CONVERGENCE OF THE VARIABLE-STEP
VARIABLE-ORDER 3-STAGE
HERMITE–BIRKHOFF ODE/DDE SOLVER
OF ORDER 5 TO 15

H. Yagoub, T. Nguyen-Ba, T. Giordano & R. Vaillancourt

July 25, 2009

To appear in Scientific Proceedings of Riga Technical University, Series 5,
Computer Science

Abstract

The ordinary and delay differential equations (ODEs/DDEs) solver HB515DDE is based on the discrete hybrid variable-step variable-order 3-stage Hermite–Birkhoff ODE solver of (consistency) order 5 to 15. The current version of the solver can handle ODEs and DDEs with state-dependent, non-vanishing, small, vanishing and asymptotically vanishing delays. Delayed values are computed using Hermite interpolation and small delays are dealt with using extrapolation. Discontinuities in DDEs are located by a bisection method. HB515DDE has proved itself superior to many other known DDE solver as Matlab’s ddesd. This article presents the theory behind HB515DDE including convergence proofs for the ODE and DDE parts separately. It is shown that the method is convergent of order 5 when seen as an ODE method and of order 3 when seen as a DDE solver under some assumptions.

Résumé

Le solveur d’équations différentielles avec ou sans retard, HB515DDE, provient d’un solveur hybride à pas et ordre variables du type Hermite–Birkhoff à 3 étages d’ordre 5 à 15. HB515DDE peut résoudre des ODEs et DDEs avec retards spatio-variables, infiniment petits ou non, et asymptotiquement infiniment petits. On calcule les retards au moyen d’interpolation d’Hermite. HB515DDE est supérieur à plusieurs solveurs de DDEs connus, tel ddesd de Matlab. On présente la théorie de HB515DDE et l’on démontre sa convergence d’ordre 5 pour les ODEs et 3 pour les DDEs.

Keywords: ODE/DDE solver, convergence, state dependent and small/vanishing delays, Hermite–Birkhoff method
1 INTRODUCTION

In [10], the authors built the variable-step variable-order 3-stage Hermite–Birkhoff ODE solver of (consistency) order 5 to 15 (HB(5-15)3 or HB515 for short). We also provided the absolute stability regions for the different consistency orders. Later, we transformed the ODE solver HB515 to handle delay differential equations (DDEs). We called it HB515DDE and presented its main features along with numerical results in [13].

We mention that a similar transformation was done on the variable-step constant-order 7-stage Hermite–Birkhoff–Taylor ODE solver HBT(8)7 of order 8 which was transformed into the DDE solver HBT8DDE presented in [15] and on the variable-step variable-order 3-stage Hermite–Birkhoff–Obrechkoff ODE solver HBO(4-14)3 with order varying between 4 and 14 constructed in [11] and the resulting DDE solver HBO414DDE was exhibited in [14].

In this article, we prove that the ODE solver HB515 is convergent of order 5 and the DDE solver HB515DDE is convergent of order 3 under some assumptions. We will review some important properties of the ODE/DDE solver HB515DDE but knowledge of [10], [13] and of the properties of ODEs and DDEs is assumed.

2 REVIEW OF THE MAIN PROPERTIES OF HB515DDE

When it is started, our ODE/DDE solver calls the Dormand–Prince DP(4,5)7M (or DP45 for short) to compute at least three initial steps after which HB515/HB515DDE takes over. It then integrates with (consistency) order 5 and advances using three of the backsteps computed by DP45. The acceptance of a step depends on one error estimator. After each accepted step, three extra error estimators are computed (with no extra function evaluations) to decide whether the order should be decreased (if it is greater than 5), increased (if it is lower than 15) or kept unchanged. The main procedures for stepsize selection and order variation were adapted from the work of Shampine and Gordon in [12].

In [13], we explained the main steps of the transformation of HB515 into HB515DDE. Firstly, because HB515 is discrete, we added a Hermite interpolant to compute delayed values. In the case of small delays, we chose to use extrapolation for simplicity. Moreover, to deal with some asymptotically vanishing delays, we provided an optional method that was adapted from a procedure constructed and justified in [9] (see [13] for more details). Because it is an optional method, we chose not to consider it in our proofs below. Secondly, we followed the idea of Enright and Hayashi in [5] to suspect the presence of a discontinuity only after a step was rejected. Moreover, we chose to use a bisection method for the location of the discontinuity. Finally, even though asking the solver to satisfy both the absolute and relative tolerances for each step seemed costly, it actually was numerically favorable so we kept it as our error control for step acceptance.
Numerical tests provided in [10] showed that the ODE solver HB515 is most efficient at stringent tolerances for problems where the function \( y' = f(t, y) \) is expensive to evaluate as in the cubic wave problem where it was superior to DP(8,7)13M. In [13], we exhibited numerical results for seven test problems where HB515DDE was compared with other DDE solvers like SYSDEL and Matlab’s \texttt{ddesd}. For all test problems HB515DDE presented the best number of function evaluations over maximum relative error ratio over all tested solvers which include Matlab’s \texttt{ddesd} or Karoui and Vaillancourt’s SYSDEL [8], [9]. The remarkable performance of HB515DDE can be explained by its variable-order and combined Runge–Kutta and multistep structure which distinguish it from many DDE solvers. These properties give our method a great amount of freedom and adaptation ability to the different behaviors of DDEs.

Hence, our goal now is to show that these encouraging experimental results stand on firm theoretical foundations. We begin with the convergence of the ODE solver HB515. We note that, in [10], we constructed the ODE solver HB515 and we provided the absolute stability regions for the different consistency orders. Still, we did not prove its convergence. Hence, we first prove that the ODE solver HB515 is convergent of order 5 and then we show that the DDE solver HB515DDE is convergent of order 3 under some assumptions. Note that HB515DDE can also solve ODEs using the code of HB515 so we will use the names HB515 and HB515DDE interchangeably when speaking about the ODE solver.

3 CONVERGENCE OF HB515

Following the notation of [2], we want to prove the convergence of the ODE solver HB515 when applied to the initial value problem

\[
y'(t) = f(t, y(t)), \quad t_0 \leq t \leq t_f, \quad y(t_0) = y_0,
\]

where \( f(t, y) \in C^0([t_0, t_f] \times \mathbb{R}^d, \mathbb{R}^d) \) is globally Lipschitz continuous with respect to \( y \) in a given norm \( \| \cdot \| \) of \( \mathbb{R}^d \), i.e. there exists some \( L > 0 \) such that

\[
\| f(t, y_1) - f(t, y_2) \| \leq L \| y_1 - y_2 \| \quad \forall t \in [t_0, t_f] \quad \text{and} \quad \forall y_1, y_2 \in \mathbb{R}^d.
\]

Let \( \Delta = \{t_0, \ldots, t_N = t_f\} \) be a mesh and let the stepsize \( h_{n+1} \) be given by \( h_{n+1} = t_{n+1} - t_n \) for \( n = 0, \ldots, N - 1 \). Next, we define the general \( k \)-step method

\[
y_{n+1} = \alpha_{n,1} y_n + \cdots + \alpha_{n,k} y_{n-(k-1)} + h_{n+1} \Phi(y_n, \ldots, y_{n-(k-1)}; f, \Delta_n),
\]

with \( n \geq k - 1, \Delta_n = \{t_{n-(k-1)}, \ldots, t_n, t_{n+1}\} \) and where the increment function \( \Phi \) satisfies a global Lipschitz condition with respect to the \( y \) arguments.

Convergence theory for the general method (3) used as an ODE solver or a DDE solver was developed in [2] so we will first make sure that we satisfy the conditions of the theorems in [2] before we prove the convergence of our method.
The first step is to rewrite, for \( p \in \{5, \ldots, 15\} \), the \((p-3)\)-step \( HB_p \) method in the form of equation (3). Hence, let \( p \in \{5, \ldots, 15\} \) and \( n \geq p - 4 \). We get

\[
y_{n+1} = \alpha_{n,1} y_n + \alpha_{n,2} y_{n-1} + \cdots + \alpha_{n,p-3} y_{n-(p-4)} + h_{n+1} \Phi(y_n, \ldots, y_{n-(p-4)}; f, \Delta_n)
\]

where \( n \geq p - 4 \) and \( \Delta_n = \{t_{n-(p-4)}, \ldots, t_n, t_{n+1}\} \). In the notation of [10], \( \alpha_{n,i} = 0 \) for all \( i \in \{3, \ldots, p-3\} \), \( \alpha_{n,1} \) and \( \alpha_{n,2} \) correspond to \( \alpha_{10} \) and \( \alpha_{11} \), respectively, and \( \Phi \) is given by

\[
\Phi(y_n, \ldots, y_{n-(p-4)}; f, \Delta_n) = b_{11} f_n + b_{12} F_{n+c_2} + b_{13} F_{n+c_3} + \sum_{j=1}^{p-4} \beta_{1j} f_{n-j}
\]

where

\[
F_{n+c_2} = f \left( \alpha_{20} y_n + \alpha_{21} y_{n-1} + h_{n+1} \left( a_{21} f_n + \sum_{i=1}^{p-4} \beta_{2j} f_{n-j} \right) \right)
\]

and

\[
F_{n+c_3} = f \left( \alpha_{30} y_n + \alpha_{31} y_{n-1} + h_{n+1} \left( a_{31} f_n + a_{32} F_{n+c_2} + \sum_{i=1}^{p-4} \beta_{3j} f_{n-j} \right) \right)
\]

with \( f_i = f(t_i, u) \) for all \( i = n - (p-4), \ldots, n \).

Secondly, the 1-step 7-stage \( DP_{45} \) can also be written in the form of equation (3) as follows. Let \( n \geq 0 \) be a step index then

\[
y_{n+1} = y_n + h_{n+1} \Phi'(y_n; f, h_{n+1})
\]

where \( \Phi' \) is the increment function for \( DP_{45} \) for which the coefficients are given in [4] (in the cited reference, \( DP_{45} \) is called \( RK5(4)7 \)).

### 3.1 CONSTANT-STEP AND VARIABLE-ORDER

We will now deeply analyze the convergence of \( HB_{515} \) in the constant-step and variable-order case (CSVO). Firstly, we need to prove that the increment function \( \Phi \) of (4) satisfied a Lipschitz condition with respect to the \( y \) argument. To do that, we must bound the coefficients of our integration method.

#### 3.1.1 BOUNDEDNESS OF THE COEFFICIENTS OF HB515 AND DP45

Since we are considering CSVO, it is possible to find explicit values for the coefficients of the predictors \( P_2, P_3 \), the integration formula \( IF \) and the step control predictor \( P_4 \). Indeed, these coefficients are independent of \( n \) (in CSVO) but depend on the integration order.

