Plant-wide Controllability and Structural Optimization of Plants with Recycles
Groenendijk, A.J.

Citation for published version (APA):
Chapter 5

Large Scale Model Reduction

5.1 Introduction

Waste minimization and flexibility are key issues in the design of a new chemical plant. The development of a 'zero discharge' plant often leads to a complex structure with many recycle loops included. Flowsheeting is a well-established technique reflecting the complexity of these systems by means of an accurate steady state simulation model. Through dynamic flowsheeting even more design and operation knowledge is obtained. In addition, large dynamic plant models give access to a whole class of controllability tools. A method for the combination of steady state and dynamic modeling with linear controllability analysis tools was described in chapter 2.

By applying this systems approach on a case study about the handling of impurities in a balanced VCM process, as described in chapter 3 and 4, we observed that the derivation of a linear model for a large, nonlinear dynamic plant is far from trivial. When a dynamic model is available, it is easy to generate a linear state space description by performing a Taylor series expansion of the nonlinear functions around the nominal point, neglecting second and higher order terms. Dynamic simulation programs like SPEEDUP™ posses such features. However, an automatic linearization procedure will inevitably include all states from the complex nonlinear model and the linear state space realization may become very large and difficult to handle. Therefore there is a real need to reduce a state space realization to lower order, which is the subject of this chapter.
5.2 Dynamic modeling

A description of dynamic models and linearization was already given in chapter 2. Let's recall this for a better understanding of the remainder of this chapter.

5.2.1 Dynamic model

A dynamic plant model contains material and energy balances (including reactions), phase equilibrium relations and volume equations. It is a combination of differential equations for the material and energy holdup relations and additional algebraic equations. In general, the model can be described as:

\[
\dot{x}(t) = f(x(t), u(t)) \quad y(t) = g(x(t), u(t))
\]

(5.1)

where \( \dot{x} = dx/dt \), \( x \) denotes a vector of \( n \) state variables, \( u \) a vector of \( m \) input variables and \( y \) a vector of \( l \) output variables, while \( f \) and \( g \) are general nonlinear functions.

The first stage of model reduction already takes place during the development of the dynamic model. An important issue is the number of components to be included, since component balances have to be solved for all trays in every column. By including only components being important for the dynamic behavior and, in addition, lumping impurities and trace components, the number of state variables can be reduced considerably. Another issue is the separation of equipment into units determining the overall dynamic behavior of the plant and units with minor influence on this behavior, to be omitted. Describing the fast responding units with instantaneous models reduces the number of state variables further, while the overall dynamic behavior is almost not affected. However, doing this the final dynamic model will normally still left with a large number of state variables, all being included in the linear state space realization.

5.2.2 Linearization

A well-known method for the development of a linear model is the curve fitting of experimental data with standard functions. Simple first and second order transfer functions often give already acceptable descriptions of the input-output behavior of complex systems, while high order polynomials can be used to give an accurate description. Extensive dynamic simulations can be used to generate the input-output data. As an advantage, this method leads
to small linear systems being easy to deal with. However, this procedure may be very time consuming, especially when experimental data have to be generated by extensive dynamic simulations.

More conveniently a linear state space realization is derived from a dynamic model. Performing a Taylor series expansion of the nonlinear functions around the nominal operating point and neglecting second and higher order terms gives:

\[
\frac{dx(t)}{dt} = \frac{\partial f}{\partial x} dx(t) + \frac{\partial f}{\partial u} du(t)
\]

\[
y(t) = \frac{\partial g}{\partial x} dx(t) + \frac{\partial g}{\partial u} du(t)
\]

or

\[
x(t) = Ax(t) + Bu(t)
\]

\[
y(t) = Cx(t) + Du(t)
\]

(5.2)

Now \( x, u \) and \( y \) are deviations from the nominal point and \( A, B, C \) and \( D \) are real matrices.

With a dynamic process simulator like SPEEDUP\( ^{TM} \), the user may specify the input and output variables to be included in the linear model. The program will carry out a linearization around the nominal point and generates the matrices \( A, B, C \) and \( D \). Such a state space realization contains only input and output variables of interest to the user, but the model inevitably contains all state variables. For control studies of complex plants, typically only 10 to 20 input and output variables are concerned, but the number of states easily amounts to thousands. Thus, the state space description has a large size and is difficult to handle. In addition, the matrix \( A \) may be numerically ill conditioned. However, realizing that states are not equally important, the possibility exists to reduce the size of the state space realization while retaining sufficient accuracy in the description of the dynamic input-output behavior. This requires special model reduction techniques.

5.3 Reducing a state space realization

A number of algorithmic procedures exist to reduce a linear state space realization to lower order. The simplest method is to discard part of the states from the model, called truncation. Another method is residualization, where part of the model is simulated steady state. A description of these methods will be given below and demonstrated on a case. The first step is to select states to be truncated or residualized. This may be performed manually for small models, but this is not feasible for models containing over a thousand states. For this reason a certain ordering of the states is required before the realization can be reduced.
5.3.1 Jordan form

A well-known method for the ordering of a state space realization is based on its corresponding time constants. To this end the A-matrix is transformed into its Jordan form, where the eigenvalues of the system are put on the main diagonal in descending order.

Consider the transformation matrix $T$ with the eigenvectors $t_i$ of the $n \times n$ matrix $A$ as its columns and the corresponding eigenmatrix $\Lambda$:

$$T = [t_1 \ t_2 \ \cdots \ t_n], \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix} \quad (5.3)$$

Now, introduce the new states $z$:

$$z = T^{-1}x \Leftrightarrow x = Tz \quad (5.4)$$

Substituting these in the state space description (5.2) will give the ordered realization:

$$Tz = ATz + Bu \Leftrightarrow \dot{z} = \Lambda z + T^{-1}Bu$$
$$y = CTz + Du \quad (5.5)$$

where $\Lambda = T^{-1}AT$ is the A-matrix in Jordan form. Note that this also produces a new B-matrix $T^{-1}B$ and a new C-matrix $CT$. The D-matrices are identical. The variables $x$, $y$, $z$ and $u$ are still time dependent, but for simplicity the notation $(t)$ is omitted here.

5.3.2 Splitting

If the A-matrix is in Jordan form, the states are ordered such that the first state is connected to the fastest process and the last state to the slowest one. The states are then readily partitioned into two groups, the first containing fast and the second slow modes.

