Chapter 5

Blended HBVMs

We shall now consider some computational aspects concerning HBVM(k, s). In more details, we now show how its cost depends essentially on s, rather than on k, in the sense that the nonlinear system to be solved, for obtaining the discrete solution, has (block) dimension s [3, 6, 8].

This could be inferred from the fact that the silent stages (1.11) depend on the fundamental stages: let us see the details. In order to simplify the notation, we shall fix the fundamental stages at $\tau_1, \ldots, \tau_s$, since we have already seen that, due to the use of an orthonormal basis, they could be in principle chosen arbitrarily, among the abscissae $\{\tau_i\}$. With this premise, we have, from (1.9), (1.17)–(1.18), and by using the notation (1.21),

$$y_i = y_0 + h \sum_{j=1}^{s} a_{ij} \sum_{\ell=1}^{k} \omega_\ell P_j(\tau_\ell) f_\ell, \qquad i = 1, \ldots, s. \tag{5.1}$$

This equation is now coupled with that defining the silent stages, i.e., from (1.6) and (1.11),

$$y_i = y_0 + h \sum_{j=1}^{s} \gamma_j \int_0^{\tau_i} P_j(t) \mathrm{d}t, \qquad i = s+1, \ldots, k. \tag{5.2}$$

Let us now partition the matrices $\mathcal{I}_s, \mathcal{P}_s \in \mathbb{R}^{k\times s}$ in (1.24)–(1.25) into

$$\mathcal{I}_{s1}, \mathcal{P}_{s1} \in \mathbb{R}^{s\times s}, \qquad \mathcal{I}_{s2}, \mathcal{P}_{s2} \in \mathbb{R}^{k-s\times s},$$

containing the entries defined by the fundamental abscissae and the silent abscissae, respectively. Similarly, we partition the vector $\boldsymbol{y}$ into $\boldsymbol{y}_1$, containing the fundamental stages, and $\boldsymbol{y}_2$ containing the silent stages and, accordingly, let

$$\Omega_1 \in \mathbb{R}^{s\times s}, \qquad \Omega_2 \in \mathbb{R}^{k-s\times k-s},$$

be the diagonal matrices containing the corresponding entries in matrix Ω. Finally, let us define the vectors

$$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_s)^T, \qquad e = (1, \ldots, 1)^T \in \mathbb{R}^s, \qquad u = (1, \ldots, 1)^T \in \mathbb{R}^{k-s}.$$

Consequently, we can rewrite (5.1) and (5.2), as

$$\begin{aligned} \boldsymbol{y}_1 &= e \otimes y_0 + h\mathcal{I}_{s1} \left(\mathcal{P}_{s1}^T \; \mathcal{P}_{s2}^T\right) \begin{pmatrix} \Omega_1 & \\ & \Omega_2 \end{pmatrix} \otimes I_{2m} \begin{pmatrix} f(\boldsymbol{y}_1) \\ f(\boldsymbol{y}_2) \end{pmatrix}, \quad (5.3) \\ \boldsymbol{y}_2 &= u \otimes y_0 + h\mathcal{I}_{s2} \otimes I_{2m}\boldsymbol{\gamma}, \quad (5.4) \end{aligned}$$

respectively. The vector $\boldsymbol{\gamma}$ can be obtained by the identity (see (1.16))

$$\boldsymbol{y}_1 = e \otimes y_0 + h\mathcal{I}_{s1} \otimes I_{2m}\boldsymbol{\gamma},$$

thus giving

$$\begin{aligned} \boldsymbol{y}_2 &= \left(u - \mathcal{I}_{s2}\mathcal{I}_{s1}^{-1}e\right) \otimes y_0 + \mathcal{I}_{s2}\mathcal{I}_{s1}^{-1} \otimes I_{2m}\boldsymbol{y}_1 \\ &\equiv \hat{u} \otimes y_0 + A_1 \otimes I_{2m}\boldsymbol{y}_1, \quad (5.5) \end{aligned}$$

in place of (5.4), where, evidently,

$$\hat{u} = \left(u - \mathcal{I}_{s2}\mathcal{I}_{s1}^{-1}e\right) \in \mathbb{R}^{k-s}, \qquad A_1 = \mathcal{I}_{s2}\mathcal{I}_{s1}^{-1} \in \mathbb{R}^{k-s\times s}. \quad (5.6)$$

