Model order reduction

Ha Binh Minh[†] (ha.b.minh@gmail.com)
and
Carles Batlle[‡] (carles.batlle@upc.edu)

[†] Hanoi University of Science and Technology
[‡] Universitat Politècnica de Catalunya—BarcelonaTech

Institut d'Organització i Control — February 2014

Summary

- Review of some state-space results: controllability, observability, balanced truncation.
- Some new results
  - A new lower bound for balanced truncation.
  - $\mathcal{H}_2$ optimal model reduction.
  - "Block reduction" (work in progress).

Input-output map

$$u \to \boxed{x} \to y$$

- Order of system $=$ dimension of x $=$ complexity.
- **Model reduction** aims to produce simpler model (i.e. lower-dimension of x) that approximates the original one.

Input-state-output description

$$\dot{x} = Ax + Bu$$
$$y = Cx + Du \tag{1}$$
$$x \in \mathbb{R}^n, u \in \mathbb{R}^m, y \in \mathbb{R}^p$$

- Order of system $= n$.
- For compactness, (1) is denoted by $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$

Questions

- Finding the lowest-order i/s/o description for given i/o map?
- Finding the lower-order model aproximating the i/o map?

Controllability

Controllable subspace

$$\begin{aligned}
\mathcal{C} &= \{\text{all states which are ``controlled'' by the input}\} \\
&= \{x_f \mid \exists\, u(t), \tau > 0, \quad x_f = \int_0^\tau A^{\tau - t} B u(t) dt\} \\
&= Im[B \quad AB \quad \cdots \quad A^{n-1}B]
\end{aligned}$$

If $\mathcal{C} = \mathbb{R}^n$, i.e. every point in the plane can be controlled, we say that (A, B) is controllable.

If $\mathcal{C} \subsetneq \mathbb{R}^n$ we can split the system into controllable part and non-controllable part.

Idea: Find similarity transformation

$$\tilde{A} = TAT^{-1} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix} \qquad \tilde{B} = TB = \begin{bmatrix} \tilde{B}_1 \\ 0 \end{bmatrix}$$

with $(\tilde{A}_{11}, \tilde{B}_1)$ controllable.

$$\begin{aligned} \dot{x}_1 &= \tilde{A}_{11}x_1 + \tilde{A}_{12}x_2 + \tilde{B}_1 u \\ \dot{x}_2 &= \tilde{A}_{22}x_2 \end{aligned}$$

- Dynamics involving $\begin{pmatrix} x_1 \\ 0 \end{pmatrix}$: $\dot{x}_1 = \tilde{A}_{11}x_1 + \tilde{B}_1 u.$

  $(\tilde{A}_{11}, \tilde{B}_1)$ controllable $\Rightarrow$ input has influence on $\begin{pmatrix} x_1 \\ 0 \end{pmatrix}$.

- Dynamics involving $\begin{pmatrix} 0 \\ x_2 \end{pmatrix}$: $\dot{x}_2 = \tilde{A}_{22}x_2$

  No input $\Rightarrow$ no influence to i/o map.

$$\begin{aligned} u &\rightarrow x_1 \\ u &\nrightarrow x_2 \end{aligned}$$

Observability

Unobservable subspace

$$\begin{aligned}\mathcal{N} &= \{\text{all initial states that do not influence} \\ &\quad \text{the output when input} = 0\} \\ &= \{\text{all initial states} \mid \text{output} = 0 \text{ when input} = 0\} \\ &= Ker \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}\end{aligned}$$

If $\mathcal{N} = \{0\}$ we say that (C, A) is observable.

If $\mathcal{N} \supsetneq \{0\}$, we can split the system into an observable part and a non-observable one.