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (5.6)$$

The corresponding state space description becomes:

$$\dot{x}_1 = A_{11}x_1 + A_{12}x_2 + B_1u$$
$$\dot{x}_2 = A_{21}x_1 + A_{22}x_2 + B_2u$$
$$y = C_1x_1 + C_2x_2 + Du \quad (5.7)$$
where the matrices $A$, $B$, and $C$ are split up as follows:

$$
A = \begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix}, \quad
B = \begin{bmatrix}
B_1 \\
B_2
\end{bmatrix}, \quad
C = \begin{bmatrix}
C_1 & C_2
\end{bmatrix}
$$

(5.8)

### 5.3.3 Truncation

When regarding fast dynamics the slow part of the system can be discarded, yielding the state space description:

$$
\begin{align*}
\dot{x}_T &= A_T x_T + B_T u \\
y &= C_T x_T + D_T u
\end{align*}
$$

(5.9)

where

$$
x_T = x_1, \quad A_T = A_{11}, \quad B_T = B_1, \quad C_T = C_1, \quad D_T = D
$$

(5.10)

Discarding states from a system that is ordered by its eigenvalues (Jordan form) is called modal truncation.

### 5.3.4 Residualization

Regarding slow processes, it should be realized that the fast processes are almost instantaneous as relative to the slow ones and therefore the states associated with the fast part of the system may be modeled as steady state:

$$
\begin{align*}
0 &= A_{11} x_1 + A_{12} x_2 + B_1 u \\
\dot{x}_2 &= A_{21} x_1 + A_{22} x_2 + B_2 u \\
y &= C_1 x_1 + C_2 x_2 + D u
\end{align*}
$$

(5.11)

From the above system $x_1$ can be solved, resulting in

$$
\begin{align*}
\dot{x}_R &= A_R x_R + B_R u \\
y &= C_R x_R + D_R u
\end{align*}
$$

(5.12)

where

$$
\begin{align*}
\dot{x}_R &= \dot{x}_2 \\
A_R &= A_{22} - A_{21} A_{11}^{-1} A_{12} \\
B_R &= B_2 - A_{21} A_{11}^{-1} B_1 \\
C_R &= C_2 - C_1 A_{11}^{-1} A_{12} \\
D_R &= D - C_1 A_{11}^{-1} B_1
\end{align*}
$$

(5.13)

Residualization of a system in Jordan form is called modal residualization.
A typical and important property of residualization is that it retains the steady state gain of the system by putting derivatives to zero, which obviously represents steady state. This is in sharp contrast to modal truncation that on the contrary retains the systems behavior at infinite frequency. For this reason modal truncation should be used when accuracy is required at high frequencies, whereas modal residualization is preferred for low frequency modeling.

**Example 5.1; Modal Truncation**

Consider a dynamic vapor-liquid flash drum with pressure and level control, where a 3-component mixture is separated into a vapor and a liquid. The feed flowrate and composition may fluctuate and therefore the partial component flowrates are selected as input variables of the linear model. The purity of the top and bottom products have to be controlled and therefore the fraction of component 1 in the outlet streams are selected as the output variables of the linear model. The pressure and temperature of the flash are manipulated to control the output variables so they also serve as input variables.

The state space realization, generated with the CDI interface from SPEEDUP™, contains 5 input variables, two output variables and 10 state variables; the material holdup of each of the three component, the total energy holdup and three states for each PID controller. The states in this realization can be ordered by transforming the A-matrix into its Jordan form, where the eigenvalues are put on the diagonal in descending order. These eigenvalues are:

\[
\text{eig}(A) = \{-3215, -62.0, -18.9, -10.5, -10.0, -10.0, -5.0, -1.0, -1.0, -0.001\}
\]

Since the last eigenvalue is much smaller than the other ones, this state might be truncated when considering fast dynamics. Regarding slow dynamics only, the first state might be residualized.

Unfortunately, little is known a priori about the effect of these reductions on the quality of the description of the overall input-output behavior. In a large and complex plant the input-output behavior is often determined by a combination of fast and slow processes, so the Jordan form is not optimal in ordering states. Apparently, it is more appropriate to order the states according to their contribution to the input-output behavior. This is the outline of the balanced realization to be described next.
5.4 Balanced realizations

A balanced realization is an asymptotically stable minimal realization with controllability and observability Gramians being equal and diagonal [Skogestad and Postlethwaite, 1996]. This is to be explained as follows. Firstly, a system is stable if it reaches a new steady state within finite time after a disturbance. This occurs when the A-matrix of the state space realization only contains negative eigenvalues, as was the case in the flash drum introduced above. Then, a minimal realization only contains states that do contribute to the input-output behavior. Non-controllable and/or non-observable states are omitted. Following this line of thought is seems proper to introduce the concepts of state controllability and state observability now.

5.4.1 State controllability

The dynamic system \( \dot{x}(t) = Ax(t) + Bu(t) \), or equivalently the pair \((A,B)\), is said to be state controllable if, for any initial state \( x(0) = x_0 \), any time \( t_f > 0 \) and any final state \( x_f \), an input \( u(t) \) exists such that \( x(t_f) = x_f \). Otherwise the system is state uncontrollable.

In other words, if a system is state controllable it can be moved from any initial state to any final state within a given finite time by using its inputs \( u \).

Many ways exist for checking whether a system is state controllable. One option is based on the controllability matrix \( \mathcal{C} \) for the system pair \((A,B)\), being defined as:

\[
\mathcal{C} = \begin{bmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{bmatrix}
\]

(5.14)

The system \((A,B)\) is state controllable if and only if the controllability matrix has rank \( n \) (full row rank), where \( n \) is the number of states.

According to an alternative method of checking controllability, the input profile \( u(t) \) is considered moving a state from its initial value \( x_0 \) to a value \( x_1 \) in time \( t_1 \), given by:

\[
u(t) = -B^T e^{A^T (t_1 - t)} W_c(t_1)^{-1} \left( e^{At_1} x_0 - x_1 \right)
\]

(5.15)

where \( W_c(t) \) represents the Gramian matrix at time \( t \),

\[
W_c(t) = \int_0^t e^{As} B B^T e^{A^T s} ds
\]

(5.16)
and $^T$ is used to specify the transpose of a matrix. The system $(A,B)$ is state controllable if and only if the Gramian matrix has full rank for any $t > 0$.

As regards stable systems only the Gramian matrix at infinite time needs to be considered,

$$W_c(\infty) = P = \int_0^\infty e^{At}BB^Te^{A^Tt}dt \tag{5.17}$$

$P$ is called the controllability Gramian and plays an important role in the balancing of a state space realization. Matrix $P$ may also be obtained as solution to the following Lyapunov equation:

$$AP + PA^T + BB^T = 0 \tag{5.18}$$

Solving this Lyapunov equation is not always a trivial task. Standard algorithms may fail for large systems, so special techniques are required. This will be discussed in section 5.5.

### 5.4.2 State observability

The dynamic system $\dot{x}(t) = Ax(t) + Bu(t)$, $y(t) = Cx(t) + Du(t)$ or the pair $(A,C)$, is said to be state observable if, for any time $t_1 > 0$, the initial state $x(0) = x_0$ can be determined from the time history of the input $u(t)$ and the output $y(t)$ in the interval $[0,t_1]$. Otherwise the system is state unobservable.

Hence, a system is state observable if the value of all individual states can be obtained by measuring the output $y(t)$ over a period of time.

