Structured Projection-Based Model Reduction with Application to Stochastic Biochemical Networks

Aivar Sootla, Member, IEEE, and James Anderson, Member, IEEE

Abstract—The Chemical Master Equation (CME) is well known to provide the highest-resolution models of a biochemical reaction network. Unfortunately, even simulating the CME can be a challenging task. For this reason, simpler approximations to the CME have been proposed. In this work, we focus on one such model, the Linear Noise Approximation. Specifically, we consider implications of a recently proposed LNA time-scale separation method. We show that the reduced-order LNA converges to the full-order model in the mean square sense. Using this as motivation we derive a network structure-preserving reduction algorithm based on structured projections. We discuss when these structured projections exist and we present convex optimisation algorithms that describe how such projections can be computed. The algorithms are then applied to a linearised stochastic LNA model of the yeast glycolysis pathway.

Index Terms—Model Reduction, Structured Model Reduction, Linear Noise Approximation, Chemical Master Equation, Stochastic Differential Equations.

I. INTRODUCTION - MODELS

Stochasticity is inherent in biochemical networks. The most general (and unfortunately, most complex) model that best encapsulates the behaviour of such a network is the Chemical Master Equation (CME). The CME is a continuous-time, infinite-dimensional Markov process that describes the evolution of a probability mass function of the number of the species of all reactants in a given biochemical reaction.

The CME models a reaction network comprising of \( R \) reactions and \( \alpha \) species evolving in a compartment of fixed volume \( \Omega \) and takes the form

\[
\frac{\partial \mathcal{P}(N,t)}{\partial t} = \Omega \sum_{i=1}^{R} \left( f(N - S_i, \Omega) - f(N, \Omega) \right) \mathcal{P}(N,t) \quad (1)
\]

where the vector \( N = [N_1, \ldots, N_\alpha]^* \) indicates the total number of molecules of each species in the volume \( \Omega \), \([\cdot]^*\) denotes transposition, \( f \) is the flux-vector, and \( S \in \mathbb{R}^{\alpha \times R} \) is the stoichiometry matrix (the \( i \)th column of \( S \) is denoted by \( S_i \)). Finally, \( \mathcal{P}(N,t) \) is the probability that at time \( t \) the number of molecules of each species is given by \( N \).

It is fairly clear from the form of (1) that the complexity of solving the CME quickly becomes intractable for all but the most simple of networks. Some approaches were designed to solve the CME approximately (cf. [26]), a great deal of research has focused on efficient methods for simulating the CME. The most popular approach is Gillespie’s stochastic simulation algorithm [14] and the computationally more efficient version known as the \( \tau \)-leaping algorithm [15]. The reader is directed to [15] for a detailed description of both algorithms and their derivations. In one of the extensions of \( \tau \)-leaping [9] the authors propose to replace propensities which correspond to reactions that are in some sense “fast”, by their averages and simulate only “slow” reactions. The idea of averaging (or integrating) out a part of a stochastic process to reduce simulation time dates back to Khasminskii (cf. [11]). This type of approach is typically called time-scale separation or averaging.

Recent developments in systems and synthetic biology revived the interest in time-scale separation of stochastic biochemical networks. In [38], the authors derived a time-scale separation method for the so-called Linear Noise Approximation (LNA) of a CME. The LNA is a Gaussian process, which approximates a CME under the assumption of a large number of reactions occurring in a large volume. We will revisit this method in what follows. In [18], the authors applied the classical Tikhonov theorem (cf. [22]) to the moments of the Chemical Langevin Equation (a nonlinear SDE approximation of the CME). It can be argued that the averaging is implicitly applied while computing these moments. Finally, stochastic averaging for CMEs has been recently proposed in [21]. The authors of [10] point out that there is not always a clear distinction between fast and slow species in a stochastic model and thus finding a transformation to the standard form is not trivial. In [13] a method for identifying distinct fast and slow variables with a view to model reduction is presented.

In the control literature, time-scale separation methods were phased out by the use of so-called projection-based methods (cf. [4]). Projection methods were specifically derived for input-output systems and typically provide better approximations (in terms of the \( L_2 \) gain). The most effective projections are computed by first mapping a system to a balanced state space [43, Ch. 3] (i.e. coordinates in which the controllability and observability ellipsoids of the realisation are aligned). In this case, it can be guaranteed that the input-output behaviour of the approximated system is similar to the original one. In the context of linear stochastic differential equations (SDEs) we can use the intrinsic noise (the Brownian motion driving...
the process) as an input [17]. This idea also appears in the so-called low-noise limit results in stochastic calculus (cf. [11]).

A caveat in using balancing is that the projections typically destroy any physical interpretation of the state space, which is not desirable in many applications, especially when analysing networked systems where structure is encoded in the system matrices. One cure for this problem relies on graph partitioning and clustering algorithms [1], [25], which unfortunately do not provide error bounds if the nodes have dynamics beyond simple integrators. In an alternative approach, called structured balanced truncation (cf. [30]), only a part of the state-space can be projected to a lower-dimensional space. Even though it is still an open question as to when structured projections exist and can be computed, in some cases existence and polynomial-time computation can be guaranteed. For instance, in the case of positive directed networks, model reduction can be performed with trivial projections [37]. A more sophisticated projection approach was recently proposed in [20]. Related work appears in [23], where structured coprime factorisations are used for model reduction, and in [8], where a similar approach is used for reduction of discrete-time LFT models.

Computing an approximation using a structured truncation approach is typically computationally expensive even in the linear time-invariant case. Furthermore, the simulation time can be reduced by other means, for example through non-structured model reduction. Therefore, the primary goal of structured balanced truncation is, typically, to simplify the analysis and the controller design of networked systems. In particular, a group of states (in our case, species) may be left intact, while the rest of the states can be approximated through a projection. This can be used in synthetic biology, for example, to study the loss of information when studying a genetic circuit independently of the entire biochemical network. The algorithms we describe are very much in line with the ethos of modularisation in systems and synthetic biology [6]. We believe they provide a valuable toolbox that allows for directed analysis of biological models while maintaining the physical meaning of the simplified model.

Contributions: In previous work ([34], [35]) we derived the main idea of the algorithm and showed that it can be applied to positive and monotone systems. In this paper, we extend these results in several directions, providing new theoretical results and implementable algorithms. In particular the main contributions of this paper are: a) We show that the averaged (reduced-order) LNA model converges to the full order LNA in the mean square sense. This extends the original result [38], where only convergence in distribution is argued; b) The structured balanced singular perturbation results are then used to derive an algorithm for providing a structure preserving reduced-order LNA model with error bounds; c) We identify a broad class of systems for which we can guarantee that structured projections can be computed for. The property we exploit is related to diagonal dominance of the drift matrix (see [42] for control theoretic implications).

Organisation: The paper is organised as follows. In Section II, we briefly introduce singular perturbation techniques for linear and stochastic systems. We also discuss balancing methods for linear, structured and nonlinear model order reduction. In Section III, we present the averaging result for the LNA and provide the convergence proof. In Section IV we apply structured model reduction techniques to the LNA and illustrate the application on examples in Section V. Auxiliary results and discussions are found in the Appendix.

Notation: $A^*$ denotes the complex conjugate transpose of the matrix $A$. The norm $\| \cdot \|_2$ is the standard induced matrix norm. $A \geq 0$ ($A \succ 0$) denotes that $a_{ij} \geq 0$ (resp., $a_{ij} > 0$) for all $i, j$. When $A$ is square symmetric, $A \succeq 0$ denotes that $A$ is positive semidefinite. $\mathbb{E}(\xi)$ and $\text{cov}(\xi)$ stand for the mean and the covariance of the random variable $\xi$, respectively, while $\xi \in \mathcal{N}(\mu, \Sigma)$ indicates that $\xi$ is drawn from a Gaussian distribution with mean $\mu$ and covariance matrix $\Sigma$.

