Bergische Universität Wuppertal

Fakultät für Mathematik und Naturwissenschaften

Institute of Mathematical Modelling, Analysis and Computational Mathematics (IMACM)

Preprint BUW-IMACM 21/04

M.W.F.M. Bannenberg, A. Ciccazzo, M. Günther

Reduced Order Multirate Schemes for Coupled Differential-Algebraic Systems

January 22, 2021

http://www.math.uni-wuppertal.de

Reduced Order Multirate Schemes for Coupled Differential-Algebraic Systems

M.W.F.M. Bannenberg[a,b], A. Ciccazzo[b], M. Günther[a]

[a] *IMACM, Chair of Applied Mathematics and Numerical Analysis (AMNA), Bergische Universität Wuppertal, Gaußstraße 20, 42119 Wuppertal, Germany.*
[b] *STMicroelectronics, Str. Primsole 50, Catania, Italy*

Abstract

In the context of time-domain simulation of integrated circuits, one often encounters large systems of coupled differential-algebraic equations. Simulation costs of these systems can become prohibitively large as the number of components keeps increasing. In an effort to reduce these simulation costs a twofold approach is presented in this paper. We combine maximum entropy snapshot sampling method and a nonlinear model order reduction technique, with multirate time integration. The obtained model order reduction basis is applied using the Gauß-Newton method with approximated tensors reduction. This reduction framework is then integrated using a coupled-slowest-first multirate integration scheme. The convergence of this combined method verified numerically. Lastly it is shown that the new method results in a reduction of the computational effort without significant loss of accuracy.

Keywords:
Mulitrate, Model Order Reduction, Differential-Algebraic Equations, Snapshot Sampling.

1. Introduction

During the previous decades the number of transistors on a chip has grown exponentially, whilst the chip size remained nearly constant or even decreased. As the number of transistors increases and the area in between them diminishes, very detailed effects have to taken into account. Due to these developments, the manufacturing of integrated circuits became increasing complex. Before a chip is produced, it needs to be analysed whether or not it behaves according to desired specifications.

With the increased complexity, it has become unfeasible to do this analysis through prototype experiments. Therefore computer-aided design (CAD) using modelling and simulation tools has become a crucial and necessary element in the industrial optimization flow. Mathematical models of the integrated circuits are derived from the network topology and natural phenomena occurring inside these chips. Now however, due to the ever increasing complexity of the circuits, we even run into limitations using CAD as the mathematical models become prohibitively large.

In the context of time-domain simulation of multiphysical integrated circuits, one often encounters large systems of coupled differential-algebraic equations (DAEs). To keep the simulation times of these systems feasible, a multitude of techniques can be applied exploiting different characteristics of the underlying systems. As there are different natural phenomena occurring at once inside these circuits, one of the exploited characteristics is the difference of time scales for each of these phenomena. This is done through multirate (MR) time integration, [8].

Another way of drastically improving the feasibility of these simulations is by incorporating nonlinear model order reduction (MOR) techniques, for which the Maximum Entropy Snapshot Sampling (MESS) method is used, [11]. This MOR technique directly reduces the snapshot matrix according to an estimate of the second-order Rényi entropy, instead of creating a basis according to information based on linear transformations, such as done by Proper Orthogonal Decomposition (POD) through singular values, which is currently an industry standard. Besides preserving the nonlinear characteristics of the snapshot matrix the

MESS method also has reduced memory constraints, as the QR-decomposition is called after the snapshot matrix has been reduced. Finally, the MESS method relies solely on pairwise distance computations, and its performance can be improved through the use of CPU/GPU parallelism.

This technique can be supplemented with hyper-reduction methods for the reduction of nonlinear function evaluations. Without this extension, the nonlinear function evaluations of the back-transformed reduced state vectors can still become a dominant factor in the computational effort. These are methods such as the Discrete Empirical Interpolation Method (DEIM) [5, 6] or, as used for our hyper-reduction, data reconstruction through the gappy POD method [16], where in our case a MESS constructed basis is used for the reconstruction.

In this paper a twofold approach is presented to efficiently simulate coupled nonlinear DAEs by combining these two techniques, into a reduced order multirate (ROMR) scheme. This is is a generalisation of [1] by also considering the possibility of DAEs in the slower subsystems. In the next section the mathematical problem is formulated and preliminaries are discussed. Section 3 is related to the numerical analysis of the ROMR method. Then in section 4 numerical experiments are performed and results presented. In the final section conclusions are drawn.

2. Problem Formulation

Consider the following coupled system of two semi explicit DAE systems, where the subscripts $\{F, S\}$ indicate a fast or slow time-scale, respectively, and independent transient sources have been omitted for notational convenience:

$$\frac{d}{dt} y_{\mathrm{F}} = f_{\mathrm{F}}(t, y_{\mathrm{F}}, z_{\mathrm{F}}, y_{\mathrm{S}}, z_{\mathrm{S}}), \quad y_{\mathrm{F}}(t_0) = y_{F_0}, \tag{1a}$$

$$0 = g_{\mathrm{F}}(t, y_{\mathrm{F}}, z_{\mathrm{F}}, y_{\mathrm{S}}, z_{\mathrm{S}}), \quad z_{\mathrm{F}}(t_0) = z_{F_0}, \tag{1b}$$

$$\frac{d}{dt} y_{\mathrm{S}} = f_{\mathrm{S}}(t, y_{\mathrm{F}}, z_{\mathrm{F}}, y_{\mathrm{S}}, z_{\mathrm{S}}), \quad y_{\mathrm{S}}(t_0) = y_{S_0}, \tag{1c}$$

$$0 = g_{\mathrm{S}}(t, y_{\mathrm{F}}, z_{\mathrm{F}}, y_{\mathrm{S}}, z_{\mathrm{S}}), \quad z_{\mathrm{S}}(t_0) = z_{S_0}, \tag{1d}$$