In order to check whether a system is state observable, the observability matrix $O$ for the system pair $(A,C)$ has to be considered:

$$O^A = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \tag{5.19}$$

The system $(A,C)$ is state observable if and only if the observability matrix has rank $n$ (full column rank), where $n$ is the number of states.
On the other hand, analogously to state controllability it is found that the Gramian matrix $W_0(t)$ at time $t$ is equal to,

$$W_0(t) = \int_0^t e^{A^T \tau} C^T Ce^{A\tau} d\tau$$  \hspace{1cm} (5.20)

The system $(A,C)$ is state observable if and only if the Gramian matrix has full rank in the interval $[0,t]$.

Again regarding stable systems one should consider:

$$W_0(\infty) = Q = \int_0^\infty e^{A^T \tau} C^T Ce^{A\tau} d\tau$$  \hspace{1cm} (5.21)

$Q$ is called the observability Gramian and may also be obtained as solution to the following Lyapunov equation:

$$A^T Q + QA + C^T C = 0$$  \hspace{1cm} (5.22)

### 5.4.3 Minimal realization

It should be noted that only controllable and observable states contribute to the input-output behavior of the system. If a state is uncontrollable, it is not affected by the inputs $u$. If a state is unobservable, is does not affect any output $y$. This leads to the statement that a state space realization is minimal if and only if $(A,B)$ is state controllable and $(A,C)$ is state observable. This means that $A$ has the smallest possible dimension (i.e. the fewest number of states).

**Example 5.2; Minimal realization**

Consider the state space realization of the flash problem from above. Since the eigenvalues of the $A$-matrix are all negative, the realization is stable and only the controllability and observability Gramians have to be considered. These are obtained by solving the Lyapunov equations 5.18 and 5.22. Both, the controllability Gramian $P$ and the observability Gramian $Q$ are of full rank, so the state space realization is minimal.

### 5.4.4 Hankel norm

Before describing the balanced realization in a more formal way, one additional concept has to be introduced, the Hankel norm and the associated Hankel singular values.
The Hankel norm of a stable system $G(s)$ is obtained when an input $u(t)$ is applied up to $t = 0$ and the output $y(t)$ is measured for $t > 0$. Furthermore $u(t)$ is selected to maximize the ratio of the 2-norms of the output- and input-signal:

$$\|G(s)\|_H = \max_{u(t)} \frac{\int_{0}^{\infty} \|y(\tau)\|^2_2 d\tau}{\int_{-\infty}^{0} \|u(\tau)\|^2_2 d\tau}$$  \hspace{1cm} (5.23)

The Hankel norm may be interpreted as an induced norm from past inputs to future outputs and is defined as the maximum Hankel singular value $\sigma_1$

$$\|G(s)\|_H = \sigma_1$$  \hspace{1cm} (5.24)

It may be shown (Skogestad and Postlethwiate, 1996) that the Hankel norm is equal to

$$\|G(s)\|_H = \sqrt{\rho(PQ)}$$  \hspace{1cm} (5.25)

where $\rho$ is the spectral radius (maximum eigenvalue), $P$ is the controllability Gramian and $Q$ the observability Gramian. The corresponding Hankel singular values are the positive square roots of the eigenvalues of the product $PQ$, so one obtains:

$$\sigma_i = \sqrt{\lambda_i(PQ)}$$  \hspace{1cm} (5.26)

The name "Hankel" is used because the matrix $PQ$ has the special structure of a Hankel matrix (which has identical elements along the "wrong-way" diagonals).

The Hankel norm has a special property; it is less than or equal to twice the sum of the Hankel singular values:

$$\|G(s)\|_H \leq 2\sum_{i=1}^{n} \sigma_i$$  \hspace{1cm} (5.27)

This property can be used as error criterion for model reduction of a balanced realization.

### 5.4.5 Balanced realization

Now, the definition of a balanced realization is revisited in a more formal way. Let $(A,B,C,D)$ be a minimal realization of a stable, rational transfer function $G(s)$, then $(A,B,C,D)$ is called a balanced realization if the solutions to the related Lyapunov equations 5.18 and 5.22 are equal and therefore diagonal, so
Large Scale Model Reduction

\[
P = Q = \begin{bmatrix}
\sigma_1 & 0 & \cdots & 0 \\
0 & \sigma_2 & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & \sigma_n
\end{bmatrix} \Delta = \Sigma
\]

(5.28)

with \( \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0 \).

\( P \) and \( Q \) are the controllability and observability Gramians respectively, while \( \sigma_i \) are the ordered Hankel singular values, given by 5.26. Since the controllability and observability Gramians of a balanced system are equal, we simply refer to the Gramian of the system by \( \Sigma \).

Typically, in a balanced realization each state is just as observable as controllable. Furthermore, \( \sigma_i \) is a relative measure for the contribution of the related state \( x_i \) of the balanced realization to the input-output behavior of the system. States corresponding to small singular values have a small contribution and may be discarded by truncation.

Since the Hankel norm is always less than or equal to twice the sum of the singular values, this may be used as a measure of the error introduced by discarding states. The Hankel norm of the difference between the full and the reduced model is less than or equal to twice the sum of the Hankel singular values that correspond to the discarded states.

\[
\| F_n(s) - F_m(s) \|_H \leq 2 \sum_{i=m+1}^{n} \sigma_i
\]

(5.29)

5.4.6 Balancing

Next, the balancing of a state space realization is discussed. Normally the controllability Gramian \( P \) and observability Gramian \( Q \) of a stable minimal realization \( (A,B,C,D) \) are not equal. It is known that the joint controllability and observability Gramian of the balanced system should be a special matrix with the Hankel singular values on the main diagonal and with its non-diagonal entries being all zero. Now, if \( P \) and \( Q \) are the solutions to the corresponding Lyapunov equations 5.18 and 5.22, they may be decomposed into their Cholesky factors \( R \) and \( S \), defined as:

\[
P = R^T R \quad \text{and} \quad Q = S^T S
\]

(5.30)

Then equation 5.26 can be written as:
where $\sigma_{SR^T}$ are the singular values of $SR^T$. These values may be obtained by Singular Value Decomposition:

$$SR^T = u_{SR^T} \sigma_{SR^T} v_{SR^T}^T$$  \hfill (5.32)

Now, the following operations will balance the system:

$$A_{bal} = u_{SR^T}^T S A R^T v_{SR^T}$$

$$B_{bal} = u_{SR^T}^T S B$$

$$C_{bal} = C R^T v_{SR^T}$$

$$D_{bal} = D$$  \hfill (5.33)

**Example 5.3; Balanced truncation**

The balanced truncation is demonstrated on the flash drum example. Since the controllability and observability Gramians of the minimal realization are known, the system may be balanced. Table 5.1 shows the Hankel singular values that are positioned on the main diagonal of the joint controllability and observability Gramian of the balanced realization. They form a measure of the contribution of the corresponding states to the input-output behavior. The table also shows the maximum error if the system is truncated after the state concerned. According to equation 5.29 this is equal to twice the sum of the discarded Hankel singular values.