II. PRELIMINARIES

A. Linear Model Reduction by Singular Perturbation

Consider a stable linear time-invariant input-output system depending on a positive parameter $\varepsilon$

$$
\dot{x}_1 = A_{11}x_1 + A_{12}x_2 + B_1u, \\
\dot{x}_2 = \varepsilon^{-1}(A_{21}x_1 + A_{22}x_2 + B_2u), \\
y = C_1x_1 + C_2x_2 + Du,
$$

where $x_1 \in \mathbb{R}^r$, $x_2 \in \mathbb{R}^{n_2}$, $u \in \mathbb{R}^{n_u}$, $y \in \mathbb{R}^{n_y}$ and the matrices $A_{ij}, B_i, C_j, D$ have appropriate dimensions. It is straightforward to check that for a small parameter $\varepsilon$, the variable $x_2$ varies on a faster time-scale than the variable $x_1$.

Therefore, if $x_2$ converges to a constant value and the speed of convergence is fast (that is, $\varepsilon \ll 1$), then we can assume that $x_2$ is at steady state with respect to the slower dynamics and thus approximate the system by eliminating $x_2$. Formally, the reduced-order model is obtained by setting $\varepsilon = 0$ and $x_2 = -A_{22}^{-1}A_{21}x_1 - A_{22}^{-1}B_2u$, which yields:

$$
\dot{z} = (A_{11} - A_{12}A_{22}^{-1}A_{21})z + (B_1 - A_{12}A_{22}^{-1}B_2)u, \\
y = (C_1 - C_2A_{22}^{-1}A_{21})z + (D - C_2A_{22}^{-1}B_2)u.
$$

This reduced-order model approximates the full-order model, if $A_{11} - A_{12}A_{22}^{-1}A_{21}$ is a stable matrix. This singular perturbation method can be applied to autonomous systems (that is, systems with $u = 0$ for all $t$), can be extended to non-linear systems, and preserves a physical interpretation of the states $x_1$. Furthermore, an extension to stochastic differential equations can also be derived. Consider a system

$$
\dot{x}_1 = A_{11}x_1 + A_{12}x_2 + B_1w, \\
\dot{x}_2 = \varepsilon^{-1}(A_{21}x_1 + A_{22}x_2 + B_2w) \\
y = C_1x_1 + C_2x_2 + Dw,
$$

where $w$ is an $n_w$-dimensional Wiener process. In this case, the reduced-order model is also obtained as

$$
\dot{z} = (A_{11} - A_{12}A_{22}^{-1}A_{21})z + (B_1 - A_{12}A_{22}^{-1}B_2)\dot{w}, \\
y = (C_1 - C_2A_{22}^{-1}A_{21})z + (D - C_2A_{22}^{-1}B_2)\dot{w}.
$$

However, the derivation and the intuition here are slightly different. In this case, we cannot directly set $\varepsilon$ equal to zero and assume that $A_{21}x_1 + A_{22}x_2 + B_2\dot{w}$ is equal to zero for all times $t$, since $\dot{w}$ has infinite variation. Therefore, the variable
\( x_2 \) is integrated out instead, but the procedure yields the same result. Due to integrating out the variables, this approach is sometimes referred to as averaging.

Whilst singular perturbation methods are widely and successfully used, however, there are some important limitations:

- the singular perturbation method does not take into account the actual magnitude of \( \varepsilon \);
- time-scale separation may not exist, that is, we may not assume the existence of a small \( \varepsilon \);
- if time-scale separation does not exist, then the variable \( x_2 \) cannot be eliminated in a unique way (the matrix \( A_{22} \) may not be invertible), or the dynamics of the reduced system (that is, the matrix \( A_{11} - A_{12}A_{22}^{-1}A_{21} \)) may be unstable.

In light of these limitations, other methods, such as balanced truncation, were developed that were seen to be more suitable for control theory.

**B. Projection-Based Model Order Reduction by Balancing**

Let us now consider a stable linear time-invariant input-output system with \( n_i \) inputs, \( n_o \) outputs, and \( n \) states:

\[
\begin{align*}
\dot{x} &= Ax + Bu, \\
y &= Cx + Du.
\end{align*}
\]

We will adopt the following shorthand notation for a realisation of a system

\[
G = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \quad \Leftrightarrow \quad G(s) = C(sI - A)^{-1}B + D, \tag{2}
\]

where \( s \in \mathbb{C} \) and \( G(s) \) is a representation of the system in the frequency domain subject to the Laplace transformation (cf. [43]). The essential step in projection-based methods is the computation of the so-called projectors \( V \in \mathbb{R}^{r \times n} \) and \( W \in \mathbb{R}^{r \times n}, \) where \( r < n \) is a pre-defined order of the reduced model. Given the projectors the approximate model is computed as follows

\[
G_r = \begin{bmatrix} VAW & VB \\ CW & D \end{bmatrix}, \tag{3}
\]

where the superscript indicates the state dimension. Note that for the full-order \( n \)-dimensional system we omit the superscript \( n \). In particular we would like to minimise the norm of the error system given by

\[
\| G - G_r \|_{\mathcal{H}_\infty},
\]

where

\[
\| G \|_{\mathcal{H}_\infty} := \sup_{Re(s) > 0} \sigma[G(s)] = \text{ess sup}_{\omega \in \mathbb{R}} \sigma[G(j\omega)].
\]

The projectors \( V \) and \( W \) can be computed using interpolation methods based on Krylov subspace techniques (cf. [4]). However, here we will employ balancing tools (cf. [4]) because in this case, we can compute structured projectors (introduced in the sequel) that can be applied to networked systems.

First, we cover the celebrated balanced truncation method and consider the Lyapunov equations

\[
\begin{align*}
AP + PA^* + BB^* &= 0, \tag{4a} \\
A^*Q + QA + C^*C &= 0, \tag{4b}
\end{align*}
\]

which are associated with a realisation of \( G \). If \( A \) is asymptotically stable then there exist unique solutions \( P \geq 0, \) \( Q \geq 0 \) to (4), which are called controllability and observability Gramians respectively. If additionally \( (A, B) \) is controllable and/or \( (A, C) \) is observable, the respective Gramian will be positive definite. The eigenvalues \( \sigma_i \) of the matrix \( (PQ)^{1/2} \) are referred to as Hankel singular values of \( G \). We call a realisation balanced if \( P = Q = \Sigma = \text{diag} \{\sigma_1, \ldots, \sigma_n\} \). The following proposition summarises two key balancing methods for model reduction and their respective error bounds.

**Proposition 1:** Consider a realisation of \( G \) from (2) and assume there exist positive definite matrices \( P \) and \( Q \) satisfying (4). Let \( \sigma_i \) be the Hankel singular values of \( G \) and let \( r \) be such that \( \sigma_i \neq \sigma_j \) for all \( i \leq r, j > r \). Let \( T \) be such that \( TPT^* = (T^*)^{-1}QT^{-1} = \Sigma = \text{diag} \{\sigma_1, \ldots, \sigma_r\} \) and consider the realisation with the following partitioning

\[
\hat{G} = \begin{bmatrix} TAT^{-1} & TB \\ CT^{-1} & D \end{bmatrix} = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} & \hat{B}_1 \\ \hat{A}_{21} & \hat{A}_{22} & \hat{B}_2 \\ \hat{C}_1 & \hat{C}_2 & D \end{bmatrix}, \tag{5}
\]

where \( \hat{A}_{11} \in \mathbb{R}^{r \times r}, \hat{B}_1 \in \mathbb{R}^{r \times n}, \hat{C}_1 \in \mathbb{R}^{n \times r} \). Then the reduced-order realisations \( \hat{G}_1, \hat{G}_2 \) defined as

\[
\hat{G}_1 = \begin{bmatrix} \hat{A}_{11} & \hat{B}_1 \\ \hat{C}_1 & D \end{bmatrix}, \quad \hat{G}_2 = \begin{bmatrix} \hat{A}_{11} - \hat{A}_{12}\hat{A}_{22}^{-1}\hat{A}_{21} & \hat{B}_1 - \hat{A}_{12}\hat{A}_{22}^{-1}\hat{B}_2 \\ \hat{C}_1 - \hat{C}_2\hat{A}_{22}^{-1}\hat{A}_{21} & D - \hat{C}_2\hat{A}_{22}^{-1}\hat{B}_2 \end{bmatrix},
\]

are both asymptotically stable, and satisfy the error bounds:

\[
\| G - \hat{G}_1 \|_{\mathcal{H}_\infty} \leq 2 \sum_{i=r+1}^{n} \sigma_i, \quad \| G - \hat{G}_2 \|_{\mathcal{H}_\infty} \leq 2 \sum_{i=r+1}^{n} \sigma_i.
\]

