Perturbation results for exponential Rosenbrock-type methods

J. Schweitzer

ETH Zürich

Abstract. In this note we consider exponential Rosenbrock-type methods that have been introduced in [1, 2, 3]. Since these methods involve several terms, which can not be computed exactly in praxis it is interesting to understand, how additional errors in the method affect the overall accuracy of the scheme. We will give a perturbation result for general perturbations and study the effect of inexact evaluation of the Jacobian.

Keywords: exponential Rosenbrock-type method, perturbed method, inexact Jacobian, error analysis
PACS: 02.60.Cb, 02.60.Lj

EXPONENTIAL ROSENBROCK-TYPE METHODS

In [2] a new class of numerical methods for the time integration of large systems of stiff or oscillatory differential equations

$$u'(t) = F(u(t)), \qquad u(t_0) = u_0 \ . \tag{1}$$

is proposed. The scheme reads

$$U_{ni} = e^{c_i \tau_n J_n} u_n + \tau_n \sum_{j=1}^{i-1} a_{ij}(\tau_n J_n) g_n(U_{nj}), \qquad 1 \le i \le s, \tag{2a}$$

$$u_{n+1} = e^{\tau_n J_n} u_n + \tau_n \sum_{i=1}^{s} b_i(\tau_n J_n) g_n(U_{ni}), \tag{2b}$$

with the Jacobian evaluated at the numerical solution at the current time t_n, where $t_{n+1} = t_n + \tau_n$, $n = 0, 1, 2, \ldots$,

$$J_n = \mathrm{D}F(u_n) = \frac{\partial F}{\partial u}(u_n) \quad \text{and} \quad g_n(u(t)) = F(u(t)) - J_n u(t) \ .$$

The coefficients a_{ij} and b_i of the method are linear combinations of entire functions related to the exponential function. In [2] stiff order conditions were derived and an error and stability analysis for the unperturbed scheme with variable step sizes was given.

GENERAL PERTURBATION

First we consider general perturbations P_{ni} in the inner stages and P_n in the final stage of the method,

$$\widetilde{U}_{ni} = e^{c_i \tau_n J_n} \widetilde{u}_n + \tau_n \sum_{j=1}^{i-1} a_{ij}(\tau_n J_n) g_n(\widetilde{U}_{nj}) + P_{ni} \ , \qquad 1 \le i \le s \tag{3a}$$

$$\widetilde{u}_{n+1} = e^{\tau_n J_n} \widetilde{u}_n + \tau_n \sum_{i=1}^{s} b_i(\tau_n J_n) g_n(\widetilde{U}_{ni}) + p_{n+1} \tag{3b}$$

which for example can result from the approximation of the matrix functions by Krylov methods or by replacing the exponential functions by rational functions.

Theorem 1. *If the problem* (1) *and the unperturbed method* (2) *satisfy the assumptions of [2], Theorem 4.1, then the solution of* (3) *satisfies the error bound*

$$\|\widetilde{u}_{n+1} - u(t_{n+1})\| \le C \sum_{j=0}^{n} \Big(\tau_j^2 \sum_{i=1}^{s} \|P_{ji}\| + \|p_{j+1}\| \Big) + C \sum_{j=0}^{n} \tau_j^{p+1}$$

uniformly on $t_0 \le t_{n+1} \le T$. The constants C are independent of the chosen step size sequence.

Proof. We split the error using the solution of the unperturbed scheme,

$$\|\widetilde{u}_{n+1} - u(t_{n+1})\| \le \|\widetilde{u}_{n+1} - u_{n+1}\| + \|u_{n+1} - u(t_{n+1})\| \,.$$

The second term on the right hand side then is the error of the unperturbed method known from [2]. For the difference of the perturbed and the unperturbed method we obtain

$$\varepsilon_{ni} = \widetilde{U}_{ni} - U_{ni} = e^{c_i \tau_n J_n} \varepsilon_n + \tau_n \sum_{j=1}^{i-1} a_{ij}(\tau_n J_n) \big(g_n(\widetilde{U}_{nj}) - g_n(U_{nj}) \big) + P_{ni} \quad \text{and,} \qquad 1 \le i \le s, \tag{4a}$$

$$\varepsilon_{n+1} = \widetilde{u}_{n+1} - u_{n+1} = e^{\tau_n J_n} \varepsilon_n + \tau_n \sum_{i=1}^{s} b_i(\tau_n J_n) \big(g_n(\widetilde{U}_{ni}) - g_n(U_{ni}) \big) + p_{n+1} \,. \tag{4b}$$