with the functions $f_A : \mathbb{R} \times \mathbb{R}^a \times \mathbb{R}^b \times \mathbb{R}^c \times \mathbb{R}^d \to \mathbb{R}^a$, with $A \in \{\mathrm{F}, \mathrm{S}\}$, where $\{a, b, c, d\} \in \mathbb{N}$ are the respective dimensions, and equivalent definitions for g_A. Consistent initial conditions are assumed, which means that Equations (1b) and (1d) are satisfied at initial time t_0. The quantities $y_{\{F,S\}} : I \to \mathbb{R}^{\{a,b\}}$ and $z_{\{F,S\}} : I \to \mathbb{R}^{\{c,d\}}$ denote the differential and algebraic variables defined on the time interval $[t_0, t_1]$. Both subsystems and the joint system are guaranteed to be index-1 by the assumption that the Jacobians

$$\frac{\partial g_{\mathrm{F}}}{\partial z_{\mathrm{F}}}, \frac{\partial g_{\mathrm{S}}}{\partial z_{\mathrm{S}}} \text{ and } \begin{pmatrix} \frac{\partial g_{\mathrm{F}}}{\partial z_{\mathrm{F}}} & \frac{\partial g_{\mathrm{F}}}{\partial z_{\mathrm{S}}} \\ \frac{\partial g_{\mathrm{S}}}{\partial z_{\mathrm{F}}} & \frac{\partial g_{\mathrm{S}}}{\partial z_{\mathrm{S}}} \end{pmatrix} \text{ are regular} \tag{2}$$

in the neighbourhood of the solution of the system. From this assumption the algebraic variables $z_{\{F,S\}}$ can be solved locally by using the implicit function theorem

$$z_{\{F,S\}} = G_{t,\{F,S\}}(y_{\mathrm{F}}, z_{\{S,F\}}, y_{\mathrm{S}}), \tag{3}$$

where the second z subscript is the opposite of the first z subscript. The partition of the system into subsystems can originate from different physical systems, such as temperature diffusion and electric currents. However, differences in time scale can also be identified by different orders of time derivatives. Here the partition is considered to be fixed during the time integration.

3. Overview of the Reduced Order Multirate Method

To keep this paper as self-contained as possible, this section provides an overview of each individual method that is used in a ROMR scheme. First a description of the Maximum Entropy Snapshot Sampling method proposed in [11], and the subsequent gappy data reconstruction, [16], for the approximation of the nonlinear functions are given. Second, the multirate implicit Euler scheme is described using a Coupled-Slowest-First approach.

3.1. Maximum Entropy Snapshot Sampling

Let m and n be positive integers and $m \gg n > 1$. Define a finite sequence $X = (x_1, x_2, ..., x_n)$ of numerically obtained states $x_j \in \mathbb{R}^m$ at time instances $t_j \in \mathbb{R}$, with $j \in \{1, 2, ..., n\}$, of a dynamical system governed by either ODEs or DAEs. Provided probability distribution p of the states of the system, the second-order Rény entropy of the sample X is

$$H_p^{(2)}(X) = -\log \sum_{j=1}^{n} p(x_j)^2 = -\log \mathbb{E}(p(x_j)), \tag{4}$$

with $\mathbb{E}(p(X))$ the expected value of the probability distribution p with respect to p itself. When n is large enough, according to the law of large numbers, the average of $p_1, p_2, ..., p_n$ almost surely converges to their expected value,

$$\frac{1}{n} \sum_{j=1}^{n} p(x_j) \to \mathbb{E}(p(X)) \quad \text{as } n \to \infty, \tag{5}$$

thus each $p(x_j)$ can be approximated by the sample's average sojourn time or relative frequency of occurrence. To obtain this frequency of occurrence, considering a norm $\|\cdot\|$ on $\mathbb{R}^m$. Then the notion of occurrence can be translated into a proximity condition. In particular, for each $x_j \in \mathbb{R}^m$ define the open ball that is centred at x_j and whose radius is $\epsilon > 0$,

$$B_\epsilon(x) = \{y \in \mathbb{R}^m \mid \|x - y\| < \epsilon\}, \tag{6}$$

and introduce the characteristic function with values

$$\chi_i(x) = \begin{cases} 1, & \text{if } x \in B_\epsilon(x_i), \\ 0, & \text{if } x \notin B_\epsilon(x_i). \end{cases} \tag{7}$$

Under the aforementioned considerations, the entropy of X can be estimated by

$$\hat{H}_p^{(2)}(X) = -\log \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \chi_i(x_j). \tag{8}$$

Provided that the limit of the evolution of $\hat{H}_p^{(2)}$ exists, for n large enough, and measures the sensitivity of the evolution of the system itself [3, §6.6], a reduced sequence $X_\mathrm{r} = (\bar{x}_{j_1}, \bar{x}_{j_2}, \ldots, \bar{x}_{j_r})$, with $r \leq n$, is sampled from X, by requiring that the entropy of X_r is a strictly increasing function of the index $k \in \{1, 2, \ldots, r\}$ [10]. The state vector $\bar{x}_{j_k}$ added to sampled snapshot space is the average value of all states in the selected ϵ-ball. A reduced basis is then generated from X_r with any orthonormalization process. The MESS procedure is outlined in Algorithm 1. It has been shown [11] that, depending on the recurrence properties of a system, any such basis guarantees that the Euclidean reconstruction error of each snapshot is bounded from above by ϵ, while a similar bound holds true for future snapshots, up to a specific time-horizon.