Unlike the singular perturbation method in the previous subsection, both balancing approaches can be applied to any stable system, while preserving stability and guaranteeing a certain quality of approximation. The application of a transformation \( T \) to \( G \) can be seen as a transformation of the state-space variable \( \dot{x} = Tx \). The balancing transformation is computed as \( T = \Sigma^{1/2}UR^{-1} \), where \( R \) is a lower triangular matrix such that \( R^*R \), while \( \Sigma \) and \( U \) are obtained from the singular value decomposition \( QR^* = US \Sigma^2U^* \). Note that we do not require \( \sigma_i \geq \sigma_{i+1} \) for all \( i = 1, \ldots, n-1 \), which may be confusing in the standard formulation but is useful in the structured extension. The proof for the error bounds still holds in this case.

The algorithm to compute \( \hat{G}_1 \) is usually called balanced truncation, while the algorithm to compute \( \hat{G}_2 \) is called balanced singular perturbation (cf. [43], [24]). The balanced singular perturbation approach matches the full-order system at the zero frequency point, while the balanced truncation matches the full system at infinite frequency. Note also that
balanced singular perturbation can be derived as the balanced truncation of a system with $\tilde{A} = A^{-1}$, $\tilde{B} = A^{-1}B$, $\tilde{C} = CA^{-1}$, $\tilde{D} = D - CA^{-1}B$. Finally, the balanced truncation is called projection-based, since we set

$$V = (I_r \ 0) T, \quad W^* = (I_r \ 0) (T^*)^{-1}$$

such that the reduced-order model $\hat{G}_i^*$ has the realisation as in (3).

C. Structured Balanced Reduction

The balancing transformation $T$ is typically a full matrix, and hence any physical meaning in $x$ is not preserved in the new variables $\hat{x} = Tx$. Additionally the sparsity of the drift matrix is lost under the transformation $TAT^{-1}$.

In some cases, the state $x$ is naturally partitioned into two groups of states $x_1, x_2$ such that $x = (x_1^* \ x_2^*)^*$, as for example in closed-loop systems with a controlled system having the states $x_1$ and a controller with the states $x_2$. In this case, it is not desirable for the transformation $T$ to mix $x_1$ and $x_2$ as to do so would destroy the controlled system-controller structure. Hence, $T$ has to be block-diagonal, and consequently so should the Gramians $P$ and $Q$. In order to find block-diagonal Gramians we need to consider the so-called Lyapunov inequalities instead of equations:

$$AP + PA^* + BB^* \preceq 0, \quad (7a)$$
$$A^*Q + QA + C^*C \preceq 0. \quad (7b)$$

The positive semidefinite solutions $P$ and $Q$ to (7) are called generalised Gramians. In what follows, we will consider generalised Gramians with a certain sparsity pattern $S$, and write $P \in S$, if $P$ has the sparsity pattern $S$. Note that these Gramians may not exist in general, but we discuss their existence in Appendix B and in further detail in [36].

If $P = Q = \Sigma = \text{diag} \{\sigma_1, \ldots, \sigma_n\}$, then the realisation is called balanced in the generalised sense, while $\sigma_i$’s are called generalised Hankel singular values. The model reduction procedure is the same, and the error bounds are instead given in the form of generalised Hankel singular values. We refer to the reduction procedure, which consists of computing structured Gramians and projections, as structured balanced reduction. The following result generalises Proposition 1 to structured balanced reduction [3].

Proposition 2: Consider a realisation $G$ in (2) and let there exist positive definite $P \in S$ and $Q \in S$ satisfying inequalities in (7). Then the statement of Proposition 1 holds, while the Hankel singular values and balancing are understood in the generalised sense.

Since there are infinitely many solutions to the matrix inequalities (7) (in fact the solutions define a subspace), there are infinitely many combinations of the generalised Hankel singular values $\{\sigma_1, \ldots, \sigma_n\}$. We want, however, to find such $P$ and $Q$ that the smallest $\sigma_i$’s are close to zero in order to achieve minimal reduction error. In this case, the following heuristic is usually proposed:

$$\min_{P \in S, \ P \succeq 0} \text{trace}(P)$$
$$\text{subject to: } AP + PA^* + BB^* \preceq 0. \quad (8)$$

Note that if the constraint $P \in S$ is dropped we obtain a solution that matches the Lyapunov equation solution (4a). The programme to compute $Q$ is derived in a similar manner. Due to symmetry of the decision variable, the trace minimisation here acts as a rank minimisation programme on $P$ and $Q$, thus minimising the smallest generalised Hankel singular values [28].

D. Extensions to Nonlinear Systems

Some of the balancing and projection techniques were extended to nonlinear systems, however, we will consider the most straightforward extensions only. Consider the system

$$\dot{x} = g(x, u),$$
$$y = h(x, u),$$

and linearise it around a fixed point $x^0$ with $u = 0$. After the computation of a constant state-space transformation $T$ for the linearised model, we can compute a change of variables $\hat{x} = Tx$ in order to balance the linearised model. The transformed nonlinear equation takes the form:

$$\dot{\hat{x}} = Tg(T^{-1}\hat{x}, u) = \hat{g}(\hat{x}, u),$$
$$y = h(T^{-1}\hat{x}, u) = h(\hat{x}, u).$$

Next, the reduced-order model can be obtained by truncation

$$\dot{\hat{z}} = Vg(Wz, u),$$
$$y = h(Wz, u),$$
or methods similar to singular perturbation approach

$$\dot{\hat{z}} = Vg(Wz + W_2z, u),$$
$$0 = Vg(W_2z + W_2z, u),$$
$$y = h(Wz + W_2z, u),$$

where $V$, $W$ satisfy (6) and

$$V_t = (0 \ I_{n-r}) T, \quad W_t^* = (0 \ I_{n-r}) (T^*)^{-1}.$$

The transformation $T$ does not in general create a time-scale separation, hence this approach should be considered a heuristic. However, around the steady state $x^0 = Tz^0$ the system admits a time-scale separation locally in terms of the input-output relationship, if the reduced states correspond to sufficiently small Hankel singular values in comparison with other values. In the rest of the paper, we discuss an extension of this heuristic to stochastic biochemical networks while employing structured projections.