By setting

$$B_1 = \mathcal{I}_{s1}\mathcal{P}_{s1}^T\Omega_1 \in \mathbb{R}^{s\times s}, \qquad B_2 = \mathcal{I}_{s1}\mathcal{P}_{s2}^T\Omega_2 \in \mathbb{R}^{s\times k-s}, \quad (5.7)$$

substitution of (5.5) into (5.3) then provides, at last, the system of block size s to be actually solved:

$$\begin{aligned} F(\boldsymbol{y}_1) \equiv\; & \boldsymbol{y}_1 - e \otimes y_0 - h\left[B_1 \otimes I_{2m}f(\boldsymbol{y}_1) + \right. \quad (5.8) \\ & \left. B_2 \otimes I_{2m}f\left(\hat{u} \otimes y_0 + A_1 \otimes I_{2m}\boldsymbol{y}_1\right)\right] = \mathbf{0}. \end{aligned}$$

By using the simplified Newton method for solving (5.8), and setting

$$C = B_1 + B_2A_1 \in \mathbb{R}^{s\times s}, \quad (5.9)$$

one obtains the iteration:

$$\begin{aligned} \left(I_s \otimes I_{2m} - hC \otimes J_0\right)\boldsymbol{\delta}^{(n)} &= -F(\boldsymbol{y}_1^{(n)}) \equiv \boldsymbol{\psi}_1^{(n)}, \quad (5.10) \\ \boldsymbol{y}_1^{(n+1)} &= \boldsymbol{y}_1^{(n)} + \boldsymbol{\delta}^{(n)}, \qquad n = 0, 1, \ldots, \end{aligned}$$

where J_0 is the Jacobian of $f(y)$ evaluated at y_0. Because of the result of Theorem 4, the following property of matrix C holds true [8].

Theorem 6. *The eigenvalues of matrix C in (5.9) coincide with those of matrix (4.2), i.e., with the eigenvalues of the matrix of the Butcher array of the Gauss-Legendre method of order $2s$.*

Proof Assuming, as usual for simplicity, that the fundamental stages are the first s ones, one has that the discrete problem

$$\boldsymbol{y} = \begin{pmatrix} e \\ u \end{pmatrix} \otimes y_0 + hA \otimes I_{2m} f(\boldsymbol{y}),$$

which defines the Runge-Kutta formulation of the method, is equivalent, by virtue of (5.3), (5.5), (5.6), (5.7), to

$$\begin{pmatrix} I_s & O_{s\times r} \\ -A_1 & I_r \end{pmatrix} \otimes I_{2m} \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{pmatrix} e \\ \hat{u} \end{pmatrix} \otimes y_0 + h \begin{pmatrix} B_1 & B_2 \\ O_{r\times s} & O_{r\times r} \end{pmatrix} \otimes I_{2m} \begin{pmatrix} f(\boldsymbol{y}_1) \\ f(\boldsymbol{y}_2) \end{pmatrix},$$

where, as usual, $r = k - s$. Consequently, the eigenvalues of the matrix A defined in (4.1) coincides with those of the pencil

$$\left(\begin{pmatrix} I_s & O_{s\times r} \\ -A_1 & I_r \end{pmatrix}, \begin{pmatrix} B_1 & B_2 \\ O_{r\times s} & O_{r\times r} \end{pmatrix} \right). \tag{5.11}$$

That is,

$$\mu \in \sigma(A) \;\Leftrightarrow\; \mu \begin{pmatrix} I_s & O_{s\times r} \\ -A_1 & I_r \end{pmatrix} \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} = \begin{pmatrix} B_1 & B_2 \\ O_{r\times s} & O_{r\times r} \end{pmatrix} \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix},$$

for some nonzero vector $(\boldsymbol{u}^T, \boldsymbol{v}^T)^T$. By setting $\boldsymbol{u} = \boldsymbol{0}$, one obtains the r zero eigenvalues of the pencil. For the remaining s (nonzero) ones, it must be $\boldsymbol{v} = A_1\boldsymbol{u}$, so that:

$$\mu\boldsymbol{u} = (B_1\boldsymbol{u} + B_2\boldsymbol{v}) = (B_1\boldsymbol{u} + B_2A_1\boldsymbol{u}) = C\boldsymbol{u} \;\Leftrightarrow\; \mu \in \sigma(C). \square$$

Remark 10. *From the result of Theorem 6, it follows that the spectrum of C doesn't depend on the choice of the s fundamental abscissae, within the nodes $\{\tau_i\}$. On the contrary, its condition number does: the latter appears to be minimized when the fundamental abscissae are symmetrically distributed and approximately evenly spaced in the interval $[0, 1]$. As a practical "*rule of thumb*", the following algorithm appears to be almost optimal:*