Hence, we solved, for each order, the systems \( M^1 u^1 = r^1 \), \( M^2 u^2 = r^2 \), \( M^3 u^3 = r^3 \) and \( M^4 u^4 = r^4 \) defined in [10] (note that since we are considering
constant stepsize, we have \( h_{n+1}/h_n = 1 \) for all \( n \) and hence \( \eta_i = 1 - i \) for \( i \in \{2, \ldots, 12\} \). See [10]). Next, we computed the maximum of the absolute value of each coefficient over all orders \( p \in \{5, \ldots, 15\} \) and these bounds are given in Tables 1 and 2. Hence, it is trivial to find a uniform bound for all those coefficients and such a bound is \( K_c = 185 \).

For DP45, it is easy to see that 15 is an upper bound for all the coefficients appearing in its Butcher Tableau. Now, it is possible to deduce the required Lipschitz condition on the increment function \( \Phi \).

**Table 1: Coefficients of \( P_2 \) and \( P_4 \).**

<table>
<thead>
<tr>
<th>Pred.</th>
<th>( P_2 )</th>
<th>Max Value</th>
<th>Pred.</th>
<th>( P_4 )</th>
<th>Max Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha_{20} )</td>
<td>7.373789131192e+00</td>
<td>( \alpha_{41} )</td>
<td>1.49790779087e+00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha_{21} )</td>
<td>1.267725442851e+01</td>
<td>( \alpha_{43} )</td>
<td>1.407796044440e+00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha_{21} )</td>
<td>5.354154046659e+00</td>
<td>( \beta_{41} )</td>
<td>2.529842718423e+00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_{21} )</td>
<td>1.229323234583e+01</td>
<td>( \beta_{42} )</td>
<td>5.148749023996e+00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_{22} )</td>
<td>1.010198025735e+01</td>
<td>( \beta_{43} )</td>
<td>8.353585022761e+00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_{23} )</td>
<td>1.304142129637e+01</td>
<td>( \beta_{44} )</td>
<td>1.045591752561e+01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_{24} )</td>
<td>1.455088215848e+01</td>
<td>( \beta_{45} )</td>
<td>1.001944562465e+01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_{25} )</td>
<td>1.300209462918e+01</td>
<td>( \beta_{46} )</td>
<td>7.287053301238e+00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_{26} )</td>
<td>9.020822150849e+00</td>
<td>( \beta_{47} )</td>
<td>3.955385239997e+00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_{27} )</td>
<td>7.619237643929e+00</td>
<td>( \beta_{48} )</td>
<td>1.553646762142e+00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_{28} )</td>
<td>8.665078399532e+00</td>
<td>( \beta_{49} )</td>
<td>4.175722252213e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_{29} )</td>
<td>9.69174891211e+00</td>
<td>( \beta_{410} )</td>
<td>6.876437545209e-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_{210} )</td>
<td>1.070152716252e+01</td>
<td>( \beta_{411} )</td>
<td>8.333333333333e-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_{211} )</td>
<td>1.169621195316e+01</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**3.1.2 Lipschitz Condition with Respect to \( y \) for HB515 and DP45**

We establish Lipschitz conditions methods used in their paper. Let \( h \) be a stepsize. In the following arguments we will use \( h_{n+1} \) and \( h \) interchangeably to designate the stepsize of the integration interval \([t_n, t_{n+1}]\) because the argument will generalize to the variable-step case that we will discuss later.

Let \( \Delta = \{t_0, \ldots, t_N = t_f\} \) be a mesh and let \( y_i, \tilde{y}_i \in \mathbb{R}^d \) be approximations to \( y(t_i) \) for \( i = n - p + 4, \ldots, n \). Put \( \tilde{f}_i = f(t_i, \tilde{y}_i), i = n - p + 4, \ldots, n \). In the same way, \( \tilde{F}_{n+c_2}, \tilde{F}_{n+c_3} \) and \( \tilde{f}_{n+1} \) are given by the same formulas as \( F_{n+c_2}, F_{n+c_3} \) and \( f_{n+1} \), respectively, with \( y_i \) being replaced by \( \tilde{y}_i \), \( i = n - p + 4, \ldots, n \). Finally, let \( \|\tilde{y} - y\| = \max\{\|\tilde{y}_n - y_n\|, \ldots, \|\tilde{y}_{n-p+4} - y_{n-p+4}\|\} \) where \( \|\cdot\| \) is the
Table 2: Coefficients of IF and P₃.

<table>
<thead>
<tr>
<th>IF</th>
<th>Max Value</th>
<th>Pred. P₃</th>
<th>Max Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>α₁₀</td>
<td>8.573388203018e-01</td>
<td>α₃₀</td>
<td>8.362133720521e+01</td>
</tr>
<tr>
<td>α₁₁</td>
<td>8.604881900714e-01</td>
<td>α₃₁</td>
<td>1.360367556699e+02</td>
</tr>
<tr>
<td>b₁₁</td>
<td>1.21790396358e+00</td>
<td>a₃₁</td>
<td>5.001681866079e+01</td>
</tr>
<tr>
<td>b₁₂</td>
<td>3.703703703704e-01</td>
<td>a₃₂</td>
<td>1.407796044440e+00</td>
</tr>
<tr>
<td>b₁₃</td>
<td>5.715592135345e-02</td>
<td>β₃₁</td>
<td>1.398225448007e+02</td>
</tr>
<tr>
<td>β₁₁</td>
<td>4.888761867360e-01</td>
<td>β₃₂</td>
<td>1.230691444360e+02</td>
</tr>
<tr>
<td>β₁₂</td>
<td>1.33175570059e-01</td>
<td>β₃₃</td>
<td>1.627609277082e+02</td>
</tr>
<tr>
<td>β₁₃</td>
<td>9.019177097007e-02</td>
<td>β₃₄</td>
<td>1.840209095766e+02</td>
</tr>
<tr>
<td>β₁₄</td>
<td>6.209491672050e-02</td>
<td>β₃₅</td>
<td>1.65823452998e+02</td>
</tr>
<tr>
<td>β₁₅</td>
<td>3.766439117974e-02</td>
<td>β₃₆</td>
<td>1.157201905028e+02</td>
</tr>
<tr>
<td>β₁₆</td>
<td>1.890113756238e-02</td>
<td>β₃₇</td>
<td>6.547435896226e+01</td>
</tr>
<tr>
<td>β₁₇</td>
<td>7.506321537931e-03</td>
<td>β₃₈</td>
<td>7.863567181544e+01</td>
</tr>
<tr>
<td>β₁₈</td>
<td>2.250885074152e-03</td>
<td>β₃₉</td>
<td>9.231737401470e+01</td>
</tr>
<tr>
<td>β₁₉</td>
<td>4.769913115603e-04</td>
<td>β₃₁₀</td>
<td>1.064691762642e+02</td>
</tr>
<tr>
<td>β₁₁₀</td>
<td>6.352277030229e-05</td>
<td>β₃₁₁</td>
<td>1.210528308429e+02</td>
</tr>
<tr>
<td>β₁₁₁</td>
<td>3.992914194077e-06</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

norm given in (2). Hence,

\[
\|F_{n+c_2} - F_{n+c_2}\| \leq \left\| f \left( t_n + c_2 h_{n+1}, \alpha_{20} \tilde{y}_n + \alpha_{21} \tilde{y}_{n-1} + h_{n+1} \left( a_{21} \tilde{f}_n + \sum_{j=1}^{p-4} \beta_{2j} \tilde{f}_{n-j} \right) \right) \right. \\
- f \left. \left( t_n + c_2 h_{n+1}, \alpha_{20} y_n + \alpha_{21} y_{n-1} + h_{n+1} \left( a_{21} f_n + \sum_{j=1}^{p-4} \beta_{2j} f_{n-j} \right) \right) \right\| \\
\leq L \left( ||\alpha_{20}|| + ||\alpha_{21}|| \|\tilde{y} - y\| + h_{n+1} \left\| a_{21} (\tilde{f}_n - f_n) + \sum_{j=1}^{p-4} \beta_{2j} (\tilde{f}_{n-j} - f_{n-j}) \right\| \right) \\
\leq L(2K_c + h_{n+1}(p-3)K_c L) \|\tilde{y} - y\| \\
\leq L K_c (2 + (t_f - t_0)(p-3)L) \|\tilde{y} - y\| \equiv \Psi_2 \|\tilde{y} - y\| \quad (7)
\]
where $\Psi_2 = LK_c(2 + (t_f - t_0)(p - 3)L) > 0$. Also, using (7) we have

$$
\|\tilde{F}_{n+c} - F_{n+c}\| \leq \|f\left(t_n + h_{n+1}, \alpha_30\tilde{y}_n + \alpha_31\tilde{y}_{n-1} + h_{n+1}\left(a31\tilde{f}_n + a32\tilde{F}_{n+c_2} + \sum_{j=1}^{p-4} \beta_{3j}\tilde{f}_{n-j}\right)\right)
$$

$$
- f\left(t_n + h_{n+1}, \alpha_30y_n + \alpha_31y_{n-1} + h_{n+1}\left(a31f_n + a32F_{n+c_2} + \sum_{j=1}^{p-4} \beta_{3j}f_{n-j}\right)\right)\| 
$$

$$
\leq L\left(2K_c\|\tilde{y} - y\| + h_{n+1}\|a31(\tilde{f}_n - f_n) + a32(\tilde{F}_{n+c_2} - F_{n+c_2})
+ \sum_{j=1}^{p-4} \beta_{3j}(\tilde{f}_{n-j} - f_{n-j})\|\right) 
$$

$$
\leq L(2K_c + h_{n+1}K_c[(p - 3)L + \Psi_2])\|\tilde{y} - y\|
\leq L(2K_c + (t_f - t_0)K_c[(p - 3)L + \Psi_2])\|\tilde{y} - y\| \equiv \Psi_3\|\tilde{y} - y\| \quad (8)
$$

where $\Psi_3 = L(2K_c + (t_f - t_0)K_c[(p - 3)L + \Psi_2]) > 0$.

Hence, from (7) and (8), we see that $\Phi$ satisfies

$$
\|\Phi(\tilde{y}_n, \tilde{y}_{n-1}, \ldots, \tilde{y}_{n-p+4}; f, \Delta_n) - \Phi(y_n, y_{n-1}, \ldots, y_{n-p+4}; f, \Delta_n)\|
\leq K_c((p - 3)L + \Psi_2 + \Psi_3)\|\tilde{y} - y\| \equiv \Psi_1\|\tilde{y} - y\|,
$$

where $\Psi_1 = K_c((p - 3)L + \Psi_2 + \Psi_3) > 0$ and we are done for the Hermite–Birkhoff solver.

Next, DP45 being a 7-stage Runge–Kutta method and taking 15 as the uniform bound for the coefficients, we get

$$
\|\Phi(y_n; f, h_{n+1}) - \Phi(\tilde{y}_n; f, h_{n+1})\| \leq \sum_{i=1}^{7} \frac{15^i}{15(i+1)}\|y_n - \tilde{y}_n\| \leq 7\frac{15^8(t_f - t_0)^8 - 1}{15(t_f - t_0) - 1}\|y_n - \tilde{y}_n\|. \quad (9)
$$

Putting

$$
L' = 7\frac{15^8(t_f - t_0)^8 - 1}{15(t_f - t_0) - 1}
$$
as the Lipschitz constant ends the Lipschitz condition for DP45. Now, convergence will follow from the two important concepts of consistency and 0-stability.