III. REDUCTION OF THE LINEAR NOISE APPROXIMATION

A. Problem Formulation

In this paper, we consider an approximation of the CME (1): the Linear Noise Approximation or LNA (cf. [38]). This approximation is valid if a large number of reactions occur per unit time and additionally the volume $\Omega$, where the reactions occur, is sufficiently large. In this case let

$$\frac{n}{\Omega} = x + \Omega^{-1/2} \eta,$$
where \( x \) is the vector of macroscopic concentrations of the species and \( \eta \) is a vector of stochastic fluctuations about \( x \). Now, by applying a Taylor expansion to the CME, it can be shown (cf. [38]) that the fluctuations \( \eta \) and macroscopic concentrations \( x \) obey the following equations

\[
\dot{x} = g(x), \quad (9a) \\
\dot{\eta} = A(x)\eta + B(x)\dot{\omega}, \quad (9b)
\]

where \( g(x) = Sf(x) \), \( A(x) \) is the Jacobian of \( Sf(x) \), \( B(x) = \Omega^{-1/2}\text{Sdiag}\left\{\sqrt{f(x)}\right\} \). \( S \) is the stoichiometric matrix from (1), \( f(x) \) is an approximation of \( f(N,\Omega) \) and \( w \) is a \( R \)-dimensional Wiener process. Note that the macroscopic fluctuation \( f \) approximating the microscopic rate functions \( f \) for the four fundamental reactions as well as some more complex reactions are given in [39]. In fact it is shown that \( \lim_{\eta \rightarrow \infty} f(N,\Omega) = f(N) \). We are now in a position to formulate the main problem this paper addresses:

**Problem formulation:** Given a partitioning of the states

\[
[x^* \ \eta^*]^T = [x_1^* \ x_2^* \ \eta_1^* \ \eta_2^*]^T, \quad (10)
\]

where \( x_1, \eta_1 \in \mathbb{R}^k \), and \( x_2, \eta_2 \in \mathbb{R}^{n-k} \), find a reduced-order representation of (9), while keeping the vectors \( x_1, \eta_1 \) intact and reducing \( n-k-r \) states with \( n-k-r > 0 \) in the vectors \( x_2, \eta_2 \).

**B. Averaging of Linear Noise Approximation**

To streamline presentation we will drop the dependence on \( x \) from the notation when referring to \( A(x) \) and \( B(x) \). An important observation is that the matrices \( A, B \) do not depend on the fluctuations \( \eta \) but depend only on the macroscopic concentrations \( x \), which is computed using deterministic differential equations. Therefore the fluctuation dynamics constitute a linear time-varying SDE and the mean and the covariance of \( \eta \) can be computed as follows (cf. [11]).

**Proposition 3:** The covariance \( \text{cov}(\eta(t)) = P(t) \) and the mean \( \mathbb{E}(\eta(t)) = m(t) \) of the solution to the SDE (9b) satisfy

\[
\frac{dm}{dt} = Am, \quad m(t_0) = m_0, \\
\frac{dP}{dt} = AP + PA^* + BB^*, \quad P(t_0) = P_0, \quad (12)
\]

where \( X(t_0) \sim \mathcal{N}(m_0, P_0) \).

We start the discussion of time-scale separation of the LNA by assuming that the state vector of (9) has been appropriately permuted and partitioned as in (10). The vector field \( g \) and the matrices \( A \) and \( B \) can then be conformally partitioned according to (10). Note that the *true state* of the system is given by \( x + \eta \). In some cases, we can assume that \( x_1 + \eta_1 \) varies on a time-scale which is \( \varepsilon \) times slower than the time-scale of \( x_2 + \eta_2 \). In this case the LNA (9) can be written as follows [38]:

\[
\dot{x}_1 = g_1(x_1, x_2), \quad (13a) \\
\dot{x}_2 = -g_2(x_1, x_2), \quad (13b) \\
\dot{\eta}_1 = A_{11}\eta_1 + \varepsilon^{-1/2}A_{12}\eta_2 + B_1\dot{\omega}, \quad (13c) \\
\dot{\eta}_2 = \varepsilon^{-1/2}A_{21}\eta_1 + \varepsilon^{-1/2}A_{22}\eta_2 + \varepsilon^{-1/2}B_2\dot{\omega}. \quad (13d)
\]

Under standard conditions on time-scale separation, the reduced-order model is given by the following equations [38]:

\[
\dot{z} = g_1(z, \tilde{\varepsilon}), \quad (14a) \\
\dot{\tilde{\varepsilon}} = A_r(z, \tilde{\varepsilon})\tilde{\varepsilon} + B_r(z, \tilde{\varepsilon})\dot{\omega}, \quad (14b)
\]

where \( \tilde{\varepsilon} \) is the unique root of the equation \( g_2(x_1, x_2) = 0 \) solved with respect to \( x_2 \), and

\[
A_r = A_{11} - A_{12}A_{22}^{-1}A_{21}, \\
B_r = B_1 - A_{12}A_{22}^{-1}B_2. \quad (15)
\]

We can now present the first main result of the paper:

**Theorem 1:** Consider the system (13), where \( g_1(x), g_2(x) \) are continuously differentiable functions with bounded derivatives and \( A_{22} \) is invertible along the trajectory of the full-order model (13) and locally exponentially stable for all \( x \). Let the system (13a-13b) satisfy standard assumptions on time-scale separation in [22, pp. 9–11]. Then there exists \( \varepsilon_1 \) such that for all \( \varepsilon \) satisfying \( \varepsilon_1 \geq \varepsilon \geq 0 \) we have

\[
\sup_{0 \leq t \leq T} \|z - x_1\|_2 = O(\varepsilon), \quad \sup_{0 \leq t \leq T} \mathbb{E}\|\eta - \eta_1\|_2^2 = O(\varepsilon),
\]

where \( z, \varepsilon \) are solutions to (14).

Before providing the proof we note that this reduced-order model was derived in [38], however, the authors argued that there is convergence in distribution with \( \varepsilon \to 0 \). This method can be seen as a type of stochastic averaging, since fast variables \( x_2 + \eta_2 \) are essentially integrated out. We also note that a similar convergence result to the reduced-order model (14a,14b) can be shown if the fluctuations \( \eta_1, \eta_2 \) evolve according to the following model

\[
\dot{\eta}_1 = A_{11}\eta_1 + A_{12}\eta_2 + B_1\dot{\omega}, \\
\dot{\eta}_2 = \varepsilon^{-1}(A_{21}\eta_1 + A_{22}\eta_2 + B_2\dot{\omega}).
\]

This model can be obtained from (13c,13d) by a change of variables \( \tilde{\eta}_1 = \eta_1, \tilde{\eta}_2 = \varepsilon^{-1/2}\eta_2 \).

**Proof:** According to [22] the root \( \tilde{\varepsilon} \) exists and for all \( t \) such that \( 0 \leq t \leq T \) we have that \( x_1(t) - z(t) = O(\varepsilon) \), therefore we need to only prove convergence of the fluctuation dynamics. We can rewrite the equation for the slow perturbation variable as follows:

\[
\dot{\eta}_1 = A_{11}\eta_1 + \varepsilon^{-1/2}A_{12}\eta_2 + B_1\dot{\omega}, \\
= A_r(x_1, x_2)\eta_1 + B_r(x_1, x_2)\dot{\omega}, \\
+ A_{12}A_{22}^{-1}(\varepsilon^{-1/2}A_{22}\eta_2 + A_{21}\eta_1 + B_2\dot{\omega}).
\]