Idea: Find similarity transformation

$$\tilde{A} = TAT^{-1} = \begin{bmatrix} \tilde{A}_{11} & 0 \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix} \qquad \tilde{C} = CT^{-1} = \begin{bmatrix} \tilde{C}_1 & 0 \end{bmatrix}$$

with $(\tilde{C}_1, \tilde{A}_{11})$ observable.

$$\begin{aligned} \dot{x}_1 &= \tilde{A}_{11}x_1 \\ \dot{x}_2 &= \tilde{A}_{21}x_1 + \tilde{A}_{22}x_2 \\ y &= \tilde{C}_1 x_1 \end{aligned}$$

- Dynamics involving $\begin{pmatrix} x_1 \\ 0 \end{pmatrix}$:

$$\begin{aligned} \dot{x}_1 &= \tilde{A}_{11} x_1 \\ y &= \tilde{C}_1 x_1 \end{aligned}$$

$(\tilde{C}_1, \tilde{A}_{11})$ observable $\Rightarrow \begin{pmatrix} x_1 \\ 0 \end{pmatrix}$ has influence on output.

- Dynamics involving $\begin{pmatrix} 0 \\ x_2 \end{pmatrix}$: $\qquad \dot{x}_2 = \tilde{A}_{22} x_2$

No output $\Rightarrow$ no influence to i/o map.

$$\begin{aligned} x_1 &\rightarrow y \\ x_2 &\nrightarrow y \end{aligned}$$

Kalman decomposition

Idea: Find similarity transformation

$$\left(\begin{array}{c|c} TAT^{-1} & TB \\ \hline CT^{-1} & D \end{array}\right) = \left(\begin{array}{cccc|c} \tilde{A}_{11} & 0 & \tilde{A}_{13} & 0 & \tilde{B}_1 \\ \tilde{A}_{21} & \tilde{A}_{22} & \tilde{A}_{23} & \tilde{A}_{24} & \tilde{B}_1 \\ 0 & 0 & \tilde{A}_{33} & 0 & 0 \\ 0 & 0 & \tilde{A}_{43} & \tilde{A}_{44} & 0 \\ \hline \tilde{C}_1 & 0 & \tilde{C}_3 & 0 & D \end{array}\right)$$

with $\begin{pmatrix} x_1 \\ x_2 \\ 0 \\ 0 \end{pmatrix}$ controllable states and $\begin{pmatrix} x_1 \\ 0 \\ x_3 \\ 0 \end{pmatrix}$ observable states.

Minimal realization

$$
\begin{array}{lll}
u & \to & x_1 \quad \to \quad y \\
u & \to & x_2 \quad \not\to \quad y \\
u & \not\to & x_3 \quad \to \quad y \\
u & \not\to & x_4 \quad \not\to \quad y
\end{array}
$$

Conclusion: Only x_1 has influence on i/o map, so the minimal realization of the i/o map is

$$
\left(\begin{array}{c|c} \tilde{A}_{11} & \tilde{B}_1 \\ \hline \tilde{C}_1 & D \end{array}\right)
$$

Observability gramian

How to quantify observability? $x(0) = x_0$, input $= 0$,

$$\|y\|^2 = \int_0^\infty y^\top(t)y(t)dt = x_0^\top \left(\int_0^\infty e^{A^\top t} C^\top C \; e^{At} dt \right) x_0 = x_0^\top Q x_0$$

$$\Rightarrow \frac{\|y\|^2}{\|x_0\|^2} = \frac{x_0^\top Q x_0}{x_0^\top x_0} \text{ quantifies observability of states in direction } x_0$$

$$\boxed{\text{Observability} \Leftrightarrow Q > 0}$$

In geometric terms, Q defines an "observability ellipsoid" in the state space.

