mu\,(1-\Theta)}\right)A^{\ell}-1\right)-\hat{u}\right]-\delta A^{m_j}\right].
\end{aligned}$$

Using $\frac{1-\beta}{A}+\beta=\frac{(1-\beta)(\varepsilon-h\mu\Theta)}{\varepsilon+h\mu(1-\Theta)}+\beta=\frac{\varepsilon-h\mu(\Theta-\beta)}{\varepsilon+h\mu(1-\Theta)}$, we derive an expression for u_{n+1} from (4.10.10),

$$\begin{aligned}
u_{n+1}&\leqslant \tilde{Y}_{n+1}^{(j)}\frac{\varepsilon+h\mu\,(1-\Theta)}{\varepsilon-h\mu\,(\Theta-\beta_j)}A^{m_j}\\
&+\frac{\sigma}{\mu}\left[(\varepsilon+h\mu\,(1-\Theta))\,\frac{D\delta}{\mu}\left(\left(1-\frac{h\mu\chi}{\varepsilon+h\mu\,(1-\Theta)}\right)A^{\ell}-1\right)-\hat{u}-\delta\frac{\varepsilon+h\mu\,(1-\Theta)}{\varepsilon-h\mu\,(\Theta-\beta_j)}A^{m_j}\right]
\end{aligned} \tag{4.10.11}$$

Rewrite this as

$$\begin{aligned}
\left(\frac{\varepsilon+h\mu\,(1-\Theta)}{\varepsilon-h\mu\,(\Theta-\beta_j)}A^{m_j}\right)^{-1}u_{n+1}&\leqslant \tilde{Y}_{n+1}^{(j)}-\frac{\sigma}{\mu}\delta+\left(\frac{\varepsilon+h\mu\,(1-\Theta)}{\varepsilon-h\mu\,(\Theta-\beta_j)}A^{m_j}\right)^{-1}\\
&\times\frac{\sigma}{\mu}\left[(\varepsilon+h\mu\,(1-\Theta))\,\frac{D\delta}{\mu}\left(\left(1-\frac{h\mu\chi}{\varepsilon+h\mu\,(1-\Theta)}\right)A^{\ell}-1\right)-\hat{u}\right]
\end{aligned} \tag{4.10.12}$$

We will leave this equation for now and go back to it later on. Let us first derive a corresponding expression for the case $\ell > m_j$. In this case instead of (4.10.11) we obtain

$$\begin{aligned}\frac{1-\beta_j}{A}\sum_{i=n-m_j}^{n} A^{n-i}\tilde{v}_i &+ \beta_j \sum_{i=n-m_j+1}^{n} A^{n-i}\tilde{v}_i \\ &= \frac{\sigma}{\mu}\Big[-\left(\frac{1-\beta_j}{A}+\beta_j\right)\left(\hat{u}+\left(\varepsilon+h\mu\left(1-\Theta\right)\right)\frac{D\delta}{\mu}\right) \\ &\quad + \left(\hat{u}-D\delta h\left(m_j+1-\beta_j\right)+\left(\varepsilon+h\mu\left(1-\Theta\right)\right)\frac{D\delta}{\mu}\right)A^{m_j}\Big]\end{aligned} \tag{4.10.13}$$

Using $\frac{1-\beta}{A}+\beta = \frac{(1-\beta)(\varepsilon-h\mu\Theta)}{\varepsilon+h\mu(1-\Theta)}+\beta = \frac{\varepsilon-h\mu(\Theta-\beta)}{\varepsilon+h\mu(1-\Theta)}$, we derive an expression for u_{n+1} from (4.10.10),

$$\begin{aligned}u_{n+1} \leqslant \hat{u}\frac{\varepsilon+h\mu\left(1-\Theta\right)}{\varepsilon-h\mu\left(\Theta-\beta_j\right)}A^{m_j} &+ \frac{\sigma}{\mu}\Big[-\hat{u}-\left(\varepsilon+h\mu\left(1-\Theta\right)\right)\frac{D\delta}{\mu} \\ &+ \left(\hat{u}-D\delta h\left(m_j+1-\beta_j\right)+\left(\varepsilon+h\mu\left(1-\Theta\right)\right)\frac{D\delta}{\mu}\right)\frac{\varepsilon+h\mu\left(1-\Theta\right)}{\varepsilon-h\mu\left(\Theta-\beta_j\right)}A^{m_j}\Big]\end{aligned}$$

Rewrite this to get the same left hand side as in (4.10.12)

$$\begin{aligned}\left(\frac{\varepsilon+h\mu\left(1-\Theta\right)}{\varepsilon-h\mu\left(\Theta-\beta_j\right)}A^{m_j}\right)^{-1} u_{n+1} &\leqslant \tilde{Y}_{n+1}^{(j)} + \frac{\sigma}{\mu}\left[\hat{u}-D\delta h\left(m_j+1-\beta_j\right)+\left(\varepsilon+h\mu\left(1-\Theta\right)\right)\frac{D\delta}{\mu}\right] \\ &- \frac{\sigma}{\mu}\left[\hat{u}+\left(\varepsilon+h\mu\left(1-\Theta\right)\right)\frac{D\delta}{\mu}\right]\left(\frac{\varepsilon+h\mu\left(1-\Theta\right)}{\varepsilon-h\mu\left(\Theta-\beta_j\right)}A^{m_j}\right)^{-1}\end{aligned} \tag{4.10.14}$$

Now we have two expressions in (4.10.12) and (4.10.14) depending on how ℓ compares with m_j. Since there are two m_j's, then there are four cases to consider.

Definition 4.10.2. Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let $\Theta \in [0,1]$ and let the stepsize $h > 0$ be such that $\frac{\varepsilon+h\mu(1-\Theta)}{\varepsilon-h\mu\Theta} > 0$. For any $\delta \in \left(0, \left|\frac{a}{c}\right|\right)$ and $\hat{u} \in \left[-\delta, -\frac{\mu}{\sigma}\delta\right]$, define the function

$$\mathcal{S}\left(\Theta,h\right)\left(\hat{u},\delta,c,2\right) = \begin{cases} \mathcal{S}_{2,2}\left(\Theta\right)\left(\hat{u},\delta\right), & \text{if } \ell \leqslant m_1 \text{ and } \ell \leqslant m_2, \\ \mathcal{S}_{2,1}\left(\Theta\right)\left(\hat{u},\delta\right), & \text{if } \ell \leqslant m_1 \text{ and } \ell > m_2, \\ \mathcal{S}_{1,2}\left(\Theta\right)\left(\hat{u},\delta\right), & \text{if } \ell > m_1 \text{ and } \ell \leqslant m_2, \\ \mathcal{S}_{1,1}\left(\Theta\right)\left(\hat{u},\delta\right), & \text{if } \ell > m_1 \text{ and } \ell > m_2. \end{cases} \tag{4.10.15}$$

where m_1, m_2, β_1, β_2, ℓ and χ are given by

$$
\begin{aligned}
m_1 &= \left\lceil \frac{a+cu_*}{h} \right\rceil, \quad \beta_1 = \frac{a+cu_*}{h} - m_1, \quad u_* = \left(\frac{\varepsilon - h\mu\Theta}{\varepsilon + h\mu\left(1-\Theta\right)}\right)\delta - \frac{h\sigma}{\varepsilon + h\mu\left(1-\Theta\right)}\hat{u} \\
m_2 &= \left\lfloor \frac{a+c\delta}{h} \right\rfloor, \quad \beta_2 = m_2 + 1 - \frac{a+c\delta}{h}, \\
\ell &= \left\lceil \frac{\hat{u}+\delta}{D\delta h} \right\rceil, \quad \chi = \ell - \frac{\hat{u}+\delta}{D\delta h}.
\end{aligned}
$$

We derive the expressions $\mathcal{S}_{ij}$ from (4.10.12) and (4.10.14). First let

$$
B = \left[\left(1-\Theta\right)\left(\frac{\varepsilon + h\mu\left(1-\Theta\right)}{\varepsilon - h\mu\left(\Theta-\beta_1\right)}A^{m_1}\right)^{-1} + \Theta\left(\frac{\varepsilon + h\mu\left(1-\Theta\right)}{\varepsilon - h\mu\left(\Theta-\beta_2\right)}A^{m_2}\right)^{-1}\right]^{-1}.
$$

For the first case in (4.10.15), take $(1-\Theta)$ times (4.10.12) with $j=1$ plus Θ times (4.10.12) with $j=2$. This yields

$$
\begin{aligned}
\frac{1}{B}u_{n+1} &\leqslant \hat{u} - \frac{\sigma}{\mu}\delta + \frac{\sigma}{\mu B}\left[\left(\varepsilon + h\mu\left(1-\Theta\right)\right)\frac{D\delta}{\mu}\left(\left(1 - \frac{h\mu\chi}{\varepsilon + h\mu\left(1-\Theta\right)}\right)A^{\ell} - 1\right) - \hat{u}\right], \\
u_{n+1} &\leqslant B\left(\hat{u} - \frac{\sigma}{\mu}\delta\right) + \frac{\sigma}{\mu}\left[\left(\varepsilon + h\mu\left(1-\Theta\right)\right)\frac{D\delta}{\mu}\left(\left(1 - \frac{h\mu\chi}{\varepsilon + h\mu\left(1-\Theta\right)}\right)A^{\ell} - 1\right) - \hat{u}\right].
\end{aligned}
$$

Define $\mathcal{S}_{2,2}$ to be the right hand side of this expression.

$$
\begin{aligned}
\mathcal{S}_{2,2}\left(\Theta\right)\left(\hat{u},\delta\right) = &\left(B - \frac{\sigma}{\mu}\right)\hat{u} \\
&+ \frac{\sigma}{\mu}\delta\left[\frac{\left(\varepsilon + h\mu\left(1-\Theta\right)\right)D}{\mu}\left(\left(1 - \frac{h\mu\chi}{\varepsilon + h\mu\left(1-\Theta\right)}\right)A^{\ell} - 1\right) - B\right]
\end{aligned}
\tag{4.10.16}
$$

The second and third cases in (4.10.15) are similar so we just perform the derivation for the second case. Take $(1-\Theta)$ times (4.10.12) with $j=1$ plus Θ times (4.10.14) with $j=2$. This yields

$$
\begin{aligned}
\frac{1}{B}u_{n+1} \leqslant\ & \hat{u} - \left(1-\Theta\right)\frac{\sigma}{\mu}\delta + \Theta\frac{\sigma}{\mu}\left[\hat{u} - D\delta h\left(m_2 + 1 - \beta_2\right) + \frac{\left(\varepsilon + h\mu\left(1-\Theta\right)\right)D\delta}{\mu}\right] \\
&+ (1-\Theta)\left[\frac{\varepsilon + h\mu\left(1-\Theta\right)}{\varepsilon - h\mu\left(\Theta-\beta_1\right)}A^{m_1}\right]^{-1}\frac{\sigma}{\mu}\left[\frac{\left(\varepsilon + h\mu\left(1-\Theta\right)\right)D\delta}{\mu}\left(\left(1 - \frac{h\mu\chi}{\varepsilon + h\mu\left(1-\Theta\right)}\right)A^{\ell} - 1\right) - \hat{u}\right] \\
&+ \Theta\left[\frac{\varepsilon + h\mu\left(1-\Theta\right)}{\varepsilon - h\mu\left(\Theta-\beta_2\right)}A^{m_2}\right]^{-1}\frac{\sigma}{\mu}\left[-\frac{\left(\varepsilon + h\mu\left(1-\Theta\right)\right)D\delta}{\mu} - \hat{u}\right].
\end{aligned}
$$

Rearranging and solving for u_{n+1} yields

$$\begin{aligned} u_{n+1} \leqslant{} & B\hat{u} - \frac{\sigma}{\mu}\hat{u} - B\left(1-\Theta\right)\frac{\sigma}{\mu}\delta + B\Theta\frac{\sigma}{\mu}\left[\hat{u} - D\delta h\left(m_2+1-\beta_2\right) + \frac{\left(\varepsilon + h\mu\left(1-\Theta\right)\right)D\delta}{\mu}\right] \\ & + B\left(1-\Theta\right)\left[\frac{\varepsilon + h\mu\left(1-\Theta\right)}{\varepsilon - h\mu\left(\Theta-\beta_1\right)}A^{m_1}\right]^{-1}\frac{\sigma\left(\varepsilon + h\mu\left(1-\Theta\right)\right)D\delta}{\mu^2}\left(1-\frac{h\mu\chi}{\varepsilon + h\mu\left(1-\Theta\right)}\right)A^{\ell} \\ & - \frac{\sigma\left(\varepsilon + h\mu\left(1-\Theta\right)\right)D\delta}{\mu^2}. \end{aligned}$$

Define $\mathcal{S}_{2,1}$ to be the right hand side of this expression. After a bit more rearranging this can be written as

$$\begin{aligned} \mathcal{S}_{2,1}\left(\Theta\right)\left(\hat{u},\delta\right) ={} & \left(B + \frac{\sigma}{\mu}\left(B\Theta - 1\right)\right)\hat{u} - B\frac{\sigma}{\mu}\delta\left(1-\Theta+\Theta Dh\left(m_2+1-\beta_2\right)\right) \\ & + \frac{\sigma\left(\varepsilon + h\mu\left(1-\Theta\right)\right)D\delta}{\mu^2}\Bigg(B\Theta - 1 + B\left(1-\Theta\right)\left[\frac{\varepsilon + h\mu\left(1-\Theta\right)}{\varepsilon - h\mu\left(\Theta-\beta_1\right)}A^{m_1}\right]^{-1} \\ & \times\left(1-\frac{h\mu\chi}{\varepsilon + h\mu\left(1-\Theta\right)}\right)A^{\ell}\Bigg). \qquad (4.10.17) \end{aligned}$$

The derivation of the third case is very similar to the second case. Define $\mathcal{S}_{1,2}$ as

$$\begin{aligned} S_{1,2}\left(\Theta\right)\left(\hat{u},\delta\right) ={} & \left(B + \frac{\sigma}{\mu}\left(B\left(1-\Theta\right) - 1\right)\right)\hat{u} \\ & - B\frac{\sigma}{\mu}\delta\left(\Theta + \left(1-\Theta\right)Dh\left(m_1+1-\beta_1\right)\right) + \frac{\sigma\left(\varepsilon + h\mu\left(1-\Theta\right)\right)D\delta}{\mu^2}\Bigg(B\left(1-\Theta\right) - 1 \\ & + B\Theta\left[\frac{\varepsilon + h\mu\left(1-\Theta\right)}{\varepsilon - h\mu\left(\Theta-\beta_2\right)}A^{m_2}\right]^{-1}\left(1-\frac{h\mu\chi}{\varepsilon + h\mu\left(1-\Theta\right)}\right)A^{\ell}\Bigg). \qquad (4.10.18) \end{aligned}$$

For the fourth case in (4.10.15), take $(1-\Theta)$ times (4.10.14) with $j=1$ plus Θ times (4.10.14) with $j=2$.

$$\begin{aligned} \frac{1}{B}u_{n+1} \leqslant{} & \hat{u} + \frac{\sigma}{\mu}\left[\hat{u} + \left(\varepsilon + h\mu\left(1-\Theta\right)\right)\frac{D\delta}{\mu}\right] - \frac{\sigma}{\mu}D\delta h\left[\left(1-\Theta\right)\left(m_1+1-\beta_1\right) + \Theta\left(m_2+1-\beta_2\right)\right] \\ & - \frac{1}{B}\frac{\sigma}{\mu}\left[\hat{u} + \left(\varepsilon + h\mu\left(1-\Theta\right)\right)\frac{D\delta}{\mu}\right] \end{aligned}$$

Note that $h\left[(1-\Theta)(m_1+1-\beta_1)+\Theta(m_2+1-\beta_2)\right] = (1-\Theta)(a+cu_*+h)+\Theta(a+c\delta)$. Define $\mathcal{S}_{1,1}$ to be the right hand side of the inequality. After some rearranging,

$$\begin{aligned}\mathcal{S}_{1,1}(\Theta)(\hat{u},\delta) = \left(B+\frac{\sigma}{\mu}(B-1)\right)\hat{u}+\frac{\sigma}{\mu}D\delta\Big[&\frac{\varepsilon+h\mu(1-\Theta)}{\mu}(B-1)\\ &-B\left[(1-\Theta)(a+cu_*+h)+\Theta(a+c\delta)\right]\Big] \quad (4.10.19)\end{aligned}$$

Recall $\mathcal{S}_1(\hat{u},\delta)$ and $\mathcal{S}_2(\hat{u},\delta)$, the functions derived for backward Euler in Section 4.5. If we set $\Theta=1$ in the expressions above then $\mathcal{S}_{1,1}(\Theta)(\hat{u},\delta)$ and $\mathcal{S}_{2,1}(\Theta)(\hat{u},\delta)$ coincides with $\mathcal{S}_1(\hat{u},\delta)$, and $\mathcal{S}_{2,2}(\Theta)(\hat{u},\delta)$ and $\mathcal{S}_{1,2}(\Theta)(\hat{u},\delta)$ coincides with $\mathcal{S}_2(\hat{u},\delta)$.