Taking into account this representation we obtain

\[
\xi(t) - \eta_1(t) = \int_0^t (B_r(z, \xi) - B_r(x_1, x_2))\dot{\omega}d\tau + \int_0^t (A_r(z, \xi) - A_r(x_1, x_2))\xi d\tau + \int_0^t A_r(x_1, x_2)(\xi - \eta_1) d\tau + \int_0^t A_{12}A_{22}^{-1}(\varepsilon^{-1/2}A_{22}\eta_2 + A_{21}\eta_1 + B_2\dot{\omega})d\tau, \quad (16)
\]

where the matrices \( A_{ij}, B_i \) depend on \( x(\tau) \). We prove the main result by showing that the mean squares of each of the
terms on the right hand side of (16) are of order $O(\varepsilon)$. Due to the Itô isometry rule (cf. [29]) we have

$$
\mathbb{E} \left\| \int_0^t \left( B_r(z, \hat{z}) - B_r(x_1, x_2) \right) w \, d\tau \right\|_2^2 \\
= \int_0^t \left\| B_r(z, \hat{z}) - B_r(x_1, x_2) \right\|_F^2 \, d\tau \\
\leq L_1 \int_0^t \left\| B_r(z, \hat{z}) - B_r(x_1, x_2) \right\|_F^2 \, d\tau \leq O(\varepsilon),
$$

where $L_1 \geq 0$ and the last inequality is due to Lemma 1 (See Appendix A). By using consecutively Cauchy-Schwartz inequality, and the bounds in Lemma 1, we have

$$
\mathbb{E} \left\| \int_0^t (A_r(z, \hat{z}) - A_r(x_1, x_2)) \xi \, d\tau \right\|_2^2 \\
\leq E \int_0^t \left\| (A_r(z, \hat{z}) - A_r(x_1, x_2)) \xi \right\|_2^2 \, d\tau \\
= \int_0^t \mathbb{E} \left\| (A_r(z, \hat{z}) - A_r(x_1, x_2)) \xi \right\|_2^2 \, d\tau \\
\leq L_2 \int_0^t \left\| A_r(z, \hat{z}) - A_r(x_1, x_2) \right\|_E^2 \, d\tau \int_0^t \mathbb{E} \left\| \xi \right\|_2^2 \, d\tau \\
\leq L_2 \int_0^t \left\| A_r(z, \hat{z}) - A_r(x_1, x_2) \right\|_E^2 \, d\tau \int_0^t \mathbb{E} \left\| \xi \right\|_2^2 \, d\tau \leq O(\varepsilon),
$$

for some $L_2 \geq 0$. Similarly we can show that

$$
\mathbb{E} \left\| \int_0^t A_r(x_1, x_2)(\eta_1 - \xi) \, d\tau \right\|_2^2 \leq K_1 \int_0^t \mathbb{E} \left\| \eta_1(t) - \xi(t) \right\|_2^2 \, d\tau
$$

for some positive $K_1$. Using Lemma 4 (see Appendix A) it is shown that the final term in (16) denoted by $C_1(t)$ satisfies $E\|C_1(t)\|^2 = O(\varepsilon)$. Finally, let $m_e(t) = \mathbb{E} \left\| \eta_1(t) - \xi(t) \right\|_2^2$. By applying the previous bounds to (16), we obtain

$$
m_e(t) \leq K_1 \int_0^t m_e(\tau) \, d\tau + O(\varepsilon).
$$

Therefore by Gronwall’s Lemma (Appendix A, Lemma 2) we have that $m_e(t) \leq O(\varepsilon) e^{K_1 t}$. As this inequality is considered only on a finite interval it follows that $m_e(t) = O(\varepsilon)$ for all $t \in [0, T]$.

Theorem 1 has shown that the reduced-order LNA model converges to the full-order LNA in the mean square sense. As in the deterministic case, there are limitations to this approach:

- we do not take into account the magnitude of $\varepsilon$;
- we reduce all the states $x_2$, $\eta_2$;
- the assumptions on the separation of time-scales may not be fulfilled.

In order to alleviate these limitations and obtain reduced-order models, we will employ structured balancing as a heuristic. Furthermore, we can guarantee stability and approximation quality of the reduced-order model locally around a fixed point of (9a).

IV. PROJECTION-BASED MODEL REDUCTION OF A LINEAR NOISE APPROXIMATION

A. A Reduced-Order Model

In this section, we describe how structured projection-based methods can be applied to model reduction of LNA as introduced in Section III. By order of the LNA, we mean the dimension of the vector $x + \eta$. It is assumed that we want to preserve the physical interpretation (and hence network structure) of the first $k$ out of $n$ states (chemical species) and that the full-order model takes the following form

$$
\dot{x} = g(x), \\
\dot{\eta} = A(x)\eta + B(x)\omega, \\
y = C(x + \eta), \\
x(0) = x_0, \quad \eta(0) = 0,
$$

where $y$ is an artificially introduced “output” of our Gaussian process with $C = (I_k \ 0_{n-k})$. Note that other choices of $C$ are perfectly valid: in particular, replacing the identity matrix with an arbitrary full matrix may be desirable. In the context of experimental biology, $C$ may be chosen to represent outputs for which data can be produced. In this setting standard model-reduction techniques aim to synthesise a model that approximates (17) in the input-output sense, i.e. the map from $\omega$ to $y$ but with fewer states. In this work we show how the physical structure of the first $k$ states can be preserved whilst the remaining $n - k$ states form a subsystem which is then reduced to order $r$ where $r < n - k$ but with no physical interpretation.

First we linearise the process (17) $\eta$ around a steady-state $x_{ss}$ of the mean dynamics and obtain the following SDE

$$
\ddot{\eta} = A\eta + B\dot{\omega}, \\
y = C\eta,
$$

where $A = A(x_{ss})$, $B = B(x_{ss})$. Let $G$ denote the realisation of the stochastic system (18). Our goal is to choose the transformation $T : (A, B, C) \rightarrow (TA^{-1}, TB, CT^{-1})$ such that by applying averaging to the transformed model, we obtain a reduced-order SDE $\dot{G}^{k+r}$ which is stable, while $\|G - \dot{G}^{k+r}\|$ is minimised in some norm. As a suitable criterion, we consider the standard $\mathbb{L}_2$ and $\mathbb{H}_\infty$ norms. The $\mathbb{L}_2$ norm has the interpretation of the integral of the trace of the covariance matrix of the process $y(t)$. The $\mathbb{H}_\infty$ norm is maximum over frequencies of the largest eigenvalue values of the spectral density of the process $y(t)$ [7]. In order to preserve the structure of the first $k$ state equations we will introduce a structured transformation $T = \text{diag} \{ I_k, T_2 \}$.

Once the transformation matrix $T$ has been constructed, we obtain the new states $\hat{x} = Tx$, $\hat{\eta} = T\eta$, and consider the following full-order model:

$$
\begin{align*}
\dot{\hat{x}} &= Tg(T^{-1}\hat{x}), \\
\dot{\hat{\eta}} &= TA(T^{-1}\hat{x})T^{-1}\hat{\eta} + TB(T^{-1}\hat{x})\dot{\omega}.
\end{align*}
$$

Given the system (19), which is equivalent to the system (17), a reduced-order model can then be computed as in [38]. Indeed, assume that

$$
\begin{align*}
V &= \text{diag} \{ I_{k+r}, 0_{k+r,n-k-r} \} T, \\
V_t &= \text{diag} \{ 0_{n-k-r,k+r}, I_{n-k-r} \} T, \\
W^* &= \text{diag} \{ I_{k+r}, 0_{k+r,n-k-r} \} (T^*)^{-1}, \\
W^*_t &= \text{diag} \{ 0_{n-k-r,k+r}, I_{n-k-r} \} (T^*)^{-1}.
\end{align*}
$$
Let $A^*(\cdot) = A_{11}(\cdot) - A_{12}(\cdot)A_{22}^{-1}(\cdot)A_{21}(\cdot)$, and $B^*(\cdot) = B_1(\cdot) - A_{12}(\cdot)A_{22}^{-1}(\cdot)B_2(\cdot)$, then

$$\dot{z} = Vg(Wz + Wtzt),$$

$$0 = V_ig(Wz + W_iz),$$

$$\dot{\xi} = VA^*(Wz + W_iz)\xi + VB^*(Wz + W_iz)\dot{w}.$$ (21c)

In practice we do not compute the Lipschitz constants or verify the separation of time-scales. In this sense our approach should be seen as a heuristic. If an ODE solver fails to solve (21) due to existence of multiple roots to (21b), and/or stability issues, we can set $z_i = V_ia^0$, and discard the algebraic equation.