Controllability gramian

How to quantify controllability? There are many inputs that bring $x(-\infty) = 0$ to $x(0) = x_0$. One of them, denoted by u_{opt}, is optimal, i.e. $\|u_{opt}\|^2$ is minimal.

$$\|u_{opt}\|^2 = x_0^\top \left(\int_0^\infty e^{At} B B^\top e^{A^\top t} dt \right) x_0 = x_0^\top P^{-1} x_0$$

$$\Rightarrow \frac{\|x_0\|^2}{\|u_{opt}\|^2} = \frac{x_0^\top P x_0}{x_0^\top x_0} \text{ quantifies controllability of states in direction } x_0$$

$$\boxed{\text{Controllability} \Leftrightarrow P > 0}$$

In geometric terms, P defines an "controllability ellipsoid" in the state space.

Balanced realization

Idea: Transformation to re-scale the ellipsoid axes as well as to rotate them. Find similarity transformation

$$\left(\begin{array}{c|c} \tilde{A} & \tilde{B} \\ \hline \tilde{C} & D \end{array}\right) = \left(\begin{array}{c|c} TAT^{-1} & TB \\ \hline CT^{-1} & D \end{array}\right)$$

such that

$$\tilde{P} = TPT^{\top} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n) = (T^{-1})^{\top} Q T^{-1} = \tilde{Q}$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$$

Geometrically, the two ellipsoids are identical and their principal axes concide with the coordinate axes of the state space.

Truncation

Truncation = delete the less important states.

$$\left(\begin{array}{c|c} \tilde{A} & \tilde{B} \\ \hline \tilde{C} & D \end{array}\right) = \left(\begin{array}{cc|c} \tilde{A}_{11} & \tilde{A}_{12} & \tilde{B}_1 \\ \tilde{A}_{21} & \tilde{A}_{22} & \tilde{B}_2 \\ \hline \tilde{C}_1 & \tilde{C}_2 & D \end{array}\right) \rightarrow \left(\begin{array}{c|c} \tilde{A}_{11} & \tilde{B}_1 \\ \hline \tilde{C}_1 & D \end{array}\right)$$

$$\Sigma := diag(\sigma_1, \sigma_2, \ldots, \sigma_n) \rightarrow diag(\sigma_1, \sigma_2, \ldots, \sigma_k) =: \Sigma_1$$

Properties preserved: Balanced truncation preserves stability, controllability and observability.

Error bound

Transfer matrix: $G(\lambda) := D + C(\lambda I - A)^{-1}B$

Size of system: If system stable,

$$\|G\|_{\mathcal{H}_\infty} := \sup_{\omega \in \mathbb{R}} \bar{\sigma}(G(j\omega)) = \sup_{u} \frac{\|y\|^2}{\|u\|^2}$$

Reduced system: $\hat{G}(\lambda) := D + C_1(\lambda I - A_{11})^{-1}B_1$

By estimating $\|G - \hat{G}\|_{\mathcal{H}_\infty}$, we can know how good the approximation is.

Global error bound

- [Glover, Enns, 1984]:

$$\sigma_{k+1} \leq \|G - \hat{G}\|_{\mathcal{H}_\infty} \leq 2(\sigma_{k+1} + \cdots + \sigma_n)$$

- [Minh, Carles, Enric, 2014]: For SISO systems,

$$\|G - \hat{G}\|_{\mathcal{H}_\infty} \geq \max(\sigma_{k+1}, 2(s_{k+1}\sigma_{k+1} + \cdots + s_n\sigma_n))$$

Global error bound

$$\boxed{\|G - \hat{G}\|_{\mathcal{H}_\infty} \geq \max(\sigma_{k+1}, 2(s_{k+1}\sigma_{k+1} + \cdots + s_n\sigma_n))}$$

Where do the s_i come from? $s_i = 1$, or $s_i = -1$ comes from the fact that, in balanced realization of SISO systems

$$b_i = c_i, \quad \text{or} \quad b_i = -c_i.$$

Consequence

In standard balanced truncation: Truncating states corresponding to the smallest Hankel singular values.

Our claim: This does NOT ALWAYS yield the best result (for fixed order of the approximating system).