Definition 4.10.3. Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let $\Theta \in [0,1]$ and let the stepsize $h>0$ be such that $\frac{\varepsilon+h\mu(1-\Theta)}{\varepsilon-h\mu\Theta} > 0$. For any $\delta \in \left(0,\left|\frac{a}{c}\right|\right)$ define

$$P_{\Theta,h}(\delta,c,2) = \sup_{\hat{u}\in\left[-\delta,-\frac{\mu}{\sigma}\delta\right]}\Big(\mathcal{S}(\Theta,h)(\hat{u},\delta,c,2)\vee\mathcal{S}(\Theta,h)(\hat{u},\delta,-c,2)\Big),$$

where $a \vee b = \max\{a,b\}$.

For the case $\Theta=1$ we know that $\mathcal{S}(\Theta=1,h)(\hat{u},\delta,-|c|,2) \leqslant \mathcal{S}(\Theta=1,h)(\hat{u},\delta,|c|,2)$ (refer to Lemma 4.5.8). Thus in Section 4.5, we simply set $P_{\Theta=1,h}(\delta,c,2) = \mathcal{S}(\Theta=1,h)(\hat{u},\delta,|c|,2)$. We cannot do this for $\Theta \neq 1$.

Lemma 4.10.4. *Let $\varepsilon, a, c > 0$, $\sigma \leqslant \mu < \frac{\varepsilon}{a}$ and $\sigma < -\mu$. Let $\Theta \in [0,1]$ and let the stepsize $h>0$ be such that $\frac{\varepsilon+h\mu(1-\Theta)}{\varepsilon-h\mu\Theta} > 0$. Let $\delta \in \left(0,\left|\frac{a}{c}\right|\right)$, $M = \left\lceil\frac{a+|c|\delta}{h}\right\rceil$, $\delta_2 = \left(\frac{\varepsilon+h\mu(1-\Theta)-h\sigma}{\varepsilon-h\mu\Theta}\right)^{-2M-1}\delta$ and $(\mu,\sigma) \in \cap_{\delta_3\in\left[\left(\frac{\varepsilon+h\mu(1-\Theta)-h\sigma}{\varepsilon-h\mu\Theta}\right)^{-1}\delta,\delta\right]}\{P_{\Theta,h}(\delta_3,c,2)<\delta_3\}$. Then if $\varphi(t)\in[-\delta_2,\delta_2]$ for $t\in[-a-|c|\delta,0]$ then the Θ method solution to (4.1.1) $\{u_n\}_{n\geqslant 0}$ satisfies $u_n\in[-\delta,\delta]$ for $n\geqslant 0$.*

Proof. The proof is similar to that of Lemma 4.5.10. Let $\delta \in \left(0,\left|\frac{a}{c}\right|\right)$, $\delta_1 = \left(\frac{\varepsilon+h\mu(1=\Theta)-h\sigma}{\varepsilon-h\mu\Theta}\right)^{-1}\delta$ and $\delta_2 = \left(\frac{\varepsilon+h\mu(1-\Theta)-h\sigma}{\varepsilon-h\mu\Theta}\right)^{-2M-1}\delta$. By Lemma 4.10.1, if $\varphi(t)\in[-\delta_2,\delta_2]$ for $t\in[-a-c\delta,0]$ then $u_n\in[-\delta_1,\delta_1]$ for $n=0,...,2M$. Assume the solution first exits this interval at the $(n+1)$st step through the upper bound with $u_{n+1}=\delta_3>\delta_1$. Then from our discussion, $u_{n+1}\leqslant P_{\Theta,h}(\delta_3,c,2)$. If $(\mu,\sigma)\in\{P_{\Theta,h}(\delta_3,c,2)<\delta_3\}$ then we get a contradiction. Thus the BE solution cannot exit $[-\delta_1,\delta_1]$ for the first time through the upper bound.

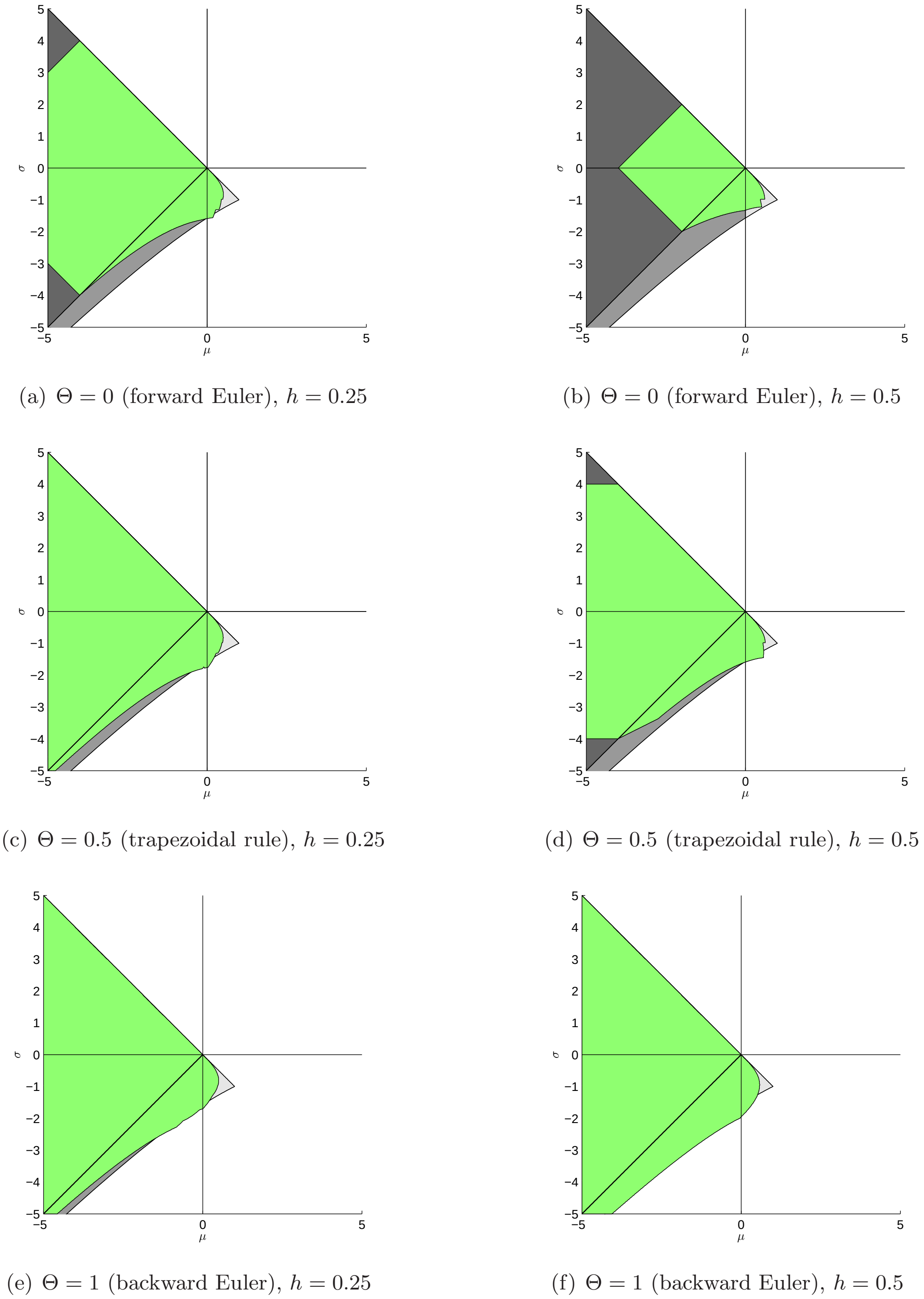
Recall from (4.8.6) that if the history function and the Θ method solution up to n are bounded inside $[-\delta_1, \delta_1]$, $\varepsilon - h\mu\Theta > 0$, $\mu + \sigma < 0$, $\sigma < 0$ and $\varepsilon + h\mu\left(1 - \Theta\right) > 0$ then

$$|u_{n+1}| \leqslant \frac{|\varepsilon + h\mu\left(1 - \Theta\right)| + |h\sigma|}{|\varepsilon - h\mu\Theta|}\delta_1 = \delta.$$

All these conditions apply because of the requirements of this lemma. Thus, $u_{n+1} = \delta_3 > \delta_1$ is only possible if $\delta_3 \in [\delta_1, \delta]$. Thus, if $(\mu, \sigma) \in \cap_{\delta_3 \in [\delta_1, \delta]}\{P_{\Theta,h}\left(\delta_3, c, 2\right) < \delta_3\}$ and $\varphi(t) \in \left[-\delta_2, \bar{\delta}_2\right]$ then u_n cannot escape $[-\delta_1, \delta_1] \subseteq [-\delta, \delta]$ through the upper bound.

Now consider the case in which the Θ method solution leaves $[-\delta, \delta]$ through the lower bound by considering the system $v\left(t\right) = -u\left(t\right)$. This yields effectively the same system (4.1.1) except with c replaced by $-c$. By our discussion above and the definition of $P_{\Theta,h}$, if $\varphi\left(t\right) \in [-\delta_2, \delta_2]$ for $t \in [-a - c\delta, 0]$ and $(\mu, \sigma) \in \cap_{\delta_3 \in [\delta_1, \delta]}\{P_{\Theta,h}\left(\delta_3, c, 2\right) < \delta_3\}$ then the Θ method solution to this system cannot escape $[-\delta, \delta]$ through the upper bound. Thus the Θ method solution to (4.1.1) cannot escape $[-\delta, \delta]$ through the lower bound either. □

Sample plots of the sets $\{P_{\Theta,h}\left(\delta, c, 2\right) < \delta\}$ are shown in Figure 4–9.



(a) $\Theta = 0$ (forward Euler), $h = 0.25$

(b) $\Theta = 0$ (forward Euler), $h = 0.5$

(c) $\Theta = 0.5$ (trapezoidal rule), $h = 0.25$

(d) $\Theta = 0.5$ (trapezoidal rule), $h = 0.5$

(e) $\Theta = 1$ (backward Euler), $h = 0.25$

(f) $\Theta = 1$ (backward Euler), $h = 0.5$

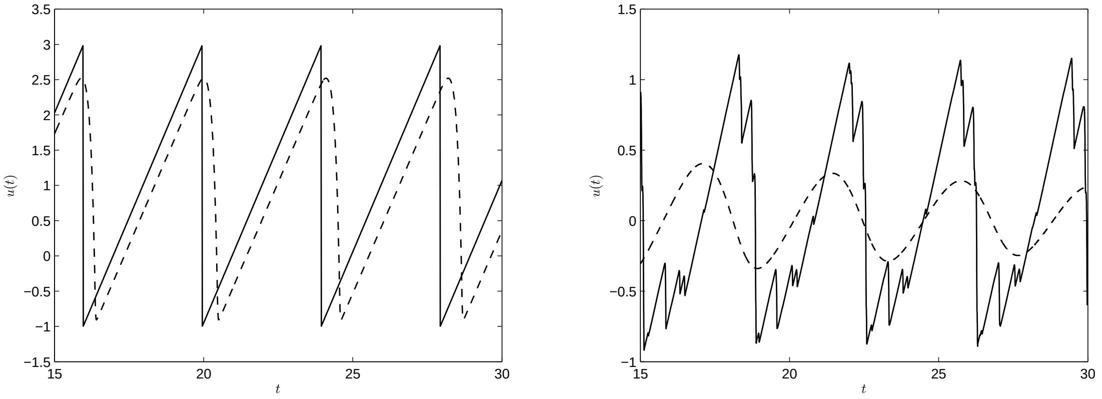
Figure 4–9: The sets $\left\{\left|\frac{|\varepsilon+h\mu(1-\Theta)|+|h\sigma|}{|\varepsilon-h\mu\Theta|}\right| < 1, \mu < 0\right\} \cup \{P_{\Theta,h}(\delta, c, 2) < \delta\}$ for different values of Θ and h are shaded green and plotted with $\varepsilon = a = c = 1$ and $\delta = 0.01$. On these sets the Θ method solution to the model problem (4.1.1) $\{u_n\}_{n\geqslant 0}$ satisfy $u_n \in [-\delta, \delta]$.

CHAPTER 5
Implementation of a class of SDIRK methods

In this chapter we discuss an implementation of a class of RK methods to solve scalar DDEs with multiple state dependent delays. Recall our model N-delay DDE from Section 2.2,

$$\begin{aligned} &\varepsilon \dot{u}\left(t\right) = -\gamma u\left(t\right) - \sum_{i=1}^{N} \kappa_i u\left(t - a_i - c_i u\left(t\right)\right), && t \geqslant 0, \\ &u\left(t\right) = \varphi\left(t\right), && t \leqslant 0. \end{aligned} \tag{5.0.1}$$

where ε, a_i, c_i, γ and $\kappa_i > 0$ for $i = 1, ..., N$. The $N = 1$ case is the model DDE we have been considering with $\mu = -\gamma$ and $\sigma = -\kappa_1$. This is a scalar problem and we know that this is a *stiff problem* when ε is small. There is currently no standard formal definition of a stiff problem even for ODEs. What we really mean when we say this is that many numerical schemes require very small stepsizes to solve (5.0.1) when ε is small. In particular, explicit numerical schemes have this difficulty. The stepsize generally needs to be smaller than ε which is impractical if we would like to observe the behaviour of the solution as $\varepsilon \to 0$.



(a) $N = 1$, $a = c = 1, \gamma = 3$, $\kappa_1 = 9$

(b) $N = 2$, $a_1 = c_i = \kappa_1 = 1$, $a_2 = \gamma = 2 = \kappa_2 = 3$

Figure 5–1: Sample solutions of (5.0.1) with $\varepsilon = 1$ (dashed) and $\varepsilon = 0.01$ (solid).