Equation (4b) is similar to the error recursion arising in the proofs of [2]. Using Taylor-series expansion and exploiting the special form of g_n we obtain

$$\big\| g_n(\widetilde{U}_{ni}) - g_n(U_{ni}) \big\| \le C \big(\tau_n + \|\varepsilon_{ni}\| \big) \|\varepsilon_{ni}\| \,.$$

From this and equation (4a), we derive

$$\|\varepsilon_{ni}\| \le C\|\varepsilon_n\| + \|P_{ni}\| + \tau_n^2 \sum_{j=1}^{i-1} \|P_{nj}\| \,.$$

Solving the recursion (4b) and using $\varepsilon_0 = 0$ yields

$$\varepsilon_{n+1} = \sum_{j=0}^{n} \tau_j e^{\tau_n J_n} \cdot \ldots \cdot e^{\tau_{j+1} J_{j+1}} \big(\widetilde{\rho}_j - \tau_j^{-1} p_{j+1} \big) \quad \text{with} \quad \widetilde{\rho}_j = \sum_{i=1}^{s} b_i(\tau_j J_j) \big(g_j(\widetilde{U}_{ji}) - g_j(U_{ji}) \big) \,.$$

Combining the stability result from [2] with the previous estimates and applying a discrete Gronwall-Lemma bounds the difference of the perturbed and the unperturbed method and thus yields the desired result. □

This estimate is a worst case estimate. If we consider the linear homogeneous parabolic problem $F(u) = Au$, we have

$$\varepsilon_{n+1} = \sum_{j=0}^{n} e^{(\tau_n + \ldots + \tau_{j+1})A} p_{j+1} \,.$$

hence most perturbations are damped. To demonstrate this, we consider the test equation

$$\frac{\partial}{\partial t} u = \frac{\partial^2}{\partial x^2} u + \frac{1}{1+u^2} + k(t) \,, \quad x \in [0,1] \,, \quad t \in [0, 0.5]$$

with homogeneous Dirichlet boundary conditions and $k(t)$ chosen such that the exact solution is $u(t,x) = x(x-1)e^t$. The problem is discretized in space by finite differences with 200 nodes. We solve it with `exprb43`, [2], employing constant step sizes and introduce artificial perturbations.

In the left picture of Fig. 1, we choose $P_{ji} = 0$ and $p_{j+1} = \tau^l r_j$ for a normalized random vector r_j (curves with circles) and $p_{j+1} = \tau^l w_j$, where w_j is the eigenvector to the largest eigenvalue of the discretized linear operator plus a small perturbation (curves with squares), $l = 1, 2, 3$ and 4 for the blue, red, green and cyan curves, respectively. In the right picture, we perturbed the inner stages in the same way as above and do not change the final stage. It can be seen, that for the random perturbation, we almost always obtain better results than the theory predicts. In the worst case, where the perturbations are chosen to be an eigenvector corresponding to the largest eigenvalue of the linear part plus a random perturbation of the size of the machine precision, the errors are not damped, but sum up in the predicted way. Also, these bounds lead to the conclusion, that it is more important to compute the final stage accurately than the inner stages since it contributes more dominantly to the overall error.

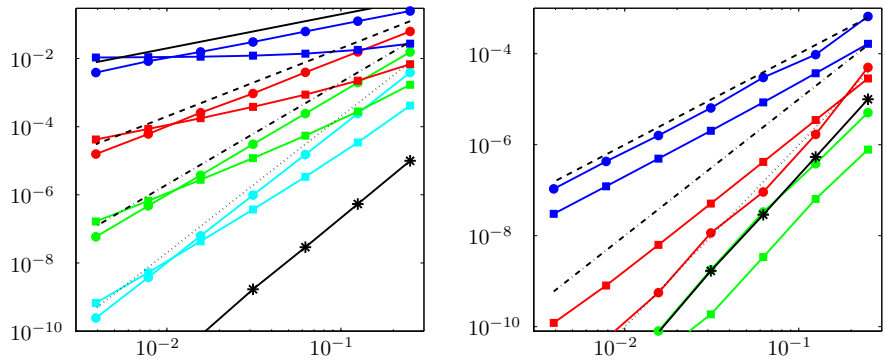

FIGURE 1. The error of the `exprb43` method with different types of perturbations are plotted versus the step size. **Left:** Only the final stages are perturbed with $p_{j+1} = \mathcal{O}(\tau^l)$, $P_{ji} = 0$. **Right:** Only the internal stages are perturbed with $P_{ji} = \mathcal{O}(\tau^l)$, $p_{j+1} = 0$. In both pictures $l = 1,2,3$ and 4 for the blue, red, green and cyan curves, respectively, where the squares represent the worst case perturbations and the circles the random perturbation. The black curve with stars is the error of the unperturbed solution. The black lines represent order curves (solid: order 1, dashed: order 2, dash-dotted: order 3, dotted: order 4).