The Estimation of ϵ: The open ball parameter ϵ, which is directly responsible for the degree of reduction within the MESS framework, can be chosen arbitrarily, much like the number of selected basis vectors provided by a POD approach. For a ballpark estimate of this parameter the following optimisation approach is provided [12]. The quantity within the logarithm in the entropy estimate (8) is often referred to as the sample's correlation sum and can be written as

$$C_\epsilon = \frac{1}{n^2} \|R_\epsilon\|_\mathrm{F}^2, \tag{9}$$

with $R_\epsilon \in \{0, 1\}^{n \times n}$ being the recurrence matrix whose entries are unity, when $\|x_i - x_j\| < \epsilon$, and $\|\cdot\|_\mathrm{F}^2$ being the Frobenius norm. In terms of probability theory, C_ϵ is a cumulative distribution function of ϵ, and hence, its derivative $\mathrm{d}C_\epsilon/\mathrm{d}\epsilon$ is the associated probability density function of ϵ. A commonly justified hypothesis is that the correlation sum scales as ϵ^D [13, Chapter 1], with $D \geq 0$ being the so-called correlation dimension

Algorithm 1: Maximum Entropy Snapshot Sampling

input : Snapshot matrix $X \in \mathbb{R}^{m\times n}$, tolerance ϵ.
output: Reduced basis $V \in \mathbb{R}^{m\times r}$.

$P_{i,j} \leftarrow ||x_i - x_j||$, $\forall i, j \in \{1, ..., n\}$;
$P \leftarrow P/max(P)$;
$R \leftarrow P < \epsilon$;
$Y \leftarrow [\,]$;
for *j = 1,..., n* **do**
  idx $\leftarrow \{i \in [1, ..., m] \mid R_{i,j} = 1\}$;
  if |idx| $\neq 0$ **then**
    $Y \leftarrow [Y \ \text{mean}(X_{:,\text{idx}})]$;
    $R_{:,\text{idx}} \leftarrow 0$, $R_{\text{idx},:} \leftarrow 0$;
  end
end
$[V, -] \leftarrow \text{qr}(Y)$;

of the manifold that is formed in $\mathbb{R}^m$ by the terms of X. Under this power law assumption, the maximum likelihood estimate [14, Chapter 8] of the correlation dimension is estimated as follows. We find a sample $\{\epsilon_i\}$, with $\epsilon_i \in [0, 1]$ for all $i \in \{1, 2, \ldots, q\}$, of a random variable E that is sampled according to C_ϵ. Then, the probability of finding a sample in $(\epsilon_i, \epsilon_i + \mathrm{d}\epsilon_i)$ in a trial is

$$\prod_{i=1}^{q} D\epsilon^{D-1}\mathrm{d}\epsilon_i. \tag{10}$$

To calculate the ϵ value for which this expression is maximized, we take the logarithm

$$q \cdot \ln D + (D-1)\sum_{i=1}^{q} \ln \epsilon_i, \tag{11}$$

and note that the maximum of this expression is attained when

$$\frac{q}{D} + \sum_{i=1}^{q} \ln \epsilon_i = 0. \tag{12}$$

This results in the most likely value $D_* = -1/\langle \ln E\rangle$, and ϵ can be estimated by

$$\epsilon_* = \text{argmin}(|D_* - \ln C_\epsilon / \ln \epsilon|). \tag{13}$$

The algorithm to calculate this most likely value for a given snapshot matrix X is described in Algorithm 2.

3.2. The Gauß-Newton with approximated tensors method

Unfortunately, a direct application of MESS is not feasible in practice, [15, Section 7.6], therefore a simplified Gauß-Newton with Approximated Tensors (GNAT), equipped with a function-sampling-hyper-reduction scheme is used. Firstly, a direct Galerkin projection may yield an unsolvable reduced system for DAEs. Secondly, the computational effort required to solve this reduced system and the full system is about the same in the nonlinear cases. This is due to the fact that the evaluation costs of the reduced system are not reduced at all because the projection basis will be a dense matrix in general.

Considering a general DAE in the form

$$\dot{\phi}(t, u) + \psi(t, u) = 0, \tag{14}$$

Algorithm 2: Epsilon estimation for a given snapshot matrix X

input : Snapshot matrix $X \in \mathbb{R}^{m \times n}$.
output: Estimated tolerance value ϵ_*.

$P_{i,j} \leftarrow ||x_i - x_j||$, $\forall i, j \in \{1, ..., n\}$;
$P \leftarrow P/max(P)$;
$\{\epsilon^i_{\text{cdf}}\} \leftarrow LinearSpace(0, 1, n_\epsilon)$;
for $i = 1, ..., n_\epsilon$ **do**
  $C_{\epsilon^i_{\text{cdf}}} \leftarrow \frac{1}{n_\epsilon^2} ||R_{\epsilon^i_{\text{cdf}}}||^2_{\text{F}}$, Equation(9);
end
$\{\epsilon_i\}_{i=0}^q \leftarrow RandomFromCDF(q, \{\epsilon^j_{\text{cdf}}\}, \{C_{\epsilon^j_{\text{cdf}}}\})$, for $j = 1, ..., n_\epsilon$;
$D_* \leftarrow -\frac{1}{\langle \ln \epsilon_i \rangle}$;
$\epsilon_* \leftarrow \epsilon^i_{\text{cdf}}$, for which $i = \text{argmin}_j(|D_* - \ln C_{\epsilon^j_{\text{cdf}}} / \ln \epsilon^j_{\text{cdf}}|)$;

where ϕ and ψ are functions of time t and some state vector u. In the discrete case, we assume that the numerical scheme exactly solves the following nonlinear system for each time step t_i,

$$R(u) = 0, \tag{15}$$

where $u \in \mathbb{R}^N$, u^0 the initial condition and the residual $R : \mathbb{R}^N \to \mathbb{R}^N$. Note that for ease of notation, the relevant time subscripts have been omitted, as this equation is solved for each individual time step. For the reduction of the dimension of Equation (15), a projection is used to search the approximated solution in the incremental affine trial subspace $u^0 + \mathcal{V} \subset \mathbb{R}^N$. Thus $\tilde{u}$ is given by

$$\tilde{u} = u^0 + V_u u_r, \tag{16}$$

where $V_u \in \mathbb{R}^{N \times n_u}$ is the n_u-dimensional projection basis for $\mathcal{V}$, and u_r denotes the reduced incremental vector of the state vector. Now deviating from the direct Galerkin projection process, Equation (16), is substituted into Equation (15). This results in an overdetermined system of N equations and n_u unknowns. Because V_u is a matrix with full column rank, it is possible to solve this system by a minimisation in least-squares sense through