1. *let the k abscissae $\{\tau_i\}$ be chosen according to a Gauss-Legendre distribution of k nodes;*

2. *then, let us consider s equidistributed nodes in $(0,1)$, say $\{\hat\tau_1,\dots,\hat\tau_s\}$;*
3. *select, as the fundamental abscissae, those nodes among the $\{\tau_i\}$ which are the closest ones to the $\{\hat\tau_j\}$;*
4. *define matrix C in (5.9) accordingly.*

Clearly, for the above algorithm to provide a unique solution (resulting in a symmetric choice of the fundamental abscissae), the difference $k-s$ has to be even which, however, can be easily accomplished.

In order to give evidence of the effectiveness of the above algorithm, in Figure 5.1 we plot the condition number of matrix $C = C(k,s)$, for $s = 2,\dots,5$, and $k \geq s$. As one can see, the condition number of $C(k,s)$ turns out to be nicely bounded, for increasing values of k, which makes the implementation (that we are going to analyze in the next section) effective also when finite precision arithmetic is used. For comparison, in Figure 5.2 there is the same plot, obtained by fixing the fundamental abscissae as the first s ones. In such a case, the condition number of $C(k,s)$ grows very fast, as k is increased.

5.1 Blended implementation

We observe that, since C is nonsingular, we can recast problem (5.10) in the *equivalent form*

$$\gamma\left(C^{-1}\otimes I_{2m} - hI_s\otimes J_0\right)\boldsymbol{\delta}^{(n)} = -\gamma C^{-1}\otimes I_{2m}\, F(\boldsymbol{y}_1^{(n)}) \equiv \boldsymbol{\psi}_2^{(n)}, \qquad (5.12)$$

where $\gamma > 0$ is a free parameter to be chosen later. Let us now introduce the *weight (matrix) function*

$$\theta = I_s\otimes\Phi^{-1}, \qquad \Phi = I_{2m} - h\gamma J_0 \in \mathbb{R}^{2m\times 2m}, \qquad (5.13)$$

and the *blended formulation* of the system to be solved,

$$\begin{aligned} M\boldsymbol{\delta}^{(n)} &\equiv [\theta\,(I_s\otimes I_{2m} - hC\otimes J_0) + \\ &\quad (I-\theta)\gamma\left(C^{-1}\otimes I_{2m} - hI_s\otimes J_0\right)]\,\boldsymbol{\delta}^{(n)} \\ &= \theta\boldsymbol{\psi}_1^{(n)} + (I-\theta)\boldsymbol{\psi}_2^{(n)} \equiv \boldsymbol{\psi}^{(n)}. \end{aligned} \qquad (5.14)$$

The latter system has again the same solution as the previous ones, since it is obtained as the *blending*, with weights θ and $(I-\theta)$, of the two equivalent forms (5.10) and (5.12). For iteratively solving (5.14), we use the corresponding *blended iteration*, formally given by [2, 10, 11, 12, 13, 14, 15, 16, 18, 30, 32]:

$$\boldsymbol{\delta}^{(n,\ell+1)} = \boldsymbol{\delta}^{(n,\ell)} - \theta\left(M\boldsymbol{\delta}^{(n,\ell)} - \boldsymbol{\psi}^{(n)}\right), \qquad \ell = 0,1,\dots. \qquad (5.15)$$