### 3.1.3 Consistency of HB515 and DP45

This section also applies to the variable-step case so we will keep our notation of $h_{n+1}$ which designates the stepsize on the integration interval $[t_n, t_{n+1}]$.

**Definition 1** We say that the ODE method (3) is consistent of order (or, equivalently, has order) $p$ if $p \geq 1$ is the largest integer such that, for all $C^p$-continuous functions $f$ in (1) and for all mesh points, we have that

$$
\|y(t_{n+1}) - y_{n+1}\| = O(h^{p+1}_{n+1})
$$


7
uniformly with respect to \( n = 0, 1, \ldots, N - 1 \), where \( y(t) \) is the exact solution to (1) and

\[
\dot{y}_{n+1} = \alpha_{n,1} y(t_n) + \cdots + \alpha_{n,k} y(t_{n-k+1}) + h_{n+1} \Phi(y(t_n), \ldots, y(t_{n-k+1}); f, \Delta_n).
\]

To avoid confusion, we point out that there are two different notions of order used in this article: consistency and convergence orders (convergence order will be introduced below). As emphasized in the definition and unless otherwise specified, the term order refers to the consistency order.

In [10], we constructed the 3-stage variable-step variable-order Hermite-Birkhoff ODE solver HB(5-15)3 to satisfy the consistency order conditions for orders 5 to 15. Hence, we do not need to prove it here. On the other hand, DP45 was also constructed to satisfy the Runge–Kutta order conditions for order 5 and hence is consistent of order 5.

### 3.1.4 0-STABILITY OF HB515 AND DP45

Let \( n \in \{0, \ldots, N\} \) be the step index and \( p \in \{5, \ldots, 15\} \) be the order of the integration formula at step \( n \). Define the first characteristic polynomial \( p_n(x) \) as follows:

\[
p_n(x) = x^p - 3 - \sum_{i=1}^{p-3} \alpha_{n,i} x^{p-3-i}
\]

where \( \alpha_{n,i} \) are the coefficients appearing in (4). Since we are analyzing constant stepsize, then for a given order, all \( p_n(x) \) are the same polynomial which we call \( p(x) \). Since \( \alpha_{n,i} = 0 \) for all \( i \geq 3 \) and \( \alpha_{n,2} + \alpha_{n,1} = 1 \) (see [10]), a simple factorization gives

\[
p(x) = x^{p-5}(x-1)(x + \alpha_{n,2}).
\]

Now, we want to prove that \( p(x) \) satisfies the well-known root condition.

**Definition 2** The polynomial \( p(x) \) satisfies the root condition if

1. all roots \( r \) of \( p(x) = 0 \) lie inside the unit disk, i.e. \( |r| \leq 1 \) and
2. if \( r \) is a root of \( p(x) = 0 \) of absolute value 1, then \( r \) is simple.

We use the fact that \( \alpha_{n,2} \) is \( \alpha_{11} \) in the notation of [10] and get from Table 2 that \( \alpha_{n,2} \) is always smaller than 1. Hence, all the roots of \( p(x) = 0 \) being of absolute value smaller or equal to 1 and 1 being a simple root, we get that \( p(x) \) satisfies the root condition.

Next, define the matrix

\[
C_n = \begin{bmatrix}
\alpha_{n,1} & \alpha_{n,2} & \alpha_{n,3} & \cdots & \alpha_{n,p-4} & \alpha_{n,p-3} \\
1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 1 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & 0
\end{bmatrix} = \begin{bmatrix}
\alpha_{n,1} & \alpha_{n,2} & 0 & \cdots & 0 & 0 \\
1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 1 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & 0
\end{bmatrix}
\]

(10)
as the companion matrix of the polynomial
\[ p_n(x) = x^{p-3} - \sum_{i=1}^{p-3} \alpha_{n,i} x^{p-3-i}. \]

We now introduce the following definition adapted from [2].

**Definition 3** The method HB515 in (4) satisfies the 0-stability condition if there exists a norm \( \| \cdot \| \) on \( \mathbb{R}^{p-3} \) independent of both \( n \) and \( \Delta \) such that the operator norm on the matrix \( C_n \) satisfies
\[ \| C_n \| \leq 1. \]

Using the same argument as for the independence of \( p(x) \) with respect to \( n \), we realize that for a fixed order all \( C_n \) are equal to the same matrix \( C \) which is the companion matrix of \( p(x) \). Solving our systems for the different orders, we found that \( 0 < \alpha_{n,2} < 1 \) is always true and hence \( 0 < \alpha_{n,1} < 1 \) because \( \alpha_{n,1} + \alpha_{n,2} = 1 \). Thus, it is easy to see that if we choose the infinity norm, we get
\[ \| C \|_\infty = \sup_{\| x \|_\infty = 1} \| Cx \|_\infty \leq 1 \]
which implies that our method is 0-stable.

More generally, Bellen and Zennaro remark in [2] that proving the first characteristic polynomial of method (3) satisfies the root condition is equivalent, for the constant-step case, to proving the 0-stability described in Definition 3.

On the other hand, we usually do not consider 0-stability for Runge-Kutta methods but because we use general theorems to prove the convergence of both HB515 and DP45, we need to say a word about it. Hence, we remark that the 0-stability is trivial for 1-step methods because the companion matrix of the first characteristic polynomial is a the trivial 1 by 1 identity matrix and thus the infinity norm can also be used to get that DP45 is 0-stable.

### 3.1.5 CONVERGENCE OF DP45 AND HB515

Using the above information, we can state the convergence Theorem 1 which is adapted from a general theorem in [2].

**Theorem 1** Let the constant \( k \)-step ODE method (5) be consistent of order \( p = 5 \), 0-stable and Lipschitz in \( y \). Then, the following two conditions:

1) the function \( f(t, y) \) in (1) is \( C^5 \)-continuous;

2) the set of starting values \( y_0, \ldots, y_{k-1} \) approximate the exact solution to order 5

imply that the ODE method is convergent of order 5 on \([t_0, t_f]\), that is
\[ \max_{1 \leq n \leq N} \| y(t_n) - y_n \| = O(h^5) \]
where \( h \) is the integration stepsize.
Proof. See the proof of Theorem 2 which is strictly more general than the one for the theorem above.

Thus, we get convergence of the discrete constant-order ODE solvers HB5 and DP45 with convergence order 5. Moreover, using the fact that HB5i15 is consistent of order $p \in \{5, \ldots, 15\}$, it is easy to see that the convergence order cannot decrease because the local error is smaller when the order is increased. We also remark that the convergence order cannot increase because, unlike the consistency which is a local property, convergence considers the maximum error over all integration steps. Hence, the smallest order is the dominant one. Thus, the ODE solver DP45-HB515 is convergent of order 5.

3.2 VARIABLE-STEP AND VARIABLE-ORDER

At first sight, the only difference between the variable-step variable-order (VSVO) case and the CSVO case is that the stepsize is not constant. Still, this small change makes the whole convergence theory a lot more difficult to put in place.

We recall that the stepsize control formulae used for our method are based on the well-known work of Shampine and Gordon in [12]. The two authors argued that close to constant stepsize was obtained when using this type of formulae. Hence, if we suppose that we have almost constant stepsize, the theory of constant stepsize could be generalized heuristically to the variable stepsize. However, we will try to be as rigorous as possible and avoid heuristics even though this will cost us some efficiency. We note that all the preliminary theory for DP45 still applies for the VSVO case because its coefficients are constant and hence no effect is noticed when we generalize from the CSVO to the VSVO case.

3.2.1 BOUNDEDNESS OF THE COEFFICIENTS OF HB515

When we look at the systems used to compute the coefficients of method (4) as they appear in [10], we see that they are functions of the ratios of preceding stepsizes. Hence, we need to study the stepsize change rates in order to have an idea on the size of the coefficients. First, when the stepsize $h_n$ is accepted for some $n \geq 0$, we choose $h_{n+1}$ using a formula of the type

$$h_{n+1} = \min \left\{ h_n \cdot \text{stfac} \cdot \left( \frac{\text{TOL}}{\text{EST}} \right)^{1/(\kappa-1)}, c_1 h_n, h_{\text{max}} \right\} \quad (11)$$

where stfac = 0.81 is a safety factor, EST is the local error estimate at step $n$, $\kappa$ is the chosen order for the next integration step, $h_{\text{max}}$ is the maximum allowed stepsize and $c_1 > 1$ is a constant. On the other hand, when, at the step $n$, a stepsize is rejected (call it $h_{n+1}^{(0)}$), we lower the stepsize using a formula of the type

$$h_{n+1}^{(i)} = \max \left\{ h_{n+1}^{(i-1)} \cdot 0.7 \cdot \text{stfac} \cdot \left( \frac{\text{TOL}}{\text{EST}} \right)^{1/\kappa}, 0.5 h_{n+1}^{(i-1)}, h_{\text{min}} \right\} \quad (12)$$
where \( i \geq 1 \), \( \kappa \) is the actual integration order and \( h_{\text{min}} = c_2 h_n \) \( (h_n \) being the last accepted step, \( c_2 \in [0, 1] \) is a constant and \( h_{\text{min}} \equiv \text{TOL} \) when \( n = 0 \)). Hence, we see from (11) and (12) that \( c_2 \leq h_{n+1}/h_n \leq c_1 \). After long experimental tests, we chose \( c_1 = 1.9 \) and \( c_2 = 0.5^3 \) giving us, for all \( n \geq 0 \),

\[
0.5^3 \leq h_{n+1}/h_n \leq 1.9.
\] (13)

The choice of \( c_2 = 0.5^3 \) comes from the fact that we allow the solver to halve the stepsize at most three times. On the other hand, the systems involved in finding the coefficients were very ill conditioned when using \( c_1 \approx 2 \) so we took 1.9 to be on the safe side.

Since we only need some constant uniform bound on the coefficients of the method, we will not explicitly use the values \( c_1 \) and \( c_2 \) to bound our coefficients. Indeed, it remains a very difficult task to get an absolute bound for all the coefficients by analyzing the worst possible cases when solving the systems. Instead, we did some experimentations using the available information and we chose a large constant uniform bound \( K = 2E25 \) on the coefficients and we “forced” the solver to only accept steps for which the coefficients were smaller than \( K \). Note that this bound is extremely big from a numerical point of view but to get the Lipschitz condition, we merely need to provide some constant upper bound for the coefficients. Hence, we found no reason to constrain our solver with a smaller \( K \).

Next, we mention that proving that the method is consistent and satisfies a Lipschitz condition with respect to the \( y \) argument is the same as in the CSVO case so we will not repeat the procedure. The only difference is in the value of the uniform bound \( K \) which replaces the \( K_c \).