There are many DDE solvers that are currently available. In MATLAB there are DDE23 by Shampine and Thompson [53] and DDESD by Shampine [52] for state dependent problems. The code RADAR5 written in FORTRAN by Guglielmi and Hairer [23] is one of the best known solvers and solves stiff DDEs. Other DDE solvers are DKLAG6 by Corwin, Sarafyan and Thompson [13], DDVERK by Enright and Hayashi [16], DDE_SOLVER by Thompson and Shampine [56] and ARCHI by Paul [49]. Many of these solvers work very well in solving (5.0.1). However, to study the $\varepsilon \to 0$ case, we are interested in building an efficient MATLAB solver specialised for scalar problems with the possibility of some stiffness.

5.1 An example using backward Euler

Recall our discussion in Section 4.2 on implementing backward Euler with linear interpolation to solve the $N = 1$ case of (5.0.1). Using a constant stepsize h, at time step t_{n+1} the update u_{n+1} is a root of the function $g_{n+1}(v)$ given in (4.2.4),

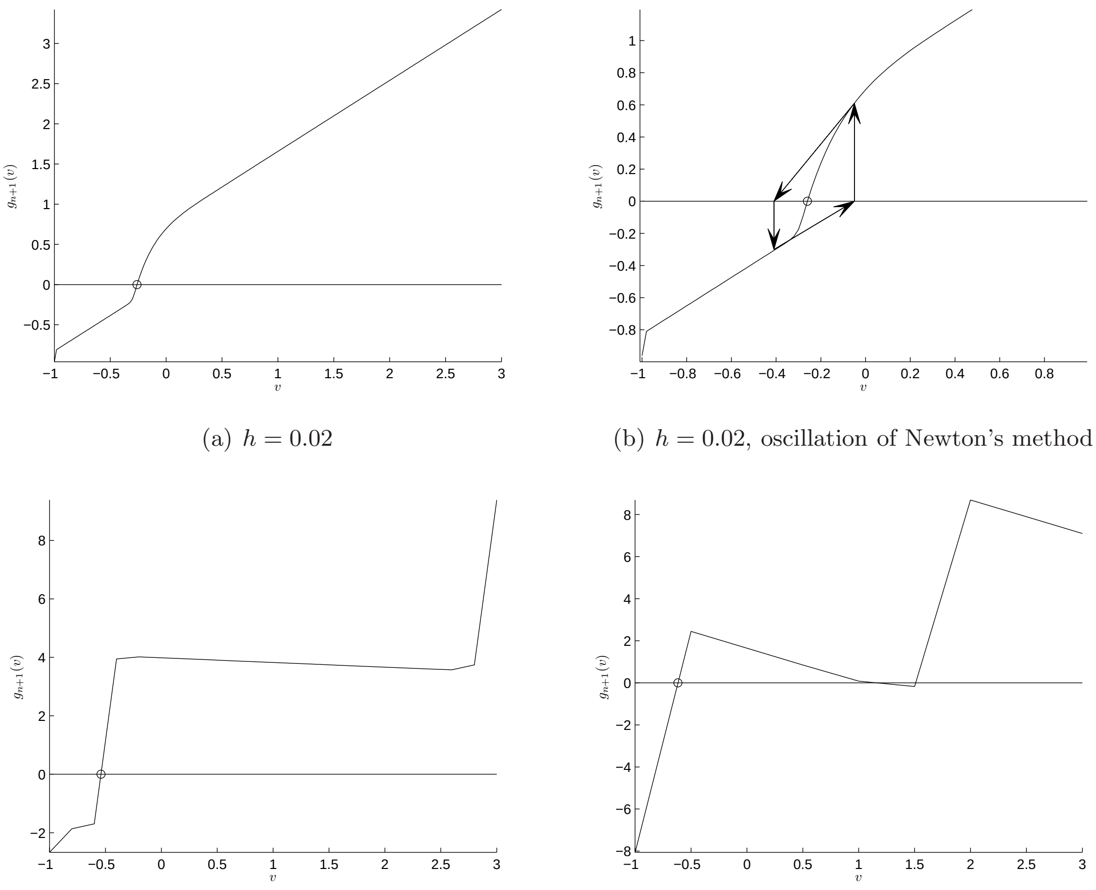
$$g_{n+1}(v) = v - u_n + \frac{h}{\varepsilon}\left(\gamma v + \kappa_1 \tilde{Y}_{n+1}(v)\right),$$

where $\tilde{Y}_{n+1}(v) = \eta_v(t_{n+1} - a_1 - c_1 v)$, $\eta(t)$ is the continuous extension and

$$\eta_v(t) = \begin{cases} \eta(t), & \text{if } t \leqslant t_n, \\ (1-\theta)\, u_n + \theta v, \quad \theta = \frac{t-t_n}{h}, & \text{if } t > t_n. \end{cases}$$

Solving for the root of $g_{n+1}(v)$ using a fixed point iteration is not recommended because this iteration requires the stepsize to be small enough in order to converge. With such a requirement we might as well have used an explicit method instead of backward Euler. A better choice would be to use a Newton iteration. Since the DDE and the delay terms are linear and the interpolation is piecewise linear, $g_{n+1}(v)$ is piecewise linear in every subinterval $\left[\frac{-a+mh}{c}, \frac{-a+(m+1)h}{c}\right]$, $m = 1, ..., n$. Consequently, the Newton method converges in one step to the correct solution if the starting point is in the same subinterval as the root. For larger stepsizes however, we run into the problems when we are solving for the turning points of the solutions. Figure 5–2 shows sample plots of $g_{n+1}(v)$ at a trough point of the $\varepsilon = 1$ solution in Figure 5–1(a) ($t \approx 20$). In this plot the root is located inside an interval where the $g_{n+1}(v)$ jumps. Figure 5–2(b) shows how the iterates of the Newton method may exhibit oscillations between the sections before and after the jump, and never actually converge to the actual root. Notice that this problem already occurs for $\varepsilon = 1$ and $h = 0.02$. The jump is sharper for larger h and smaller ε. Examples of these are shown in Figures 5–2(c) and (d). In these cases the Newton method fails in finding the roots unless the starting point is in the same subinterval as the root. It is difficult to choose such a starting point for this problem because the jumps in

$g_{n+1}(v)$ correspond to jumps in the solution. A starting guess for u_{n+1} based on extrapolation from previous mesh values will not be close to the root.



(a) $h = 0.02$

(b) $h = 0.02$, oscillation of Newton's method

(c) $h = 0.2$

(d) $h = 0.3$

Figure 5–2: Illustration of $g_{n+1}(v)$ for the $\varepsilon = 1$ solution in Figure 5–1(a) at a trough point ($t \approx 20$). Newton method by itself fails at finding the roots in these cases unless the starting point is sufficiently close to the root.

The problems shown in Figure 5–2 go away with smaller stepsizes. But since we are using an implicit method so that we may use a large stepsize, it does not make sense to reduce the stepsize for the root-finder. The bisection method is more reliable than the Newton method in finding the roots in the examples in Figure 5–2. Even better, a Newton-bisection algorithm may be used in which the default iteration is the Newton method but the algorithm switches to the bisection method if it starts to detect oscillations. Since there is no simple multidimensional analogue to the bisection algorithm, we cannot use this fix for non-scalar problems.

In more sophisticated solvers like RADAR5 [23], variable stepsize selection is used to accurately solve for the roots. This is beyond the scope of our current work. We instead use a fixed stepsize which we only change when we have to include a discontinuity point in the mesh (refer to Section 5.5).

5.2 Definitions

The general form of a scalar, state dependent, multiple delay problem that we would like to solve is

$$\begin{aligned} &\dot{u}(t) = f(t, u(t), u(\alpha_1(t, u(t))), ..., u(\alpha_N(t, u(t)))), && t_0 \leqslant t \leqslant t_f, \\ &u(t) = \varphi(t), && t \leqslant t_0, \end{aligned} \tag{5.2.1}$$

where $f(t, u, v_1, ..., v_N) \in C([t_0, t_f] \times \mathbb{R} \times \mathbb{R} \times ... \times \mathbb{R}, \mathbb{R})$. Advances are not allowed so throughout the entire solution we require $\alpha_i(t, u(t)) \leqslant t$ for all $t \in [t_0, t_f]$ and $i = 1, ..., N$. Sample problems, some with exact solutions, are found in Section 5.6. If for every compact subset R of $\mathbb{R}$ there exist constants $L_0, ..., L_N > 0$ such that

$$\|f(t, u, v_1, ..., v_N) - f(t, \tilde{u}, \tilde{v}_1, ..., \tilde{v}_N)\| \leqslant L_0 \|u - \tilde{u}\| + L_1 \|v_1 - \tilde{v}_1\| + ... + L_N \|v_N - \tilde{v}_N\|$$

for all $t \in [t_0, t_f]$ and $u, v_1, ..., v_N, \tilde{u}, \tilde{v}_1, ..., \tilde{v}_N \in R$ then the local existence and uniqueness result in Theorem 1.1.3 by Driver [14] applies.

Recall the notation introduced in Section 1.2. An RK method extended to solve DDEs is given by its Butcher tableau with matrix A, abscissae c_i and weight polynomial functions $b_i(\theta)$. Given a mesh $\Delta = \{t_n\}_{n=0}^{n_f}$ of discrete time values, the approximation u_n to the solution of (5.2.1) at time t_n is obtained by setting $u_0 = \varphi(t_0)$ and solving

$$\begin{aligned} &K_{n+1}^{(i)} = f\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}, \tilde{Y}_{n+1,1}^{(i)}, \tilde{Y}_{n+1,2}^{(i)}, ..., \tilde{Y}_{n+1,N}^{(i)}\right), \\ &u_{n+1} = u_n + h_{n+1} \sum_{i=1}^{s} b_i(1) K_{n+1}^{(i)}, \end{aligned} \tag{5.2.2}$$

where $t_{n+1}^{(i)} = t_n + c_i h_{n+1}$, $Y_{n+1}^{(i)} = u_n + h_{n+1} \sum_{j=1}^{s} a_{ij} K_{n+1}^{(j)}$ and $\tilde{Y}_{n+1,k}^{(i)} = \eta\left(\alpha_k\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}\right)\right)$ for $i = 1, ..., s$ and $k = 1, ..., N$. The function η is the continuous extension of the discrete method defined by

$$\eta(t) = \begin{cases} \varphi(t), & t \leqslant t_0 \\ \text{interpolation using mesh and stage values found in previous steps,} & t_0 \leqslant t \leqslant t_n, \\ \text{interpolation using mesh and stage values used in the current step,} & t \geqslant t_n. \end{cases}$$

The continuous extension $\eta(t)$ for $t \in [t_m, t_{m+1}]$ is found by first calculating $\theta = \frac{t-t_m}{h_{m+1}}$ so that $t = t_m + \theta h_{m+1}$ and

$$\eta(t) = u_m + h_{m+1} \sum_{i=1}^{s} b_i(\theta) K_{m+1}^{(i)}. \tag{5.2.3}$$

In this chapter, the equations are written in K-notation and solving for the stages means solving for $K_{n+1}^{(i)}$ instead of $Y_{n+1}^{(i)}$. As discussed in Section 1.2, this is more convenient when there is overlapping. Recall that the overlapping occurs if we are currently solving for the update u_{n+1} and for some $i \in \{1, ..., s\}$ and $k \in \{1, ..., N\}$, $\alpha_k\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}\right) \in [t_n, t_{n+1}]$. In this case the equation for the continuous extension (1.2.5) becomes an implicit equation even if the method itself is explicit. For sufficiently small stepsizes, these equations are known to be well-posed problems [8]. For sufficiently small stepsizes, Theorem 1.2.2 also states that a method of order p with a continuous extension of order q can be implemented so that the overall global error is $\min\{p, q+1\}$ [8]. In both cases, this sufficiently small stepsize is very small for stiff problems.

5.3 SDIRK methods extended to solve DDEs

Diagonally implicit RK methods (DIRKs) are RK methods where the A matrix is lower triangular. Singularly diagonally implicit RK methods (SDIRKs) are DIRK methods were the entries in the diagonal all have the same value. These methods are sometime called semi-implicit methods because the stages are solved for in order, one at a time. For scalar ODE problems this means that one would only need to solve a scalar equation at each stage, a more tractable problem than solving an s-dimensional system all at once. When SDIRK schemes are extended to solve DDEs the need to find the spurious stages destroys this special property of DIRK methods when there is overlapping in the lower stages. For instance, suppose that at $i < s$, $\alpha_k\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}\right) \in [t_n, t_{n+1}]$ for some $k \in \{1, ..., N\}$. Then in order to solve the first stage we require the value of $\tilde{Y}_{n+1,k}^{(i)} = \eta\left(\alpha_k\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}\right)\right)$. But we see from (5.2.3) that η on the current interval requires $K_{n+1}^{(j)}$, $j = 1, ..., s$. Thus in the overlapping case the first stage cannot be solved without the other stages and in this case we might as well have used a fully implicit method.

For non-stiff problems, it is not a big deal to solve for all the stages all at once. However for stiff problems such as (5.0.1) with small ε, the multidimensional iteration may come across a multidimensional version of the problems we observed in Figure 5–2. Due to this it is difficult to get the Newton iterations to converge and there is no simple multidimensional bisection algorithm to help it out. Because of this we would like to preserve the SDIRK property when extended to DDEs with possible overlapping.

The problem is circumvented by employing continuous extensions that do not need to use higher stages before they have been found, i.e.

$$b_i\left(\theta\right) = 0, \quad \text{for } \theta < c_{i-1} \tag{5.3.1}$$

With this property, if $\theta \in [0, c_i]$ the continuous extension (5.2.3) can be written as

$$\eta\left(t_m + \theta h_{m+1}\right) = u_m + h_{m+1} \sum_{j=1}^{i} b_j\left(\theta\right) K_{m+1}^{(j)}. \tag{5.3.2}$$

To our knowledge, such continuous extensions have not been used before, at least in the numerical solution DDEs. Thus the continuous extensions we used for our solver had to be derived using the order and continuity conditions. Examples of such continuous extensions are derived in the next section.

Definition 5.3.1. An s-stage RK method given by its Butcher tableau with matrix A, abscissae c_i and piecewise polynomial weight functions $b_i\left(\theta\right)$ has a DIRK-type continuous extension, and the weight functions are called DIRK-type weight functions if they satisfy (5.3.1) for $i = 1, ..., s$. In this case, we may write the weight functions as

$$b_i\left(\theta\right) = B_{j,i}\left(\theta\right), \text{ if } \theta \in \left[c_{j-1}, c_j\right], \tag{5.3.3}$$

where $c_0 = 0$. Since $B_{j,i}\left(\theta\right) = 0$ if $j < i$, the matrix B is lower triangular.

Theorem 1.2.2 and indeed most of other results on RK methods for DDEs are stated using polynomial weight functions. The results however can be easily extended to RK methods with DIRK-type continuous extensions. This result is given in Theorem 5.3.2.