### B. Computation of Structured Transformations

Consider an SDE (18) and fix the dimension of the reduced-order model to be equal to $k + r$. We will use the structured balanced singular perturbation approach, which results in the following reduced-order realisation

$$\hat{G}^{k+r} = \begin{bmatrix} \hat{A}_{11} & -\hat{A}_{12}\hat{A}_{22}^{-1}\hat{A}_{21} & \hat{B}_1 - \hat{A}_{12}\hat{A}_{22}^{-1}\hat{B}_2 \\ \hat{C}_1 & 0 \end{bmatrix},$$

where the matrices $V$, $W$, $V_t$, $W_t$ are given in (20) and

$$\hat{A}_{11} = VA_{11}W, \quad \hat{A}_{12} = VA_{12}W_t, \quad \hat{A}_{21} = V_tA_{21}W, \quad \hat{A}_{22} = V_tA_{22}W_t, \quad \hat{B}_1 = VB, \quad \hat{B}_2 = V_tB, \quad \hat{C}_1 = C_1.$$ (22)

To make comparisons with $\hat{G}_2$ in Proposition 1, note that $C_2$ is equal to zero in our case, hence $\hat{C}_2 = C_2W_t$ is equal to zero as well. All that remains is to compute a state-space transformation $T = \text{diag}\{I_k, T_2\}$ – if it exists. This will be the focus of the next section.

**$H_\infty$ balancing:** In this section we will assume that structured Gramians exist, i.e. $P, Q \in S$ satisfy (7). When the drift matrix $A$ is diagonally stable, which by definition means that

$$\exists \Sigma > 0, \quad \text{s.t.} \quad \Sigma A + AA^* < 0$$

where $\Sigma$ is diagonal, then existence of the generalised Gramian is guaranteed [34]. We refer the reader to Appendix B for results that broaden the class of systems which admit structured Gramians.

In order to compute the projections consider the following semidefinite programmes

$$\min_{\Sigma_P,P_2} \text{trace}(P)$$

s. t.: $AP + PA^* + BB^* < 0$\hspace{1cm}(23a)

$$P = \begin{pmatrix} \Sigma_P & 0_{e,n-k} \\ 0_{n-k,k} & P_2 \end{pmatrix} > 0$$

$$\min_{\Sigma_Q,Q_2} \text{trace}(Q)$$

s. t.: $A^*Q + QA + C^*C < 0$\hspace{1cm}(23b)

$$Q = \begin{pmatrix} \Sigma_Q & 0_{e,n-k} \\ 0_{n-k,k} & Q_2 \end{pmatrix} > 0$$

where $\Sigma_P, \Sigma_Q$ are diagonal matrices. Now we are in a position to state the following result.

**Theorem 2:** Let $A$ be diagonally stable and the matrices $P$ and $Q$ satisfy LMIs (23a)–(23b). Next, let $T$ be such that $TPT^* = T^{-1}Q(T^*)^{-1} = \Sigma$, let $\sigma_i$ denote the eigenvalues of $(\Sigma_2Q_2)^{1/2}$ such that $\sigma_i \geq \sigma_{i+1}$, and let $\sigma_r > \sigma_{r+1}$ for some integer $r < n - k$. Consider the projections defined in (20) and reduced-order model $\hat{G}_r^{k+r}$ defined in (22). Then the reduced system $\hat{G}_r^{k+r}$ is diagonally stable and

$$\|G - \hat{G}_r^{k+r}\|_{H_\infty} \leq 2 \sum_{i=r+1}^{n-k} \sigma_i.$$ (24)

**Proof:** The proof of this theorem is by application of the results in Subsection II-B and Lemma 5 in Appendix C. If we are not interested in preserving diagonal stability, then we can relax the structure of the matrices $\Sigma_P$ and $\Sigma_Q$ to be full positive definite matrices. Additionally, if the system is not diagonally stable, but block-diagonal Gramians $P, Q$ exist (that is $\Sigma_P$ and $\Sigma_Q$ are full positive definite matrices), then Theorem 2 still holds with $\hat{G}_r^{k+r}$ being a stable realisation.

**$H_2$ balancing:** An arguably better way of measuring the norm of a stochastic process is the $H_2$ norm, defined as

$$\|G\|_{H_2}^2 := \frac{1}{2\pi} \int_{-\infty}^{\infty} \text{trace}(G^*(j\omega)G(j\omega))d\omega.$$ (25)

There are many methods for model reduction in the $H_2$ norm (cf. [12], [16]), however, none of them can easily be extended to the structured projection techniques. However, there exists a simple heuristic, according to which we only diagonalise the generalised controllability Gramian $P$ computed by the programme (23a). In this case, the matrix $P$, over-approximates the covariance of $\eta$ around the linearisation point $x_s$. We were not able to obtain any meaningful error bounds for this heuristic, but the computational results are satisfactory and are demonstrated in what follows. Again the proof is a straightforward application of the results in Subsection II-B and Appendix C.

**Theorem 3:** Let $A$ be diagonally stable and $P$ be a solution of (23a). Let $T$ be such that $TPT^* = \Sigma$, let $\sigma_i$ denote the eigenvalues of $P_2$ such that $\sigma_i \geq \sigma_{i+1}$, and let $\sigma_r > \sigma_{r+1}$ for some integer $r < n - k$. Consider the projections defined in (20) and reduced-order model $\hat{G}_r^{k+r}$ defined in (22). Then the realisation $\hat{G}_r^{k+r}$ is diagonally stable.

We have shown that the structured transformations always exist if the drift matrix $A$, which was found by linearising (19) about a stable fixed point, is diagonally stable. It is hard to formally define a class of biochemical networks admitting a diagonally stable drift matrix, but our numerical computations indicate that for many such networks this condition holds. This observation has also been made by other authors c.f. [33]. We elaborate on this observation in Appendix B.

### V. Examples

We now provide two examples that illustrate the theory developed. The first is a simple example that demonstrates the importance of respecting the network interaction structure when using structured model reduction techniques. The second model is a more realistic biological example that demonstrates how multiple parts of a network can be preserved.
Fig. 1. Toy Example. In the configuration in panel (a), species $S_1$ and $S_3$ are grouped together and state is removed. In the configuration in panel (b), species $S_1$, $S_2$, and $S_3$, $S_4$ are grouped together, with one state from each group being reduced.

**A. Comparison of the models**

We compare separately the error in the macroscopic dynamics (mean) and the fluctuations (variance). The error $E(y - y_r)$ in macroscopic dynamics is computed by perturbing the initial state $x_i$ from the steady-state $x_{ss}$, and is measured in $L_1$, $L_2$ and $L_\infty$ norms.

A comparison in terms of the fluctuations $\eta$ is performed by computing the covariance matrix of the outputs $y$ and $y_r$. For the full-order model this matrix is computed as

$$\text{cov}(y) = C \text{cov}(\eta \eta^*) C^* = CPC^*,$$

where $P$ satisfies the Lyapunov equation (12). Similarly, the covariance matrix for the reduced-order models $\text{cov}(y_r)$ can be computed. Note that the $L_2$ error of the outputs serves as a lower bound on the $H_2$ norm computation.

**B. Toy Example.**

The first network we consider consists of only four species, see Figure 1. One can interpret the species $B$. Toy Example. For the full-order model this matrix is computed as $\text{cov}(y) = CPC^*$, where $P$ satisfies the Lyapunov equation (12). Similarly, the covariance matrix for the reduced-order models $\text{cov}(y_r)$ can be computed. Note that the $L_2$ error of the outputs serves as a lower bound on the $H_2$ norm computation.