**Table 5.1; Hankel singular values and error bound for the balanced state space realization of the 3-component flash with level and pressure control.**

<table>
<thead>
<tr>
<th>State</th>
<th>Hankel singular value</th>
<th>Max. modeling error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.03e+01</td>
<td>8.32e+00</td>
</tr>
<tr>
<td>2</td>
<td>2.80e+00</td>
<td>2.71e+00</td>
</tr>
<tr>
<td>3</td>
<td>1.35e+00</td>
<td>3.58e-03</td>
</tr>
<tr>
<td>4</td>
<td>1.45e-03</td>
<td>6.74e-04</td>
</tr>
<tr>
<td>5</td>
<td>3.37e-04</td>
<td>3.88e-14</td>
</tr>
<tr>
<td>6</td>
<td>1.71e-14</td>
<td>4.60e-15</td>
</tr>
<tr>
<td>7</td>
<td>1.68e-15</td>
<td>1.23e-15</td>
</tr>
<tr>
<td>8</td>
<td>5.49e-16</td>
<td>1.31e-16</td>
</tr>
<tr>
<td>9</td>
<td>6.45e-17</td>
<td>2.05e-18</td>
</tr>
<tr>
<td>10</td>
<td>1.03e-18</td>
<td>0</td>
</tr>
</tbody>
</table>
It is clearly seen that the last five states have hardly any contribution and may be discarded without affecting the quality of the input-output description significantly. If less accuracy is required, the balanced system might even be truncated after the third state.

This example clearly illustrates that balancing is a more elegant way of ordering the states than the Jordan form. Especially for large and complex systems, where the input-output behavior depends on processes operating at different time scales, the Jordan form is not useable while discarding states from a balanced system proceeds in a straightforward manner.

5.5 Solving Lyapunov equations

It should be recalled that the solution to the Gramians might be obtained by solving the Lyapunov equations. In case of low dimensional problems ($n \leq 50$) numerous solution techniques are available. One of the most efficient and widely used procedures is the Bartles-Steward algorithm (Bartles and Steward, 1972). In this algorithm the A-matrix is transformed into a real Schur form $H = U^T A U$ in which $U$ is orthogonal and $H$ is quasi upper triangular. This gives the modified Lyapunov equation:

$$H \tilde{P} + \tilde{P} H + U^T B B^T U = 0$$

in which $\tilde{P} = U^T P U$, and this equation is easily solved by back substitution.

Unfortunately, this method is unsuitable for solving large Lyapunov equations, since the computation time and storage requirements become prohibitive, while the method suffers from numerical instability too. In order to resolve this problem, Jaimoukha [Jaimoukha et al., 1992; Jaimoukha and Kasenally 1994] presented an algorithm to compute low rank approximate solutions to the Lyapunov equations for large systems using classical Krylov subspace methods.

5.5.1 Low rank approximations

The solutions of the Lyapunov equations are normally used to balance the state space realization in order to select the states to be discarded from the model. This does not require an exact solution. Since the dominant eigenspace of $P$ is known to be associated with the dominant modes of the system, only the dominant eigenspace of the solution $P$ is of interest,
rather than $P$ itself. Therefore, in practice, solutions to large Lyapunov equations frequently admit good low rank approximations.

The idea is to approximate the solution to the Lyapunov equation (5.18) by projecting the state space matrices $A$ and $B$ on an orthogonal matrix $V_m \in \mathbb{R}^{n \times m}$ of low rank (Oblique projection) and calculating the exact solution $X_m$ to the reduced order Lyapunov equation:

$$(V_m^T A V_m) X_m + X_m (V_m^T A^T V_m) + V_m^T B B^T V_m = 0$$

The estimate of $P$ is then given by $P_m = V_m X_m V_m^T$.

In the same manner, the approximate solution to the Lyapunov equation (5.22) can be found by projecting $A^T$ and $C^T$ on an orthogonal basis $W_m \in \mathbb{R}^{r \times m}$ and solving the reduced order Lyapunov equation:

$$(W_m^T A^T W_m) Y_m + Y_m (W_m^T A W_m) + W_m^T C^T C W_m = 0$$

The estimate of $Q$ is then given by $Q_m = W_m Y_m W_m^T$.

The remainder of this section is focussed on the approximation $P_m$ and the basis $V_m$, but this also is a valid procedure for the approximation $Q_m$ and the basis $W_m$.

### 5.5.1.1 The basis

The problem of obtaining a good approximate solution $P_m$ to the full order Lyapunov equation 5.18 is focussed on the selection of the orthogonal matrix $V_m$ serving as a basis for the projection of the state space matrices $A$ and $B$. In order to obtain the dominant modes of the system, this basis should contain at least the dominant eigenspace of $P$.

To this end the $mp$-dimensional Krylov space $K_m$ is defined as a subspace of the controllability matrix $\mathcal{E}$ for the system pair $(A, B)$:

$$K_m = \text{span}\{[B \ AB \ A^2 B \ \cdots \ A^{m-1} B]\}$$

The matrix $V_m$ may be selected then to be an orthonormal basis of this subspace.

### 5.5.1.2 Arnoldi process

The orthonormal basis $V_m$ for the Krylov subspace $K_m$ may be calculated with the well-established Arnoldi algorithm. This algorithm starts with the orthogonalization of the state
space matrix B by a standard QR factorization and then adds directions that are associated
with AB, A²B, ... Aᵐ⁻¹B. It is basically outlined as follows:

- Compute B = Q₁R₁ and set p₁ := number of columns of Q₁. (QR factorization)
- Do j = 1, ..., m
  (a) Set Vₖ = [Q₁ Q₂ ... Qⱼ]
  (b) Compute \[ \begin{bmatrix} A_{1j} \\ A_{2j} \\ \vdots \\ A_{jj} \end{bmatrix} = Vₖ^T A Q₂ \]
  (c) Qₖ₊₂Aₖ₊₂ = AQₖ − \[ \sum_{k=1}^{j} Q_k A_{kj} \], (QR factorization) and
    \[ p_{j+1} := \text{number of columns of } Q_{j+1} \]
- End Do.

The QR factorization in the Arnoldi algorithm may be performed, for example, according to
the modified Gram-Smidt orthogonalization process.