Theorem 5.3.2. *Let $f\left(t,u,v_1,...,v_N\right) \in C^p([t_0,t_f] \times \mathbb{R}^{(N+1)d},\mathbb{R}^d)$, let the deviated arguments $\alpha_k\left(t,u(t)\right) \leqslant t$ be C^p-continuous in $[t_0,t_f]$ for $k=1,...,N$ and $\varphi\left(t\right)$ be C^p-continuous. Assume that the discrete mesh $\{t_0,...,t_{n_f}\}$ includes all the discontinuity points of order $\leqslant p$ in $[t_0,t_{n_f}]$. If the underlying CRK method with DIRK-type continuous extension has discrete order p and uniform order q, then the DDE method given by* (5.2.2) *and* (5.3.2) *applied to the DDE* (5.2.1) *has discrete global order and uniform global order $q' = \min\{p,q+1\}$; that is*

$$\max_{1\leqslant n\leqslant n_f} \|u(t_n)-u_n\| = O\left(h^{q'}\right), \quad \max_{t_0\leqslant t\leqslant t_{n_f}} \|u(t)-\eta(t)\| = O\left(h^{q'}\right)$$

where $h = \max_{1\leqslant n\leqslant n_f} h_n$.

Proof. In the proof of Theorem 1.2.2 (Theorem 6.1.2 in Bellen and Zennaro [8]), the continuous extension is only required to have order q and satisfy the endpoint conditions $b_i\left(0\right)=0$ and $b_i\left(1\right)=b_i$. There are no continuity or differentiability conditions. Thus the same proof can be applied to RK methods with DIRK-type polynomial functions. The extension to multiple delays is straightforward. □

Recall from Definition 1.2.1 that the order of a method is only defined for $h\to 0$. So the order of the method is really only relevant for small stepsizes. For stiff problems, stability of a method is more important. The SDIRK methods that we consider are all L-stable methods (A-stable with its stability functions $R\left(z\right)$ satisfying $R\left(\infty\right)=0$). Requiring $u_{n+1}=Y_{n+1}^{(s)}$ (the last stage) results in the $R\left(\infty\right)=0$ property [26]. We consider four methods, of orders one through four with a default continuous extension that uses polynomial weight functions. When there is overlapping at some stage $i<s$, the continuous extension switches to a DIRK-type weight function. The continuous extensions chosen are not natural (the $\tilde{Y}_{n+1}^{(i)}$ do not necessarily coincide with $\eta\left(t_{n+1}^{(i)}\right)$) unless otherwise specified.

SDIRK1 (backward Euler with linear interpolation)

In our discussion of backward Euler in Chapter 4 it was more convenient to use the Y-notation because the only stage of the method is also the update u_{n+1}. Here is the form of backward Euler with linear interpolation using the K-notation.

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}, \qquad b_1\left(\theta\right)=\theta$$

SDIRK2

The two-stage, L-stable, second order method is given by the following Butcher tableau with $\gamma = 1 - \frac{1}{\sqrt{2}}$. This method is derived in [26].

$$\begin{array}{c|cc} \gamma & \gamma & \\ 1 & 1-\gamma & \gamma \\ \hline & 1-\gamma & \gamma \end{array}$$

There are several choices for continuous extensions of this method, some of which satisfy (5.3.1). We use the notation given in (5.3.3).

I. Default: First order continuous extension using linear interpolation between adjacent mesh values.

$$\begin{aligned} b_1(\theta) &= (1-\gamma)\,\theta \\ b_2(\theta) &= \gamma\theta \end{aligned}$$

II. Natural, first order DIRK-type continuous extension that is C^1 in the $[0,1]$. The polynomials are degree one for $\theta \in [0,\gamma]$ and degree two for $\theta \in [\gamma,1]$.

$$\begin{aligned} B_{1,1}(\theta) &= \theta \\ B_{2,1}(\theta) &= -5 + \frac{7\sqrt{2}}{2} + \left(7 - 4\sqrt{2}\right)\theta + \left(-2 + \sqrt{2}\right)\theta^2 \\ B_{2,2}(\theta) &= 5 - \frac{7\sqrt{2}}{2} + \left(4\sqrt{2} - 6\right)\theta + \left(2 - \sqrt{2}\right)\theta^2 \end{aligned}$$

III. Natural, first order DIRK-type continuous extension that is degree one in the entire interval $[0,1]$. This is equivalent to linear interpolation between the stages $Y_{n+1}^{(i)}$.

$$\begin{aligned} B_{1,1}(\theta) &= \theta \\ B_{2,1}(\theta) &= \frac{\gamma^2 - 2\gamma\theta + \theta}{1-\gamma} \\ B_{2,2}(\theta) &= \frac{\gamma(\theta-\gamma)}{1-\gamma} \end{aligned}$$

SDIRK3

The three-stage, L-stable, third order method can be derived using order condition and the requirement that $a_{si} = b_i$.

$$\begin{array}{c|ccc} \gamma & \gamma & & \\ c & c-\gamma & \gamma & \\ 1 & 1-b-\gamma & b & \gamma \\ \hline & 1-b-\gamma & b & \gamma \end{array}$$

The way that the method has been written up, there are only three more order conditions to satisfy third order. Solving these equations for b, c and γ yields

$$c = \gamma\left(1 + \frac{\frac{1}{6} - \frac{\gamma}{2}}{\frac{1}{6} - \gamma + \gamma^2}\right), \quad b = \frac{\left(\frac{1}{6} - \gamma + \gamma^2\right)^2}{\gamma^2\left(\frac{1}{6} - \frac{\gamma}{2}\right)},$$

and γ as the root of the cubic equation

$$\gamma^3 - 3\gamma^2 + \frac{3}{2}\gamma - \frac{1}{6} = 0.$$

The value of γ that makes this an L-stable method is the middle root of that polynomial (approximately 0.435866521508459) according to the analysis of the stability function of the method by Hairer and Wanner [26] and Owren and Simonsen [47]. Using the polynomial equation for γ, the expressions for b and c can also be simplified to

$$b = \frac{2}{3}\frac{(1-3\gamma)}{(1-\gamma)^2}, \quad c = \frac{1}{2}(1+\gamma).$$

I. Default: Second order Natural Continuous Extension (NCE). NCEs are discussed in Chapter 5 of Bellen and Zennaro [8]. It can be shown that an NCE of a p-th order RK method has order $q \geqslant \lfloor \frac{p+1}{2} \rfloor$ and that one may be obtained by requiring the following conditions:

$$b_i(0) = 0, \qquad \int_0^1 \theta^r b_i'(\theta)\, d\theta = b_i c_i^r, \quad \text{for } r = 0, ..., q-1.$$

Applying these conditions with $q = 2$ yields the following NCE for the third order method

$$b_i(\theta) = \theta(4 - 6c_i)\, b_i + \theta^2(6c_i - 3)\, b_i.$$

II. Second order DIRK-type continuous extension that utilizes u_n as an extra stage. The polynomial function $b_0(\theta) = B_{j,0}(\theta)$ is the DIRK-type weight function corresponding to this stage.

$$
\begin{aligned}
B_{1,0}(\theta) &= \theta - \frac{\theta^2}{2\gamma} \\
B_{1,1}(\theta) &= \frac{\theta^2}{2\gamma} \\
B_{2,0}(\theta) &= \frac{\theta}{2}\left(1 - \frac{\theta-\gamma}{c-\gamma}\right) \\
B_{2,1}(\theta) &= \frac{\theta}{2}\left(1 + \gamma\frac{\theta-\gamma}{(c-\gamma)^2}\right) \\
B_{2,2}(\theta) &= \frac{\theta}{2}\frac{(\theta-\gamma)(c-2\gamma)}{(c-\gamma)^2} \\
B_{3,1}(\theta) &= 1 - b - \gamma + (\theta-1)\left(1 - b - \gamma + \frac{2\gamma-1}{2(c-\gamma)}\theta\right) \\
B_{3,2}(\theta) &= b + (\theta-1)\left(b + \frac{1-4\gamma}{2(c-\gamma)}\right) \\
B_{3,3}(\theta) &= \gamma\left[1 + (\theta-1)\left(1 + \frac{\theta}{1-c}\right)\right]
\end{aligned}
$$

III. Natural, first order DIRK-type continuous extension that is degree one in the entire interval $[0,1]$. Using this continuous extension lowers the order of the overall method to two, but unlike the first option, it does not require the use of u_n as an extra stage (which may affect the stability properties of the method).

$$
\begin{aligned}
B_{1,1}(\theta) &= \theta \\
B_{2,1}(\theta) &= \frac{\gamma^2 - 2\gamma\theta + c\theta}{1-\gamma} \\
B_{2,2}(\theta) &= \frac{\gamma(\theta-\gamma)}{c-\gamma} \\
B_{3,1}(\theta) &= \frac{(1-\theta)(c-\theta) + (\theta-c)(1-b-\gamma)}{1-c} \\
B_{3,2}(\theta) &= \frac{\gamma(1-\theta) + (\theta-c)b}{1-c} \\
B_{3,3}(\theta) &= \frac{(\theta-c)\gamma}{1-c}
\end{aligned}
$$

SDIRK4

There is no four-stage fourth order L-stable SDIRK method [2, 47]. Instead we use a five-stage fourth order method is presented in Chapter IV.6 of Hairer and Wanner [26].

$$\begin{array}{c|ccccc}
\frac{1}{4} & \frac{1}{4} & & & & \\
\frac{3}{4} & \frac{1}{2} & \frac{1}{4} & & & \\
\frac{11}{20} & \frac{17}{50} & -\frac{1}{25} & \frac{1}{4} & & \\
\frac{1}{2} & \frac{371}{1360} & -\frac{137}{2720} & \frac{15}{544} & \frac{1}{4} & \\
1 & \frac{25}{24} & -\frac{49}{48} & \frac{125}{16} & -\frac{85}{12} & \frac{1}{4} \\
\hline
 & \frac{25}{24} & -\frac{49}{48} & \frac{125}{16} & -\frac{85}{12} & \frac{1}{4}
\end{array}$$

I. Default: Hairer and Wanner [26] presented a continuous extension with polynomial weight functions for this method that is third order.

$$b_1(\theta) = \tfrac{11}{3}\theta - \tfrac{463}{72}\theta^2 + \tfrac{217}{36}\theta^3 - \tfrac{20}{9}\theta^4$$

$$b_2(\theta) = \tfrac{11}{2}\theta - \tfrac{385}{16}\theta^2 + \tfrac{661}{24}\theta^3 - 10\theta^4$$

$$b_3(\theta) = -\tfrac{125}{18}\theta + \tfrac{20125}{432}\theta^2 - \tfrac{8875}{216}\theta^3 + \tfrac{250}{27}\theta^4$$

$$b_4(\theta) = -\tfrac{85}{4}\theta^2 + \tfrac{85}{6}\theta^3$$

$$b_5(\theta) = -\tfrac{11}{9}\theta + \tfrac{557}{108}\theta^2 - \tfrac{359}{54}\theta^3 + \tfrac{80}{27}\theta^4$$

II. First order DIRK-type continuous extension that is degree one in the entire interval $[0, 1]$. Using this continuous extension lowers the global order of the method to two.

$$\begin{aligned}
B_{1,1}(\theta) &= \theta \\
B_{2,1}(\theta) &= B_{3,1}(\theta) = \frac{\theta}{2} - \frac{1}{8} \\
B_{2,2}(\theta) &= B_{3,2}(\theta) = \frac{\theta}{2} + \frac{1}{8} \\
B_{3,3}(\theta) &= B_{4,1}(\theta) = B_{4,2}(\theta) = B_{4,3}(\theta) = B_{4,4}(\theta) = 0 \\
B_{5,1}(\theta) &= \frac{13\theta}{6} - \frac{9}{8} \\
B_{5,2}(\theta) &= -\frac{61\theta}{12} + \frac{65}{16} \\
B_{5,3}(\theta) &= \frac{25}{6}\left(\theta - \frac{3}{4}\right) \\
B_{5,4}(\theta) &= -\frac{85}{3}\left(\theta - \frac{3}{4}\right) \\
B_{5,5}(\theta) &= \theta - \frac{3}{4}
\end{aligned}$$

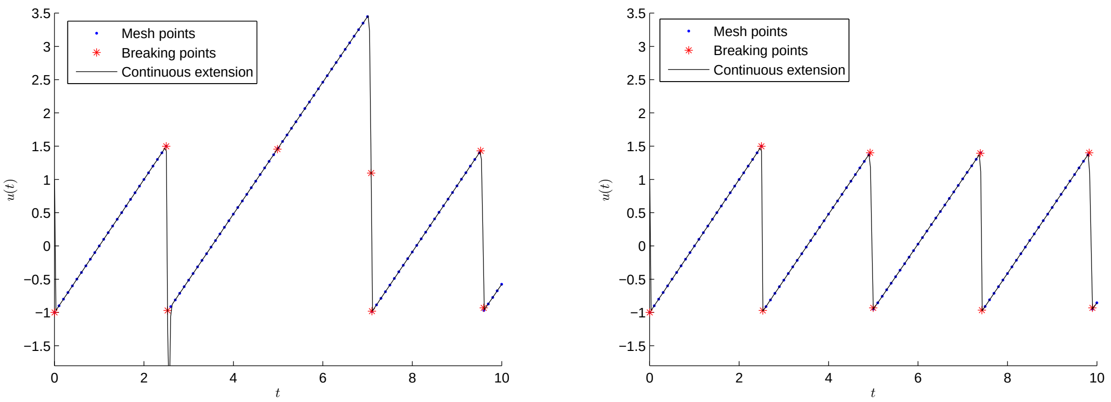
To always be clear about what continuous extensions are being used, we use the following notation: SDIRKp (x, y) stands for the p-th order SDIRK method with continuous extension x for the non-overlapping case and y for the overlapping case. So SDIRK2 (I,II) stands for the second order SDIRK method using continuous extension I when there is no overlapping and II for the overlapping case. For SDIRK1, there is only one continuous extension so it is not necessary to specify this. For higher order methods, the continuous extension is always I for the non-overlapping case.

By Theorem 5.3.2, an RK method with order p requires a continuous extension of at least order $p - 1$ to retain its order. SDIRK1, SDIRK2 and SDIRK3 are each provided with one DIRK-type continuous extensions that accommodates this. For SDIRK4 we currently only have a DIRK-type continuous extension of order one, yielding a global order of two. Due to this it would be preferable to use SDIRK3 to SDIRK4 when overlapping cases and vanishing delays are expected. In cases where the delay terms are bounded away from zero (i.e. there exists a $\bar{\tau} > 0$ such that $\tau_i(t, u(t)) = t - \alpha_i(t, u(t)) \geqslant \bar{\tau}$ for all $t \geqslant t_0$) then choosing h to be small enough ($h < \bar{\tau}$) eliminates overlapping cases. Thus for these problems SDIRK4(I,II) is still of order four.

Applied to stiff problems, it is possible to find a further reduction in the order of DDE methods analogous to the known reduction of order of ODE methods for stiff ODEs [26]. For example, the code RADAR5 [23] uses s-stage Radau IIA methods that are order $2s$ for ODEs. For DDEs, these methods have global order $s + 1$ which may reduce to order s for stiff DDEs [23].