The first network we consider consists of only four species, see Figure 1. One can interpret the species $B$. Toy Example. For the full-order model this matrix is computed as $\text{cov}(y) = CPC^*$, where $P$ satisfies the Lyapunov equation (12). Similarly, the covariance matrix for the reduced-order models $\text{cov}(y_r)$ can be computed. Note that the $L_2$ error of the outputs serves as a lower bound on the $H_2$ norm computation.

We can observe that $m_i$ representing the mRNA concentrations are faster than $p_i$ representing the protein concentrations, hence species $S_1$, and $S_3$ can be reduced using the results of [38]. We can view this method as using (21) with a trivial projection $T = I_4$ (subject to a permutation of states), $k = 2$, $r = 0$. We also obtain reduced-order models by using non-trivial projections according to the configurations in Figure 1. In the configuration depicted in Figure 1(a) we reduce only one species, while in the configuration depicted in Figure 1(b) we reduce two species. We initiate simulations from $x_0 = (1, 10, 1, 1)^*$, which lies in the domain of attraction of the steady-state $x_{ss} = (0.2889, 3.4611, 0.0578, 0.6922)^*$. We compute the initial covariance $P_0$ from the linearisation around $x_0$ and then compute $P(t)$ and $m(t)$ from (11)-(12). We compute the projections with respect to the linearised model at the fixed point $x_{ss}$. This linearised model has a Metzler drift matrix (see Appendix B for definition and properties), hence the diagonal Gramian always exists and we can test different configurations for model reduction. The reduced-order models are then computed as in (21).

Comparisons between the various reduced and full models are depicted in Figure 2. It can be seen that the trivial projection approach (i.e., $T = I_4$) always performs worse than the $H_\infty$ balancing method for the configuration in Figure 1(b). This happens since the species vary on time-scales which are not separated well enough (i.e., the magnitude of $\varepsilon$ is large). This example also highlights the importance of selecting which parts of the system to reduce: in the configuration depicted in Figure 1(a) we reduce one species, but doing so is worse than reducing two species according to the configuration depicted in Figure 1(b). This happens since we do not respect the topology. The $H_\infty$ balancing does poorly compared to the trivial projection approach even though reduction for the linearised models gave a similar result to $H_\infty$ balancing.

**C. Kinetic Model of Yeast Glycolysis.**

This 12-state model was published in [40] and is provided online in the supplementary material to [27]. This model consists of twelve metabolites, four boundary fluxes, and it is deterministic, however, we obtain a Gaussian process by
setting the diffusion term to $S_{\text{diag}} \left\{ \sqrt{f_i(x)} \right\}$. We set the state-space of the model to:

$$x = \left( \begin{array}{c} \text{GLCi} \ G6P \ F6P \ F16P \ TRIO \ BPG \ldots P3G \ P2G \ PEP \ PYR \ ACALD \ NADH \end{array} \right)^T$$

We model the network’s response to the change of glucose in the system. We treat levels of ATP and glycose GLCo as control inputs. At time zero we change the levels of ATP and GLCo from 3 to 1.5 and 0.25 to 5 respectively. Let $x_0$ be the steady state with ATP = 3 and GLCo = 0.25, while $x_s$ is the steady state ATP = 1.5 and GLCo = 5.

We refer the reader to [40] for a complete description of the model, but we mention that the drift matrices of the linearised dynamics around the model, but we mention that the drift matrices of the steady state with $ATP = 3$ and $GLCo = 0.25$, while $x_s$ is the steady state ATP = 1.5 and GLCo = 5.

We refer the reader to [40] for a complete description of the model, but we mention that the drift matrices of the linearised dynamics around $x_0$ and $x_s$, are stable $\mathcal{H}$ matrices (see Appendix B for definition and properties). Hence we can compute diagonal Lyapunov matrices by solving semi-definite programmes. We pick two groups of species to reduce (\{BPG, P3G, P2G, PEP\} and \{GLCi, G6P, F6P\}); see Figure 3, however, we consider them separately, meaning that the Gramians $P$ and $Q$ have three blocks, not two. As demonstrated above violating topological constraints in the graph (mixing these two groups) can result in reduced-order models of worse quality. Even though in the described configuration we have two groups of “lumped” states and a group of “intact” states, we can obtain a reduced-order model in a straightforward manner similar to the procedure in (21). In this case, the projection matrices will have three blocks on the diagonal.

### Table I

**Reduction of the glycolysis model. The error** $\mathbb{E}[\|y(t) - y_r(t)\|]$ **in different norms, where $y$ and $y_r$ are the trajectory of the full and reduced-order models, respectively.**

<table>
<thead>
<tr>
<th>Reduced States \ Error</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$L_\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F6P, 2PG, PEP</td>
<td>1.2143</td>
<td>0.7400</td>
<td>0.9782</td>
</tr>
<tr>
<td>F6P, 3PG, 2PG, PEP</td>
<td>1.5740</td>
<td>1.0674</td>
<td>1.5582</td>
</tr>
</tbody>
</table>

**Table I-A. Approximation results obtained by using (21) with $T = I_{12}$ and $r = 0$.**

<table>
<thead>
<tr>
<th>Lumpred Region(s)</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$L_\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>{G6P, F6P}</td>
<td>1.1816</td>
<td>0.7864</td>
<td>1.0118</td>
</tr>
<tr>
<td>{G6P, F6P}</td>
<td>1.4176</td>
<td>0.7273</td>
<td>0.8702</td>
</tr>
<tr>
<td>{G6P, F6P}</td>
<td>0.3818</td>
<td>0.2527</td>
<td>0.3243</td>
</tr>
<tr>
<td>{G6P, F6P}</td>
<td>1.4129</td>
<td>0.7242</td>
<td>0.8651</td>
</tr>
</tbody>
</table>

**Table I-B. Reduction by $\{l_1, l_2\}$ states in every region using the $L_\infty$ balancing.**

<table>
<thead>
<tr>
<th>Lumpred Region(s)</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$L_\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>{G6P, F6P}</td>
<td>1.0990</td>
<td>0.6857</td>
<td>0.8815</td>
</tr>
<tr>
<td>{G6P, F6P}</td>
<td>1.1002</td>
<td>0.6936</td>
<td>0.8954</td>
</tr>
<tr>
<td>{G6P, F6P}</td>
<td>0.2935</td>
<td>0.1587</td>
<td>0.1979</td>
</tr>
<tr>
<td>{G6P, F6P}</td>
<td>0.2104</td>
<td>0.1080</td>
<td>0.1313</td>
</tr>
</tbody>
</table>

The error is computed by simulating the resulting reduced-order models and comparing them as described at the beginning of the section. The results are presented in Table I for various reduction configurations. We apply (21) with $T = I_{12}$, and $r = 0$ to metabolite concentrations. We note that this approach is equivalent to the method outlined in [38]. Using the proposed method we lump those metabolites in one state so that the number of reduced states is similar in both cases. The first two rows of each sub-table in Table I can be compared directly, and it is clear that the proposed method performs better in terms of quality than using (21) with $T = I_{12}$.

The proposed methods are also more flexible in terms of reduction choices. In the fourth row of Table I-B, the region \{BPG-PEP\} contains four metabolites; however, we reduced only two states after computing the state-space transformation. In the fifth row, in the region \{GLCi-F6P\}, which contains three metabolites, we reduce just one state and this provides us with the best model among all the reduction attempts. Finally, the results in Table I-C indicate that the $L_2$ balancing outperforms the $L_\infty$ balancing on this example.

### VI. Conclusion

In this paper, we studied model order reduction of the Linear Noise Approximation of the Chemical Master Equation. We showed that a time-scale separation method results in a reduced-order model, which converges in the mean-square sense to the slow dynamics of the LNA. We then considered the application of structure-preserving, projection-based model reduction to the LNA. One of the bottlenecks of projection-based methods is existence of the projectors, which cannot be always guaranteed. We were able to provide sufficient conditions that describe when such projectors exist. Furthermore, these are spectral conditions on the drift matrix of the linearised dynamics, hence they are easy to check.