5.5.1.3 Rank deficiency

The QR factorization of B ∈ \( \mathbb{R}^{mn} \) will normally produce a unitary matrix Q ∈ \( \mathbb{R}^{mn} \) and an
upper triangular matrix R ∈ \( \mathbb{R}^{mn} \). Problems might occur when B does not have full column
rank. In that case, let the QR factorization be given by:

\[ B \Pi = \begin{bmatrix} \hat{Q}_1 \\ \hat{Q}_2 \end{bmatrix} \begin{bmatrix} \hat{R}_{11} & \hat{R}_{12} \\ 0 & 0 \end{bmatrix} \]

(5.38)

where \( \Pi \) is a permutation matrix so that \( \hat{R}_{11} \) is upper triangular. Now set \( Q₁ := \hat{Q}_1 \) and
\( R₁ := \begin{bmatrix} \hat{R}_{11} & \hat{R}_{12} \end{bmatrix} \Pi^T \). Then \( p₁ = \text{rank}(Q₁) = \text{rank}(B) \) and \( V₁ \) has only \( p₁ \) columns. The upper-
triangular structure of \( R₁ \) is lost because of the permutation with \( \Pi \), but this does not affect
the Arnoldi process. In the same manner a rank drop of \( A Q_j − \sum_{k=1}^{j} Q_k A_{kj} \) in the main loop of
the Arnoldi process is resolved.
Values on the main diagonal of $R$ being lower than a certain tolerance may also be treated as zero. This may prevent the Arnoldi process from adding vectors to the basis $V_m$ that do not correspond to directions of the Krylov subspace, but rather to directions of the residue being a result of numerical errors.

### 5.5.1.4 Residual error

After $m$ steps of the Arnoldi process associated with $\mathcal{Z}_m(A, B)$, an orthonormal basis $V_m$ is produced. Then the exact solution $X_m$ of the reduced order Lyapunov equation (5.35) can be computed and an approximate solution $P_m$ to the original Lyapunov equation (5.18) can be found. The question then rises whether this low rank approximation is acceptable.

This regards the residual error of the full order Lyapunov equation (5.18) in terms of the orthonormal basis $V_m$ and the exact solution $X_m$ of the reduced order Lyapunov equation (5.35), given by:

$$ R_m(X_m) = A(V_mX_mV_m^T) + (V_mX_mV_m^T)A^T + V_mB_mB_m^TV_m^T \tag{5.39} $$

where $B_m:=\begin{bmatrix} R_1 \\ 0_{nx} \end{bmatrix}$, $i$ is the number of inputs and $k = \sum_{j=2}^{m} p_j$, so $B = V_mB_m$.

Now, with the $mpxmp$ block upper Hessenberg matrix $A_m$, defined as:

$$ A_m = \begin{bmatrix} A_{11} & A_{12} & \cdots & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & \cdots & \vdots \\ 0 & A_{32} & A_{33} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & A_{m,m-1} & A_{mm} \end{bmatrix} \tag{5.40} $$

where $A_{11}, \ldots, A_{mm}$ originate form step b of the Arnoldi Process, it is easy to verify that

$$ AV_m = V_{m+1}A_{m+1,m} = V_mA_m + Q_{m+1}A_{m+1,m}E_m^T \tag{5.41} $$

and

$$ A_m = V_m^TAV_m \tag{5.42} $$

where $E_m$ is a matrix of the last $p$ columns of an identity matrix with the same size as $V_m$ and $V_{m+1} = [V_m Q_{m+1}]$. Substituting (5.41) into (5.39) gives:

$$ R_m(X_m) = V_{m+1} \begin{bmatrix} A_mX_m + X_mA_m^T + B_mB_m^T & X_mE_mE_m^T \\ A_{m+1,m}E_m^TX_m & 0 \end{bmatrix} V_{m+1}^T \tag{5.43} $$
With $A_m = V_m^TAV_m$ and $B_m = V_m^TB$, it is verified that the upper left element between the brackets equals zero, since $X_m$ is the exact solution to the reduced order Lyapunov equation (5.35).

Now it is interesting to take the Frobenius norm of this residue, defined as:

$$\|R_m\|_F = \sqrt{\text{tr}(R_mR_m^T)}$$

(5.44)

Since $V_m$ is part of an orthonormal basis, the Frobenius norm of the residual error becomes:

$$\|R_m(X_m)\|_F = \sqrt{2\|A_{m+1,m}E_m^TX_m\|_F^2}$$

(5.45)

This residual error norm does not require the computation of the approximate solution $P_m$ at each iteration, but it is computable using low dimensional matrix products. This provides a useful stopping criterion in a practical implementation of the algorithm. It allows evaluating the quality of the low rank approximation in an economic way.

The key points of the above method are summarized as follows. Firstly, that the Lyapunov equation 5.35 is of low dimension and can be solved accurately using the Bartels-Stewar algorithm. Secondly $P_m$ may be efficiently stored as the product of low matrices. Thirdly, the residual error norm is calculated via low dimensional matrix products. However, a drawback of the method is that manipulations with $V_m$ become expensive and storage requirements excessive with increasing $m$.

**Example 5.4; Approximating Gramians**

In the previous example, the Bartels-Steward algorithm is used to solve the Lyapunov equations of the flash system, which yield $P$ and $Q$. They may also be approximated with the Oblique projection method yielding $P_m$ and $Q_m$. In this example the reduced order Gramians and residual errors have been calculated after each Arnoldi step to show the progress of the approximations (Table 5.2 and Table 5.3). For large systems several Arnoldi steps (typically 3 or 4) may be executed before calculating the reduced order Gramian and residual error. Five input variables are to be selected for the flash problem, but the B-matrix does not have full column rank. This rank drop is tackled in the way described above, so the approximation of the controllability Gramian starts in iteration 1 with an order 4. In the second Arnoldi step, another rank drop occurs while the residual error already becomes small. In the third step two
more directions are added to the basis $V_m$ and the residual error is now sufficiently small to ensure a good approximation of the controllability Gramian.

The approximation of the observability Gramian starts with an order 2, which is the rank of matrix $C$ that contains two output directions. Again, in the second Arnoldi step a rank drop occurs. The residual error is already small this time. Adding another direction (the third step), the residual error norm is almost zero implying that the approximated solution to the observability Gramian has been found.

Now the approximated Gramians $P_m$ and $Q_m$ may be compared with the exact solutions $P$ and $Q$, calculated with the Bartels-Steward algorithm. Taking the 2-norm of the differences yields $\|P-P_m\|_2 = 1.77e-6$ and $\|Q-Q_m\|_2 = 2.76e-11$, which proves that the low rank approximations are acceptable solutions to the full order Lyapunov equations.