For SDIRK2 and SDIRK3 we have two choices of DIRK-type continuous extensions. The second option for overlapping (given as continuous extension III for both methods) is a continuous extension of order and degree one. The only DIRK-type continuous extension we currently have for SDIRK4 is also of order and degree one. Using a weight function with a low degree has the advantage of avoiding the problem of "spiking" which is displayed by SDIRK2 (I,II) in Figure 5–3(a). Spiking occurs when there are sharp changes in the solution corresponding to large changes in $K_{n+1}^{(i)}$ values. There is a sharp jump down in the solution at $t \approx 2.5$. If such a jump is to be reflected in going from u_n to u_{n+1}, some or all of the stage values $K_{n+1}^{(i)}$ must be relatively large. As a result of these large $K_{n+1}^{(i)}$ values, the continuous extension spikes too far down first before coming up to the value of u_{n+1}. The $Y_{n+1}^{(i)}$ stages themselves are not on the downwards spike which occurs in the time interval with $\theta \in [c_1, 1]$, the second interval of the interpolation where the $B_{2,i}(\theta)$ are degree two polynomials. The spiking of the continuous extension affects the solution at $t \approx 5$ to 7 because at these times the deviated argument maps back to $t \approx 2.5$. Figure 5–3(b) shows that switching to SDIRK2 (I,III) fixes this problem.

By Theorem 5.3.2, using a continuous extension of order and degree one does not affect the global order of the second order method as long as the discontinuity points are found accurately (we discuss this issue in Section 5.5). For the third order method, using a continuous extension of order one results in a reduction of the global order to two. This reduction of order is necessary for stiff problems because allowing for the spiking of continuous extension would be worse. Once again, stability is more important than order when using "large" stepsizes.



(a) SDIRK2 (I,II)

(b) SDIRK2 (I,III)

Figure 5–3: Numerical solutions of $0.001u(t) = -2u(t) - 3u(t-1-u(t))$ using SDIRK2 with $h = 0.1$ and different continuous extensions for the overlapping cases. This shows the spiking of continuous extension II at a trough point.

5.4 Solving for the stages

In this section we discuss the Newton-bisection algorithm used to solve for the stages. For convenience consider a DDE with a single deviated argument $\alpha(t, u(t))$. Suppose that the solution is already known up to t_n for some $n \geqslant 0$. To determine the update u_{n+1} at time t_{n+1}, the stages have to be solved for in order. For $i = 1$ to s, the stage $K_{n+1}^{(i)}$ is the root of the following function

$$G^{(i)}\left(K_{n+1}^{(i)}\right) = K_{n+1}^{(i)} - f\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}, \tilde{Y}_{n+1}^{(i)}\right), \tag{5.4.1}$$

where $Y^{(i)}_{n+1} = y_n + h_{n+1} \sum\limits_{j=1}^{i} a_{ij} K^{(j)}_{n+1}$ and $\tilde{Y}^{(i)}_{n+1} = \eta \left(\alpha \left(t^{(i)}_{n+1}, Y^{(i)}_{n+1} \right) \right)$ and η given by (5.3.2). In Section 5.1 we gave an example of how Newton's method by itself may not converge in finding the root of these functions. A better option is to use a combined Newton and bisection method. An outline of this is given in Algorithm 1.

```
Tolerance: TolX and TolY
Maximum iterate count: MaxCount
Input: $i$, $t_{n+1}$, $u_n$, $K^{(j)}_{n+1}$ for $j = 1, ..., i$, $h_{n+1}$, an interval $[L, M]$ in which the root must line
on and a starting point $K \in [L, M]$.
    1. If $G^{(i)}(L)\, G^{(i)}(M) > 0$ then
        • Look for an interval in which $G^{(i)}$ changes sign. This can be done using fminbnd in
          MATLAB fminbnd which uses parabolic interpolation and the golden section search
          (MATLAB references [11] and [19]).
        • Choose the leftmost interval and return it as $[L, M]$.
        • If no such interval is found then exit with an error.
    2. Set signleft← sign $\left(G^{(i)}(L)\right)$, IterateCount=0.
    3. While $\left|G^{(i)}(K)\right|$ >TolY or $M - L$ >TolX and IterateCount<MaxCount
        • IterateCount←IterateCount+1.
        • If sign $\left(G^{(i)}(K)\right)$ =signleft then $L \leftarrow K$. Otherwise $M \leftarrow K$.
        • Take a Newton step,
              $K \leftarrow K - \frac{G(K)}{\frac{d}{dK^{(i)}_{n+1}} G^{(i)}(K)}.$
          If $K \notin (L, M)$ then instead assign
              $K \leftarrow \frac{L + M}{2}.$
```

Algorithm 1: Root-finding algorithm using Newton and bisection methods

The derivative of $G^{(i)}$ is necessary to perform the Newton step. At any stage $i = 1, ..., s$, if $\alpha \left(t^{(i)}_{n+1}, Y^{(i)}_{n+1} \right) \leqslant t_n$ (no overlapping at this stage) then

$$\begin{aligned} \frac{\partial G^{(i)}}{\partial K^{(j)}_{n+1}} = \delta_{ij} &- \partial_2 f \left(t^{(i)}_{n+1}, Y^{(i)}_{n+1}, \tilde{Y}^{(i)}_{n+1} \right) h_{n+1} a_{ij} \\ &- \partial_3 f \left(t^{(i)}_{n+1}, Y^{(i)}_{n+1}, \tilde{Y}^{(i)}_{n+1} \right) \eta' \left(\alpha \left(t^{(i)}_{n+1}, Y^{(i)}_{n+1} \right) \right) \partial_2 \alpha \left(t^{(i)}_{n+1}, Y^{(i)}_{n+1} \right) h_{n+1} a_{ij}, \end{aligned} \tag{5.4.2}$$

where δ_{ij} is the Kronecker delta. If there is overlapping then we first solve for

$$\theta^{(i)}_{n+1} = \frac{\alpha \left(t^{(i)}_{n+1}, Y^{(i)}_{n+1} \right) - t_n}{h_{n+1}}, \tag{5.4.3}$$

and use this in finding the derivative

$$\begin{aligned}\frac{\partial G^{(i)}}{\partial K^{(j)}_{n+1}} &= \delta_{ij} - \partial_2 f\left(t^{(i)}_{n+1}, Y^{(i)}_{n+1}, \tilde{Y}^{(i)}_{n+1}\right) h_{n+1} a_{ij} \\ &- \partial_3 f\left(t^{(i)}_{n+1}, Y^{(i)}_{n+1}, \tilde{Y}^{(i)}_{n+1}\right) h_{n+1} \left(b_j\left(\theta^{(i)}_{n+1}\right) + \left(\sum_{k=1}^{s} b'_k\left(\theta^{(i)}_{n+1}\right) K^{(k)}_{n+1}\right) \partial_2 \alpha\left(t^{(i)}_{n+1}, Y^{(i)}_{n+1}\right) a_{ij}\right).\end{aligned}$$

This can be rewritten to appear more similar to (5.4.2).

$$\begin{aligned}\frac{\partial G^{(i)}}{\partial K^{(j)}_{n+1}} &= \delta_{ij} - \partial_2 f\left(t^{(i)}_{n+1}, Y^{(i)}_{n+1}, \tilde{Y}^{(i)}_{n+1}\right) h_{n+1} a_{ij} \\ &\quad - \partial_3 f\left(t^{(i)}_{n+1}, Y^{(i)}_{n+1}, \tilde{Y}^{(i)}_{n+1}\right) \eta'\left(\alpha\left(t^{(i)}_{n+1}, Y^{(i)}_{n+1}\right)\right) \partial_2 \alpha\left(t^{(i)}_{n+1}, Y^{(i)}_{n+1}\right) h_{n+1} a_{ij} \\ &\quad - \partial_3 f\left(t^{(i)}_{n+1}, Y^{(i)}_{n+1}, \tilde{Y}^{(i)}_{n+1}\right) h_{n+1} b_j\left(\theta^{(i)}_{n+1}\right)\end{aligned} \tag{5.4.4}$$

These derivatives may be used in multidimensional Newton iterations. However, our solver only needs $j = i$. To extend these results to multiple delays, the arguments $\left(t^{(i)}_{n+1}, Y^{(i)}_{n+1}, \tilde{Y}^{(i)}_{n+1}\right)$ should be replaced with $\left(t^{(i)}_{n+1}, Y^{(i)}_{n+1}, \tilde{Y}^{(i)}_{n+1,1}, ..., \tilde{Y}^{(i)}_{n+1,N}\right)$, the $\partial_3 f$ with $(\partial_3, ..., \partial_{N+2}) f$, the α with $(\alpha_1, ..., \alpha_N)$ and $\eta'(\alpha)$ with $(\eta'(\alpha_1), ..., \eta'(\alpha_N))$. All the products involving these terms in (5.4.4) should now be changed to dot products.

```
1. Find a starting value for $K^{(i)}_{n+1}$ for $i = 1, ..., s$. It could be its value from the previous step
   or a guess based on the values of $K^{(j)}_{n+1}$ for $j = 1, ..., i$.
2. Set the weight polynomials to be the default polynomial weight functions.
3. For $i = 1, ..., s$
      • Derive an interval $[L, M]$ such that $K^{(i)}_{n+1}$ must lie in this interval.
      • Solve for $K^{(i)}_{n+1}$ using Algorithm 1.
      • If $\alpha\left(t_n + c_i h_{n+1}, Y^{(i)}_{n+1}\right) \geqslant t_n$ exit this for loop.
4. If $i < s$ then
      • Set the weight polynomials to be the DIRK-type polynomial weight functions.
      • For $i = 1, ..., s$
           – Derive an interval $[L, M]$ such that $K^{(i)}_{n+1}$ must lie in this interval.
           – Solve for $K^{(i)}_{n+1}$ using Algorithm 1.
All the $K^{(i)}_{n+1}$ values need to be saved along with a marker that states whether or not there was
overlapping in that interval.
```

Algorithm 2: Solving for the stages

When solving for $K^{(i)}_{n+1}$, the interval $[L, M]$ in Algorithm 2 should be derived from the DDE that is being solved numerically. For (5.0.1) we know from Section 2.2 that the solution may be bounded inside $[L_0, M_0]$ from (2.2.3). Using these, one may say that the bounds on the derivative of the solution has to be in $[-M, M]$ where $M = \frac{\gamma + \sum_{i=1}^{N} \kappa_i}{\varepsilon} \max\{L_0, M_0\}$. However, these bounds may be very large for small ε. A more accurate lower bound can be found by requiring all delays to not become advances

$$\alpha\left(t_{n+1}^{(i)}, u_n + h_{n+1}\sum_{j=1}^{i-1} a_{ij}K_{n+1}^{(j)} + h_{n+1}a_{ii}K_{n+1}^{(i)}\right) \leqslant t_{n+1}^{(i)}. \tag{5.4.5}$$

For (5.0.1), this yields an effective lower bound

$$L = \frac{-\frac{a_1}{c_1} - u_n - h\sum_{j=1}^{i-1} A_{ij}K_{n+1}^{(j)}}{hA_{ii}}. \tag{5.4.6}$$

Similar bounds may be found for other equations.

5.5 Tracking discontinuities

Theorem 4.3.8 of Bellen and Zennaro [8] state that if a DDE has a smooth solution apart from a finite number of discontinuities in the derivative, then for any choice of mesh an RK method of order $p \geqslant 1$ with a continuous extension of uniform order $q \geqslant 1$ performs as a DDE method with global order $\min\{p, 2\}$. Thus, we can implement SDIRK1 and SDIRK2 without discontinuity tracking and expect no loss of order for the non-stiff problems. This is a useful option when the discontinuities are hard to find. In general, we would prefer to include all discontinuities up to order p in the mesh. To do this, first check to see if there is a discontinuity in the derivative at the starting point t_0. If there is one, store this as $\tilde{\xi}_1 = t_0$ with order $\zeta_1 = 1$. Say we have already solved for all the discontinuities up to time t_n and these are saved in the array $\{\tilde{\xi}\}_{i=1}^{\ell}$ with corresponding array $\{\zeta\}_{i=1}^{\ell}$ which stores the order of the discontinuity points. An outline of how to check and capture discontinuities in $[t_n, t_{n+1}]$ is given in Algorithm 3.

If there is no overlapping in the current time step $[t_n, t_{n+1}]$ and Algorithm 3 detects a discontinuity point in this interval that maps $\alpha_k(t, \eta(t))$ back to a previously calculated discontinuity point ξ, it uses a Newton step to determine the next correct stepsize. The function associated with this Newton step is

$$H(h_{n+1}) = \alpha_k(t_n + h_{n+1}, w) - \xi, \tag{5.5.1}$$

```
Tolerance: TolX and TolY
Maximum iterate count: MaxCount
At every time step, set flag$\leftarrow 1$, DiscCountn$\leftarrow 0$, $h \leftarrow h_{\max}$, $h_{\text{left}} \leftarrow 0$ and $h_{\text{right}} \leftarrow h_{\max}$.
While flag=1 and DiscCountn<DiscCountMax do the following:
  1. DiscCountn$\leftarrow$DiscCountn$+1$.
  2. Solve for the continuous approximation $\eta(t)$ on $[t_n, t_{n+1}]$ where $t_{n+1} = t_n + h$.
  3. If DiscCountn=1 then
       • For $k = 1, .., N$,
           – For every $j$, check if there is an $x \in [t_n, t_n + h_{\max}]$ such that

                 $\alpha_k\left(x, \eta\left(x\right)\right) = \tilde{\xi}_i, \quad \text{for some } i = 1, ..., \ell \text{ and } k = 1, ..., N,$

                 $\zeta_i \leqslant p - 1.$

             This may be done by checking the values at the endpoints. For every pair of $x$
             and $\tilde{\xi}_i$ found, if $|x - t_n| <$ TolX then ignore this point.
  4. If there are no such discontinuity points found then set flag=0. Otherwise choose the set
     of $x$, $k$ and $\tilde{\xi}_i$ values with the minimum $x$.
  5. signleft$\leftarrow \operatorname{sign}\left(\alpha_k\left(t_n, u_n\right) - \xi_i\right)$.
  6. If flag=1
       • If $\left|\alpha_k\left(t_n + h, \eta\left(t_n + h\right)\right) - \tilde{\xi}_i\right|$ <TolY
           – $\tilde{\xi}_{\ell+1} \leftarrow t_n + h$, $\zeta_{\ell+1} \leftarrow \zeta_i + 1$, flag$\leftarrow 0$.
       • Else, if $|h_{\text{left}} - h_{\text{right}}|$ <TolX
           – $h \leftarrow h_{\text{left}}$, $\tilde{\xi}_{\ell+1} \leftarrow t_n + h_{\text{left}}$, $\zeta_{\ell+1} \leftarrow \zeta_i + 1$, flag$\leftarrow 0$.
       • Otherwise reject the current step and the approximation $\eta(t)$ on that interval.
           – If $\operatorname{sign}\left(\alpha_j\left(t_n + h, w\right) - \xi_i\right) =$signleft then $h_{\text{left}} \leftarrow h$. Otherwise $h_{\text{right}} \leftarrow h$.
           – If there is no overlapping, $\tilde{h} \leftarrow h - \frac{H(h)}{\frac{dH(h)}{dh_{n+1}}}$ (from (5.5.1),(5.5.4)). Otherwise,
             $\tilde{h} \leftarrow \frac{h_{left} + h_{right}}{2}$.
           – If $\tilde{h} \in [h_{\text{left}}, h_{\text{right}}]$ then $h \leftarrow \tilde{h}$. Otherwise $h \leftarrow \frac{h_{\text{left}} + h_{\text{right}}}{2}$.
           – flag$\leftarrow 1$.
```

Algorithm 3: Discontinuity tracking

where

$$w = \eta\left(t_n + h_{n+1}\right) = u_n + h_{n+1} \sum_{i=1}^{s} b_i\left(1\right) K_{n+1}^{(i)}. \tag{5.5.2}$$