### 3.2.2 0-STABILITY OF HB515

We now get to one of the major difficulties that we encountered in reconciling a rigorous theory with the experimentation. First, it is easy to see that the 0-stability (as it is given in Definition 3) is satisfied under the infinity norm if we impose

\[
\alpha_{n,2} \in [0, 1].
\] (14)

Now, we must clarify that (14) is a sufficient but definitely not a necessary condition for 0-stability. Indeed, we had pleasing results when we only asked for \( |\alpha_{n,2}| \leq 1 \) (see [10]). Actually, we already obtained good results when no constraint was imposed on \( \alpha_{n,2} \). Still, it is very difficult to prove that big \( \alpha_{n,2} \) values are compensated by smaller ones and by corrections taking place along the integration.

Another way to see how difficult it is to get minimal conditions for 0-stability is to look at the notion from the theory of Hairer et al. in [7]. Indeed, they link the 0-stability to finding a bound for finite products of successive matrices \( C_n \) defined in (10). Hence, it is now obvious that one cannot predict the norm of those products when no simple explicit formulae are available for the coefficients \( \alpha_{n,1} \) and \( \alpha_{n,2} \) let alone when no explicit formulae are available in the first place.
Hence, imposing (14) was a very difficult choice to make and a major constraint to impose but we chose to abide by it in order to be certain that we satisfy the theoretical 0-stability condition.

### 3.2.3 CONVERGENCE OF DP45–HB515

For sake of completeness, we restate and prove the convergence theorem for the general variable-step case which is adapted from [2]. Before that, we state and prove a preliminary lemma.

**Lemma 1** Let $C$ be a real $k \times k$ matrix and $\| \cdot \|_\ast$ be a vector norm on $\mathbb{R}^k$. Let $d \geq 1$ be an integer. We define a vector norm on $\mathbb{R}^{dk}$ as follows. Suppose $B = [B^{(1)}, \ldots, B^{(k)}]^T$ is a stacking of $k$ vectors $B^{(i)}$ each of them belonging to $\mathbb{R}^d$. Then define

$$
\|B\| = \max_{1 \leq i \leq d} \left\| \begin{bmatrix} B^{(1)}_i, \ldots, B^{(k)}_i \end{bmatrix}^T \right\|_\ast. \tag{15}
$$

For both norms, the induced matrix norm will be denoted in the same way as the vector norm to avoid confusion. Now, if $\|C\|_\ast \leq 1$, then

$$
\|C \otimes I_d\| \leq 1,
$$

where $\otimes$ is the Kronecker product (see [6, §4.5.5]) and $I_d$ is the $d \times d$ identity matrix.

**Proof.** We have $\|C\|_\ast \leq 1$. Then,

$$
\sup_{z \neq 0} \frac{\|Cz\|_\ast}{\|z\|_\ast} \leq 1.
$$

Now, we want to prove

$$
\|C \otimes I_d\| = \sup_{x \neq 0} \frac{\|(C \otimes I_d)x\|}{\|x\|} \leq 1.
$$

We have the following property on Kronecker products (see [6, p.180]):

$$
y = (C \otimes I_d)x \iff Y = I_dXC^T = XC^T
$$

where $y, x \in \mathbb{R}^{dk}$ are stackings of $k$ vectors $y^{(1)}, \ldots, y^{(k)}$ and $x^{(1)}, \ldots, x^{(k)}$, respectively, with $y^{(i)}, x^{(i)} \in \mathbb{R}^d$, and $Y, X \in \mathbb{R}^{d \times k}$ are matrices whose columns are $y^{(1)}, \ldots, y^{(k)}$ and $x^{(1)}, \ldots, x^{(k)}$, respectively. Define on $\mathbb{R}^{d \times k}$ the norm

$$
\|A\|' = \max_{1 \leq i \leq d} \|A(i, :)^T\|_\ast.
$$

Hence,

$$
\sup_{x \neq 0} \frac{\|(C \otimes I_d)x\|}{\|x\|} \leq 1 \iff \sup_{X \neq 0} \frac{\|XC^T\|'}{\|X\|'} \leq 1.
$$
Let \( 0 \neq X \in \mathbb{R}^{d \times k} \). We have

\[
\|XC^T\|_*' \leq 1 \iff \frac{\max_{1 \leq i \leq d} \|(XC^T)(i,:)\|_*}{\max_{1 \leq i \leq d} \|X(i,:)\|_*} \leq 1 \iff \frac{\max_{1 \leq i \leq d} \|C(X(i,:))^T\|_*}{\max_{1 \leq i \leq d} \|X(i,:)\|_*} \leq 1.
\]

Next, let \( k \) be such that \( \max_{1 \leq i \leq d} \|(XC^T)(i,:)\|_* = \|((X^T)(k,:))\|_* \) and set \( z = (X(k,:))^T \) (\( z \) cannot be the 0-vector since it is the row of \( X \) with biggest norm and if it were 0 then \( X \) would be the 0 matrix which is not the case by the choice of \( X \)). Hence,

\[
\frac{\max_{1 \leq i \leq d} \|C(X(i,:))^T\|_*}{\max_{1 \leq i \leq d} \|X(i,:)\|_*} = \frac{\|Cz\|_*}{\max_{1 \leq i \leq d} \|X(i,:)\|_*} \leq \|Cz\|_* \leq 1
\]

by hypothesis. Therefore,

\[
\|C \otimes I_d\| = \sup_{x \neq 0} \frac{\|(C \otimes I_d)x\|}{\|x\|} \leq 1
\]

which proves the statement.

**Theorem 2** Let the variable \( k \)-step ODE method (3) be consistent of order 5, 0-stable and Lipschitz in \( y \). Then, the following two conditions:

1) the function \( f(t, y) \) in (1) is \( C^5 \)-continuous;

2) the set of starting values \( y_0, \ldots, y_{k-1} \) approximate the exact solution to order 5;

imply that the ODE method is convergent of order 5 on \([t_0, t_f]\), that is

\[
\max_{1 \leq n \leq N} \|y(t_n) - y_n\| = O(h^5)
\]

where \( h = \max_{1 \leq n \leq N} \{h_n\} \).

**Proof.** To avoid a conflict of notation, we will use \( \Phi \) as the increment function for the \( k \)-step method at hand which will apply to both the increment functions of HB515 and DP45. We use the fact that the method is consistent of order 5 and get

\[
y_{n+1} = \alpha_{n,1} y(t_n) + \cdots + \alpha_{n,k} y(t_{n-k+1}) + h_{n+1} \Phi(y(t_n), \ldots, y(t_{n-k+1}); f, \Delta_n) + \epsilon_{n+1},
\]

with

\[
\|\epsilon_{n+1}\| \leq ch_{n+1}^6,
\]

for some constant \( c > 0 \) for all \( n = k - 1, \ldots, N - 1 \).

We define the \( y_n = [y_n, y_{n-1}, \ldots, y_{n-k+1}]^T \) and \( y(t_n) = [y(t_n), y(t_{n-1}), \ldots, y(t_{n-k+1})]^T \) both of dimensions \( dk \times 1 \). Then, subtracting (3) from (16) gives us

\[
y(t_{n+1}) - y_{n+1} = \zeta_{n}(y(t_n) - y_n) + h_{n+1} \Gamma_n + E_{n+1},
\]

with

\[
\|E_{n+1}\| \leq ch_{n+1}^6.
\]

13
\( n = k - 1, \ldots, N - 1 \), where \( \zeta_n = C_n \otimes I_d \),

\[
\Gamma_n = [\Phi(y(t_n); f, \Delta_n) - \Phi(y_n; f, \Delta_n), 0, \ldots, 0]^T
\]

and \( E_{n+1} = [\epsilon_{n+1}, 0, \ldots, 0]^T \), with 0 is the zero vector of \( \mathbb{R}^d \).

Let \( \| \cdot \| \) be the norm on \( \mathbb{R}^{kd} \) defined in Lemma 1 (\( C \) and \( \| \cdot \|_\infty \) in the lemma are \( C_n \) and \( \| \cdot \|_\infty \), respectively, in this proof) and get that \( \| \zeta_n \| \leq 1 \). Therefore, by (18) we get

\[
\| y(t_{n+1}) - y_{n+1} \| \leq \| y(t_n) - y_n \| + h_n + 1 \| \Gamma_n \| + \| E_{n+1} \|,
\]

\( n = k - 1, \ldots, N - 1 \). Next, since \( \Phi \) is Lipschitz with respect to the \( y \) argument and by the equivalence of the norms in finite dimensional spaces, there exists a constant \( Q > 0 \) such that

\[
\| \Gamma_n \| \leq Q \| y(t_n) - y_n \|.
\]

Again, using the equivalence of the norms in finite dimensional spaces, there exists \( c' > 0 \) such that (17) becomes

\[
\| y(t_{n+1}) - y_{n+1} \| \leq (1 + h_{n+1}Q) \| y(t_n) - y_n \| + c'h_{n+1}h^5,
\]

\( n = k - 1, \ldots, N - 1 \).

We next define

\[
\hat{c} = \max_{k \leq i \leq n+1} \frac{\| \epsilon_i \|}{h_i^6}
\]

and get

\[
\| y(t_{n+1}) - y_{n+1} \| \leq \left[ \prod_{i=k}^{n+1} (1 + h_iQ) \right] \| y(t_{k-1}) - y_{k-1} \| + \left( \sum_{i=k}^{n+1} \left[ \prod_{j=i+1}^{n+1} (1 + h_jQ) \right] h_i \right) \hat{c}h^5
\]

\[
\leq \left[ \prod_{i=k}^{n+1} e^{h_iQ} \right] \| y(t_{k-1}) - y_{k-1} \| + \left( \sum_{i=k}^{n+1} \left[ \prod_{j=i+1}^{n+1} e^{h_jQ} \right] h_i \right) \hat{c}h^5
\]

\[
\leq e^{Q(t_j - t_0)} \| y(t_{k-1}) - y_{k-1} \| + e^{Qt_j} \left( \sum_{i=k}^{n+1} e^{-Qt_j} h_i \right) \hat{c}h^5
\]

\[
\leq e^{Q(t_j - t_0)} \| y(t_{k-1}) - y_{k-1} \| + e^{Qt_j} \left( \int_{t_0}^{t_j} e^{-Qt} dt \right) \hat{c}h^5
\]

\[
= e^{Q(t_j - t_0)} \| y(t_{k-1}) - y_{k-1} \| + \frac{e^{Q(t_j - t_0)} - 1}{Q} \hat{c}h^5,
\]

\( n = k - 1, \ldots, N - 1 \). Thus, again by equivalence of the norms, since the starting values approximate the solution \( y(t) \) up to order 5, the proof is complete.