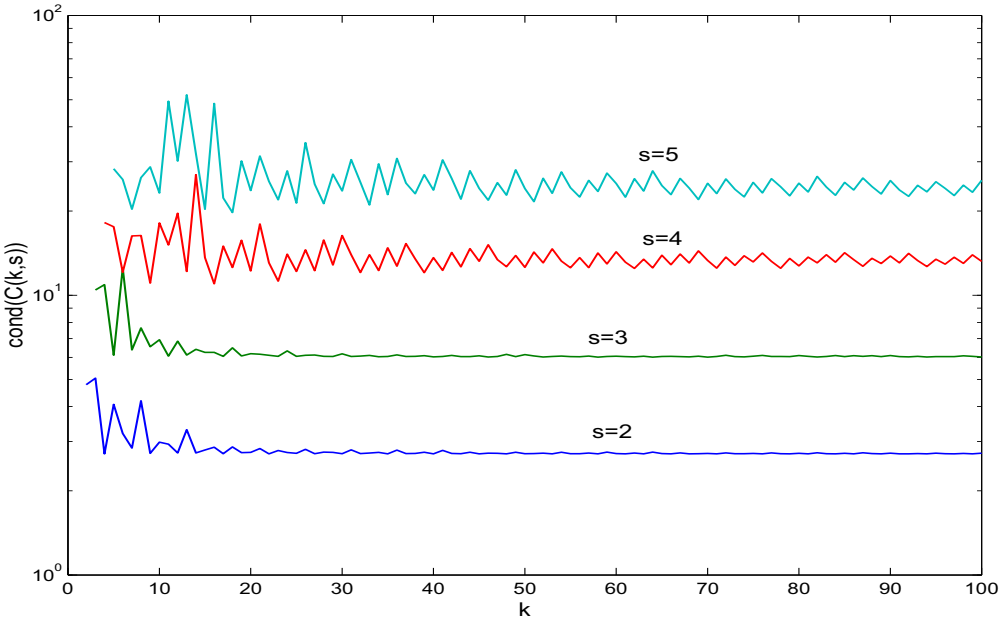

Figure 5.1: Condition number of the matrix $C = C(k, s)$, for $s = 2, 3, 4, 5$ and $k = s, s + 1, \ldots, 100$, with the fundamental abscissae chosen according to the algorithm sketched in Remark 10.

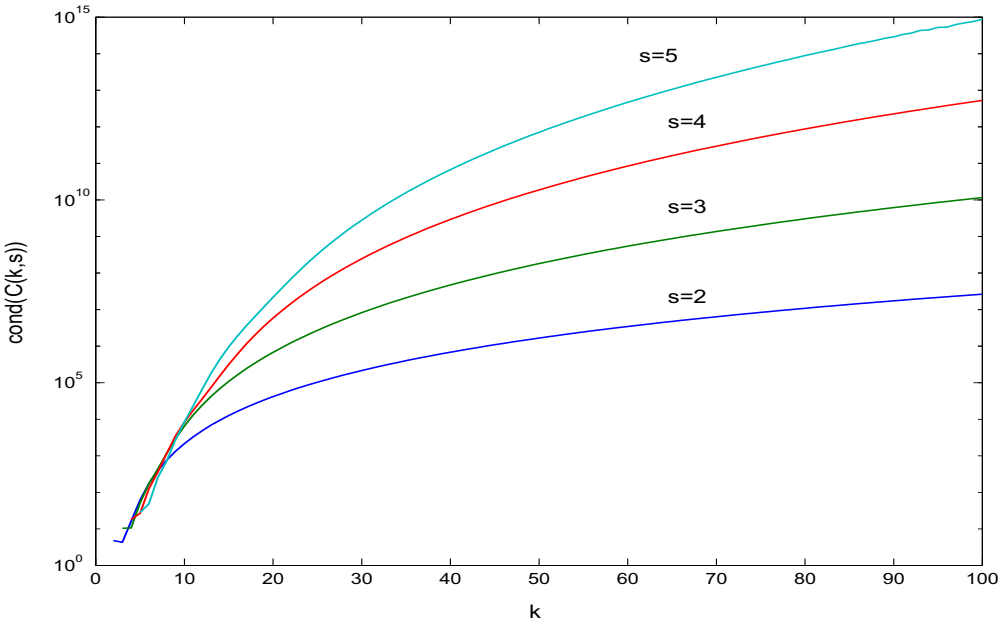

Figure 5.2: Condition number of the matrix $C = C(k, s)$, for $s = 2, 3, 4, 5$ and $k = s, s + 1, \ldots, 100$, with the fundamental abscissae chosen as the first s ones.

Remark 11. *A nonlinear variant of the iteration (5.15) can be obtained, by starting at* $\boldsymbol{\delta}^{(n,0)} = \mathbf{0}$ *and updating* $\boldsymbol{\psi}^{(n)}$ *as soon as a new approximation is available. This results in the following iteration:*

$$\boldsymbol{y}^{(n+1)} = \boldsymbol{y}^{(n)} + \theta\boldsymbol{\psi}^{(n)}, \qquad n = 0, 1, \ldots. \tag{5.16}$$

Remark 12. *We observe that, for actually performing the iteration (5.13)–(5.15), as well as (5.16), one has to factor* only *the matrix* Φ *in (5.13), which has the same size as that of the continuous problem.*

We end this section by observing that the above iterations (5.15) and (5.16) depend on a free parameter γ. It will be chosen in order to optimize the convergence properties of the iteration, according to a linear analysis of convergence, which is sketched in the next section.

