Tensor-train approximation of the chemical master equation and its application for parameter inference

Ion Gabriel Ion [*,**] **Christian Wildner** [**]
Dimitrios Loukrezis [*,**] **Heinz Koeppl** [*,**,***]
Herbert De Gersem [*,**]

[*] *Centre for Computational Engineering, Technische Universität Darmstadt (e-mail: ion@temf.tu-darmstadt.de).*
[**] *Department of Electrical Engineering, Technische Universität Darmstad.*
[***] *Centre for Synthetic Biology, Technische Universität Darmstadt.*

1. INTRODUCTION

Traditional chemical kinetic models use ordinary differential equations (ODEs) to predict the concentrations of the involved molecule types (Gillespie (1992)). The evolution of the corresponding probability distribution is given by the chemical master equation (CME) which, in principle, can be solved by numerical integration. Unfortunately, the computational cost grows exponentially with the number of species, due to the fact that the system states must be labeled explicitly to cast the CME into a ODE. A framework for performing Bayesian inference tasks for the parameter-dependent CME is suggested, by exploiting the so called tensor-train (TT) decomposition to approximate the joint distribution over the CME states and parameters (see Ion et al. (2021)). For that purpose, we construct an explicit representation of the evolution operator in the TT format and show that it can be constructed without ever assembling the corresponding matrix. The TT format has the advantage that the storage requirement scales linearly with respect to the number of dimensions, while at the same time being a numerically robust tensor decomposition. To that end, we combine the state space and the parameter space into a higher-dimensional tensor-product space. The parameter dependence is expressed by means of a B-spline basis. Since typically every reaction is governed by an individual rate constant, the parameters can be seamlessly included in the tensor representation, thus allowing for efficiently solving the joint system. In practice, however, the system parameters are often unknown. Therefore, we develop a framework for filtering, smoothing, and parameter inference based on the efficient TT representation of the joint system.

2. TENSOR-TRAIN DECOMPOSITION FOR THE CHEMICAL MASTER EQUATION

2.1 Chemical master equation

The chemical master equation describes the time evolution of the probability mass function (PMF) of well-mixed reaction system with d species (Gillespie (1992)). After the state space truncation of the PMF $\mathbf{p}$ to a $n_1 \times \cdots \times n_d$ tensor, the CME is represented as a large linear system of ODEs

$$\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = \mathbf{A}\mathbf{p}(t),$$

where $\mathbf{A}$ is a tensor operator with size $(n_1 \times \cdots \times n_d) \times (n_1 \times \cdots \times n_d)$. Parameters that govern the reactions can be included in the framework, leading to the following parameter dependent CME

$$\frac{\mathrm{d}\mathbf{p}(t,\boldsymbol{\theta})}{\mathrm{d}t} = \mathbf{A}(\boldsymbol{\theta})\,\mathbf{p}(t,\boldsymbol{\theta}). \tag{1}$$

For the joint state-parameter density together with the time dependency, a tensor-product basis representation in used:

$$\mathsf{p}_{\boldsymbol{i}}(t,\boldsymbol{\theta}) \approx \sum_j \sum_{\boldsymbol{l}} \mathbf{p}_{\boldsymbol{i}\boldsymbol{l}j} b_j(t) L_{\boldsymbol{l}}(\boldsymbol{\theta}), \tag{2}$$

where $\mathbf{p}$ is $d + N_p + 1$ dimensional tensor, $\{b_j\}_j$ is the basis for the time dependency (Chebyshev polynomials) and $\{L_{\boldsymbol{l}}\}_{\boldsymbol{l}}$ is a tensor-product basis for the parameter to accommodate parameter dependency (product of univariate B-splines). Galerkin projection is then used to derive an extended multilinear system for $\mathbf{p}$. Since the unknown $\mathbf{p}$ is high-dimensional, the storage requirements grow exponentially and therefore compression schemes are employed for the unknown $\mathbf{p}$ and the multilinear system.

2.2 Tensor-trains

An array $\mathbf{x}$ of shape $n_1 \times \cdots \times n_d$ is said to be in the TT format if it can be elementwise written as

$$\mathbf{x}_{\boldsymbol{i}} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \cdots \sum_{r_{d-1}=1}^{R_{d-1}} \mathbf{g}^{(1)}_{1 i_1 r_1} \mathbf{g}^{(2)}_{r_1 i_2 r_2} \cdots \mathbf{g}^{(d)}_{r_{d-1} i_d 1}, \tag{3}$$

where the three-dimensional tensors $\mathbf{g}^{(k)}$ are called the TT-cores and $\boldsymbol{R} = (1, R_1, ..., R_{d-1}, 1)$ are called the TT-ranks (Oseledets (2011) provides a detailed look). The storage complexity becomes linear with respect to d and once the tensors are converted in the TT-format, the basic operations (elementwise addition, multiplication, summing over indices) can be efficiently performed without building the full d-dimensional object. The construction of a low-rank TT-decomposition provides an error bound and rank reduction can be also performed within a given accuracy (see Oseledets (2011) for more details). In addition