The above theorem gives us convergence of order 5 for the variable-step constant-order HB5. Hence, we can deduce the convergence of order 5 for the variable-order HB515 because the local error does not increase when the order is raised. Finally, assuming the consistency of DP45, we can adapt the theorem to get that DP45 is convergent of order 5. Thus, the variable-step variable-order ODE solver DP45-HB515 is convergent of convergence order 5.
In this section, we prove, under some assumptions, the convergence of the general VSVO DDE solver HB515DDE for the problem

\[ y'(t) = f(t, y(t), y(t - \tau(t, y(t)))), \quad t_0 \leq t \leq t_f; \quad y(t) = \phi(t), \quad t \leq t_0, \quad (19) \]

where \( f(t, u, v) \in C^0([t_0, t_f] \times \mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}^d) \) is globally Lipschitz continuous with respect to \( u \) and \( v \) in a given norm \( \| \cdot \| \) of \( \mathbb{R}^d \) with Lipschitz constants \( L > 0 \) and \( M > 0 \) respectively.

### 4.1 EFFECT OF DISCONTINUITIES

The appearance and propagation of discontinuities is a major threat for the convergence of an ODE method. Hence, a lot of attention is required when modifying an ODE code to solve DDEs. Indeed, knowing which discontinuities the solver can detect enables us to determine the smoothness of the integrated function and the actual order of convergence.

Recall that the error estimate (EST) is given by

\[ \text{EST} = \| y_{n+1} - \tilde{y}_{n+1} \|_2 \]

where \( y_{n+1} \) and \( \tilde{y}_{n+1} \) are given by the integration formula IF and the corrector \( P_4 \), respectively. Then a special routine for detecting a discontinuity at \( \xi \in (t_n, t_{n+1}] \) is called only if a step is rejected by the error test \( \text{EST} \leq TOL \). Hence, we must study the effect of crossing a discontinuity on the error estimate. This includes knowing how the size and the order of the jump discontinuity affect the error estimate and which coefficients of the method are involved in signaling the appearance of the discontinuity.

We first give the following definition adapted from [2].

**Definition 4** A discontinuity point \( \xi \) is said to be of order \( q \) if \( y^{(v)}(\xi) \) exists for \( v = 0, \ldots, q \) and \( y^{(q)} \) is Lipschitz continuous at \( \xi \).

To illustrate the effect of discontinuities, suppose, for simplicity, that we have the following scalar DDE:

\[ y'(t) = f(t, y(t), y(\alpha(t))), \quad t_0 \leq t \leq t_f; \quad y(t) = \phi(t), \quad t \leq t_0, \]

with \( \alpha \) being a strictly increasing state independent delay. Suppose also that we are integrating at order \( p \in \{5, \ldots, 15\} \) on \([t_n, t_{n+1}]\) and that (only) one jump discontinuity \( \xi \) lies on the integration interval. Let the order of the discontinuity be \( q \geq 0 \) (i.e. the discontinuity is in \( y^{(q+1)} \)) and the size of the jump be \( K_q \).

Suppose for now that \( t_n + c_2 h < \xi < t_n + c_3 h \). Hence, \( \xi \) does not affect \( Y_{n+c_2} \) which approximates the exact solution to order \( p - 2 \). Also, \( Y_{n+c_3} \) is not \textit{a priori} affected by \( \xi \) because it only uses values of \( y \) and \( f \) at \( t < \xi \). However, because the extrapolation does not take \( \xi \) into account, \( Y_{n+c_3} \) will lie on some smooth continuation of \( y \) and not on \( y \) itself. This will lead to an inaccurate
Because it uses $F_{n+c_3} = f(t_n + c_3h, Y_{n+c_3}, y(\alpha(t_n + c_3h)))$. Finally, $\tilde{y}_{n+1}$ will also get a share of inaccuracy because it uses the computed $y_{n+1}$. This lack of precision will appear as a huge error estimation whose size depends on $q$ and $K_q$. Following a similar argument, if the discontinuity happens to be in the interval $[t_n, t_n + c_2h]$, then we expect that both $Y_{n+c_2}$ and $Y_{n+c_3}$ will be affected by it and, again, a large EST should appear.

Taking a close look at the terms where $F_{n+c_2}$ and $F_{n+c_3}$ appear in $y_{n+1}$ and $\tilde{y}_{n+1}$ (which compose EST), we see that the main coefficients linked to the effect of the discontinuity are $b_{12}, b_{13}$ and $a_{43}$. Since the routine for locating a discontinuity is called only after a step rejection, it is useful to make sure that $b_{12}, b_{13}$ and $a_{43}$ are neither too small in which case they would absorb the effect of the discontinuity, nor too large in which case they would amplify a small $K_q$ which would have been of no nuisance to the solver. Hence, the size of the three coefficients has been monitored for the seven test problems given in [13] and lower and upper bounds for each coefficient are displayed in Table 3. These results show that $b_{12}, b_{13}$ and $a_{43}$ are close to constant and neither too small nor too big, as desired.

<table>
<thead>
<tr>
<th></th>
<th>$b_{12}$</th>
<th>$b_{13}$</th>
<th>$a_{43}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>0.51030</td>
<td>0.002342</td>
<td>0.20934</td>
</tr>
<tr>
<td>P2</td>
<td>0.52876</td>
<td>0.037879</td>
<td>0.21398</td>
</tr>
<tr>
<td>P3</td>
<td>0.52876</td>
<td>0.039220</td>
<td>0.21213</td>
</tr>
<tr>
<td>P4</td>
<td>0.52885</td>
<td>0.004365</td>
<td>0.19592</td>
</tr>
<tr>
<td>P5</td>
<td>0.52885</td>
<td>0.038814</td>
<td>0.19837</td>
</tr>
<tr>
<td>P6</td>
<td>0.52885</td>
<td>0.001481</td>
<td>0.19592</td>
</tr>
<tr>
<td>P7</td>
<td>0.52892</td>
<td>0.023461</td>
<td>0.20785</td>
</tr>
<tr>
<td>Min</td>
<td>0.51030</td>
<td>0.001481</td>
<td>0.19592</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$b_{12}$</th>
<th>$b_{13}$</th>
<th>$a_{43}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>0.74647</td>
<td>0.09918</td>
<td>0.49719</td>
</tr>
<tr>
<td>P2</td>
<td>0.68486</td>
<td>0.09605</td>
<td>0.46667</td>
</tr>
<tr>
<td>P3</td>
<td>0.68254</td>
<td>0.09605</td>
<td>0.46525</td>
</tr>
<tr>
<td>P4</td>
<td>0.74336</td>
<td>0.09603</td>
<td>0.49668</td>
</tr>
<tr>
<td>P5</td>
<td>0.68301</td>
<td>0.09603</td>
<td>0.46568</td>
</tr>
<tr>
<td>P6</td>
<td>0.74777</td>
<td>0.09603</td>
<td>0.49835</td>
</tr>
<tr>
<td>P7</td>
<td>0.71225</td>
<td>0.09602</td>
<td>0.47260</td>
</tr>
<tr>
<td>Max</td>
<td>0.74777</td>
<td>0.09919</td>
<td>0.49835</td>
</tr>
</tbody>
</table>

As pointed out by Calvo et al. in [3], a general concept for an adaptive code like ours is that a discontinuity of order $q$ usually produces an error of $O(h^q)$ (i.e. of order $q - 1$) at the error estimation. This can be deduced from the following adaptation of Taylor’s theorem.

16


Theorem 3 Let \( n \geq 0 \) be a step index and \( t_n \) and \( h_{n+1} > 0 \) be the corresponding mesh point and stepsize, respectively. Let \( t_{n+1} = t_n + h_{n+1} \) and \( r \geq 0 \) be an integer. If \( y \in C^r([t_n, t_{n+1}]) \) and \( y \in C^{r+1}([t_n, t_{n+1}]) \) then

\[
y(t_{n+1}) = y(t_n) + \frac{y'(t_n)}{1!}h_{n+1} + \cdots + \frac{y^{(r)}(t_n)}{n!}h_{n+1}^r + O(h_{n+1}^{r+1}).
\]

Because \( y_{n+1} \) is of order \( p \) and \( \hat{y}_{n+1} \) is of order \( p-2 \), then our error estimator is of order \( p-2 \). Hence, for \( y \) sufficiently smooth, EST can achieve an \( O(h^{p-1}) \) accuracy. However, if a discontinuity \( \xi \in [t_n, t_{n+1}] \) of order \( q \) appears in \( y^{(q+1)} \) then the error term is \( O(h^q) \) and the local order becomes \( q-1 \). Therefore, if \( q-1 \geq p-2 \), then the method will remain of order \( p-2 \) and we may not even notice the discontinuity, whereas for \( q \leq p-2 \) the solver usually produces an EST which is larger than the prescribed tolerance for the current step.

Following a similar argument, the local error estimator (EST') of DP45 being of order 4, a discontinuity of order \( q \) may not be noticed by the local error estimator if \( q-1 \geq 4 \) whereas for \( q \leq 4 \) the solver usually produces a large EST' that we can use to signal the presence of the discontinuity. Hence, we make the following optional assumption.

**Assumption 1:** (Optional) The error estimates of DP45 and HB515DDE detect all numerically significant discontinuities of order smaller than or equal to 3.

Note that this assumption follows from the fact that the error estimates of DP45 and HBp, \( p = 5, \ldots, 15 \), are locally of order greater than or equal to 3. Moreover, Assumption 1 is justified since it is impossible to determine an exact order \( q \) such that the solver will always detect all discontinuities of order smaller than or equal to \( q \) because for any \( q \), there exists, a jump discontinuity with a \( K_q \) small enough such that the discontinuity will not be seen by the solver.

Hence, there are always cases where the estimate does not detect the presence of a discontinuity of order smaller than or equal to the one of the local error estimator. In that case we assume that, because the error term is small, the theoretical discontinuity has no major consequence on the overall integration.

Giving exact conditions for the detection of some discontinuity in the general case is complex and not useful to the user of the solver. Indeed, the criteria for the detection of a discontinuity would involve computing data (giving sharp bounds on partial derivatives of \( f \) and on the deviated arguments, etc.) that the user would only have if he/she already knows \( y(t) \) and has a deep analysis of its properties at hand. This would obviously go against the goal of solving the DDE numerically.

### 4.2 HERMITE INTERPOLANT \( \eta(t) \)

Now, we dedicate some time for the analysis of the Hermite interpolant. As a reminder from [13], we use a Hermite interpolating polynomial \( \eta(t) \) to find the value of \( y \) at a delayed time \( t^* \) by interpolating/extrapolating past accepted \( y_j \)'s. The number of points \( m \) used for the interpolation/extrapolation is given...
by the formula

\[ m(p) = \lfloor p/2 \rfloor \]  

(20)

where \( p \in \{5, \ldots, 15\} \) is the actual integration order, \( \lfloor p/2 \rfloor \) is the smallest integer not less than \( p/2 \). Hence,

\[ 3 \leq m(p) \leq 8. \]

Suppose we are integrating at order \( p \). Let \( s \geq 0 \) and \( t_s, \ldots, t_{s+m-1} \) be \( m \) distinct mesh points where \( m \) is given by equation (20). Put \( u = s + m - 1 \leq N - 1 \). If the solution \( y \) of (19) is \( C^1([t_s, t_u]) \), then the unique polynomial \( \eta(t) \) of least degree with

1) \( \eta(t_i) = y(t_i) \),
2) \( \eta'(t_i) = f(t_i, y(t_i), z(t_i)) \),

for all \( i = s, \ldots, u \) is the Hermite polynomial of degree at most \( 2m - 1 \) given by

\[ \eta(t) = \sum_{j=s}^{u} H_{m-1,j}(t) y(t_j) + \Psi(y(t_u), \ldots, y(t_s); t, f, \Delta') \]  

(21)

where the increment is given by

\[ \Psi(y(t_u), \ldots, y(t_s); t, f, \Delta') = \sum_{j=s}^{u} \hat{H}_{m-1,j}(t) f(t_j, y(t_j), z(t_j)) \]

with \( \Delta' = \{t_s, \ldots, t_u\} \) and the coefficients are as follows

\[ H_{m-1,j}(t) = [1-2(t-t_j)L_{m-1,j}(t)]L_{m-1,j}^2(t) \]

and \( \hat{H}_{m-1,j}(t) = (t-t_j)L_{m-1,j}^2(t) \)

with

\[ L_{m-1,j}(t) = \prod_{i=s,i\neq j}^{u} \frac{t-t_i}{t_j-t_i} \]

being the \( j \)th Lagrange polynomial of degree \( m - 1 \).