**Table 5.2; Progress of the controllability Gramian approximation**

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Order of the approximation</th>
<th>Residual error norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>2.97e-01</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1.34e-04</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>2.42e-06</td>
</tr>
</tbody>
</table>

**Table 5.3; Progress of the observability Gramian approximation**

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Order of the approximation</th>
<th>Residual error norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3.46e-01</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1.38e-04</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>1.61e-11</td>
</tr>
</tbody>
</table>

### 5.6 Krylov Space Model Reduction

The Oblique projection method described in the previous section is used to find approximate solutions to the Lyapunov equations. These solutions are needed to generate a balanced realization of a state space description, which can then be reduced to lower order. For large systems this procedure may be rather expensive. Although low rank approximate solutions to the Lyapunov equations are realized, the balancing procedure is performed on the full order model. More conveniently it seems to directly use Krylov subspace techniques providing computationally efficient model reduction schemes for large-scale state space systems. Jaimoukha described such a method (Jaimoukha and Kasenally, 1995) to be discussed next.
When considering the transfer function \( G(s) \) in terms of the state space matrices \((A,B,C,D)\):

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

(5.46)

Equation 5.46 may be rewritten as \( G(s) = CG_B(s) + D = G_C(s)B + D \), where \( G_B(s) = (sI-A)^{-1}B \) and \( G_C(s) = C(sI-A)^{-1} \). Consequently, \( G_B(s) \) and \( G_C(s) \) can be considered as solutions to the coupled linear systems:

\[
(sI - A)G_B(s) = B \quad \text{and} \quad G_C(s)(sI - A) = C
\]  

(5.47)

Now \( G(s) \) is approximated by obtaining approximate solutions \( G_{B,m}(s) \) and \( G_{C,m}(s) \) to these linear systems leading to two different realizations:

\[
G_{m,1}(s) = CG_{B,m}(s) + D, \quad G_{m,2}(s) = G_{C,m}(s)B + D
\]  

(5.48)

Good approximations of \( G(s) \) should contain states contributing to the input-output behavior. These states should be both controllable and observable. In order to select controllable states, the approximate solution \( G_{B,m}(s) \) of \( G_B(s) \) should be a part of the Krylov subspace \( \mathcal{K}_m(A,B) \) of the controllability Gramian \( \mathcal{G} \). In the same manner, the approximation \( G_{C,m}(s) \) should be part of the Krylov subspace \( \mathcal{L}_m(A^T, C^T) \) of the observability Gramian \( \mathcal{G} \).

In order to include all states contributing to the input-output behavior, the residue of the approximation \( G_{B,m}(s) \), given by \( (sI - A)G_{B,m}(s) - B \), should not contain observable states and therefore should not be a part of the Krylov subspace \( \mathcal{L}_m(A^T, C^T) \), so

\[
\mathcal{L}_m(A^T, C^T) \perp (sI - A)G_{B,m}(s) - B
\]  

(5.49)

For the same reason, the residue \( G_{C,m}(s)(sI - A) - C \) should not contain controllable states and therefore should not be a part of the Krylov subspace \( \mathcal{K}_m(A,B) \), so

\[
\mathcal{K}_m(A,B) \perp G_{C,m}(s)(sI - A) - C
\]  

(5.50)

The Arnoldi process may be used to find the orthogonal basis \( V_m \) for the Krylov subspace \( \mathcal{K}_m(A,B) \) and the orthogonal basis \( W_m \) for the Krylov subspace \( \mathcal{L}_m(A^T, C^T) \). Then the problem is to find approximate solutions \( G_{B,m}(s) = V_m H_m(s) \) and \( G_{C,m}(s) = F_m(s)W_m^T \) that satisfy the Galerkin-type conditions:
\[ W_m^T[(sI - A)V_m H_m(s) - B] = 0 \quad \forall s \] (5.51)

and
\[ \{F_m(s)W_m^T(sI - A) - C\}V_m = 0 \quad \forall s \] (5.52)

It has been observed before that from the Arnoldi process associated with \( \mathcal{A}_m(A,B) \) an upper Hessenberg matrix \( H_m \) can be produced and that:

\[
B = V_m B_m \quad \text{and} \quad AV_m = V_m H_m + \tilde{V}_m \tilde{H}_m \tag{5.53}
\]

in which \( B_m := \begin{bmatrix} R_i \\ 0_{kx1} \end{bmatrix} \), where \( i \) is the number of inputs, \( k = \sum_{j=2}^{m} p_j \), \( \tilde{V}_m = Q_{m+1} \) and \( \tilde{H}_m = H_{m+1,m} E_m^T \), where \( E_m^T \) are the last \( p \) columns of the identity matrix.

Similarly, associated with \( \mathcal{L}_m(A^T,C^T) \), the Arnoldi process produces a lower Hessenberg matrix \( F_m \) and

\[
C^T = W_m C_m^T \quad \text{and} \quad A^T W_m = W_m F_m^T + \bar{W}_m \bar{F}_m^T \tag{5.54}
\]

in which \( C_m := \begin{bmatrix} S_i \\ 0_{oxk} \end{bmatrix} \), where \( o \) is the number of outputs, \( k = \sum_{j=2}^{m} q_j \), \( \bar{W}_m = O_{m+1} \) and \( \bar{F}_m = F_{m+1,m} E_m^T \).

With these results the following matrices can be defined:

\[
\hat{H}_m := T_m^{-1} W_m^T A V_m := H_m + T_m^{-1} W_m^T \tilde{V}_m \tilde{H}_m \tag{5.55}
\]

and

\[
\hat{F}_m := W_m^T A V_m T_m^{-1} := F_m + \tilde{F}_m \bar{W}_m \tilde{V}_m T_m^{-1} \tag{5.56}
\]

for nonsingular \( T_m := W_m^T V_m \) while noticing that \( \hat{H}_m \) and \( \hat{F}_m \) are upper and lower Hessenberg, respectively.

Now it is readily verified that the Galerkin-type conditions 5.51 and 5.52 are satisfied if and only if \( H_m(s) = (sI - \hat{H}_m)^{-1} B_m \) and \( F_m(s) = C_m(sI - \hat{F}_m)^{-1} \). Under these conditions the residual error norms are:
\[ \|B - (sI - A)V_mH_m(s)\|_2 = \left\| \begin{bmatrix} T_m^{-1}W_m^T\hat{V}_m \\ I_p \end{bmatrix} \hat{H}_mH_m(s) \right\|_2 \] 
\[ (5.57) \]

\[ \|C - F_m(s)W_m^T(sI - A)\|_2 = \left\| F_m(s)\hat{F}_m\left[ \hat{W}_m^TV_mT_m^{-1} I_q \right] \right\|_2 \] 
\[ (5.58) \]

and the approximations, given by:

\[ G_{m,1}(s) = CV_mH_m(s) + D = \begin{bmatrix} T_m^{-1}W_m^TAV_m \\ \frac{C_mT_m}{D} \end{bmatrix} \]
\[ (5.59) \]

\[ G_{m,2}(s) = F_m(s)W_m^TB + D = \begin{bmatrix} \frac{W_m^TA}{C_m} \frac{T_m^{-1}}{D} \end{bmatrix} \]
\[ (5.60) \]

are different low order realizations of the transfer function \( G(s) \).

5.6.1 Minimal realization

When \( m \) steps of the Arnoldi process have been taken and \( H_{m+1,m} = 0 \), then the matrix \( V_m \) forms an orthogonal basis for the controllable space. This implies that the low order realizations \( G_{m,1}(s) \) and \( G_{m,2}(s) \) are equivalent to the high order model \( G(s) \). This holds also if \( F_{m+1,m} = 0 \) and \( W_m \) forms an orthogonal basis for the observable space or when both \( H_{m+1,m} = 0 \) and \( F_{m+1,m} = 0 \). An important implication of these types of breakdowns of the model reduction schemes is that the reduced order realizations turn out to be just minimal realizations of the high order model \( G(s) \).