Example

SISO balanced realization:

$$b = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad c = \begin{bmatrix} 1 \\ 1 \\ -1 \\ 1 \\ 1 \end{bmatrix}^T, \quad \sigma = \begin{bmatrix} 5 \\ 4 \\ 3 \\ 2.5 \\ 2 \end{bmatrix},$$

$$A = \begin{bmatrix} -0.100 & -0.111 & 0.500 & -0.133 & -0.143 \\ -0.111 & -0.125 & 1.000 & -0.154 & -0.167 \\ -0.500 & -1.000 & -0.167 & 2.000 & 1.000 \\ -0.133 & -0.154 & -2.000 & -0.200 & -0.222 \\ -0.143 & -0.167 & -1.000 & -0.222 & -0.250 \end{bmatrix}.$$

If we truncate σ_4 and σ_5

$$\|G(s) - G_1(s)\|_{\mathcal{H}_\infty} = 9.000,$$

If we truncate σ_2 and σ_3

$$\|G(s) - G_2(s)\|_{\mathcal{H}_\infty} = 5.6287,$$

Why?

- **By old estimation:**

$$2.5 \leq \|G(s) - G_1(s)\|_{\mathcal{H}_\infty} \leq 9$$

$$4 \leq \|G(s) - G_2(s)\|_{\mathcal{H}_\infty} \leq 14$$

- **By our estimation:**

$$9 \leq \|G(s) - G_1(s)\|_{\mathcal{H}_\infty} \leq 9$$

$$4 \leq \|G(s) - G_2(s)\|_{\mathcal{H}_\infty} \leq 14$$

$\mathcal{H}_2$ optimal model reduction using cross Gramians

- $\mathcal{H}_2$-norm of $G(s)$ for a MIMO system

$$\|G\|_{\mathcal{H}_2}^2 := \text{trace}\left(\frac{1}{2\pi}\int_{-\infty}^{\infty} G(-j\omega)^T G(j\omega)\text{d}\omega\right).$$

- It can be computed as

$$\|G\|_{\mathcal{H}_2}^2 = \text{trace}(B^T Q B) = \text{trace}(C P C^T)$$

with P, Q the controllability and observability Gramians, which satisfy the Lyapunov matrix equations

$$AP + PA^T + BB^T = 0, \quad A^T Q + QA + C^T C = 0,$$

constructed from a minimal and stable realization

$$(A, B, C) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times n}.$$

- For square systems, *i.e.* $m = p$, one can also construct the cross-Gramian R, which satisfies the Sylvester-type equation

$$AR + RA + BC = 0.$$

- In this case one can also compute $\|G\|_{\mathcal{H}_2}^2$ in terms of R

$$\|G\|_{\mathcal{H}_2}^2 = \mathrm{trace}(CRB).$$

- For SISO systems or MIMO symmetric systems ($G^T = G$) one can show that

$$R^2 = PQ.$$

Error system

- Let
$$\widehat{G}(s) = \widehat{C}(sI - \widehat{A})^{-1}\widehat{B}$$
be the transfer function of the reduced system, with $(\widehat{A}, \widehat{B}, \widehat{C}) \in \mathbb{R}^{r\times r} \times \mathbb{R}^{r\times m} \times \mathbb{R}^{p\times r}$, with $p = m$ and $r \leq n$ fixed (ideally, $r \ll n$).
- The error system has transfer function
$$E(s) := G(s) - \widehat{G}(s),$$
with realization
$$\{A_e, B_e, C_e\} = \left\{ \begin{bmatrix} A & 0 \\ 0 & \widehat{A} \end{bmatrix}, \begin{bmatrix} B \\ \widehat{B} \end{bmatrix}, \begin{bmatrix} C & -\widehat{C} \end{bmatrix} \right\}.$$