$$\min_{\tilde{u} \in u^0 + \mathcal{V}} ||R(\tilde{u})||_2. \tag{17}$$

This nonlinear least-squares problem is solved by the Gauß-Newton method, leading to the iterative process for $k = 1, ..., K$, solving

$$s^k = \text{argmin}_{a \in \mathbb{R}^{n_u}} ||J^k V_u a + R^k||_w, \tag{18}$$

and updating the search value w_r^k with

$$w_r^{k+1} = w_r^k + s^k, \tag{19}$$

where K is defined through a convergence criterion, initial guess w_r^0, $R^k \equiv R(u_0 + V_u w_r^k)$ and $J^k \equiv \frac{\partial R}{\partial u}(u_0, V_u u_r^k)$. Here J^k is the full order Jacobian of the residual at each iteration step k. Since the computation of this Jacobian scales with the original full dimension of Equation (15) this is a computational bottleneck. This bottleneck can be circumvented by the application of hyper reduction methods, for which this paper utilises a gappy data reconstruction method.

Gappy Maximum Entropy Snapshot Sampling: The evaluation of the nonlinear function $R(u_0 + V_u w_r^k)$ has a computational complexity that is still dependent on the size of the full system. To reduce the complexity of this evaluation the gappy MESS procedure, based on gappy POD [7], is applied. Like the gappy POD approach gappy MESS uses a reduced basis to reconstruct gappy data. However, unlike the gappy POD approach the basis used is now not obtained through POD but by MESS. Gappy MESS starts by defining

a mask vector n for a solution state u as

$$n_j = 0 \text{ if } u_j \text{ is missing,}$$
$$n_j = 1 \text{ if } u_j \text{ is known,}$$

where j denotes the j-th element of each vector. The mask vector n is applied point-wise to a vector by $(n, u)_j = n_j u_j$. This sets all the unobserved values to 0. Then, the gappy inner product can be defined as $(x, y)_n = ((n, x), (n, y))$, which is the inner product of the each vector masked respectively. The induced norm is then $(||x||_n)^2 = (x, x)_n$. Considering the reduction base obtained by MESS $V_{\text{gap}} = \{v^i\}_{i=1}^r$, now we can construct an intermediate "repaired" full size vector $\tilde{g}$ from a reduced vector g with only r elements by

$$\tilde{g} \approx \sum_{i=1}^{r} b_i v^i, \tag{20}$$

where the coefficients b_i need to minimise an error E between the original and repaired vector, which is defined as

$$E = ||g - \tilde{g}||_n^2. \tag{21}$$

This minimisation is done by solving the linear system

$$Mb = f, \tag{22}$$

where

$$M_{ij} = (v^i, v^j)_n, \text{ and } f_i = (g, v^i)_n. \tag{23}$$

From this solution $\tilde{g}$ is constructed. Then the complete vector is reconstructed by mapping the reduced vectors elements to their original indices and filling the rest with the reconstructed values.

Algorithm 3: Gappy reconstruction

input : Snapshot matrix $X \in \mathbb{R}^{m \times n}$, tolerance ϵ.
output: Matrix M.

$U \leftarrow MESS(X, \epsilon)$;
$[Q, R] \leftarrow qr(U, \text{thin})$, such that $AP = QR$;
$S \leftarrow \text{triu}(R)$;
for $j = 1, ..., m$ **do**
  if $j \in S$ **then**
    $n_j^{\text{mask}} = 1$;
  else
    $n_j^{\text{mask}} = 0$;
  end
end
for $i = 1, ..., m$ **do**
  for $j = 1, ..., m$ **do**
    $M_{i,j} = ||U_{:,i}, U_{:,j}||_{n^{\text{mask}}}$;
  end
end

3.3. The Reduced System

To incorporate the previous two sections into the partitioned DAE system (1a)-(1d), we first rewrite (1c)-(1d) in a more general DAE form, to have the slow subsystem encapsulated into one equation.

$$\frac{d}{dt} y_{\text{F}} = f_{\text{F}}(t, y_{\text{F}}, z_{\text{F}}, u_{\text{S}}), \quad y_{\text{F}}(t_0) = y_{\text{F}_0}, \tag{24}$$
$$0 = g_{\text{F}}(t, y_{\text{F}}, z_{\text{F}}, u_{\text{S}}), \quad z_{\text{F}}(t_0) = z_{\text{F}_0}, \tag{25}$$
$$\frac{d}{dt} \phi(u_{\text{S}}) = F_{\text{S}}(t, y_{\text{F}}, z_{\text{F}}, u_{\text{S}}), \quad u_{\text{S}}(t_0) = (y_{\text{S}_0}, z_{\text{S}_0})^\top, \tag{26}$$

where $F_{\mathrm{S}} : \mathbb{R} \times \mathbb{R}^a \times \mathbb{R}^b \times \mathbb{R}^{m_{\mathrm{S}}} \to \mathbb{R}^{m_{\mathrm{S}}}$ and $u_{\mathrm{S}} = (y_{\mathrm{S}}, z_{\mathrm{S}})^\top$. Into these equation we incorporate the back projected reduced state $\tilde{u}_{\mathrm{S_r}} = u_{\mathrm{S}}^0 + V_u u_{\mathrm{S_r}}$

$$\frac{d}{dt} y_{\mathrm{F_r}} = f_{\mathrm{F}}(t, y_{\mathrm{F_r}}, z_{\mathrm{F_r}}, \tilde{u}_{\mathrm{S_r}}), \tag{27}$$

$$0 = g_{\mathrm{F}}(t, y_{\mathrm{F_r}}, z_{\mathrm{F_r}}, \tilde{u}_{\mathrm{S_r}}), \tag{28}$$