INEXACT JACOBIAN

It might be convenient to use the scheme with an inexact Jacobian, for instance if it is computed via numerical differentiation. Replacing $J_n = \mathrm{D}F(u_n)$ by an approximation $\widetilde{J}_n$ yields

$$U_{ni} = e^{c_i \tau_n \widetilde{J}_n} u_n + \tau_n \sum_{j=1}^{i-1} a_{ij}(\tau_n \widetilde{J}_n) \widetilde{g}_n(U_{nj}), \qquad 1 \le i \le s \tag{5a}$$

$$u_{n+1} = e^{\tau_n \widetilde{J}_n} u_n + \tau_n \sum_{i=1}^{s} b_i(\tau_n \widetilde{J}_n) \widetilde{g}_n(U_{ni}). \tag{5b}$$

The function $\widetilde{g}_n$ is given by $\widetilde{g}_n\big(u(t)\big) = F\big(u(t)\big) - \widetilde{J}_n u(t)$.

Theorem 2. *Let the problem* (1) *and the unperturbed method* (2) *satisfy the assumptions of [2], Theorem 4.1. Also we assume, that* $\Delta\widetilde{J}_n = J_n - \widetilde{J}_n$ *is sufficiently small and satisfies*

$$\sum_{j=0}^{n} \left(\sum_{k=0}^{j-1} \left(\tau_k^2 \|\Delta\widetilde{J}_k\|_{X\leftarrow X} \right) + \|\Delta\widetilde{J}_j\|_{X\leftarrow X} \right) \le C_{\mathrm{J}}. \tag{6}$$

If $\widetilde{J}_n$ *also satisfies the semigroup assumption of [2], the solution of* (5) *satisfies the error bound*

$$\|u_{n+1} - u(t_{n+1})\| \le C \sum_{j=0}^{n} \left(\tau_j^{p+1} + \tau_j^2 \|\Delta\widetilde{J}_j\|_{X\leftarrow X} \right)$$

uniformly on $t_0 \le t_{n+1} \le T$.

The analysis follows the proof of the error and stability of the unperturbed scheme from [2].

Proof. For a representation of the defects for (5), we apply the variation-of-constants formula to $u'(t) = \widetilde{J}_n u(t) + \widetilde{G}_n(t)$, where $\widetilde{G}_n(t) = \widetilde{g}_n(u(t))$. To bound them, we need

$$\|\widetilde{G}_n'(t_n)\|_{X\leftarrow X} \le \|\Delta\widetilde{J}_n\|_{X\leftarrow X} + \|G_n'(t_n)\|_{X\leftarrow X} \le \|e_n\| + \|\Delta\widetilde{J}_n\|_{X\leftarrow X},$$

where e_n is the error of the unperturbed method, which is known to be small. This leads to the following estimates for the new defects

$$\|\widetilde{\Delta}_{ni}\| \le C\tau_n^2 \|\Delta\widetilde{J}_n\|_{X\leftarrow X} + C\tau_n^3, \quad \|\widetilde{\delta}_{n+1}\| \le C\tau_n^2 \|\Delta\widetilde{J}_n\|_{X\leftarrow X} + C\tau_n^{p+1}.$$

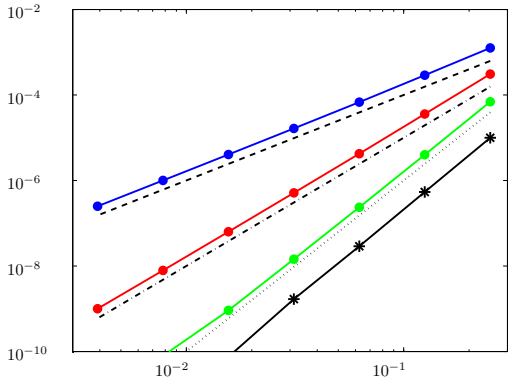

FIGURE 2. The errors of the `exprb43` method with $\Delta\widetilde{J}_j = \mathscr{O}(\tau_j^l)$ are plotted versus the step size for $l = 1,2$ and 3 for the blue, red and green curves, respectively. The black curve with stars is the error of the unperturbed solution. The black lines represent order curves (dashed: order 2, dash-dotted: order 3, dotted: order 4).