First, as it was the case with the Hermite–Birkhoff method, we need to have a uniform bound on the various coefficients of the interpolant to be able to prove the convergence of our DDE solver.

### 4.2.1 BOUNDEDNESS OF THE COEFFICIENTS FOR THE INTERPOLANT

As we did before with method (4), it is possible to bound the coefficients \( H_{m-1,j}(t) \) and \( \hat{H}_{m-1,j}(t) \) using the bounds on the ratio of consecutive step-sizes \( h_{n+1}/h_n \) given in (13). For the Hermite–Birkhoff method, we chose to study the behavior of the solver under the given bounds and then force the solver to lower the stepsize until the coefficients were bounded by the chosen constant. However, for the interpolant, we do not want to force step rejections and hence we will provide theoretical bounds for the worst case. Hence, using
where \( u \) is a mesh-independent uniform bound on all the coefficients of the Hermite interpolant.

Therefore, \( \bar{c}_r = 2^n \) for all \( n \geq 0 \). Note that \( \bar{c}_r \) gives us the freedom of halving the minimum stepsize that the solver produces at most six times. This will be useful when steps have to be inserted between two close discontinuities for high order interpolation (see Subsection 4.3).

Let \( p \in \{5, \ldots, 15\} \) and \( m \) be the number of mesh points used for interpolation when integrating at order \( p \). Put

\[
h^* = \min_{s \leq 1 \leq u-1} \{t_{l+1} - t_l\}
\]

where \( u = s + m - 1 \) and \( s \geq 0 \). Then, for \( t \in [t_s, t_{u+1}] \) we have

\[
|L_{m-1,j}(t)| \leq \prod_{i=s+1}^{u} \frac{|t - t_i|}{h^*} \leq \prod_{i=s, i \neq j}^{u} h^* \sum_{i=0}^{m-1} h^* c_1^i \leq \left( \frac{c_1^{m+1} - 1}{c_1 - 1} \right)^{m-1} \leq \left( \frac{c_1^8 - 1}{c_1 - 1} \right) \equiv \rho.
\]

Therefore,

\[
|\hat{H}_{m-1,j}(t)| = |(t - t_j)L_{m-1,j}^2(t)| \leq (t_f - t_0) \rho^2.
\]

On the other hand,

\[
|L'_{m-1,j}(t_j)| = \left| \prod_{i=s, i \neq j}^{u} \frac{1}{t_j - t_i} \left( \sum_{i=s,i \neq j}^{u} \prod_{i=s,i \neq j}^{u} (t_j - t_i) \right) \right|
\]

\[
\leq \frac{1}{(h^*)^{m-1}} \left( \sum_{i=s, i \neq j}^{u} \prod_{i=s,i \neq j}^{u} \left[ \sum_{i=0}^{m-1} h^* c_1^i \right] \right)
\]

\[
= \frac{m - 1}{h^*} \left( \frac{c_1^2 - 1}{c_1 - 1} \right)^{m-2} \leq \frac{7}{h^*} \left( \frac{c_1^8 - 1}{c_1 - 1} \right)^6 = \tilde{\rho} \frac{h^*}{h^*}.
\]

Hence,

\[
|H_{2m-1,j}(t)| = |1 - 2(t - t_j)L'_{m-1,j}(t_j)|L_{m-1,j}^2(t)
\]

\[
\leq (1 + 2|t - t_j||L_{m-1,j}(t_j)|)L_{m-1,j}^2(t)
\]

\[
\leq \left[ 1 + 2h^* \left( \frac{c_1^{m+1} - 1}{c_1 - 1} \right) \frac{\tilde{\rho}}{h^*} \right] \theta^2
\]

\[
\leq \left[ 1 + 2\tilde{\rho} \left( \frac{c_1^9 - 1}{c_1 - 1} \right) \right] \theta^2.
\]

Now, it is easy to see that

\[
K_{\text{int}} = \max \left\{ (t_f - t_0) \theta^2, \left[ 1 + 2\tilde{\rho} \left( \frac{c_1^9 - 1}{c_1 - 1} \right) \right] \theta^2 \right\}
\]

is a mesh-independent uniform bound on all the coefficients of the Hermite interpolant.
4.2.2 LIPSCHITZ CONDITION WITH RESPECT TO $y$ FOR THE INTERPOLANT

We can see from (21) that the increment function $\Psi$ of $\eta(t)$ satisfies a Lipschitz condition with respect to the $y$ argument because

1) $\Psi$ is linear in $f$;
2) $f$ is Lipschitz with respect to $y$;
3) the coefficients of $\eta(t)$ are bounded by $K_{\text{int}}$.

4.2.3 CONSISTENCY (INTERPOLANT)

In the case of a smooth $y$, we use the localizing assumption to get that the local error term of the Hermite interpolation/extrapolation at $t^* \in [t_s, t_{u+1}]$ is

$$
y^{(2m)}(c) \frac{(2m)!}{(2m)!} \prod_{i=s}^{u}(t^* - t_i)^2 = O(h^{2m})$$

where $c \in [t_s, t_{u+1}]$ and $h = \max_{s \leq i \leq u} t_{i+1} - t_i$.

The above formula tells us that the local error is of order $2m - 1$ which is greater or equal to $p - 1$ by equation (20). Having the order of the local error of the interpolant greater or equal to $p - 1$ is required for convergence of order $p$. However, we will only be able to prove a theoretical convergence of order $p - 2$ but we try to get the best convergence order numerically.

Next, when $y$ is not smooth, Assumption 1 tells us that HB515 detects all numerically significant discontinuities of order smaller or equal to 3. Upon its detection, HB515DDE includes the discontinuity in the mesh. It then calls the interpolant on intervals $[t_s, t_u]$ such that no tracked discontinuity lies in their interior. Hence, the local error of the Hermite interpolant cannot be greater than $O(h^3)$ giving us a consistency order of at least 2. Finally, since the error is $O(h^3)$ at any point $t \in [t_s, t_{u+1}]$, then the Hermite interpolant has uniform order 2 on $[t_s, t_{u+1}]$.

4.3 RESTARTING THE METHOD

As it is known in the literature, it is very important to restart a linear multistep method after crossing a discontinuity $\xi$. Indeed, the effect of the discontinuity can be disastrous for the reliability of the method when it uses values from both sides of the discontinuity. Hence, after a discontinuity $\xi$ is precisely located (see [13]) and added to the mesh, the solver is restarted and the self-starting DP45 is called on the new integration interval $[\xi, t_f]$. Indeed, DP45 does not use backsteps for its integration so it is less affected than the Hermite–Birkhoff part by a “turbulence” in the behavior of the DDE. Then, DP45 computes enough backsteps so that HB515DDE can take over without feeling the effect of the discontinuity.
Hence, the intervals on which the method is called are usually bounded by two consecutive discontinuities. Therefore, it is worth making sure that there are enough mesh points in each interval \([\xi_{i-1}, \xi_i]\) so that high-order interpolation can be performed within the interval when a delayed value is needed at some \(t \in \map{\xi_{i-1}}{\xi_i}\). Since the maximum number of interpolation points is 8, then the solver is asked to insert enough steps between every two consecutive discontinuities until the number of mesh points in each interval \([\xi_{i-1}, \xi_i]\) is at least 8.

### 4.4 CONVERGENCE

First, we prove the following two lemmas needed in the proof of Theorem 4.

**Lemma 2 (HB515)** Let \(p \in \{5, \ldots, 15\}\) and let \(f = f(t, y(t))\) be defined as in (1) along with the given norm \(\|\cdot\|\) of \(\mathbb{R}^d\). Next, let \(n \in \{p-4, \ldots, N-1\}\) and \(y_{n-p+1}, \ldots, y_n\) be nodal values associated with a mesh \(\Delta_n\). Then, there exists \(\gamma_f > 0\) such that for all \(\hat{f} \in C^0([t_{n-p+4}, t_{n+1}] \times \mathbb{R}^d, \mathbb{R}^d)\) we have

\[
\|\Phi(y_n; f, \Delta_n) - \Phi(y_n; \hat{f}, \Delta_n)\| \leq \gamma_f \sup_{t_{n-p+4} \leq t \leq t_{n+1}, \ y \in \mathbb{R}^d} \|f(t, y(t)) - \hat{f}(t, y(t))\|
\]

where \(y_n = [y_n, \ldots, y_{n-p+4}]^T\) and \(\Phi\) is the increment function defined in (5).

**Proof.** Put \(I = [t_{n-p+4}, t_{n+1}]\). Then,

\[
\|\Phi(y_n; f, \Delta_n) - \Phi(y_n; \hat{f}, \Delta_n)\| \leq \left| b_{11} + b_{12} + b_{13} + \sum_{j=1}^{p-4} \beta_{1j} \right| \sup_{t \in I, \ y \in \mathbb{R}^d} \|f(t, y(t)) - \hat{f}(t, y(t))\|
\]

\[
\leq (p-1)K \sup_{t \in I, \ y \in \mathbb{R}^d} \|f(t, y(t)) - \hat{f}(t, y(t))\|
\]

where \(K\) is the chosen uniform bound for the coefficients of the Hermite–Birkhoff method (4). Hence, taking \(\gamma_f = 14K\) gives the desired bound and ends the proof.