In this case, the $K_{n+1}^{(i)}$'s are functions of the step size h_{n+1} as well. The new stepsize using a Newton iteration should be

$$h_{n+1}^{new} = h_{n+1}^{old} - \frac{H\left(h_{n+1}^{old}\right)}{\frac{dH\left(h_{n+1}^{old}\right)}{dh_{n+1}}} \tag{5.5.3}$$

The derivative of $H\left(h_{n+1}\right)$ is given by

$$\frac{dH\left(h_{n+1}\right)}{dh_{n+1}} = \partial_1 \alpha_k\left(t_n + h_{n+1}, w\right) + \partial_2 \alpha_k\left(t_n + h_{n+1}, w\right) \frac{dw}{dh_{n+1}}, \tag{5.5.4}$$

$$\frac{dw}{dh_{n+1}} = \sum_{i=1}^{s} b_i\left(1\right) K_{n+1}^{(i)} + h_{n+1} \sum_{i=1}^{s} b_i\left(1\right) \frac{dK_{n+1}^{(i)}}{dh_{n+1}}. \tag{5.5.5}$$

We need to calculate the derivative of the stages. In this derivation we first consider the case of one deviated argument $\alpha\left(t, u\left(t\right)\right)$ for simplicity. Suppose that there is no overlapping in the stages. Then,

$$\frac{dK_{n+1}^{(i)}}{dh_{n+1}} = \frac{df\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}, \tilde{Y}_{n+1}^{(i)}\right)}{dh_{n+1}}$$

$$\begin{aligned}\frac{dK_{n+1}^{(i)}}{dh_{n+1}} = {} & \partial_1 f\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}, \tilde{Y}_{n+1}^{(i)}\right) c_i + \partial_2 f\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}, \tilde{Y}_{n+1}^{(i)}\right)\left(\sum_{j=1}^{i} a_{ij} K_{n+1}^{(i)} + h_{n+1} \sum_{j=1}^{i} a_{ij} \frac{dK_{n+1}^{(i)}}{dh_{n+1}}\right) \\ & + \partial_3 f\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}, \tilde{Y}_{n+1}^{(i)}\right) \eta'\left(\alpha\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}\right)\right)\Bigg[\partial_1 \alpha\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}\right) c_i \\ & + \partial_2 \alpha\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}\right)\left(\sum_{j=1}^{i} a_{ij} K_{n+1}^{(i)} + h_{n+1} \sum_{j=1}^{i} a_{ij} \frac{dK_{n+1}^{(i)}}{dh_{n+1}}\right)\Bigg]\end{aligned}$$

To simplify the notation, let us write f for $f\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}, \tilde{Y}_{n+1}^{(i)}\right)$ and α for $\alpha\left(t_{n+1}^{(i)}, Y_{n+1}^{(i)}\right)$. Then this equation becomes

$$\begin{aligned}\frac{dK_{n+1}^{(i)}}{dh_{n+1}} = {} & \partial_1 f c_i + \partial_2 f\left(\sum_{j=1}^{i} a_{ij} K_{n+1}^{(i)} + h_{n+1} \sum_{j=1}^{i} a_{ij} \frac{dK_{n+1}^{(i)}}{dh_{n+1}}\right) \\ & + \partial_3 f \eta'\left(\alpha\right)\left(\partial_1 \alpha c_i + \partial_2 \alpha\left(\sum_{j=1}^{i} a_{ij} K_{n+1}^{(i)} + h_{n+1} \sum_{j=1}^{i} a_{ij} \frac{dK_{n+1}^{(i)}}{dh_{n+1}}\right)\right).\end{aligned}$$

Now we solve for $\frac{dK_{n+1}^{(i)}}{dh_{n+1}}$ as a function of $\frac{dK_{n+1}^{(j)}}{dh_{n+1}}$ for $j = 1, ..., i$. Rearrange the last equation to become

$$\frac{dK_{n+1}^{(i)}}{dh_{n+1}} = \partial_1 f c_i + \partial_3 f \eta' (\alpha) \partial_1 \alpha c_i + \left(\partial_2 f + \partial_3 f \eta' (\alpha) \partial_2 \alpha\right) \left(\sum_{j=1}^{i} a_{ij} K_{n+1}^{(i)} + h_{n+1} \sum_{j=1}^{i} a_{ij} \frac{dK_{n+1}^{(i)}}{dh_{n+1}} \right),$$

and collect the $\frac{dK_{n+1}^{(i)}}{dh_{n+1}}$ terms together

$$\begin{aligned} &\left[1 - \left(\partial_2 f + \partial_3 f \eta' (\alpha) \partial_2 \alpha\right) h_{n+1} a_{ii}\right] \frac{dK_{n+1}^{(i)}}{dh_{n+1}} \\ &\quad = \partial_1 f c_i + \partial_3 f \eta' (\alpha) \partial_1 \alpha c_i + \left(\partial_2 f + \partial_3 f \eta' (\alpha) \partial_2 \alpha\right) \left(\sum_{j=1}^{i} a_{ij} K_{n+1}^{(i)} + h_{n+1} \sum_{j=1}^{i-1} a_{ij} \frac{dK_{n+1}^{(i)}}{dh_{n+1}} \right). \end{aligned}$$

This yields

$$\frac{dK_{n+1}^{(i)}}{dh_{n+1}} = \frac{\partial_1 f c_i + \partial_3 f \eta' (\alpha) \partial_1 \alpha c_i + (\partial_2 f + \partial_3 f \eta' (\alpha) \partial_2 \alpha) \left(\sum_{j=1}^{i} a_{ij} K_{n+1}^{(i)} + h_{n+1} \sum_{j=1}^{i-1} a_{ij} \frac{dK_{n+1}^{(i)}}{dh_{n+1}} \right)}{1 - (\partial_2 f + \partial_3 f \eta' (\alpha) \partial_2 \alpha) h_{n+1} a_{ii}}. \tag{5.5.6}$$

As in the previous section, the result can be extended to N-delays by replacing $\partial_3 f$ with $(\partial_3, ..., \partial_{N+2}) f$, α with $(\alpha_1, ..., \alpha_N)$ and $\eta' (\alpha)$ with $(\eta' (\alpha_1), ..., \eta' (\alpha_N))$. All the products involving these terms in (5.5.6) should now be changed into dot products. Then this expression can be used in (5.5.5) and (5.5.4) in order to get the relevant derivatives. Then the new stepsize can be found using (5.5.3).

Since the $K_{n+1}^{(i)}$ derivative expressions were only derived for the nonoverlapping case, a Newton step to find the next stepsize is only taken in Algorithm 3 if there is no overlapping. We also have to consider the differentiability requirement in using Newton's method. In our problems we assume that the history function is always continuous. Because of this the discontinuity points of a solution must all stem from the discontinuity in the derivative at the initial time. Then any discontinuity point that we would be solving for in a time interval $[t_n, t_{n+1}]$ has to be of minimum order two. Then the continuous extension $\eta(t)$ and $H(h_{n+1})$ has a continuous first derivative in this interval

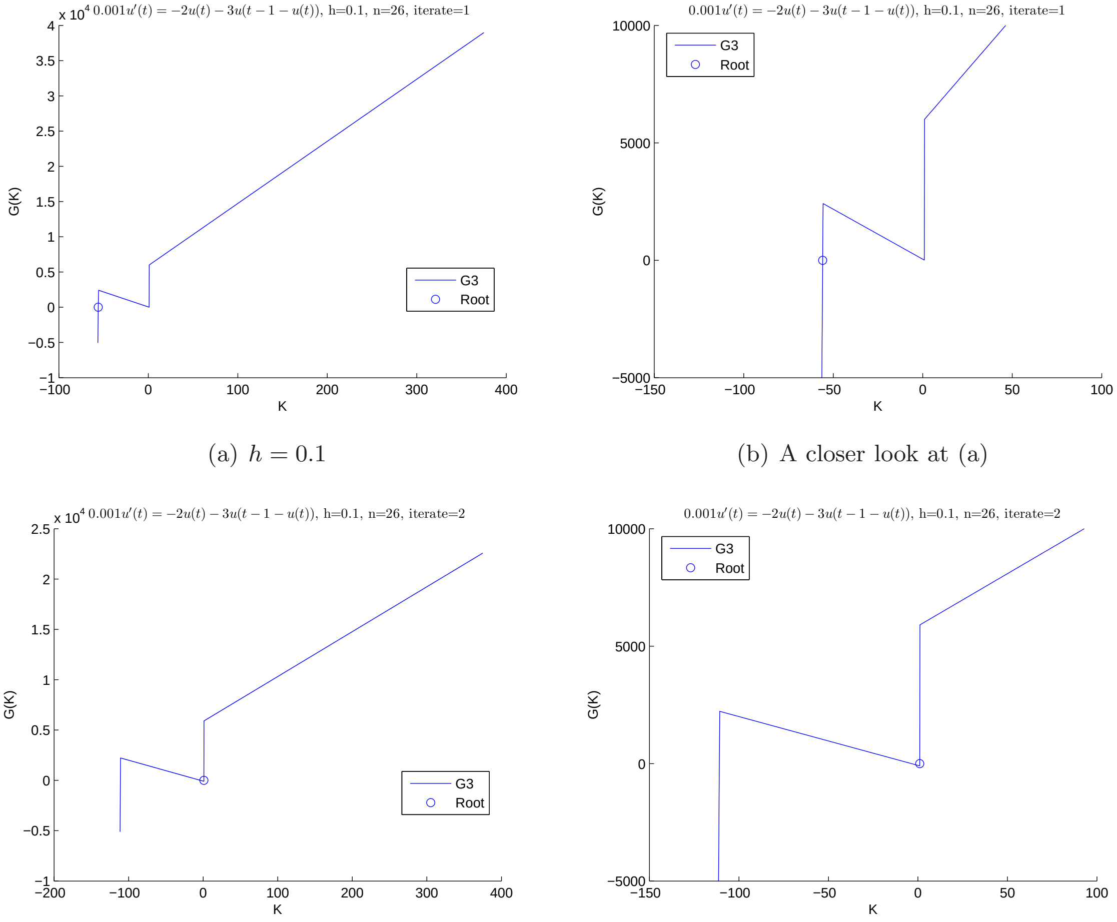
The stepsize h is kept within the interval $[h_{\text{left}}, h_{\text{right}}]$. Initially this interval is $[0, h_{\max}]$ but this changes as the iteration gives us more information on the bounds of the correct stepsize. If the stepsize found using (5.5.3) is outside a range of values $[h_{\text{left}}, h_{\text{right}}]$ then the new stepsize is found instead using a bisection step $h_{\text{new}} = \frac{h_{\text{left}}+h_{\text{right}}}{2}$. In Algorithm 3, if the interval becomes small enough ($|h_{\text{left}} - h_{\text{right}}| <$`TolX`) then the algorithm exits with $h = h_{\text{left}}$. The reason why we do this is illustrated in Figures 5–4 and 5–5. In Figure 5–4 there are multiple zeros and one disappears as we change the stepsize. In Figure 5–5 there is a nearly horizontal section of zeros. In both cases, a small change in the stepsize h will cause a small change in the function $G^{(i)}$ itself but it causes a large change in the value of the root. When there is a discontinuity point to be found then h_{left} and h_{right} can be very close together but they result in very different roots of the $G^{(i)}$ functions which results in very different updates. In the case of (5.0.1), it could be the difference between a trough point and a crest point. This is why we choose h_{left} as the stepsize. In these situations we will detect the discontinuity point again in the next time step. It makes sense to actually solve for the discontinuity point again now because this time we can solve for it with a smaller stepsize which results in much smoother $G^{(i)}$ functions and a better estimate for the discontinuity point. So far the code only finds the same discontinuity points a maximum of two times.

Other methods to find discontinuities

There are other methods to detect breaking points. Guglielmi and Hairer [24] point out that the methods for directly finding the discontinuity points (such as our method) may be expensive and some computed breaking points may not even be relevant for the actual computation. Instead, they presented an algorithm which activates the search for breaking points only when there has been a step rejection (when the local error is too large or the solver failed to converge to a value at the next step). Their method along with convergence results and proof are discussed in [24].

5.6 Test problems

A report by Paul [48] provides a collection of test scalar delay differential equations, some with their solutions. Chapter 1 of Bellen and Zennaro [8] also provides some test problems that have known stability properties. This section lists some of the problems that have been used to test our solver.



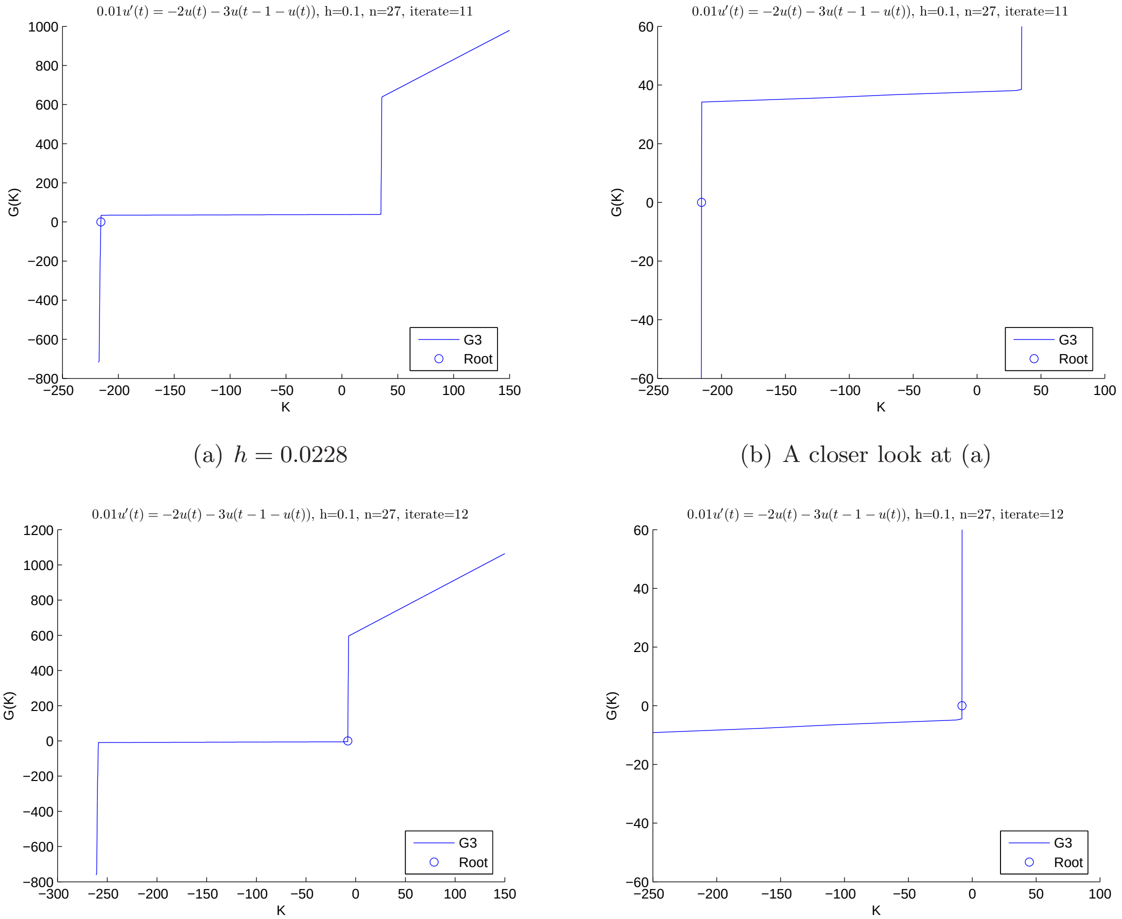
(a) $h = 0.1$

(b) A closer look at (a)

(c) $h = 0.05$

(d) A closer look at (c)

Figure 5–4: Plots of $G^{(3)}\left(K_{n+1}^{(3)}\right)$ displaying the problem when there are multiple roots and one disappears when the stepsize is changed. A small change in stepsize leads to a large change in the root. This plot is from applying SDIRK3 (I,III) to the DDE $0.001\dot{u}(t) = -2u(t) - 3u(t-1-u(t))$.