- Let $R_e := \begin{bmatrix} R & X \\ Y & -\widehat{R} \end{bmatrix}$ be the cross Gramian associated with $E(s)$, which satisfies $A_e R_e + R_e A_e + B_e C_e = 0$,
- In block form,

$$\begin{bmatrix} A & 0 \\ 0 & \widehat{A} \end{bmatrix} \begin{bmatrix} R & X \\ Y & -\widehat{R} \end{bmatrix} + \begin{bmatrix} R & X \\ Y & -\widehat{R} \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & \widehat{A} \end{bmatrix} + \begin{bmatrix} B \\ \widehat{B} \end{bmatrix} \begin{bmatrix} C & -\widehat{C} \end{bmatrix} = 0.$$

- This yields the four Sylvester-type equations

$$AR + RA + BC = 0, \qquad \widehat{A}\widehat{R} + \widehat{R}\widehat{A} + \widehat{B}\widehat{C} = 0,$$
$$AX + X\widehat{A} - B\widehat{C} = 0, \qquad \widehat{A}Y + Y A + \widehat{B}C = 0.$$

- The $\mathcal{H}_2$-norm of the error system $E(s)$ can be computed via the cross Gramian R_e as

$$\begin{aligned} \|E(s)\|_{\mathcal{H}_2}^2 &= \operatorname{trace}(C_e R_e B_e) \\ &= \operatorname{trace}(CRB + CX\widehat{B} - \widehat{C}YB + \widehat{C}\widehat{R}\widehat{B}). \end{aligned}$$

- Let

$$\mathcal{J}(\widehat{A}, \widehat{B}, \widehat{C}) := \|E(s)\|_{\mathcal{H}_2}^2.$$

- We can now consider variations $\widehat{A}$ to $\widehat{A} + \Delta_{\widehat{A}}$ and compute Δ_X, Δ_Y, $\Delta_{\widehat{R}}$ so that $X + \Delta_X$, $Y + \Delta_Y$, $\widehat{R} + \Delta_{\widehat{R}}$ satisfy the equations that define the cross-Gramian.
- Using those equations one can show that, to first order,

$$\begin{aligned} \Delta_{\mathcal{J}} &= \mathcal{J}(\widehat{A} + \Delta_{\widehat{A}}, \widehat{B}, \widehat{C}) - \mathcal{J}(\widehat{A}, \widehat{B}, \widehat{C}) \\ &= 2 \operatorname{trace}((\widehat{R}^2 + YX)\Delta_{\widehat{A}}). \end{aligned}$$

- Hence
$$\nabla_{\widehat{A}}\mathcal{J} = 2(\widehat{R}^2 + YX).$$
- This is the "matrix derivative" of the $\mathcal{H}_2$ norm of the error system under changes of $\widehat{A}$.
- In the same way, by considering variations of $\widehat{B}$ and $\widehat{C}$, one gets, respectively,
$$\nabla_{\widehat{B}}\mathcal{J} = 2(\widehat{C}\widehat{R} + CX), \quad \nabla_{\widehat{C}}\mathcal{J} = 2(\widehat{R}\widehat{B} - YB).$$
- By putting all these to zero, one gets conditions, in terms of the elements $\widehat{R}$, X, Y of the cross-Gramian of the error system, for the $\mathcal{H}_2$ norm of the later to be stationary (and hence minimal).
- The resulting equations are called Wilson's optimal conditions.

- Wilson's conditions can be solved implicitly as

$$\widehat{A} = W^T A V, \quad \widehat{B} = W^T B, \quad \widehat{C} = CV, \quad W^T V = \mathbb{I}_n,$$

where

$$W = Y^T (\widehat{R}^{-1})^T, \quad V = -X\widehat{R}^{-1}.$$

- This defines $\widehat{A}, \widehat{B}, \widehat{C}$ as functions of $\widehat{R}$, X and Y, *i.e.*

$$\begin{aligned} (\widehat{A}, \widehat{B}, \widehat{C}) &= G(\widehat{R}, X, Y) \\ &= (-\widehat{R}^{-1} Y A X \widehat{R}^{-1}, \widehat{R}^{-1} Y B, -CX\widehat{R}^{-1}). \end{aligned}$$