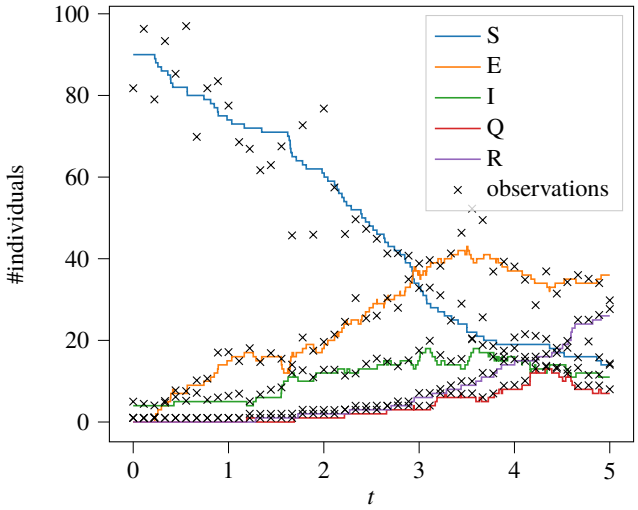

Fig. 1. Noisy observation sample for the SEIQR model (sample size is 45).

to that, multilinear systems can be solved directly in the TT-format using optimization based methods (see Dolgov and Savostyanov (2014)). The CME operator **A** presented in the previous section can be directly represented in the TT-format and the multilinear system arising from the Galerkin projection can be as well solved in the TT-format.

2.3 Inference tasks

Given a number of noisy observations of a system governed by a CME with unknown governing parameters, one can be interested in finding those parameters. A probabilistic description of the distribution over the parameter space (called posterior) can be obtained using Bayes rule. As presented in Ion et al. (2021), updating the posterior implies solving the CME and constructing the likelihood (conditional probability of observing the data given the underlying state of the system). Both of the steps are efficiently performed using the TT-format without being affected by the curse of dimensionality, since both the observation model and the CME operator can be computed directly in the TT-format. A prior PDF over the parameters can be included in the framework.

3. RESULTS

Numerical experiments have been performed to showcase the advantages of the proposed framework in terms of accuracy and computational efficiency Ion et al. (2021). Among them, we present here only he SEIQR model. It has 5 species: susceptible (S), exposed (E), infected (I), quarantined (Q) and recovered (R) involved in 9 reactions with 4 parameters assumed as unknown. From a sample path, noisy observations are generated (see Fig. 1) and the TT CME solver is used to infer the parameters with the dimension of the parameter space basis is 64 for every parameter. Marginals of the posterior are shown in Fig. 2, comparing the obtained posterior with the prior (green dashed line) and also displaying the exact parameter.

The execution time for a TT-solver is $\approx$ 55 minutes with a maximum posterior size in the QTT-format of $\approx$ 30 MB. As a comparison, the chosen state truncation

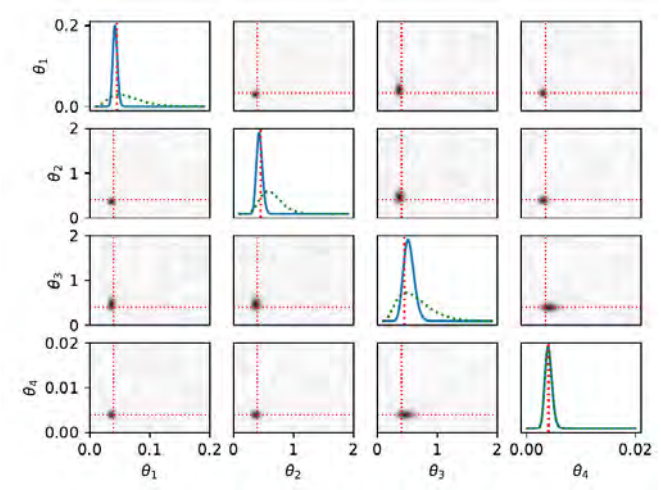

Fig. 2. Posterior marginal distributions for the four unknown reaction rates of the SEIQR model. The exact parameters are marked with the red dashed lines and the prior with green dashed lines.

of $(128, 64, 64, 32, 32)$ would require $\approx$ 4.2 GB only for storing the state for one parameter realization. The storage complexity for the parameter-dependent CME operator in the QTT format is $\approx$ 200 KB.

4. CONCLUSION

We presented a method based on the TT decomposition to solve the CME, either in its standard form or including parameter dependencies, and approximate the joint distribution over the state-parameter space, including the time dependency as well. Using the considered TT-framework, inference tasks such as parameter identification can be performed accurately and efficiently.

ACKNOWLEDGEMENTS

The work of I.G. Ion is supported by the Graduate School Computational Engineering within the Centre for Computational Engineering at Technische Universität Darmstadt. D. Loukrezis is supported by the German Federal Ministry for Education and Research (BMBF) via the research contract 05K19RDB.

REFERENCES

Dolgov, S.V. and Savostyanov, D.V. (2014). Alternating minimal energy methods for linear systems in higher dimensions. *SIAM Journal on Scientific Computing*, 36(5), A2248–A2271. doi:10.1137/140953289.

Gillespie, D.T. (1992). A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1), 404–425. doi: https://doi.org/10.1016/0378-4371(92)90283-V.

Ion, I.G., Wildner, C., Loukrezis, D., Koeppl, H., and De Gersem, H. (2021). Tensor-train approximation of the chemical master equation and its application for parameter inference. *The Journal of Chemical Physics*, 155(3), 034102. doi:10.1063/5.0045521.

Oseledets, I. (2011). Tensor-train decomposition. *SIAM J. Sci. Comput.*, 33, 2295–2317.