**Lemma 3 (Hermite interpolant)** Let \(p \in \{5, \ldots, 15\}\) and \(m = m(p)\) be the number of interpolation/extrapolation points for the Hermite interpolant associated with the order \(p\). Let \(f = f(t, y(t))\) be defined as in (1) along with the given norm \(\|\cdot\|\) of \(\mathbb{R}^d\). Next, for \(s \geq 0\) let \(y_s, \ldots, y_{s+m-1}\) be nodal values associated with some mesh. Then, there exists \(\gamma_{f, \text{int}} > 0\) such that for all \(\hat{f} \in C^0([t_s, t_{s+m}] \times \mathbb{R}^d, \mathbb{R}^d)\) we have

\[
\|\Psi(y_s, \ldots, y_{s+m-1}; t, f, \Delta') - \Psi(y_s, \ldots, y_{s+m-1}; t, \hat{f}, \Delta')\|
\]

\[
\leq \gamma_{f, \text{int}} \sup_{t_s \leq t \leq t_{s+m}, \ y \in \mathbb{R}^d} \|f(t, y(t)) - \hat{f}(t, y(t))\|
\]

(22)

where \(\Delta' = \{t_s, \ldots, t_{s+m-1}\}\) and \(\Psi\) is the increment function of the Hermite interpolant defined in (21).
Proof. This can be proved using a similar argument as the one of Lemma 2 by noticing the linearity of $\Psi$ in $f$ and using the fact that all coefficients in $\Psi$ are bounded by $K_{\text{int}}$. Thus, taking $\gamma_{f,\text{int}} = 8K_{\text{int}}$ ends the proof because $m \leq 8$.

It must be noted that, in Lemma 3, $m + s - 1$ is always smaller than the index $n$ of the last integrated step because we do not use interpolation points which are not yet computed. Also, the supremum was taken over $t \in [t_s, t_{n+m}]$ instead of $t \in [t_s, t_{s+m+1}]$ to include the extrapolation case which is done within $(t_{s+m-1}, t_{s+m}]$.

Because our solver is constructed to deal efficiently with many types of delays, many special codes were added to make the solver stable and able to adapt itself to the different behaviors of the DDE functions. Hence, for sake of rigor, we make the following assumptions:

1) When integrating on $[t_n, t_{n+h}]$, we ask for $\alpha(t, y(t)) \leq t_{n+1}$ for all $t \in [t_n, t_{n+1}]$. This constraint was also used by Baker et al. as mentioned in [1]. Then, to avoid far extrapolations, we truncate to $t_{n+1}$ any $\alpha(t, y(t))$ which exceeds it. We suppose that these truncations do not affect the order of convergence of the method.

2) Suppose there is a discontinuity at $t_i, i \geq 0$, and that an approximation to $y$ is needed at $t > t_i$ when $t_{i+1}$ is not available. Then, the solver is allowed to extrapolate over $t_i$ i.e. the solver is allowed to use the $m$ points $t_{i-m+1}, \ldots, t_i$ to extrapolate at $t$. We suppose that these extrapolations do not affect the convergence order of the method.

3) Suppose we are integrating over $[t_n, t_{n+1}]$. When a value at $t \in [t_i, t_{i+1}]$ is needed, $i \in \{0, \ldots, n\}$, the $m$-point interpolant does not always use the mesh points $t_{i-m+1}, \ldots, t_i$. Actually, these interpolating points are only chosen when $i = n$ or when there is a tracked discontinuity at $t_i$. Else, the solver picks, depending on their availability, the same number of interpolating points on both sides of $t$. This choice can only increase the accuracy of the approximation of $y$ at $t$. To avoid a long case-by-case analysis of the different ways interpolation/extrapolation is done, we suppose that the solver always picks the interpolating points $t_{i-m+1}, \ldots, t_i$ for any interpolation within $[t_i, t_{i+1}]$ with a local error of at most $O(h^3)$.

We now prove the convergence of the 2-step HB5DDE together with the 3-point Hermite interpolant. Assuming the well-posedness of the problem along with Assumption 1, we state and prove Lemma 4 and Theorem 4 adapted from a general convergence theorem in [2].

**Lemma 4** Let $\zeta(t)$ and $\eta(t)$ be the numerical solutions of the initial value problems

$$z'_{n+1} = f \left(t, z_{n+1}, u(t - \tau(t, z_{n+1}(t)))\right), \quad t_n \leq t \leq t_{n+1}; \quad z_{n+1}(t_n) = z_n,$$

and

$$w'_{n+1} = f \left(t, w_{n+1}, v(t - \tau(t, w_{n+1}(t)))\right), \quad t_n \leq t \leq t_{n+1}; \quad w_{n+1}(t_n) = w_n,$$
respectively, obtained by method (4) with starting values \( z_n = [z_n, z_{n-1}]^T \) and \( w_n = [w_n, w_{n-1}]^T \), and by interpolant (21) with nodal values \( \tilde{z}_n = [z_{n-2}, z_{n-1}, z_n]^T \) and \( \tilde{w}_n = [w_{n-2}, w_{n-1}, w_n]^T \).

Then there exist constants \( P, Q, R, S, T > 0 \), independent of the mesh \( \Delta \), such that

\[
\|z_{n+1} - w_{n+1}\| \leq (1 + h_{n+1}Q)\|z_n - w_n\| + h_{n+1}P \max_{t \leq t_{n+1}} \|u(t) - v(t)\|, \tag{23}
\]

where \( z_{n+1} = [\zeta(t_{n+1}), z_n]^T \), \( w_{n+1} = [\eta(t_{n+1}), w_n]^T \) and \( \| \cdot \| \) is the norm on \( \mathbb{R}^{2d} \) defined by (15) with \( \| \cdot \| = \| \cdot \|_\infty \), and

\[
\max_{t_n \leq t \leq t_{n+1}} \|\zeta(t) - \eta(t)\| \leq (T + h_{n+1}S) \max_{n-2 \leq s \leq n} \|z_s - w_s\| + h_{n+1}R \max_{t \leq t_{n+1}} \|u(t) - v(t)\|. \tag{24}
\]

**Proof.** By subtracting the two formulae (4) for \( z_{n+1} = \zeta(t_{n+1}) \) and \( w_{n+1} = \eta(t_{n+1}) \) we get

\[
z_{n+1} - w_{n+1} = \begin{bmatrix}
\zeta(t_{n+1}) - \eta(t_{n+1}) \\
z_n - w_n
\end{bmatrix}
= \begin{bmatrix}
\alpha_{n,1}(z_n - w_n) + \alpha_{n,2}(z_{n-1} - w_{n-1}) + h_{n+1}(\Phi(z_n; f_u, \Delta_n) - \Phi(w_n; f_v, \Delta_n)) \\
z_n - w_n
\end{bmatrix}
= (C_n \otimes I_d)(z_n - w_n) + h_{n+1}\Gamma_n,
\]

where \( \Gamma_n = [\Phi(z_n; f_u, \Delta_n) - \Phi(w_n; f_v, \Delta_n), 0]^T \) with \( f_u(t, y) = f(t, y, u(t - \tau(t, y(t)))) \)

and \( f_v(t, y) = f(t, y, v(t - \tau(t, y(t)))) \).

Next, using the fact that \( \|C_n \otimes I_d\| \leq 1 \), we get

\[
\|z_{n+1} - w_{n+1}\| \leq \|z_n - w_n\| + h_{n+1}\|\Gamma_n\|.
\]

Hence, by equivalence of the norms on finite dimensional spaces, there exists \( C > 0 \) such that

\[
\|\Gamma_n\| \leq C\|\Phi(z_n; f_u, \Delta_n) - \Phi(w_n; f_v, \Delta_n)\|
\leq C\left(\|\Phi(z_n; f_u, \Delta_n) - \Phi(w_n; f_u, \Delta_n)\| + \|\Phi(w_n; f_u, \Delta_n) - \Phi(w_n; f_v, \Delta_n)\|\right).
\]

Then, using the fact that \( \Phi \) is Lipschitz with respect to the \( y \) argument, Lemma 2 and the fact that \( f \) is Lipschitz with respect to its third argument, then there exist \( Q > 0 \), \( \gamma_f > 0 \) and \( P > 0 \) such that

\[
\|\Gamma_n\| \leq Q\|z_n - w_n\| + C_S \sup_{t_n \leq t \leq t_{n+1}, \ y \in \mathbb{R}^d} \|f_u(t, y) - f_v(t, y)\|
\leq Q\|z_n - w_n\| + P \sup_{t_n \leq t \leq t_{n+1}, \ y \in \mathbb{R}^d} \|u(t - \tau(t, y(t))) - v(t - \tau(t, y(t)))\|.
\]

Now, when integrating on \([t_n, t_{n+1}]\), we constrain the solver to give \( t - \tau(t, y(t)) \leq t_{n+1} \) for all \( t \in [t_n, t_{n+1}] \). Then, we get (23).
As for the continuous extensions, if we put $\Delta'_{n} = \{t_{n-2}, t_{n-1}, t_{n}\}$ then we have, for all $t = t_{n} + \theta h_{n+1}$ with $\theta \in [0, 1]$, 
\[
\zeta(t) - \eta(t) = \sum_{j=n-2}^{n} H_{2,j}(t)(z_{j} - w_{j}) + h_{n+1}(\Psi(\tilde{z}_{n}; \theta, f_{u}, \Delta'_{n}) - \Psi(\tilde{w}_{n}; \theta, f_{v}, \Delta'_{n})).
\]
Therefore, putting $R > 0$ and $S > 0$ which gives (24) after we put $\max_{t_{n} \leq t \leq t_{n+1}} \|\zeta(t) - \eta(t)\|$ gives
\[
\max_{t_{n} \leq t \leq t_{n+1}} \|\zeta(t) - \eta(t)\| \leq 3K_{\text{int}}D_{n} + h_{n+1} \max_{t_{n} \leq t \leq t_{n+1}} \|\Psi(\tilde{z}_{n}; \theta, f_{u}, \Delta'_{n}) - \Psi(\tilde{w}_{n}; \theta, f_{v}, \Delta'_{n})\| \leq 3K_{\text{int}}D_{n} + h_{n+1} \max_{t_{n} \leq t \leq t_{n+1}} \|\zeta(t) - \eta(t)\| + \|\Psi(\tilde{w}_{n}; \theta, f_{u}, \Delta'_{n}) - \Psi(\tilde{w}_{n}; \theta, f_{v}, \Delta'_{n})\|.
\]

Now, using the fact that $\Psi$ is Lipschitz with respect to the $y$ and $f$ arguments and that $f$ is Lipschitz with respect to its third argument, there exist constants $R > 0$ and $S > 0$ such that
\[
\max_{t_{n} \leq t \leq t_{n+1}} \|\zeta(t) - \eta(t)\| \leq (3K_{\text{int}} + h_{n+1}S)D_{n} + h_{n+1}R \max_{t_{n} \leq t \leq t_{n+1}} \|u(t - \tau(t, w_{n+1}(t))) - v(t - \tau(t, w_{n+1}(t)))\|
\]
which gives (24) after we put $T = 3K_{\text{int}}$ and use the fact that $t - \tau(t, y(t)) \leq t_{n+1}$ for all $t \in [t_{n}, t_{n+1}]$.