5.6.2 Breakdown

Suppose that after \( m \) steps of the Arnoldi process a basis \( V_m \) of order \( mp \) and a basis \( W_m \) of order \( mq \) are produced, such that \( mp \neq mq \). This means that the number of observable directions is not equal to the number of controllable directions. Then the matrix \( T_m = W_m^TV_m \) is non-square and \( T_m^{-1} \) is undefined. This type of breakdown will occur after a partial breakdown of the block Arnoldi process, or when \( p \neq q \). In that case the smaller base should be expanded executing some extra Arnoldi steps such that \( V_m \) and \( W_m \) will become equal in size.
Another breakdown of the reduction algorithm occurs if $T_m$ becomes singular. In this case there are observable directions that are uncontrollable and controllable directions that are unobservable. Expanding both $V_m$ and $W_m$ with some additional Arnoldi steps may resolve this problem.

A practical solution to deal with the above problems is replacing $T_m^{-1}$ in the equations 5.57 to 5.60 by the pseudo-inverse $T_m^* = \text{pinv}(T_m)$. The computation is based on a Singular Value Decomposition of $T_m$ and any singular value less than a certain tolerance is treated as zero. $T_m^*$ will have the dimension of $T_m^T$, $T_m^* T_m T_m^* = T_m^*$, $T_m^* T_m = T_m$ and $T_m T_m^*$ and $T_m^* T_m$ are Hermitian. We have found such realizations to be good low order approximations to the full order model $G(s)$. However, $G_{m,1}(s)$ and $G_{m,2}(s)$ are not automatically minimal realizations. If the number of non-zero singular values of $T_m$ is less then $mp$ then $G_{m,1}(s)$ will contain redundant states, and if this number is less than $mq$, $G_{m,2}(s)$ is not minimal. However, since both approximations are only different realizations of the same transfer function, the smaller one may be considered as the reduced system. Moreover, since the reduced systems are much smaller than the original set, it is often possible to use a standard method to transform the larger one into a minimal realization.

**Example 5.5; Krylov space model reduction**

The flash problem is reconsidered to demonstrate the Krylov model reduction scheme. Previously, it was shown that the controllability Gramian $P$ could be approximated with the orthogonal basis $V_m$ for the Krylov space $\mathcal{K}_m(A,B)$ of order eight, while the orthogonal basis $W_m$ for the Krylov space $\mathcal{L}_m(A^T,C^T)$ of order five was used to approximate the observability Gramian $Q$. The Krylov model reduction scheme uses the same Arnoldi process, so identical bases are developed. Table 5.4 shows these bases $V_m$ to be of order eight and $W_m$ of order five, as developed after three Arnoldi steps, and to give an approximate realization $G_{m,2}$ of order five with a residual error norm close to zero. This is comparable to the result of the balanced realization being truncated after five states.

Notice that the approximate realization $G_{m,1}$, which should be just another realization of the same transfer function, is yet of higher order than $G_{m,2}$ as a result of the different order of the bases $V_m$ and $W_m$. This implies that the realization $G_{m,1}$ contains redundant states not contributing to the input-output behavior of the system. However, the basis $V_m$ should be
developed to an order 8 before it contains all observable directions needed for a good approximate solution $G_{m,2}$. Certain controllable directions being added in the first Arnoldi steps could be omitted, as they are not observable but that option does not exist in the Arnoldi process. Therefore, in this example, $G_{m,2}$ is the realization that should be used as reduced model, although the redundant states that $G_{m,1}$ contains can be removed by transforming this approximation to a minimal realization.

Table 5.4: Progress of the Krylov model reduction of the flash system

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Order of $V_m$</th>
<th>Order of $W_m$</th>
<th>Residual error norm of $G_{m,1}$</th>
<th>Residual error norm of $G_{m,2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1.92e+03</td>
<td>2.39e-01</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>4</td>
<td>2.24e-03</td>
<td>2.97e-04</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>5</td>
<td>3.38e-03</td>
<td>5.68e-11</td>
</tr>
</tbody>
</table>

5.7 Decomposition

Usually, only the dominant modes of a system are of interest and since the dominant eigenspace of $P$ is known to be associated with the dominant modes of the system, $V_m$ is selected to be the orthogonal basis of the Krylov subspace $\mathcal{K}_m(A,B)$ and $W_m$ the orthogonal basis of the Krylov subspace $\mathcal{L}_m(A^T,C^T)$. However, situations might exist where the input-output behavior cannot be described by dominant modes only. Suppose that the state space realization describes a system with many processes, operating at different time scales. The dominant modes are related to the fast processes and they are included in the reduced model first. Processes operating on longer time scales only are included in the reduced system if the bases $V_m$ and $W_m$ are developed to further extent. However, due to numeric problems the Arnoldi scheme is probably unable to find the required directions.

Yet another situation exists where this problem could rise. If the most controllable states are hardly observable and the most observable states are hardly controllable, the bases $V_m$ and $W_m$ should be developed almost to completeness before obtaining common directions. If such directions cannot be found, matrix $T_m := W_m^TV_m$ becomes singular and it is concluded that not any state exist contributing to the input-output behavior, although in fact this is only caused by an improper Arnoldi scheme.
In order to resolve the problems mentioned, the system should be separated into smaller fragments. Each of the subsystem should contain a group of states associated to processes operating at similar time scales. The Arnoldi process may subsequently be used to develop orthogonal bases and reduce each subsystem separately. Combining the reduced subsystems will give a reduced system that describes the complete input-output behavior. In other words, the system is decomposed into a number of subsystems with similar eigenvalues.

In order to decompose the system into two parts, the stable / anti-stable decomposition algorithm will be used. The A-matrix has to be transformed to its ordered real Schur form first, so the states are grouped according to their eigenvalues. This procedure will return a transformation matrix with which new B and C matrices can be generated:

\[
\begin{bmatrix}
A_{11} & A_{12} \\
0 & A_{22}
\end{bmatrix} = U^T A U
\]

\[
\begin{bmatrix}
B_1 \\
B_2
\end{bmatrix} = U^T B
\]

\[
\begin{bmatrix}
C_1 & C_2
\end{bmatrix} = CU
\]

As regards the decomposition in itself, the Sylvester equation \( A_{11} X - X A_{22} + A_{12} = 0 \) is solved using Kronecker products. Subsequently, two subsystems can be formed.

\[
\begin{align*}
A_{s1} &= A_{11} & A_{s2} &= A_{22} \\
B_{s1} &= B_1 - X B_2 & B_{s2} &= B_2 \\
C_{s1} &= C_1 & C_{s2} &= C_1 X + C_2
\end{align*}
\]

(5.62)

If states are grouped in such a way that \( A_{11} \) contains all negative eigenvalues of A and \( A_{22} \) the non-negative ones, then the subsystem formed by \((A_{s1}, B_{s1}, C_{s1}, D)\) will contain the stable modes, while the subsystem formed by \((A_{s2}, B_{s2}, C_{s2}, D)\) will be the unstable part. Now, when \( A_{11} \) is selected to contain the large eigenvalues of \( A_{s1} \) and \( A_{22} \) to contain the smaller, the stable part may be decomposed to further extent. By repeating this procedure several times a number of smaller subsystems are created.