$$\frac{d}{dt} \phi(\tilde{u}_{\mathrm{S_r}}) = F_{\mathrm{S}}(t, y_{\mathrm{F_r}}, z_{\mathrm{F_r}}, \tilde{u}_{\mathrm{S_r}}). \tag{29}$$

and then, with the Gappy MESS complexity reduction incorporated we obtain

$$\frac{d}{dt} y_{\mathrm{F_r}} = f_{\mathrm{F}}(t, y_{\mathrm{F_r}}, z_{\mathrm{F_r}}, \tilde{u}_{\mathrm{S_r}}), \tag{30}$$

$$0 = g_{\mathrm{F}}(t, y_{\mathrm{F_r}}, z_{\mathrm{F_r}}, \tilde{u}_{\mathrm{S_r}}), \tag{31}$$

$$\frac{d}{dt} \phi(\tilde{u}_{\mathrm{S_r}}) = F_{\mathrm{S_r}}(t, y_{\mathrm{S_r}}, z_{\mathrm{F_r}}, \tilde{u}_{\mathrm{S_r}}). \tag{32}$$

Where $F_{\mathrm{S_r}}$ denotes the function F_{S} solved by the reduced least squares approach. Note that the subscript r denotes a reduction, and not the reduction factor.

3.4. Multirate Implicit Euler

The overall index-1 system (30)-(32) can be integrated with the stiffly accurate implicit Euler scheme, which automatically assures that also for $t > 0$ the quantities will remain consistent. To exploit the assumed different time scales, a multirate integration scheme is proposed. This approach is analogous to [1], but with a slow subsystem consisting of DAEs. where $h = H/m$ and the coupling variables are denoted by $\bar{y}_{\mathrm{F}}$, $\bar{z}_{\mathrm{F}}$, $\bar{\tilde{u}}_{\mathrm{S_r}}$. The coupling strategy is chosen to be the *Coupled-Slowest-First* approach as this is shown to have a consistency of order 1 for the problem posed in [9]: First the whole system is solved for the macro-step, $t_n \to t_{n+1} = t_n + H$

$$y^*_{\mathrm{F_r},n+1} = y_{\mathrm{F_r}n} + H f_{\mathrm{F}}(y^*_{\mathrm{F_r},n+1}, z^*_{\mathrm{F_r},n+1}, \tilde{u}_{\mathrm{S_r},n+1}), \tag{33}$$

$$0 = g_{\mathrm{F}}(y^*_{\mathrm{F_r},n+1}, z^*_{\mathrm{F_r},n+1}, \tilde{u}_{\mathrm{S_r},n+1}), \tag{34}$$

$$\phi(\tilde{u}_{\mathrm{S_r},n+1}) = \phi(\tilde{u}_{\mathrm{S_r},n}) + H F_{\mathrm{S_r}}(y^*_{\mathrm{F_r},n+1}, z^*_{\mathrm{F_r},n+1}, \tilde{u}_{\mathrm{S_r},n+1}). \tag{35}$$

The step size H is chosen according to the slow dynamics, whilst the full system remains solvable. From this it follows that the fast solutions, $y^*_{\mathrm{F_r},n+1}$ and $z^*_{\mathrm{F_r},n+1}$, are not accurate enough and can be discarded, as they will be computed in the last micro step. In a second step, the fast solutions are computed for the micro steps $l = 0, \ldots, m-1$,

$$y_{\mathrm{F_r},n+(l+1)/m} = y_{\mathrm{F_r},n+l/m} + h f_{\mathrm{F}}(y_{\mathrm{F_r},n+(l+1)/m}, z_{\mathrm{F_r},n+(l+1)/m}, \bar{\tilde{u}}_{\mathrm{S_r},n+(l+1)/m}), \tag{36}$$

$$0 = g_{\mathrm{F}}(y_{\mathrm{F_r},n+(l+1)/m}, z_{\mathrm{F_r},n+(l+1)/m}, \bar{\tilde{u}}_{\mathrm{S_r},n+(l+1)/m}), \tag{37}$$

$$\phi(\tilde{u}_{\mathrm{S_r},n+(l+1)/m}) = \phi(\tilde{u}_{\mathrm{S_r},n+l/m}) + h F_{\mathrm{S_r}}(\bar{y}_{\mathrm{F_r},n+(l+1)/m}, \bar{z}_{\mathrm{F_r},n+(l+1)/m}, \tilde{u}_{\mathrm{S_r},n+(l+1)/m}). \tag{38}$$

For stability reasons, the interpolated values $\bar{\tilde{u}}_{\mathrm{S_r},n+(l+1)/m}$ are obtained by constant interpolation based on $\tilde{u}_{\mathrm{S_r},n+1}$, then the *Coupled-Slowest-First* Euler approach is unconditionally A-stable.

4. Numerical Analysis

In this section, the error induced by the ROMR scheme from one macro-step $t_n \to t_{n+1} = t_n + H$ is estimated. We define the error in each variable class as

$$||y_{\mathrm{F}}(t_{n+1}) - y_{\mathrm{F_r},n+1}||, \tag{39}$$

$$||z_{\mathrm{F}}(t_{n+1}) - z_{\mathrm{F_r},n+1}||, \tag{40}$$

$$||u_{\mathrm{S}}(t_{n+1}) - \tilde{u}_{\mathrm{S_r},n+1}||. \tag{41}$$

Here $||\cdot||$ is the 2-norm in Euclidean space. The analytical solutions of a state variable is notated by with a parenthesised time argument whilst the numerical approximation is noted with subscript, e.g. $u_{\mathrm{S}}(t_n)$ and $u_{\mathrm{S},n}$. To analyse this error, it is split into two parts, the numerical approximation error and the discrete reduction error.