(a) $h = 0.0228$

(b) A closer look at (a)

(c) $h = 0.0227$

(d) A closer look at (c)

Figure 5–5: Plots of $G^{(3)}\left(K_{n+1}^{(3)}\right)$ displaying the problem when the function is close to flat near the root. A small change in stepsize leads to a large change in the root. This plot is from applying SDIRK3 (I,III) to the DDE $0.01\dot{u}(t) = -2u(t) - 3u(t-1-u(t))$.

1. Our main test problem is (5.0.1), the simplest test equation with state dependent delay. We used a constant history function for all the tests.

$$\begin{array}{ll} \varepsilon \dot{u}(t) = -\gamma u(t) - \sum_{i=1}^{N} \kappa_i u\left(t - a_i - c_i u(t)\right), & t \geqslant 0 \\ u(t) = 1, & t \leqslant 0 \end{array}$$

   As mentioned before, For small ε this is a stiff test problem.

2. The source for this test problem is Feldstein [18].

$$\begin{array}{ll} \dot{u}(t) = u\left(\frac{t}{(1+2t)^2}\right)^{(1+2t)^2}, & t \geqslant 0 \\ u(t) = 1, & t \leqslant 0 \end{array}$$

   The analytical solution of this is $u(t) = e^t$. There are no discontinuities in the solution. This is good for testing the overlapping cases. See Figure 5–6 for the order test using this test equation.

3. The source for this test problem is Neves and Feldstein [44].

$$\begin{array}{ll} \dot{u}(t) = \frac{u(t)u(\ln(u(t)))}{t}, & t \geqslant 1 \\ u(t) = 1, & t \leqslant 1 \end{array}$$

   The analytical solution of this is known for $t \in \left[1, e^2\right]$

$$u(t) = \begin{cases} t, & 1 \leqslant t \leqslant e \\ \exp\left(\frac{t}{e}\right), & e \leqslant t \leqslant e^2 \end{cases}$$

   This is good for testing discontinuity tracking. See Figure 5–7 for the order test using this test equation.

4. The source for this test problem is Neves and Feldstein [45].

$$\begin{array}{ll} \dot{u}(t) = \frac{\exp(u(u(t)-\ln(2)+1))}{t} & t \geqslant 1 \\ u(t) = 0, & t \leqslant 1 \end{array}$$

   The analytical solution of this on the interval $[1, 4]$ is

$$u(t) = \begin{cases} \ln(t), & 1 \leqslant t \leqslant 2 \\ \frac{1}{2} t \ln(2) - 1, & 2 \leqslant t \leqslant 4 \end{cases}$$

   This has discontinuities at $t = 1, 2, 4$ respectively.

5. The source for this test problem is Neves and Feldstein [45].

$$\begin{aligned} \dot{u}\left(t\right) &= \tfrac{u\left(u(t)-\sqrt{2}+1\right)}{2\sqrt{t}}, & t \geqslant 1 \\ u\left(t\right) &= 1, & t \leqslant 1 \end{aligned}$$

The analytical solution of this on the interval $[\xi_1, \xi_3]$ is

$$u\left(t\right) = \begin{cases} \sqrt{t}, & \xi_1 \leqslant t \leqslant \xi_2 \\ \frac{t}{4} + \frac{1}{2} + \left(1 - \frac{1}{\sqrt{2}}\right)\sqrt{2}, & \xi_2 \leqslant t \leqslant \xi_3 \end{cases}$$

where $\xi_1 = 1$, $\xi_2 = 2$ and $\xi_3 = 5.0294372515248$.

6. The source for this test problem is Tavernini [55].

$$\begin{aligned} &\dot{u}\left(t\right) = u\left(u\left(t\right)\right) + \left(3+a\right)t^{2+a} - t^{(3+a)^2} \quad 0 \leqslant t \leqslant 1 \\ &u\left(0\right) = 0, \end{aligned}$$

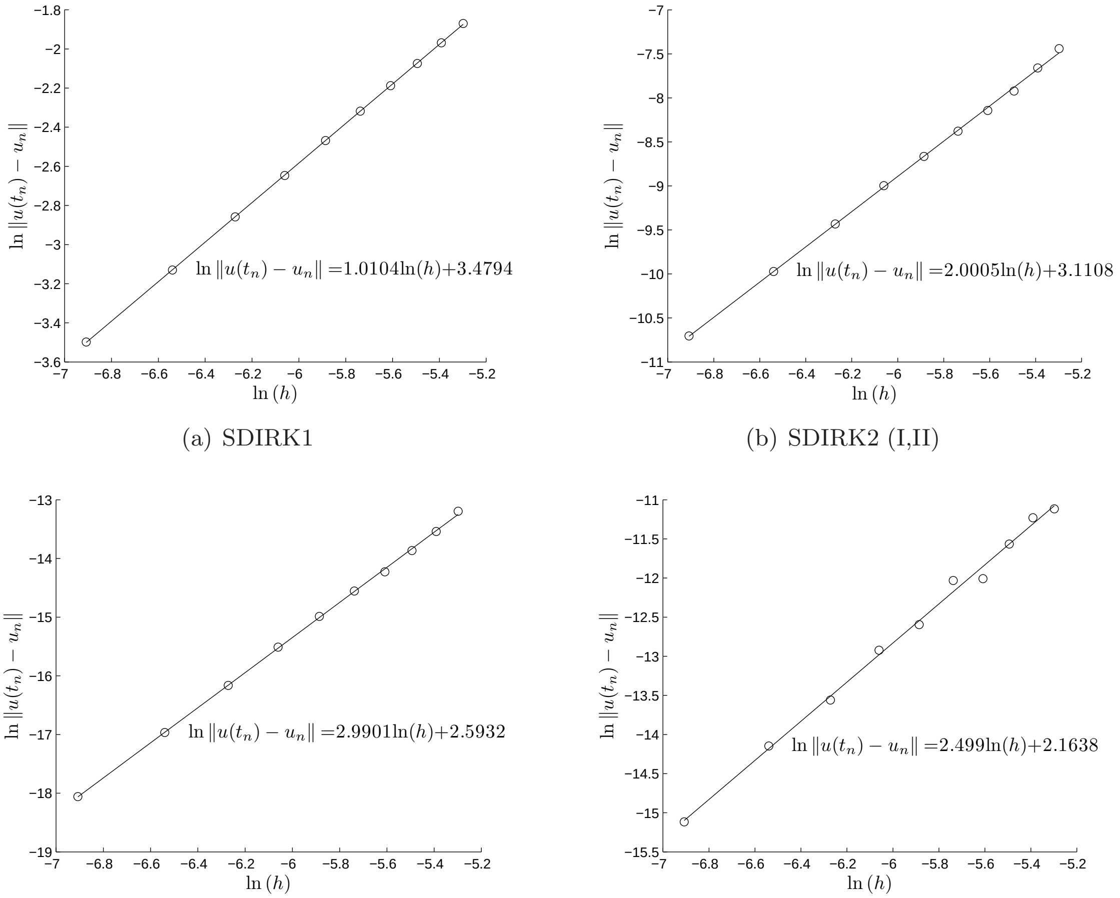
The analytical solution of this on the interval $[0, 1]$ is $u\left(t\right) = t^{3+a}$. The solution has no discontinuities and is a good test for overlapping.

5.7 Performance of the SDIRK solver

The work on the SDIRK solver is still ongoing. Here we state some results on the performance of the preliminary SDIRK solver. The schemes described in this chapter have been implemented successfully for all non-stiff test problems listed in Section 5.6. The order tests on test equations 2 and 3 are found in Figures 5–6 and 5–7 respectively.

For stiff problems, SDIRK1, SDIRK2 (I,III) and SDIRK3 (I, III) are fast and efficient as long as the discontinuity tracking is turned off. When used to integrate test equation 1, the methods are fast when using a stepsize of 0.1 even when $\varepsilon = 10^{-16}$. SDIRK4 is currently not recommended for use on stiff equations.

As expected, the methods take much longer when the discontinuity tracking is turned on. SDIRK1, SDIRK2 (I,III) and SDIRK3 (I, III) are still fairly fast when ε is about 100 times smaller than the stepsize. However, past this point the SDIRK3 becomes much slower. At the trough points of solutions to test equation 1 with small ε, the intervals have both discontinuity points and overlapping if the stepsize is not small enough. Because of this we cannot use the Newton method iteration for finding the correct stepsize to capture the discontinuity point in the mesh so we have to use bisection which is much slower.



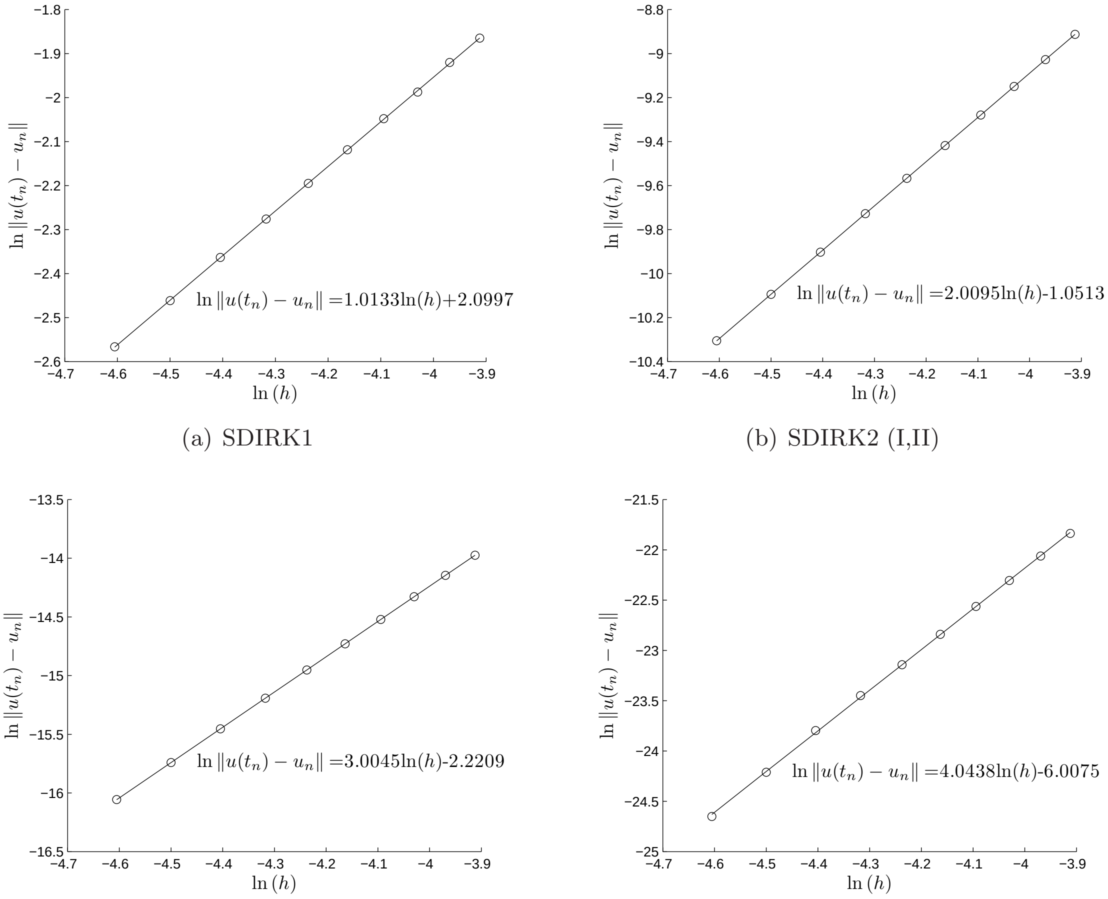
(a) SDIRK1

(b) SDIRK2 (I,II)

(c) SDIRK3 (I,II)

(d) SDIRK4 (I,II)

Figure 5–6: Plots of the error of the methods when applied to test equation 2 versus the stepsize in logarithmic scale. The order of the error is the slope of the regression line. SDIRK1, SDIRK2 (I,II), SDIRK3 (I,II) all exhibit close agreements with their respective expected orders of one, two and three. SDIRK4 (I,II) is expected to perform as a method of order two when there is overlapping, but for this problem it displays a slightly better order of 2.5. Notice that for this test equation, overlapping occurs when $t - \frac{t}{(1+2t)^2} < h$ so $4t^2\frac{1+t}{(1+2t)^2} < h$. From this we can derive that the length of overlapping interval is of order $h^{0.5}$. Thus the order of 2.5 for SDIRK4 (I,II) can be explained by carefully deriving the error term at the overlapping interval and realizing at these steps the error is the method order of h^2 multiplied by the $h^{0.5}$ because of the length of the integration interval.



(a) SDIRK1

(b) SDIRK2 (I,II)

(c) SDIRK3 (I,II)

(d) SDIRK4 (I,II)

Figure 5–7: Plots of the error of the methods when applied to test equation 3 versus the stepsize in logarithmic scale. The order of the error is the slope of the regression line. SDIRK1, SDIRK2 (I,II), SDIRK3 (I,II) and SDIRK4 (I,II) all exhibit close agreements with their respective expected orders of one, two, three and four. There is no overlapping for this test equation.

One difficulty with the current format of the solver is that to solve for a stage $K_{n+1}^{(i)}$, an interval $[L, M]$ such that $G^{(i)}\left(K_{n+1}^{(i)}\right)$ has different signs at the two endpoints is required. For non-stiff problems, it is usually not hard to pick such an interval but for stiff problems this is a problem since $\left|K_{n+1}^{(i)}\right|$ can be very large. As mentioned, one of these bounds may be found using (5.4.5). For test equation 1, this requirement yields (5.4.6). Finding such a bound would be difficult to automate for general problems.