- However, the $\widehat{R}, X, Y$, in turn, depend on $\widehat{A}, \widehat{B}, \widehat{C}$ through the 3 Sylvester-like matrix equations

$$\left.\begin{aligned} \widehat{A}\widehat{R} + \widehat{R}\widehat{A} + \widehat{B}\widehat{C} &= 0 \\ AX + X\widehat{A} - B\widehat{C} &= 0 \\ \widehat{A}Y + YA + \widehat{B}C &= 0 \end{aligned}\right\} \text{ i.e. } \quad (\widehat{R}, X, Y) = F(\widehat{A}, \widehat{B}, \widehat{C}).$$

- Putting all together one has

$$(\widehat{A}, \widehat{B}, \widehat{C}) = G(\widehat{R}, X, Y) = (G \circ F)(\widehat{A}, \widehat{B}, \widehat{C}).$$

- Therefore, the optimal approximating systems are fixed points of the map $H = G \circ F$ and, in principle, they could be computed iteratively.
- This is the same situation than the controllability and observability Gramian approach presented in [P. van Dooren *et al.*, Appl. Math. Lett. **21**, 1267 (2008)].
- In that case, however, there are 4 Lyapunov matrix equations, instead of 3 Sylvester-like, involved in each iteration, so the computational advantages of our method could be important for very large scale systems.

- Notice that, when compared with other methods such as balanced truncation, the one considered here allows one to fix an *a priori* order r of the approximation, and the error can then be computed after the optimal point has been obtained.
- One can show that $\widehat{G}$ given by $(\widehat{A}, \widehat{B}, \widehat{C})$ defines a *tangential interpolation* of G at the *mirror poles* of $\widehat{G}$, *i.e.*

$$\hat{G}(-\hat{\lambda}_i) = G(-\hat{\lambda}_i), \quad \frac{\mathrm{d}}{\mathrm{d}s}\left(G(s) - \hat{G}(s)\right)_{s=-\hat{\lambda}_i} = 0$$

for any pole $\hat{\lambda}_i$ of $\hat{G}$.

Block reduction

- Model order reduction of network systems looks for reduction methods that somehow preserve the topology of the network. Work in this direction has been presented, for instance, in [T. Ishizaki *et al*, Model reduction and clusterization of large-scale bidirectional networks, *IEEE Trans. on Autom. Control* **59**, 48 (2014)].
- Inspired by that work, we (Josep M. Olm, HBM, CB) propose a new kind of reduction idea, which aims at lumping together some variables which share a similar input-to-state dynamics. Provisionally, we call this *block model* order reduction.
- The method is based on diagonalization. Hence it is not suitable for very large scale systems, but might be useful for control design for systems up to a few thousand degrees of freedom.

- Consider a linear system with a single input and with state-output

$$\begin{aligned} \dot{x} &= \bar{A}x + \bar{B}u, \\ y &= \mathbb{I}x, \end{aligned}$$

with $\bar{A} \in \mathbb{R}^{n\times n}$ stable and symmetric, $\bar{B}, x, y \in \mathbb{R}^n$, $u \in \mathbb{R}$.

- By means of an unitary (orthogonal in fact) transformation U we can bring this to diagonal form

$$\begin{aligned} \dot{z} &= \underbrace{U^T\bar{A}U}_{A} z + \underbrace{U^T\bar{B}}_{B} u, \\ y &= \underbrace{U}_{C} z, \end{aligned}$$

with $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

- Using the cyclic property of the trace, the $\mathcal{H}_2$ norm of this system can be computed as

$$\begin{aligned} \|G(s)\|_{\mathcal{H}_2}^2 &= \operatorname{trace}(CPC^T) = \operatorname{trace}(UPU^T) \\ &= \operatorname{trace}(PU^TU) = \operatorname{trace}(P). \end{aligned}$$