$$||y_{\mathrm{F}}(t_{n+1}) - y_{\mathrm{F_r},n+1}|| \leq ||y_{\mathrm{F}}(t_{n+1}) - y_{\mathrm{F},n+1}|| + ||y_{\mathrm{F},n+1} - y_{\mathrm{F_r},n+1}||, \tag{42}$$

$$||z_{\mathrm{F}}(t_{n+1}) - z_{\mathrm{F_r},n+1}|| \leq ||z_{\mathrm{F}}(t_{n+1}) - z_{\mathrm{F},n+1}|| + ||z_{\mathrm{F},n+1} - z_{\mathrm{F_r},n+1}||, \tag{43}$$

$$||u_{\mathrm{S}}(t_{n+1}) - \tilde{u}_{\mathrm{S_r},n+1}|| \leq ||u_{\mathrm{S}}(t_{n+1}) - u_{\mathrm{S},n+1}|| + ||u_{\mathrm{S},n+1} - \tilde{u}_{\mathrm{S_r},n+1}||. \tag{44}$$

The first error term on the right-hand side of the inequality can be identified to be the error induced by a non-reduced order implicit multirate scheme. This error E_{MR} is $\mathcal{O}(H^2)$, following [9, Theorem 2]. The second error on the right-hand side, the error induced by the GNAT and hyper-reduction method, can be analysed in the following manner. For the slow subsystem, we only have to consider the macro-step and thus it holds that

$$||u_{\mathrm{S},n+1} - \tilde{u}_{\mathrm{S_r},n+1}|| \leq E_{\mathrm{macro}} \tag{45}$$

where E_{macro} is the error bound of [4, Proposition 4.1]. This, due to the fact that the macro-step of the ROMR scheme is an implicit Euler step using GNAT and hyper-reduction, identical to the prerequisites of the proposition, using the fact that the algebraic variable values are directly obtained through the implicit function theorem. The only error that now needs to be bounded such that the whole ROMR induced error is bounded, is the micro-step error. For each micro-step, again using the fact that the algebraic values are solved locally by the implicit function theorem, we only have to analyse the error in the fast dynamical variables y_{F}

$$R^n(y_{\mathrm{F},n+(l+1)/m}) = y_{\mathrm{F},n+(l+1)/m} - y_{\mathrm{F},n+l/m} - hf_{\mathrm{F}}(y_{\mathrm{full},n+(l+1)/m}), \tag{46}$$

and

$$\tilde{R}^n(\tilde{y}_{\mathrm{F_r},n+(l+1)/m}) = y_{\mathrm{F_r},n+(l+1)/m} - y_{\mathrm{F_r},n+l/m} - hf_{\mathrm{F}}(\tilde{y}_{\mathrm{full},n+(l+1)/m}). \tag{47}$$

Here $y_{\mathrm{full},n}$ is a shorthand notation for the full state $(y_{\mathrm{F},n}, z_{\mathrm{F},n}, u_{\mathrm{S},n})$ for f. Using $\zeta : (x) \to x - hf_{\mathrm{F}}(x)$ and the inverse Lipschitz constant

$$\mathcal{L}_n \equiv \sup_{x \neq y} \frac{||x - y||}{||\zeta(x) - \zeta(y)||}. \tag{48}$$

we obtain a bound for the local micro-step approximation error

$$||y_{\mathrm{F_r},n+(l+1)/m} - \tilde{y}_{\mathrm{F_r},n+(l+1)/m}|| \leq \mathcal{L}_n \left(\epsilon_{\mathrm{Newton}} + ||\tilde{R}^n(\tilde{y}_{\mathrm{F_r},n+(l+1)/m})|| + ||y_{\mathrm{F},n+l/m} - \tilde{y}_{\mathrm{F_r},n+l/m})||\right) \tag{49}$$

This then results in

$$||y_{\mathrm{F_r},n+1} - \tilde{y}_{\mathrm{F_r},n+1)/m}|| \leq \sum_{k=0}^{m-1} a^k \left(\epsilon_{\mathrm{Newton}} + ||\tilde{R}^k(\tilde{y}_{\mathrm{F_r},n+(k+1)/m})|| + ||y_{\mathrm{F},n+k/m} - \tilde{y}_{\mathrm{F_r},n+k/m})||\right) \tag{50}$$

where $a = \mathcal{L} \equiv \sup_{k \in \{0,\ldots,m-1} \mathcal{L}_k$. For h small enough, it follows that

$$||y_{\mathrm{F},n+1} - y_{\mathrm{F_r},n+1}|| \leq E_{\mathrm{micro}}. \tag{51}$$

Thus it has been shown that the cumulative micro-step error is bounded as well. Now we assume that the reduction induced error bound $E_{\{\mathrm{macro,micro}\}} \ll E_{\mathrm{MR}}$, which should always be the case as model order reduction should only be used if the reduced model is able to accurately capture the full order dynamics. So for a macro-step, the following holds,

$$||y_{\mathrm{F}}(t_{n+1}) - y_{\mathrm{F_r},n+1}|| \leq E_{\mathrm{MR}} + E_{\mathrm{micro}} \approx \mathcal{O}(H^2), \tag{52}$$

$$||z_{\mathrm{F}}(t_{n+1}) - z_{\mathrm{F_r},n+1}|| \leq E_{\mathrm{MR}} + E_{\mathrm{micro}} \approx \mathcal{O}(H^2), \tag{53}$$

$$||u_{\mathrm{S}}(t_{n+1}) - \tilde{u}_{\mathrm{S_r},n+1}|| \leq E_{\mathrm{MR}} + E_{\mathrm{macro}} \approx \mathcal{O}(H^2). \tag{54}$$

Then, for the error propagation over several macro-steps we obtain by using [9, Theorem 2] that the global error is $\mathcal{O}(H)$. To illustrate this result and deduct if the reduction of the computational effort is adequate, numerical experiments are performed in the next section.

5. Numerical Experiments

In this section the previously found analytical results are verified numerically. The reduced order multirate scheme is compared against a standard implicit Euler integration scheme. A transient analysis is performed for an academic test case and the convergence of the error is investigated. Furthermore, computational times are compared to verify the efficiency of the reduced order multirate scheme.