CHAPTER 6
Conclusions

Here we summarise the main contributions of this thesis and discuss areas of future research. In the first chapter we gave an introduction to retarded functional differential equations, a general class of equations that includes DDEs with discrete delays, distributed delays and neutral equation. However RFDE theory cannot be automatically applied to state dependent DDEs even when the delay is known to be bounded. Conditions on RFDEs are difficult to transcribe into conditions on state dependent DDEs. A separate treatment of state dependent DDEs is needed and a good starting point is the following model equation

$$\varepsilon\dot{u}\left(t\right)=\mu u\left(t\right)+\sigma u\left(t-a-cu\left(t\right)\right). \tag{6.0.1}$$

This is the simplest DDE with a state dependent delay. The state dependence is the only nonlinearity in this equation and from it interesting dynamics arise that cannot be found in the constant delay case ($c=0$), much less the ODE case ($\sigma=0$) [38, 39, 40, 41]. This model problem is a prototype for other state dependent DDE problems when the solutions are close to zero. In this thesis we also consider the generalisation of (6.0.1) to N delays,

$$\varepsilon\dot{u}\left(t\right)=-\gamma u\left(t\right)-\sum_{i=1}^{N}\kappa_i u\left(t-a_i-c_i u\left(t\right)\right). \tag{6.0.2}$$

Properties of these model equations were considered in Chapter 2 starting with the existence and uniqueness of solutions. Global existence and uniqueness results were obtained for some parameter regions. These results are summarized in Theorem 2.1.4 for (6.0.1), and Theorem 2.2.2 for (6.0.2). The conditions for which the deviated arguments $\alpha_i\left(t,u(t)\right)$ must eventually become monotonically increasing were also considered. The results are found in Theorem 2.1.8 for (6.0.1), and Theorem 2.2.4 for (6.0.2). Finally, bounds on the solutions to special cases of (6.0.1) and (6.0.2) were derived in Section 2.3 using a Gronwall argument.

In Chapter 3 we considered the stability of the zero solution to (6.0.1). For fixed ε, a and c, the parameter region in the (μ,σ) plane in which the zero solution is stable is known to be the same for both the constant delay ($c=0$) and state dependent delay ($c\neq0$) cases [25]. This region is denoted by $\Sigma_\star=\overset{\Delta}{\Sigma}\cup\overset{w}{\Sigma}\cup\overset{c}{\Sigma}$ and described in Definition 3.1.1. Different approaches were used to directly prove stability in parts of this analytic region for the state dependent case. The first approach, summed up in Theorem 3.2.6, uses a Gronwall argument to prove that if $(\mu, \sigma) \in \{r(0) \in (0,1)\}$ (refer to Definition 3.2.4) then the zero solution to (6.0.1) is asymptotically stable. The second approach, culminating in Theorem 3.4.13, uses a Razumikhin-style argument to prove that for $k \geqslant 2$, if $(\mu, \sigma) \in \{P(1,0,k) < 1\}$ (refer to Definition 3.4.11) then the zero solution to (6.0.1) is Lyapunov stable. This second approach is a generalisation of the work of Barnea [6] who considered the $\mu = c = 0$ and $k = 2$ case. The parameter regions in which stability is proven by these methods are shown in Figure 3–13 and compared in Section 3.5. These regions include the delay independent stability region $\overset{\Delta}{\Sigma}$ and a significant portion of the delay dependent stability region $\overset{w}{\Sigma} \cup \overset{c}{\Sigma}$.

The Gronwall argument used to prove stability in Chapter 3 is the same one used in Section 2.3 to derive bounds on the solutions to special cases of (6.0.1) and (6.0.2). In future work I would like to use the Razumikhin-style argument to find stricter bounds on the solutions.

The direct techniques used in Chapter 3 were extended in Chapter 4 to find parameter regions for which the backward Euler solution to (6.0.1) is stable. The results for the Gronwall argument and the Razumikhin-style argument are given in Theorems 4.4.6 and 4.5.15. Using the Razumikhin-style argument with $k = 2$, we showed that backward Euler is stable when $(\mu, \sigma) \in \{P(1,0,2) < 1\}$ for all stepsizes $h \in (0, a]$. These results were then extended to derive analytic expressions of stepsize-dependent regions for which general Θ methods applied to (6.0.1) are stable. These expressions were evaluated numerically and sample plots are shown in Figure 4–9.

Eventually I would like to extend the results on the stability of the zero solution to (6.0.1) using the Razumikhin-style technique so that more of $\overset{w}{\Sigma}$ is included. One idea on how to do this is to use the decaying oscillations that undergo several cycles over a time period a that are displayed by solutions to (6.0.1) when $(\mu, \sigma) \in \overset{w}{\Sigma}$. Future improvements on the arguments applied to the DDEs can possibly be extended to improve the results on numerical stability.

The stability of backward Euler solution to (6.0.1) was also considered when the stepsize $h > a$. In Section 4.6 it was shown that if $h > a$ then for all $(\mu, \sigma) \in \Sigma_\star$ there is a BE solution that converges to zero. The BE solution may not be unique so there may be another BE solution that is behaving in an entirely different manner. This is another thing I would like to investigate further.

In the last chapter a new scheme for numerically integrating scalar DDEs with multiple state dependent delays was presented. This scheme is based on singularly diagonally implicit Runge-Kutta (SDIRK) methods. New continuous extensions, called DIRK-type continuous extensions are chosen to accompany the SDIRK scheme. These continuous extensions consist of piecewise polynomial weight functions, unlike the usual polynomial weight functions in a continuous RK method, and they are derived to maintain the SDIRK structure even when applied to a DDE problems with the possibility of overlapping. Four SDIRK schemes (of orders one through four) were implemented and found to be successful in maintaining their order using appropriate DIRK-type continuous extensions. The methods were also tested against (6.0.1) and (6.0.2) which are stiff problems when ε is very small. The lower order methods were found to be very successful in performing this integration for very small ε when the discontinuity tracking is turned off. The solver is still currently at its preliminary stages and more work is necessary to improve the convergence of the method when the discontinuity tracking is turned on. One option to do this would be to implement stepsize control using a lower order method.

References

[1] A.N. Al-Mutib. Stability properties of numerical methods for solving delay differential equations. *Journal of Computational and Applied Mathematics*, 10:71–79, 1984.

[2] R. Alexander. Diagonally implicit Runge-Kutta methods for stiff O.D.E.'s. *SIAM Journal of Numerical Analysis*, 14(6):1006–1021, 1977.

[3] Hbid M.L. Arino, O. and E. Ait Dads. *Delay Differential Equations and Applications.* NATO Science Series. Springer, Dordrecht, 2006.

[4] C.T.H. Baker and C.A.H. Paul. Computing stability regions - Runge-Kutta methods for delay differential equations. *IMA Journal of Numerical Analysis*, 14:347–362, 1994.

[5] C.T.H. Baker, C.A.H. Paul, and D.R. Willé. Issues in the numerical solution of evolutionary delay differential equations. *Advances in Computational Mathematics*, 3:171–196, 1995.

[6] D.I. Barnea. A method and new results for stability and instability of autonomous functional differential equations. *SIAM J. Appl. Math*, 17(4):681–697, 1969.

[7] V.K. Barwell. Special stability problems for functional differential equations. *BIT Numerical Mathematics*, 15.

[8] A. Bellen and M. Zennaro. *Numerical Methods for Delay Differential Equations.* Numerical Mathematics and Scientific Computation. Oxford Science Publications, New York, 2003.

[9] A. Bellen, M. Zennaro, S. Maset, and N. Gulglielmi. Recent trends in the numerical solution of retarded functional differential equations. *Acta Numerica*, 18:1–110, 2009.

[10] R.E. Bellman and K.L. Cooke. *Differential-Difference Equations.* Academic Press, New York, 1963.

[11] R.P. Brent. *Algorithms for Minimization Without Derivatives.* Prentice Hall, New Jersey, 1973.

[12] M. Calvo and T. Grande. On the asymptotic stability of θ-methods for delay differential equations. *Numerische Mathematik*, 54:257–269, 1988.

[13] S.P. Corwin, D. Sarafyan, and S. Thompson. Dklag6: a code based on continuously imbedded sixth-order Runge-Kutta methods for the solution of state-dependent functional differential equations. *Applied Numerical Mathematics*, 24:319–330, 1997.

[14] R.D. Driver. Existence theory for a delay-differential system. *Contributions to Differential Equations*, 1:317–336, 1963.

[15] L.E. El'sgol'ts and S.B. Norkin. *Introduction to the Theory and Application of Differential Equations with Deviating Arguments.* Academic Press, New York, 1973.

[16] W.H. Enright and H. Hayashi. A delay differential equation solver based on a continuous Runge-Kutta method with defect control. *Numerical Algorithms*, 16:349–364, 1997.

[17] T. Erneux. *Applied Delay Differential Equations*. Surveys and Tutorials in the Applied Mathematical Sciences. Springer, New York, 2009.

[18] M.A. Feldstein. *Discretization methods for retarded ordinary differential equations*. PhD thesis, UCLA Mathematics Department, 1964.

[19] G.E. Fosythe, M.A. Malcolm, and C.B. Moler. *Computer Methods for Mathematical Computations*. Prentice-Hall Series in Automatic Computation. Prentice Hall, New Jersey, 1977.

[20] N. Guglielmi. On the asymptotic stability properties for Runge-Kutta methods for delay differential equations. *Numerische Mathematik*, 77:467–485, 1997.

[21] N. Guglielmi. Delay dependent stability regions of θ-methods for delay differential equations. *IMA Journal of Numerical Analysis*, 18:399–418, 1998.

[22] N. Guglielmi and E. Hairer. Order stars and stability for delay differential equations. *Numerische Mathematik*, 83:371–383, 1999.

[23] N. Guglielmi and E. Hairer. Implementing Radau IIA methods for stiff delay differential equations. *Computing*, 67:1–12, 2001.

[24] N. Guglielmi and E. Hairer. Computing breaking points in implicit delay differential equations. *Advances in Computational Mathematics*, 29(3):229–247, 2007.

[25] I. Györi and F. Hartung. Exponential stability of a state-dependent delay system. *Discrete and Continuous Dynamicals Systems - Series A*, 18:4:773–791, 2007.

[26] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II Stiff and Differential-Algebraic Problems*, volume 14 of *Springer Series in Compuational Mathematics*. Springer-Verlag, Berlin Heidelberg, 2 edition, 2002.

[27] A. Halanay. *Differential equations: stability, oscillations, time lags*, volume 23 of *Mathematics in Science and Engineering*. Academic Press Inc., New York, 1966.

[28] J. Hale. *Theory of Functional Differential Equations*, volume 99 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1993.

[29] J. Hale and S.M. Verduyn Lunel. *Introduction to Functional Differential Equations*, volume 99 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1993.

[30] F. Hartung, T. Krisztin, H.-O. Walther, and J. Wu. Functional differential equations with state-dependent delays: theory and applications. In A. Cañada, P. Drábek, and A. Fonda, editors, *Handbook of Differential Equations: Ordinary Differential Equations*, volume 3, pages 435–545. 2006.

[31] G. Hutchinson. Circular causal systems in ecology. *Annals of the New York Academy of Sciences*, 50:221–246, 1948.

[32] A. Ivanov, E. Liz, and S. Trofimchuk. Halanay inequality, Yorke 3/2 stability criterion, and differential equations with maxima. *Tohoku Math. J.*, 54:277–295, 2002.

[33] V. Kolmanovskii and A. Myshkis. *Applied theory of Functional Differential equations.* Kluwer Academic Publishers, Dordrecht, 1992.

[34] N.N. Krasovskii. *Stability of Motion Applications of Lyapunov's Second Method to Differential Equations with Delay.* Stanford University Press, Stanford, California, 1963.

[35] T. Krisztin. Stability for functional differential equations and some variational problems. *Tohoku Math. J.*, 42:402–417, 1990.

[36] Y. Kuang. *Delay Differential Equations with Applications in Population Dynamics.* Academic Press, Boston, 1993.

[37] M.Z. Liu and M.N. Spijker. The stability of the θ-methods in the numerical solution of delay differential equations. *IMA Journal of Numerical Analysis*, 10:31–48, 1990.

[38] J. Mallet-Paret and R.D. Nussbaum. Boundary layer phenomena for differential-delay equations with state-dependent time lags, I. *Archive for Rational Mechanics and Analysis*, 120:99–146, 1992.

[39] J. Mallet-Paret and R.D. Nussbaum. Boundary layer phenomena for differential-delay equations with state-dependent time lags: II. *Journal für die reine und angewandte Mathematik*, 477:129–197, 1996.

[40] J. Mallet-Paret and R.D. Nussbaum. Boundary layer phenomena for differential-delay equations with state-dependent time lags: III. *Discrete and Continuous Dynamicals Systems - Series A*, 189:640–692, 2003.

[41] J. Mallet-Paret and R.D. Nussbaum. Superstability and rigorous asymptotics in singularly perturbed state-dependent delay-differetnial equations. *Journal of Differential Equations*, 250:4037–4084, 2011.

[42] J. Mallet-Paret, R.D. Nussbaum, and P. Paraskevopoulos. Periodic solutions for functional differential equations with multiple state-dependent time lags. *Topological Methods in Nonlinear Analysis*, 3:101–162, 1994.

[43] S. Maset. Stability of Runge-Kutta methods for linear delay differential equations. *Numerische Mathematik*, 87:355–371, 2000.

[44] K.W. Neves and A. Feldstein. Characterization of jump discontinuities for state dependent delay differential equations. *Journal of Mathematical Analysis and Applications*, 56(3):689–707, 1976.

[45] K.W. Neves and A. Feldstein. High-order methods for state-dependent delay differential equations with non-smooth solutions. *SIAM Journal of Numerical Analysis*, 21(5):844–863, 1984.

[46] G. Orosz, R.E. Wilson, and G. Stépán. Traffic jams: dynamics and control. *Philosphical Transactions of the Royal Society A*, 368:4455–4479, 2010.

[47] B. Owren and H.H. Simonsen. Alternative integration methods for problems in structural dynamics. *Computer Methods in Applied Mechanics and Engineering*, 122:1–10, 1995.

[48] C.A.H. Paul. A test set of functional differential equations. Manchester center for computational mathematics numerical analysis report 243, Department of Mathematics, University of Manchester, February 1994.

[49] C.A.H. Paul. A user-guide to ARCHI: An explicit Runge-Kutta code for solving delay and neutral differential equations and parameter estimation problems. Manchester Center for Computational Mathematics Numerical Analysis Report 283, Department of Mathematics, University of Manchester, April 1997.

[50] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C, The Art of Scientific Computing.* Cambridge University Press, Cambridge, second edition edition, 1999.

[51] B.S. Razumikhin. An application of Lyapunov method to a problem on the stability of systems with a lag. *Automation and Remote Control*, 21:740–748, 1960.

[52] L.F. Shampine. Solving ODEs and DDEs with residual control. *Applied Numerical Mathematics*, 52:113–127, 2005.

[53] L.F. Shampine and S. Thompson. Solving DDEs in matlab. *Applied Numerical Mathematics*, 37:441–458, 2001.

[54] H. Smith. *An Introduction to Delay Differential Equations with Applications to the Life Sciences.* Texts in Applied Mathematics. Springer, New York, 2011.

[55] L. Tavernini. The approximate solution of Volterra differential systems with state-dependent time lags. *SIAM Journal of Numerical Analysis*, 15(5):1039–1052, 1978.

[56] S. Thompson and L.F. Shampine. A friendly FORTRAN DDE solver. *Applied Numerical Mathematics*, 56:503–516, 2006.

[57] L. Torelli. Stability of numerical methods for delay differential equations. *Journal of Computational and Applied Mathematics*, 25:15–26, 1989.

[58] H.-O. Walther. The solution manifold and C^1-smoothness for differential equations with state dependent delay. *Journal of differential equations*, 195:46–65, 2003.

[59] M. Wiederholt. Stability of multistep methods for delay differential equations. *Mathematics of Computation*, 30:283–290, 1976.

[60] E.M. Wright. A non-linear difference-differential equation. *Journal für die reine und angewandte Mathematik*, 194:66–87, 1955.

[61] M. Zennaro. P-stability properties of Runge-Kutta methods for delay differential equations. *Numerische Mathematik*, 49:305–318, 1986.