- Since A is symmetric, the controllability Gramian P obeys the Lyapunov equation $AP + PA + BB^T = 0$, or

$$P + A^{-1}PA = -A^{-1}BB^T.$$

- One has then, using that A is diagonal and that B is a column vector, $B = (b_1 \cdots b_n)^T$,

$$\begin{aligned} 2\ \operatorname{trace}(P) &= -\operatorname{trace}(A^{-1}BB^T) = -\sum_{i,j}^{n} A_{ij}^{-1} B_j B_i \\ &= -\sum_{i=1}^{n} \frac{b_i^2}{\lambda_i}. \end{aligned}$$

- Thus, for a $A^T = A < 0$ system with single input and full-state output

$$\|G\|_{\mathcal{H}_2}^2 = -\frac{1}{2}\sum_{i=1}^{n}\frac{b_i^2}{\lambda_i}.$$

- One can also show that, under the same restrictions,

$$\|G\|_{\mathcal{H}_\infty}^2 \leq -\sum_{i=1}^{n}\frac{|b_i|}{\lambda_i}.$$

- One can also compute $\|G\|_{\mathcal{H}_2}^2$ using

$$\|G(s)\|_{\mathcal{H}_2}^2 = \mathrm{trace}(B^T Q B) = \mathrm{trace}(Q B B^T).$$

- The observability Gramian matrix Q obeys, in this case

$$AQ + QA = -C^T C = -U^T U = -\mathbb{I}$$

and, for a nonsingular A, this can be solved as

$$Q = -\frac{1}{2} A^{-1}.$$

- Hence

$$\|G(s)\|_{\mathcal{H}_2}^2 = -\frac{1}{2}\mathrm{trace}(A^{-1} B B^T),$$

which coincides with the result of the first calculation.

- Assume that we select a block of m (not necessarily ordered or contiguous) states from the diagonalized system $(A, B, C = U)$, so that we have the mth order, single input, full n state output system $(\hat{A}, \hat{B}, \hat{C})$ with

$$\hat{A} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m), \quad \hat{B} = (b_1 \cdots b_m)^T, \quad \hat{C} = \hat{U},$$

  where $\hat{U}$ is the $n \times m$ matrix formed by the selected m columns of U. The λ_i, b_i are the eigenvalues and entries of B corresponding to the selected states.
- Now we want to approximate this by another mth order system of the form

$$\tilde{A} = \mathrm{diag}(\lambda, \ldots, \lambda), \quad \tilde{B} = (b \cdots b)^T, \quad \tilde{C} = \hat{U},$$

  such that the error system has an $\mathcal{H}_2$ norm as small as possible.

- Notice that $\hat{U}^T\hat{U} = \mathbb{I}_{m\times m}$ because the m columns of $\hat{U}$ are orthonormal vectors in $\mathbb{R}^n$, but $\hat{U}\hat{U}^T \neq \mathbb{I}_{n\times n}$.
- The rationale behind the system $(\tilde{A}, \tilde{B}, \tilde{C})$ is that a single dynamics with eigenvalue λ and input coefficient b yields all the states. Hence, only an integration must be performed, and then the output is recovered from m copies of this single state.
- For a given set of $(\lambda_1, \ldots, \lambda_m)$ and $(b_1, \ldots, b_m)$, the scalars λ and b must be chosen so as to make the error as small as possible. In turn, the value of this minimal error can be made small by carefully selecting the $(\lambda_1, \ldots, \lambda_m)$ and $(b_1, \ldots, b_m)$ that form a block.