5.1. The Test Model

The underlying test case consists of a large diode chain model, [2], that is very suitable for reduction due to internal redundancy, coupled to a two dimensional oscillating DAE system. This second subsystem is dependent on the large diode chain through the voltage at Φ_2.

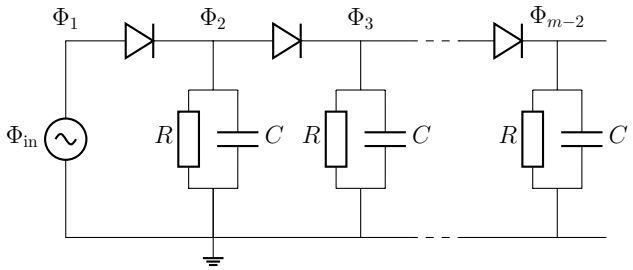

Figure 1: The diode chain

The diode chain model is described by the following differential-algebraic equations

$$\begin{aligned}
\Phi_1 - \Phi_{\text{in}} &= 0,\\
I(\Phi_1, \Phi_2) - I(\Phi_2, \Phi_3) - C\frac{\mathrm{d}\Phi_2}{\mathrm{d}t} - \frac{1}{R}\Phi_2 &= 0,\\
I(\Phi_{i-1}, \Phi_i) - I(\Phi_i, \Phi_{i+1}) - C\frac{\mathrm{d}\Phi_i}{\mathrm{d}t} - \frac{1}{R}\Phi_i &= 0,\\
I(\Phi_i, \Phi_{i+1}) - C\frac{\mathrm{d}\Phi_{i+1}}{\mathrm{d}t} - \frac{1}{R}\Phi_{i+1} &= 0,\\
i_{\text{E}} - I(\Phi_1, \Phi_2) &= 0.
\end{aligned}$$

$$I(x, y) = I_{\text{s}}\left[\mathrm{e}^{(x-y)/0.0256} - 1\right], \quad \Phi_{\text{in}} = 8\sin(7 \cdot 10^8 \cdot \frac{t}{2\pi}).$$

Where $C = 10^{-11}$ and R is a coupled resistance term. Through this term, the variables of the slow subsystems depend only weakly on the variables variables of the fast subsystem, this coupling is given by $R = R_0 + y_1 \cdot 10^2$, where y_1 is defined by a fast oscillating academic DAE system that is dependent the nodal voltage Φ_2.

$$\begin{aligned}
0 &= C_{\text{A}}\frac{\mathrm{d}y_1}{\mathrm{d}t} - y_2 - \frac{1}{R}\Phi_2\\
0 &= y_2 - \sin(7 \cdot 10^8 t).
\end{aligned}$$

Below, in Figure 2 and 3, the results of a transient analysis for a time interval from 0 to 37.5 ns is shown. The diode chain parameters are $R_0 = 10000\ \Omega$ and $C = 10^{-11}$.

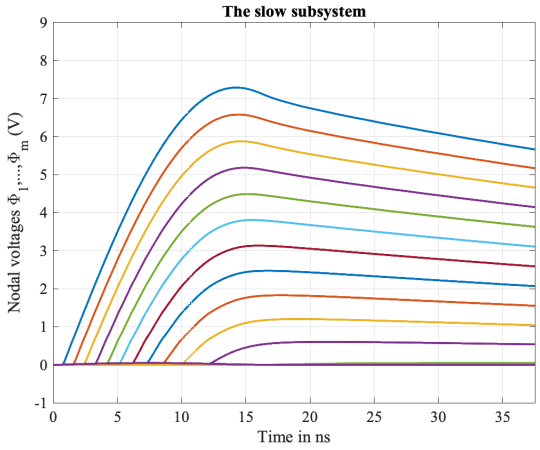

Figure 2: The evolution of the slow subsystems from the transient analysis. From top to bottom we have $\Phi_1, ..., \Phi_m$.

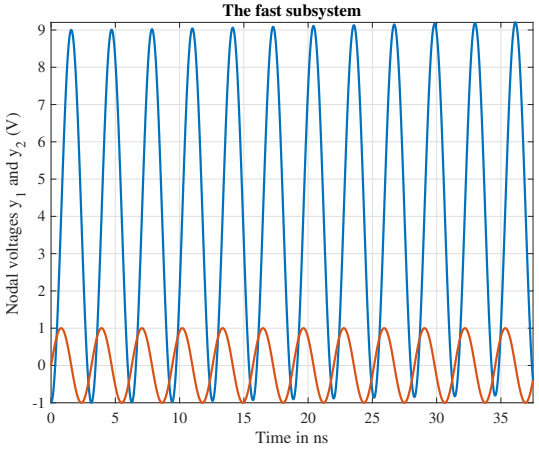

Figure 3: The evolution of the fast subsystems from the transient analysis. In blue we have y_1 and in red y_2.

For the fast subsystem the resistance is taken to be equal to that of the diode chains resistors, and the capacitance is set to $C_{\mathrm{A}} = 10^{-10}$. The dimension parameter of the diode chain is set to $m = 1000$. The snapshot matrix is provided by a high accuracy integration of the full system and snapshots are taken with $\Delta t_{\mathrm{HF}} = 0.0375$ ns. By applying the ϵ estimation procedure we obtain $\epsilon_* = 0.1928$ and this results in the reduced system size $r = 14$. The same reduced basis size is used for the gappy reconstruction.

5.2. Implicit Euler versus ROMR

Regarding the convergence of the ROMR integration scheme, Figure 4 illustrates the order 1 convergence rate. We see that the ROMR accuracy is nearly identical to that of the full order solutions. Furthermore, in Figure 5 it shows that this accuracy is achieved with a significant reduction in computational time. The computational effort is almost a order of magnitude lower for the reduced schemes, while the precision is maintained. The positive effects of model order reduction, multirate time integration and the combination of both is evident.