As a straightforward extension of this approach, we may consider model reduction for time-varying SDEs using, for
example, [31]. This may provide better quantitative approximations of LNAs. However, there are deeper issues with the LNA itself. If the underlying CME is bimodal (in some cases this implies, for example, that the deterministic model of macroscopic reaction rates is bistable), then LNA and hence our approximation procedure will not capture this phenomenon. Therefore one needs to derive projection-based reduction methods for the CME or at least for the Chemical Langevin Equation, which is a nonlinear SDE. In this case, it is perhaps possible to use nonlinear balancing tools [32], which are based on a nonlinear generalisation of the Gramians using energy functions. The controllability energy function is identical to the action functional, which is used to solve the fastest escape problem (cf. [11]). It remains to establish, however, whether this action functional can be used for model reduction.

APPENDIX A. TECHNICAL LEMMAS FOR THE PROOF OF THEOREM 1

Lemma 1: The functions $A_r(\cdot), B_r(\cdot)$ defined in (15) satisfy the following bounds

$$\int_0^t \|A_r(z, \hat{z}) - A_r(x_1, x_2)\|_2^2 dt \leq O(\epsilon),$$

$$\int_0^t \|B_r(z, \hat{z}) - B_r(x_1, x_2)\|_2^2 dt \leq O(\epsilon),$$

$$\|A_r(x_1, x_2)\|_2 \leq K_1,$$

where $x_1(t), x_2(t), z(t), \hat{z}$ are defined in (13), and (14).

Proof: According to [22] the root \( \hat{z} \) exists and for all $t$ such that $0 \leq t \leq T$ we have that $x_1(t) - z(t) = O(\epsilon)$, therefore the statement is well posed.

We begin with the final inequality in the lemma. As the elements of the matrix $A_r$ are polynomials in $z$ and $\hat{z}$ the final inequality follows from the definition on the operator norm and the fact that polynomial functions are Lipschitz on a compact domain of arbitrary size (c.f. Lemma 3 in Appendix VI). Using this bound and the equivalence of norm property, we can show the first inequality as follows:

$$\int_0^t \|A_r(z, \hat{z}) - A_r(x_1, x_2)\|_2^2 dt$$

$$\leq \int_0^t \|A_r(z, \hat{z}) - A_r(x_1, x_2)\|_2^2 dt$$

$$\leq L_1 \int_0^t \left( \|x_1(t) - z(t)\|_2^2 + \|x_2(t) - \hat{z}(t)\|_2^2 \right) dt,$$

where $\| \cdot \|_F$ is the Frobenius matrix norm.

Furthermore, $x_2(t)$ asymptotically converges to $\hat{z}(t)$ with $\epsilon \to 0$, hence there exists a small enough $\epsilon_1$ such that for all $0 \leq \epsilon \leq \epsilon_1$ we have $\int_0^t \|x_2(t) - \hat{z}(t)\|_2^2 dt \leq O(\epsilon)$. The same argument is used for the second inequality involving $B_r$.

Lemma 2 (Gronwall Lemma): Let $g(t), t \in [0, T]$ be a non-negative, continuous real-valued function that satisfies

$$g(t) \leq C + K \int_0^t g(s) ds$$

for all $t \in [0, a]$ where $C$ and $K$ are positive constants. Then it follows that for all $t \in [0, a]$,

$$g(t) \leq Ce^{Kt}.$$
the drift matrix $A$ in the realisation $G$ is diagonally stable, then we can find diagonal generalised Gramians as shown in [34]. There exist no easy parametrisations of the class of diagonally stable matrices, but there are some sufficient conditions. For example, in [5] it was shown that the so-called cacti graphs admit diagonal generalised Gramians. We will not formally define these graphs, but just mention that these graphs rule out sparsity patterns generated by reversible reactions, and thus their applicability to biochemical reaction networks is limited. Our results build upon [19] that certifies existence of diagonal Lyapunov functions for a broad class of graphs that seem to appear frequently in biochemical reaction networks. To proceed we require a few definitions:

**Definition 1:** A matrix $A = \{a_{ij}\}_{i,j=1}^{n} \in \mathbb{R}^{n \times n}$ is called Metzler if it has non-negative off-diagonal elements, that is $a_{ij} \geq 0$ for $i \neq j$.

Note that there is no implication of stability in the definition of Metzler matrices.

**Definition 2:** A matrix $M(A) = \{m_{ij}\}_{i,j=1}^{n}$ is called a comparison matrix of $A = \{a_{ij}\}_{i,j=1}^{n} \in \mathbb{R}^{n \times n}$, if $m_{ij} = \begin{cases} |a_{ij}| & i = j \\ -|a_{ij}| & i \neq j \end{cases}$.

**Definition 3:** A matrix $A \in \mathbb{R}^{n \times n}$ is called an $H$ matrix if $M(A)$ has all eigenvalues with a nonnegative real part. If additionally $a_{ii} > 0$ for all $i$ we say that $A$ is an $H_+ \ pmatrix{F_{11} & F_{12} \\ F_{21} & F_{22}} \prec 0$, where $\Sigma_1$ is diagonal, and $P_2$, $X$ are full positive definite matrices. Let $W$ be an invertible matrix partitioned $W^* = \begin{pmatrix} w_1 & w_2 \end{pmatrix}$, where $w_1$ has $r$ columns. Let also $V = W^{-1} = \begin{pmatrix} v_1 & v_2 \end{pmatrix}$, where $v_1$ has $r$ columns. Furthermore, let $V$ be such that $VP_2V^* = \Sigma_2 = \begin{pmatrix} \Sigma_{2,1} & 0 \\ 0 & \Sigma_{2,2} \end{pmatrix}$, where $\Sigma_{2,1}$ is an $r \times r$ diagonal matrix and $\Sigma_{2,2}$ is a diagonal matrix of an appropriate dimension. Then the matrix

$$F_r = \begin{pmatrix} F_{11} & F_{12}w_1 \\ v_1^TF_{21} & v_1^TF_{22}w_1 \end{pmatrix} - \begin{pmatrix} F_{12}w_2(v_2^TF_{22}w_2)^{-1}v_2 & (F_{21}w_1)^* \end{pmatrix}$$

is diagonally stable.

**Proof:** By the premise we have that

$$\text{Sym} \begin{pmatrix} (F_{11}v_1^TF_{21}) & F_{12}v_1^TF_{22}w_1 \\ (v_1^TF_{21})^TF_{12} & (v_1^TF_{22}w_1)^* \end{pmatrix} \prec -FF^* \begin{pmatrix} I & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & V \end{pmatrix} = Y.$$
REFERENCES


Aivar Sootla received his MSc from Lomonosov Moscow State University (Russia) in 2006 and his PhD from Lund University (Sweden) in 2012. He has held research positions at the Imperial College London and the University of Li`ege (Belgium), and currently he is a postdoctoral research associate at the Department of Engineering Science, University of Oxford. His research interests include linear and nonlinear positive systems, as well as nonlinear control and analysis methods for biological and biomedical applications.

James Anderson received his BSc and MSc degrees from the University of Reading in 2005 and 2006, and his DPhil in Engineering Science from the University of Oxford in 2012. Until 2016, he was a Junior Research Fellow in Engineering at St John’s College, Oxford, and a faculty member of the Department of Engineering Science. In October 2016 he joined the Department of Computing and Mathematical Sciences at the California Institute of Technology as a Senior Postdoctoral Scholar. His research interests cover robust control theory, nonlinear systems, and optimisation for networked systems, with applications in biological and power systems.