Next, we adjust the estimate for the error in the internal stages,

$$\|\widetilde{E}_{ni}\| \leq C\|\widetilde{e}_n\|\left(1+\tau_n\|\Delta\widetilde{J}_n\|_{X\leftarrow X}\right)+C\tau_n^3+C\tau_n^2\|\Delta\widetilde{J}_n\|_{X\leftarrow X}\,.$$

Since we use $\widetilde{J}_n$ to evaluate the matrix functions, we have to derive a stability bound for the discrete evolution operator involving $\widetilde{J}_n$ instead of J_n. Here we have

$$\begin{aligned}\|\widetilde{J}_n-\widetilde{J}_{n-1}\|_{X\leftarrow X} &\leq \|\Delta\widetilde{J}_n\|_{X\leftarrow X}+\|J_n-J_{n-1}\|_{X\leftarrow X}+\|\Delta\widetilde{J}_{n-1}\|_{X\leftarrow X}\\ &\leq C\left(\tau_n+\|e_n\|+\|e_{n-1}\|\right)+\|\Delta\widetilde{J}_n\|_{X\leftarrow X}+\|\Delta\widetilde{J}_{n-1}\|_{X\leftarrow X}\end{aligned}$$

which leads to

$$\left\|e^{-t\widetilde{J}_n}-e^{-t\widetilde{J}_{n-1}}\right\|_{X\leftarrow X} \leq C_L\left(\tau_{n-1}+\|e_n\|+\|e_{n-1}\|+\|\Delta\widetilde{J}_n\|_{X\leftarrow X}+\|\Delta\widetilde{J}_{n-1}\|_{X\leftarrow X}\right)e^{\widetilde{\omega}t} \quad \text{and}$$

$$\left\|e^{\tau_n\widetilde{J}_n}\cdot\ldots\cdot e^{\tau_0\widetilde{J}_0}\right\|_{X\leftarrow X} \;\leq\; Ce^{\widehat{\omega}(\tau_0+\ldots+\tau_n)+C_E\sum_{j=1}^n(\|e_j\|+\|\Delta\widetilde{J}_j\|)}\,.$$

Solving the error recursion and applying the modified stability bound yields the desired bound. The constant depends on

$$\sum_{j=0}^{n}\left(\|e_j\|+\|\Delta\widetilde{J}_j\|_{X\leftarrow X}\right) \leq \sum_{j=0}^{n}\left(\sum_{k=0}^{n-1}\left(\tau_k^{p+1}+\tau_k^2\|\Delta\widetilde{J}_k\|_{X\leftarrow X}\right)+\|\Delta\widetilde{J}_j\|_{X\leftarrow X}\right),$$

which is bounded uniformly using the assumptions on the unperturbed scheme and (6). □

It is obvious that $\|\Delta\widetilde{J}_j\|_{X\leftarrow X} \leq C\tau_j$ should be satisfied. Otherwise stability cannot be guaranteed any more. To get the same order as for the exact Jacobian, $\|\Delta\widetilde{J}_j\|_{X\leftarrow X} \leq C\tau_j^{p-1}$ is required.

For the same example as in the previous section, the errors of the perturbed method are shown in Fig. 2. We can see, that the order is reduced in the predicted way.

ACKNOWLEDGMENTS

I would like to thank Marlis Hochbruck for her advice and many helpful discussions. This work has been supported by the Deutsche Forschungsgemeinschaft through the Transregio SFB TR 18.

REFERENCES

1. M. Hochbruck, A. Ostermann, "Explicit integrators of Rosenbrock-type", *Oberwolfach Reports* **18**, 1107–1110 (2006).
2. M. Hochbruck, A. Ostermann, J. Schweitzer, "Exponential Rosenbrock-type methods", *SIAM J. Numer. Anal.* **47**, No. 1, 786–803 (2009).
3. J. Schweitzer, "Numerical Simulation of Relativistic Laser-Plasma Interaction", PhD Thesis, University of Düsseldorf, 2009.