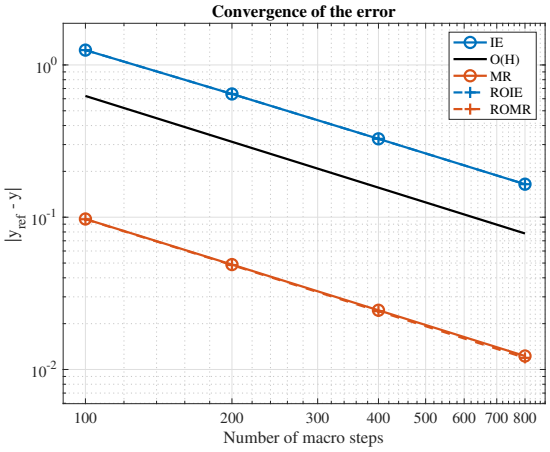

Figure 4: The order 1 convergence of the computational error descending parallel to the black reference convergence.

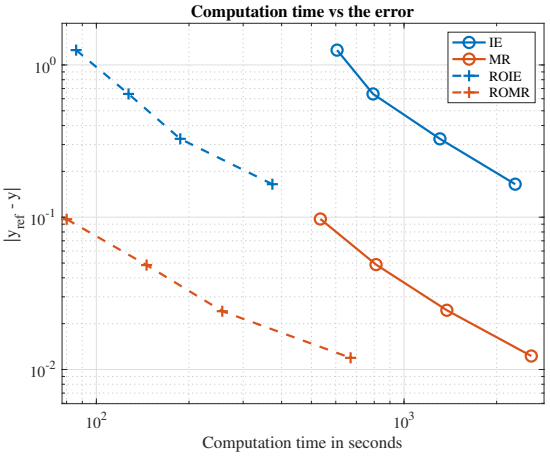

Figure 5: The effect of the different numerical methods on the computational time and accuracy.

6. Conclusions and outlook

In this work, the mathematical foundations of the reduced order multirate (ROMR) scheme have been presented. The method consists of the Gauß-Newton with approximated tensors (GNAT) nonlinear model order reduction method. This has been extended with a hyper reduction by gappy data reconstruction and a coupled slowest-first multirate integration scheme. Both the reduction and hyper reduction methods use a reduced basis obtained by the maximum entropy snapshot sampling (MESS) method followed by a QR-decomposition.

By numerical analysis, the ROMR has been shown to have an order 1 convergence rate for the error, under the assumption that the large dimensional slower time scale model can be accurately reduced. The analytical results have been verified by a numerical experiment. A diode chain model was reduced and coupled to an oscillator and a transient analysis of the resulting model has been performed. The results show that the ROMR scheme performs as predicted and is capable of outperforming a regular integration scheme.

Further research will be done to incorporate these techniques into fully functional simulation software and measure the performance with real world test cases as provided by STMicroelectronics. Another interesting topic for further research regards the optimallity of the reduction base as provided by the MESS method.

Acknowledgements

The authors are indebted to the funding given by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 765374, ROMSOC.

References

[1] MWFM Bannenberg, A Ciccazzo, and M Günther. Coupling of model order reduction and multirate techniques for coupled dynamical systems. *Applied Mathematics Letters*, 112:106780, 2020.

[2] T Bechtold, A Verhoeven, EJW Ter Maten, and T Voss. Model order reduction: an advanced, efficient and automated computational tool for microsystems. In *Applied And Industrial Mathematics In Italy II*, pages 113–124. World Scientific, 2007.

[3] H. Broer and F. Takens. *Dynamical systems and chaos*. Springer-Verlag New York, 2011.

[4] Kevin Carlberg, Charbel Farhat, Julien Cortial, and David Amsallem. The GNAT method for nonlinear model reduction: effective implementation and application to computational fluid dynamics and turbulent flows. *Journal of Computational Physics*, 242:623–647, 2013.

[5] Saifon Chaturantabut and Danny C. Sorensen. Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.*, 32(5):2737–2764, 2010.

[6] Zlatko Drmac and Serkan Gugercin. A new selection operator for the discrete empirical interpolation method—improved a priori error bound and extensions. *SIAM Journal on Scientific Computing*, 38(2):A631–A648, 2016.

[7] Richard Everson and Lawrence Sirovich. Karhunen–loeve procedure for gappy data. *JOSA A*, 12(8):1657–1664, 1995.

[8] Michael Günther, Anne Kvaernø, and Peter Rentrop. Multirate partitioned Runge-Kutta methods. *BIT Numerical Mathematics*, 41(3):504–514, 2001.

[9] Christoph Hachtel, Andreas Bartel, Michael Günther, and Adrian Sandu. Multirate implicit Euler schemes for a class of differential–algebraic equations of index-1. *Journal of Computational and Applied Mathematics*, page 112499, 2019.

[10] F. Kasolis, D. Zhang, and M. Clemens. Recurrent quantification analysis for model reduction of nonlinear transient electro-quasistatic field problems. In *International Conference on Electromagnetics in Advanced Applications (ICEAA 2019)*, pages 14–17, 2019.

[11] Fotios Kasolis and Markus Clemens. Maximum entropy snapshot sampling for reduced basis generation. *arXiv preprint arXiv:2005.01280*, 2020.

[12] F. Takens. On the numerical determination of the dimension of an attractor. In *Dynamical systems and bifurcations*, pages 99–106. Springer, 1985.

[13] Howell A.M. Tong. *Dimension estimation and models*, volume 1. World Scientific, 1993.

[14] B.L. van der Waerden. *Mathematical Statistics*. Springer, 1969.

[15] A. Verhoeven. *Redundancy reduction of IC models : by multirate time-integration and model order reduction*. PhD thesis, Department of Mathematics and Computer Science, 2008.

[16] Karen Willcox. Unsteady flow sensing and estimation via the gappy proper orthogonal decomposition. *Computers & fluids*, 35(2):208–226, 2006.