- Let $\hat{G}(s) = \hat{U}(s\mathbb{I}_m - \hat{A})^{-1}\hat{B}$ be the transfer function of the block and $\tilde{G}(s) = \hat{U}(s\mathbb{I}_m - \tilde{A})^{-1}\tilde{B}$ that of the approximating block.
- The error system $E(s) = \hat{G}(s) - \tilde{G}(s)$ has a realization of order $2m$ given by

$$\begin{aligned} A_E &= \operatorname{diag}(\lambda_1, \ldots, \lambda_m, \lambda, \overset{m)}{\ldots}, \lambda) \\ B_E &= (b_1, \ldots, b_m, b, \overset{m)}{\ldots}, b)^T \\ C_E &= (\hat{U} \quad -\hat{U}) \end{aligned}$$

- Then $\|E\|_{\mathcal{H}_2}^2 = \operatorname{trace}(B_E^T Q_E B_E)$, with Q_E the solution to

$$A_E Q_E + Q_E A_E = -C_E^T C_E = \begin{pmatrix} -\mathbb{I}_{m\times m} & \mathbb{I}_{m\times m} \\ \mathbb{I}_{m\times m} & -\mathbb{I}_{m\times m} \end{pmatrix}.$$

- It can be seen that

$$Q_E = \begin{pmatrix} -\frac{1}{2}\hat{A}^{-1} & (\hat{A}+\lambda\mathbb{I}_{m\times m})^{-1} \\ (\hat{A}+\lambda\mathbb{I}_{m\times m})^{-1} & -\frac{1}{2\lambda}\mathbb{I}_{m\times m} \end{pmatrix}.$$

- Then

$$\|E\|_{\mathcal{H}_2}^2 = \sum_{i=1}^{m}\left(-\frac{b_i^2}{2\lambda_i} + 2\frac{bb_i}{\lambda+\lambda_i} - \frac{b^2}{2\lambda}\right) \equiv F(\lambda, b).$$

- Critical points of $F(\lambda, b)$ are given by

$$b^* = \frac{2\lambda^*}{m}\sum_{j=1}^{m}\frac{b_j}{\lambda^*+\lambda_j}$$

with λ^* a fixed point of the map

$$f(x) = \frac{1}{2}\frac{\sum_{j=1}^{m}\frac{b_j}{x+\lambda_j}}{\sum_{j=1}^{m}\frac{b_j}{(x+\lambda_j)^2}}.$$

- Using these results, one can compute the value of $\|E\|_{\mathcal{H}_2}^2$ at the critical point as

$$F(\lambda^*, b^*) = -\sum_{j=1}^{m} \frac{b_j^2}{2\lambda_j} + \frac{1}{m} \left(\sum_{j=1}^{m} \frac{b_j}{\lambda^* + \lambda_j} \right)^3 \left(\sum_{j=1}^{m} \frac{b_j}{(\lambda^* + \lambda_j)^2} \right)^{-1}.$$

- As a check, if $\lambda_j = \lambda$ and $b_j = b$ for all $j = 1, \ldots, m$, then $b^* = b$, $f(x) = \frac{1}{2}(x + \lambda)$, $\lambda^* = \lambda$ and $F(\lambda^*, b^*) = 0$.

- A detailed examination of $F(\lambda^*, b^*)$ reveals that, in order to make the error as small as possible, the λ_i and b_i of a block should be, as expected, packed as close as possible. Additionally, however,
  - a block should not contain b_i of different signs.
  - the λ_i should be bounded away from 0.
- A bound on $F(\lambda^*, b^*)$ can be computed in terms of the radii of the sets of λ_i and of b_i.
- Presently, we are using the *kmeans* clustering procedure of Matlab to form the blocks, but it is not optimal. We are working on a in-house clustering algorithm.
- This method seems quite promising. In our examples, which certainly might not be representative of practical situations, errors of less than 1% are obtained when approximating systems with $n = 1000$ with just 10 clusters.

Key general references on model order reduction

 A.C. Antoulas, *Approximation of Large-Scale Dynamical Systems*, SIAM Press, 2005.

 W.H.A. Schilders, H.A. van der Vorst, J. Rommes, Eds., *Model Order Reduction: Theory, Research Aspects and Applications*, Springer, 2008.