5.2 Linear analysis of convergence

The linear analysis of convergence for the iteration (5.15) is carried out by considering the usual scalar test equation (see, e.g., [14] and the references therein),

$$y' = \lambda y, \qquad \Re(\lambda) < 0.$$

By setting, as usual $q = h\lambda$, the two equivalent formulations (5.10) and (5.12) become, respectively (omitting, for sake of brevity, the upper index n),

$$(I_s - qC)\boldsymbol{\delta} = \boldsymbol{\psi}_1, \qquad \gamma(C^{-1} - qI_s)\boldsymbol{\delta} = \boldsymbol{\psi}_2.$$

Moreover,

$$\theta = \theta(q) = (1 - \gamma q)^{-1} I_s, \tag{5.17}$$

and the blended iteration (5.15) becomes

$$\boldsymbol{\delta}^{(\ell+1)} = (I_s - \theta(q)M(q))\boldsymbol{\delta}^{(\ell)} + \theta(q)\boldsymbol{\psi}(q), \tag{5.18}$$

with

$$\begin{aligned} M(q) &= \theta(q)\,(I_s - qC) + (I_s - \theta(q))\gamma\left(C^{-1} - qI_s\right), \\ \boldsymbol{\psi}(q) &= \theta(q)\boldsymbol{\psi}_1 + (I_s - \theta(q))\boldsymbol{\psi}_2. \end{aligned} \tag{5.19}$$

Consequently, the iteration will be convergent if and only if the spectral radius $\rho(q)$ of the iteration matrix,

$$Z(q) = I_s - \theta(q)M(q), \tag{5.20}$$

is less than 1. The set

$$\Gamma = \{q \in \mathbb{C} \,:\, \rho(q) < 1\}$$

is the *region of convergence of the iteration*. The iteration is said to be:

Table 5.1: optimal values (5.23), and corresponding maximum amplification factors (5.24), for various values of s.

s	γ	ρ^*
2	0.2887	0.1340
3	0.1967	0.2765
4	0.1475	0.3793
5	0.1173	0.4544
6	0.0971	0.5114
7	0.0827	0.5561
8	0.0718	0.5921
9	0.0635	0.6218
10	0.0568	0.6467

- A-convergent, if $\mathbb{C}^- \subseteq \Gamma$;
- L-convergent, if it is A-convergent and, moreover, $\rho(q) \to 0$, as $q \to \infty$.

For the iteration (5.18) one verifies that (see (5.17), (5.19), and (5.20))

$$Z(q) = \frac{q}{(1-\gamma q)^2} C^{-1} \left(C - \gamma I_s\right)^2, \tag{5.21}$$

which is the null matrix at $q = 0$ and at ∞. Consequently, the iteration will be A-convergent (and, therefore, L-convergent), provided that *maximum amplification factor*,

$$\rho^* \equiv \max_{\Re(q)=0} \rho(q) \ \leq 1. \tag{5.22}$$

From (5.21) one has that, by setting hereafter $\sigma(C)$ the spectrum of matrix C,

$$\mu \in \sigma(C) \ \Leftrightarrow \ \frac{q(\mu-\gamma)^2}{\mu(1-\gamma q)^2} \in \sigma(Z(q)).$$

By taking into account that

$$\max_{\Re(q)=0} \frac{|q|}{|(1-\gamma q)^2|} = \frac{1}{2\gamma},$$

one then obtains that

$$\rho^* = \max_{\mu\in\sigma(C)} \frac{|\mu-\gamma|^2}{2\gamma|\mu|},$$

For Gauss-Legendre methods (and, then, for any matrix C having the same spectrum), it can be shown that (see [10, 16]) the choice

$$\gamma = |\mu_{\min}| \equiv \min_{\mu\in\sigma(C)} |\mu|, \tag{5.23}$$

minimizes ρ^*, which turns out to be given by

$$\rho^* = 1 - \cos\varphi_{\min} \ < 1, \qquad \varphi_{\min} = \mathrm{Arg}(\mu_{\min}). \tag{5.24}$$

In Table 5.1, we list the optimal value of the parameter γ, along with the corresponding maximum amplification factor ρ^*, for various values of s, which confirm that the iteration (5.18) is L-convergent.

Remark 13. *We then conclude that the* blended iteration *(5.15) turns out to be L-convergent, for any HBVM(k,s) method, for all $s \geq 1$ and $k \geq s$.*

We end this chapter, by emphasizing that the property of L-convergence has proved to be computationally very effective, as testified by the successful implementation of the codes `BiM` and `BiMD` [30, 32]. We then expect good performances also for the *blended implementation* of HBVM(k,s).