The Arnoldi algorithm more easily deals with the subsystems and they may be reduced separately. However, apart form the additional time required to create subsystems, a drawback of this method is that reduced subsystems might contain common directions. In that case the reduced model is not a minimal realization and still contains redundant states.
Example 5.6; Krylov space model reduction with decomposition

The eigenvalues of the state space realization from the flash system of example 1 are: -3215, -62.0, -18.9, -10.5, -10.0, -5.0, -1.0, -1.0 and -0.001. A very large number appears, a very small one and an intermediate part exists, so it is beforehand to decompose the system into three subsystems that are reduced separately using the Krylov model reduction scheme. This leads to a first subsystem containing the mode with eigenvalue -0.001. The Krylov model reduction scheme did not find any common directions for this subsystem, so this part can be omitted completely. The second subsystem, containing the states corresponding to the intermediate eigenvalues, between -1 and -62, are reduced to a realization with four states. The last subsystem, containing the state corresponding to the large eigenvalue and consequently describing the fastest process in the system, cannot be reduced according to the Krylov model reduction scheme. Hence combining the reduced subsystems again yield a reduced state space realization of order five. This demonstrates that the decomposition may be used in combination with the Krylov model reduction scheme without affecting results.

5.8 Reduction of a large scale model

The described concepts of model reduction are applied now on a large system, the dynamic model of the balanced VCM process. The model that is used in chapter 3 is a simplified version of the complete system. The plant contains a liquid phase reactor that is modeled rigorously, including dynamics, while the two gas phase reactors are modeled as simple steady state conversion reactors, assumed to act instantaneous. Five important distillation columns are present, described by rigorous dynamic models considering individual sieve trays. The other units are combined to three simplified separation blocks. We have selected six important components to be taken into account in the rigorous dynamic distillation column models and another seven components only regarded in reactors and simplified separation blocks. The simplified dynamic model still contains 4384 equations.

In order to perform a controllability analysis, a linear state space description has been generated with the CDI interface from SPEEDUP™. The control problem involves three impurities to be controlled and five variables are identified to be useful as manipulated variables. These variables are from different units. The linear model contains 668 states. Performing a controllability analysis with this model is expensive and storage requirements are excessive. Therefore it is an obvious need to reduce the system to lower order.
Table 5.5: VCM plant state space model reduction

<table>
<thead>
<tr>
<th>Subsystem</th>
<th>Range of eigenvalues</th>
<th>Full order</th>
<th>Reduced order</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0 : -1</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>-1 : -10</td>
<td>30</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>-10 : -50</td>
<td>64</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>-50 : -100</td>
<td>45</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>-100 : -500</td>
<td>206</td>
<td>33</td>
</tr>
<tr>
<td>7</td>
<td>-500 : -1000</td>
<td>156</td>
<td>21</td>
</tr>
<tr>
<td>8</td>
<td>-1e3 : -5e3</td>
<td>56</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>-5e3 : -1e4</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>-1e4 : -∞</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td><strong>Full system</strong></td>
<td></td>
<td><strong>668</strong></td>
<td><strong>115</strong></td>
</tr>
</tbody>
</table>

Figure 5.2: Frequency responses of full (-) and reduced (…) order models.

The state space realization of the dynamic VCM model with 668 states is reduced to a realization with 115 states using the Krylov space model reduction scheme on subsystems. The magnitudes of the input-output responses of both models as function of frequency are comparable.
Dynamic simulations showed that a broad range of time constants exist in the system, ranging from a few minutes to several days. Therefore the system will be decomposed into a number of subsystems, 10 in total, while each subsystem will be reduced separately by applying the Krylov subspace model reduction technique. The results are given in Table 5.5.

Most subsystems can be reduced considerably, while the subsystems with the highest eigenvalues, corresponding to the fastest processes, can be even skipped completely. This is not surprising since control of impurities in a large and complex plant is a slow process. The complete reduced state space realization only contains 115 states, being less than 20% of the original system, while the description of the input-output behavior is sufficiently accurate across the whole frequency range of interest (Figure 5.2). So the order of the state space description of this problem can be reduced significantly.

5.9 Conclusions

As an essential part of designing chemical plants with complex structures and intricate dynamic characteristics the performing of controllability analyses on linearized models is required. The linearization of full dynamic plant models often leads to large state space descriptions, being difficult to use as such. In order to examine the dynamic behavior in an efficient way, linear dynamic models of a reduced order have to be generated. In this chapter it is explained how a state space description may be reduced to lower order without affecting the quality in describing the input-output behavior by truncation of a balanced realization. In order to balance a state space realization, the Lyapunov equations have to be solved first. There are several techniques available for this, like the Bartles-Steward algorithm. However, they cannot be used to solve large order Lyapunov equations ($n>50$). For these systems, another method is more appropriate to find low order approximate solutions to the Lyapunov equations. An Arnoldi algorithm is used to develop low rank orthogonal bases for certain Krylov subspaces, which are subsystems of the controllability and observability matrices. The state space matrices $(A,B)$ respectively $(A^T,C^T)$ are projected on these bases to reduce the order of the corresponding Lyapunov equations. The low order Lyapunov equations may subsequently be solved by a standard algorithm and the solutions are transformed back to yield approximate solutions of the full order controllability and observability Gramians. With a dynamic flash drum example is has been shown that the approximate solutions are nearly identical to the solutions of the original system.
The orthogonal bases for the Krylov subspaces can also be used to generate low order approximate state space realizations without solving Lyapunov equations and balancing state space systems. This Krylov model reduction scheme works properly for most systems. A possible problem arising from large systems is the fact that the Arnoldi process is not able to find all directions of the Krylov subspaces needed to give a good approximation. This problem especially occurs if only weak relations exist between inputs and outputs of a state space description. Decomposition of such systems into small subsystems with similar eigenvalues is a simple way to resolve these problems. The reduction of each subsystem may be performed then with the Krylov model reduction scheme and the reduced subsystems can be combined again to form a complete state space realization of low order. Although the decomposition is relatively time-consuming, the overall process is still an economic procedure. As an extensive case study, the dynamic model of the VCM problem from chapter 3 is reduced by this approach from 668 states to 115 states, while the description of the input-output behavior is preserved with an acceptable accuracy. Calculations with this reduced realization are about 25 times faster than with the full order linear system. This proves that model reduction might be useful before a controllability analysis is applied.