**Theorem 4** Consider the state dependent DDE
\[
\begin{align*}
\begin{cases}
y' = f(t, y(t), y(t - \tau(t, y(t)))), & \quad t_{0} \leq t \leq t_{f}, \\
y(t) = \phi(t), & \quad r \leq t \leq t_{0},
\end{cases}
\end{align*}
\] (25)
where $f \in C^{3}([t_{i}, t_{i+1}] \times \mathbb{R}^{d} \times \mathbb{R}^{d})$ for all $i = 0, \ldots, N - 1$ with the mesh being $\Delta = \{t_{0}, t_{1}, \ldots, t_{n}, \ldots, t_{N} = t_{f}\}$, the delay $\tau \in C^{3}([t_{0}, t_{f}] \times \mathbb{R}^{d})$ and $\phi \in C^{3}([r, t_{0}])$. Moreover, we have that

(a) The mesh $\Delta$ includes all discontinuity points $\xi_{1} < \xi_{2} < \cdots < \xi_{s}$ of order smaller or equal to 3.

(b) The ODE method HB5DDE is restarted after each discontinuity point $\xi_{i}$, $i = 0, 1, \ldots, s$, by a method of order greater or equal to 2.

(c) The 3-point Hermite interpolant is consistent of uniform order 2.

(d) For each $n$, the interval where the interpolation takes place is included in $[\xi_{i}, \xi_{i+1}]$ for some $i \in \{0, \ldots, s\}$.
Then the resulting DDE method has discrete global order and uniform global order 3, that is

\[
\max_{1 \leq n \leq N} \|y(t_n) - y_n\| = O(h^3)
\]

and

\[
\max_{t_0 \leq t \leq t_f} \|y(t) - \eta(t)\| = O(h^3),
\]

where \( h = \max_{1 \leq n \leq N} h_n \).

**Proof.** In connection to the DDE (25), for each \( n = 1, \ldots, N - 1 \), consider the local ODE

\[
\begin{align*}
\dot{z}_{n+1}(t) &= f(t, z_{n+1}(t), y(t - \tau(t, z_{n+1}(t)))), \quad t_n \leq t \leq t_{n+1}, \quad z_{n+1}(t_n) = y(t_n), \\
y(t) &= \phi(t), \quad t \leq t_0,
\end{align*}
\]

whose solution evidently is \( z_{n+1} = y(t) \), i.e. the solution of (25). Moreover, consider the auxiliary problem

\[
\begin{align*}
\dot{w}_n(t) &= f(t, w_{n+1}(t), \eta(t - \tau(t, w_{n+1}(t)))), \quad t_n \leq t \leq t_{n+1}, \\
z_{n+1}(t_n) &= \eta(t_n), \\
\eta(t) &= \phi(t), \quad t \leq t_0,
\end{align*}
\]

where, for \( s \leq t_{n+1} \), \( \eta(s) \) is the continuous numerical solution given by the DDE method itself. Then define \( y_n = [y(t_n), y(t_{n-1})]^T \) and \( \eta_n = [\eta(t_n), \eta(t_{n-1})]^T \). By Lemma 4 with \( u(x) = y(x) \), \( v(x) = \eta(x) \), \( z_n = y_n \) and \( w_n = \eta_n \), the numerical solutions \( \zeta(t) \) and \( \eta(t) \) of (26) and (27), respectively, satisfy the inequality

\[
\|z_{n+1} - \eta_{n+1}\| \leq (1 + h_{n+1}Q)\|y_n - \eta_n\| + h_{n+1}P \max_{t \leq t_{n+1}} \|y(t) - \eta(t)\|,
\]

where \( z_{n+1} = [\zeta(t_{n+1}), y(t_n)]^T \), and

\[
\max_{t_n \leq t \leq t_{n+1}} \|\zeta(t) - \eta(t)\| \leq (T + h_{n+1}S) \max_{n-2 \leq s \leq n} \|y(s) - \eta(s)\| + h_{n+1}R \max_{t \leq t_{n+1}} \|y(t) - \eta(t)\|.
\]

Now consider the inequality

\[
\|y_{n+1} - \eta_{n+1}\| \leq \|y_{n+1} - z_{n+1}\| + \|z_{n+1} - \eta_{n+1}\|.
\]

By hypothesis (a), the solution \( z_{n+1}(t) \) of (26) is at least 4-times differentiable on \( [t_n, t_{n+1}] \). Therefore, by hypothesis (b), (28) yields

\[
\|y_{n+1} - \eta_{n+1}\| \leq M_1 h_{n+1}^4 + (1 + h_{n+1}Q)\|y_n - \eta_n\| + h_{n+1}P \max_{t \leq t_{n+1}} \|y(t) - \eta(t)\|.
\]

Therefore, with

\[
e_n = \max_{1 \leq i \leq n} \|y_i - \eta_i\|
\]

and

\[
E_n = \max_{t \leq t_{n+1}} \|y(t) - \eta(t)\|
\]

Therefore, with

\[
e_n = \max_{1 \leq i \leq n} \|y_i - \eta_i\|
\]

and

\[
E_n = \max_{t \leq t_{n+1}} \|y(t) - \eta(t)\|
\]
for $n = 1, \ldots, N$, we obtain

$$e_{n+1} \leq M_{1}h_{n+1}^4 + (1 + h_{n+1}Q)e_{n} + h_{n+1}PE_{n+1} \quad (31)$$

for $n = 1, \ldots, N - 1$. Similarly, for the interpolant consider the inequality

$$\max_{t_{n} \leq t \leq t_{n+1}} \| y(t) - \eta(t) \| \leq \max_{t_{n} \leq t \leq t_{n+1}} \| y(t) - \zeta(t) \| + \max_{t_{n} \leq t \leq t_{n+1}} \| \zeta(t) - \eta(t) \|.$$

By the smoothness of $y(t)$, hypotheses (c) and (d), and (29), we get

$$\max_{t_{n} \leq t \leq t_{n+1}} \| y(t) - \eta(t) \| \leq M_{2}h_{n+1}^3 + (T + h_{n+1}S) \max_{n-2 \leq s \leq n} \| y(t_{s}) - \eta(t_{s}) \|$$

$$+ h_{n+1}R \max_{t \leq t_{n+1}} \| y(t) - \eta(t) \|.$$

Thus, since there exists a constant $K > 0$ such that

$$\| y(t_{s}) - \eta(t_{s}) \| \leq K \| y_{s} - \eta_{s} \|,$$

for all $s$, we obtain

$$\max_{t_{n} \leq t \leq t_{n+1}} \| y(t) - \eta(t) \| \leq M_{2}h_{n+1}^3 + (T + h_{n+1}S)Ke_n + h_{n+1}RE_{n+1} \quad (32)$$

for $n = 1, \ldots, N - 1$. With $L = \max\{M_{1}, M_{2}, P, Q, R, TK, SK\}$ the inequalities (31) and (32) yield

$$e_{n+1} \leq (1 + h_{n+1}L)e_{n} + h_{n+1}LE_{n+1} + h_{n+1}Lh^3 \quad (33)$$

and

$$\max_{t_{n} \leq t \leq t_{n+1}} \| y(t) - \eta(t) \| \leq (1 + h)Le_{n} + hLE_{n+1} + Lh^3 \quad (34)$$

for $n = 1, \ldots, N - 1$. Since both $e_{n}$ and $E_{n}$ are monotone increasing, (34) implies

$$E_{n+1} \leq (1 + h)Le_{n} + hLE_{n+1} + Lh^3 \quad (35)$$

and hence, for $h < 1/L$, we have

$$E_{n+1} \leq \frac{(1 + h)L}{1 - hL}e_{n} + \frac{L}{1 - hL}h^3, \quad (36)$$

for $n = 1, \ldots, N-1$. Now assume, without any restriction, that $h \leq \min\{1, 1/(2L)\}$, and define $\Lambda = 2(L + L^2)$. By substituting (36) into (33), we get

$$e_{n+1} \leq \left[ 1 + h_{n+1} \left( L + \frac{(1 + h)L^2}{1 - hL} \right) \right] e_{n} + h_{n+1} \left( L + \frac{L^2}{1 - hL} \right) h^3$$

$$\leq (1 + \Lambda h_{n+1})e_{n} + h_{n+1}\Lambda h^3$$

$$\leq e^{\Lambda h_{n+1}}e_{n} + h_{n+1}\Lambda h^3 \quad (37)$$

26
for $n = 1, \ldots, N - 1$. Now we have

$$e_n \leq e^{\Lambda(t_n - t_1)} e_1 + \left( \sum_{i=2}^{n} e^{\Lambda(t_n - t_{i-1})} h_i \right) \Lambda h^3$$

$$\leq e^{\Lambda(t_f - t_1)} e_1 + e^{\Lambda f} \left( \int_{t_1}^{t_f} e^{-\Lambda t} dt \right) \Lambda h^3,$$

$$\leq e^{\Lambda(t_f - t_1)} e_1 + \left( e^{\Lambda(t_f - t_1)} - 1 \right) h^3,$$

and hence, since hypotheses (b), (c) and (d) imply $e_1 = O(h^3)$, the proof is complete.

Next, it is possible to adapt Theorem 4 to prove that DP45 together with a 3-point Hermite interpolant is convergent of global, discrete and uniform, (convergence) orders 3. Hence, the DP45-HB5DDE solver together with a 3-point Hermite interpolant is convergent of global, discrete and uniform, (convergence) orders 3. Again, note that variable-order does not lower the convergence order because the local error of any order $p$ greater than 5 can only be smaller than the one associated with order 5.

Hence, under the given assumptions, the combined DP45-HB515DDE with an $m$-point Hermite interpolant is convergent of global, discrete and uniform, (convergence) orders 3. Moreover, it is easy to see that we cannot get a higher convergence order because the convergence considers the error over all integration steps and hence the smallest order is the dominant one.

If we want to disregard Assumption 1 (and the location of the discontinuities as a whole), then it is possible to prove the following theorem adapted from [2] which says that we can get an approximate solution accurate to at least order 2.

**Theorem 5** If equation (19) has a smooth solution apart from a finite number of discontinuities of order 1, then the ODE method DP45-HB515 along with the $m$-point Hermite interpolant furnishes an approximate solution of uniform global order at least 2 for any choice of the mesh.

**ACKNOWLEDGMENTS**

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada and the Centre de recherches mathématiques of the Université de Montréal.

**References**


Yagoub H., M.Sc.
University of Ottawa, Department of Mathematics and Statistics
Address: 585 King Edward Ave., Ottawa, Ontario, K1N 6N5, Canada
Phone: 1-613-562-5864
Email: hy.emails@gmail.com

Nguyen-Ba T., Ph.D.
University of Ottawa, Department of Mathematics and Statistics
Address: 585 King Edward Ave., Ottawa, Ontario, K1N 6N5, Canada
Phone: 1-613-562-5864
Email: tnguyen@mathstat.uottawa.ca

Giordano., Ph.D.
University of Ottawa, Department of Mathematics and Statistics
Address: 585 King Edward Ave., Ottawa, Ontario, K1N 6N5, Canada
Phone: 1-613-562-5864
Email: giordano@uottawa.ca

Vaillancourt R., Ph.D.
University of Ottawa, Department of Mathematics and Statistics
Address: 585 King Edward Ave., Ottawa, Ontario, K1N 6N5, Canada
Phone: 1-613-562-5864
Email: remi@uottawa.ca