Empirical Gramians and Balanced Truncation for Model Reduction of Nonlinear Systems

Antoni Ras

Departament de Matemàtica Aplicada 4

Universitat Politècnica de Catalunya

Lecture goals

- To review the basic notions of controllability and observability.
- To introduce the technique of model reduction for nonlinear control systems through empirical gramians and balancing.
- To present the computational skills of this technique.

Outline

- Gramians, controllability and observability.
- Model reduction.
- Model reduction of nonlinear controlled systems: empirical gramians, balanced model and reduced model.
- Computation.
- References.

Controllability (1)

Consider a SISO linear, time-invariant system

$$\begin{aligned} \dot{\mathbf{x}}(t) &= A\,\mathbf{x}(t) + B\,u(t) \\ y(t) &= C\,\mathbf{x}(t) \end{aligned} \qquad (1)$$

where A, B, C and D are matrices, $u(t)$ is the input (the control) and $y(t)$ is the output.

Assume that A has n distinct eigenvalues, λ_1, ..., λ_n and let V be a eigenvectors matrix so that the transformation V^{-1} diagonalizes A :

$$\begin{aligned} \dot{\hat{x}}(t) &:= \overbrace{V^{-1}\mathbf{x}(t)}^{\cdot} = V^{-1}AV\,\hat{\mathbf{x}}(t) + V^{-1}B\,u(t) \\ y(t) &= CV\,\hat{\mathbf{x}}(t) + D\,u(t) \end{aligned}$$

and now $V^{-1}AV = \text{diagonal}(\lambda_1, ..., \lambda_n)$.

Controllability (2)

If some coefficient k of $V^{-1}B$ is null, the corresponding state $\hat{x}_k$ is uncontrollable, since its time behavior $\hat{x}_k = e^{\lambda_k t}\,\hat{x}_k(0)$ does not depend on the control $u(t)$.
In order to discover such a problem without computing the eigenvalues of A, the controllability matrix is introduced:

$$\mathcal{C} := [B,\ AB,\ A^2B,\ ...,\ A^{n-1}B].$$

Then, the lack of controllability of one state corresponds to a fall in the rang of $\mathcal{C}$.

Controllability (3)

Definition

A linear system like (1) is said to be controllable if its controllability matrix has full rank.

There are some equivalent alternatives for establishing controllability. One such alternatives is through the controllability gramian ([K 96]):

Definition

The controllability gramian for initial time t_0 and final time t_f is the matrix

$$G_C(t_0, t_f) := \int_{t_0}^{t_f} e^{A\,(t-t_0)}\, B\, B^T\, e^{A^T\,(t-t_0)}\, \mathrm{d}t.$$

Controllability (4)

Then:

Proposition

The system is controllable $\Leftrightarrow$
The controllability grammian is nonsingular for all $t_f > t_0$

When a system is controllable, there is a control input which transfers any initial state $x(t_0)$ to any arbitrary final state $x(t_f)$. Such an input may be evaluated in terms of the gramian as

$$u(t) = B^T \, e^{A^T \, (t - t_0)} \, G_C^{-1}(t_0, \, t_f) \, \left[e^{A \, (t_f - t_0)} \, \mathbf{x}(t_f) - \mathbf{x}(t_0) \right].$$

Observability (1)

In an analogous way, in order to check when initial states may be determined from the output, the observability matrix is introduced:

$$\mathcal{O} := [C,\, CA,\, ...,\, CA^{n-1}]^T$$

Definition

The observability gramian for initial time t_0 and final time t_f is the matrix

$$G_O(t_0,\, t_f) := \int_{t_0}^{t_f} e^{A^T\,(t-t_0)}\, C^T\, C\, e^{A\,(t-t_0)}\, \mathrm{d}t.$$

Then

Observability (2)

Proposition

The following statements are equivalent:

- The observability matrix has full rank n
- Observability gramian is nonsingular for all $t_f > t_0$

Then, system is said to be observable.

When a system is observable, its initial state may be determined from the output in terms of the observability gramian as

$$\mathbf{x}(t_0) = G_O^{-1}(t_0, t_f) \int_{t_0}^{t_f} e^{A^T (t-t_0)} C^T y(t) \, \mathrm{d}t,$$

Outline

The central idea in model reduction is **to find a low-dimensional affine subspace of the state space, in which the dynamics of interest of the original system are contained**.

Model reduction includes:

- The definition of an appropriate measure in the space of trajectories
- The search of the best approximating subspace with respect to this measure: the Karhunen-Loève decomposition
- The projection of the dynamics onto this subspace, where original system will be approximated by a small number of equations: the Galerkin projection

Definition of the best approximation

The method is similar to regression or principal component analysis in Statistics. First of all, one should collect empirical data (either form experiment or from simulation), consisting in sampled values $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$ of the trajectory $\mathbf{x}(t)$. Let S be a subspace and let

$$Q : \mathbb{R}^n \longrightarrow S$$

be the projection operator. As $|| \mathbf{x}^{(i)} - Q(\mathbf{x}^{(i)}) ||$ is the distance from $\mathbf{x}^{(i)}$ to its orthogonal projection $Q(\mathbf{x}^{(i)})$, the best approximation for the sampled solution is obtained minimizing the squared sum

$$d(Q) := \sum_{i=1}^{N} || \mathbf{x}^{(i)} - Q(\mathbf{x}^{(i)}) ||^2,$$

where the norm refers to a measure defined in the state space, selected according to physical meaning.

Minimum of the error

Let n be the number of states and let N be the number of collected points. The correlation matrix R of the sampled trajectory data $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$ is defined as

$$R_{rs} := \sum_{i=1}^{N} x_r^{(i)} \cdot x_s^{(i)}, \quad \text{for } 1 \leq r, s \leq n$$

Theorem

Let $\lambda_1 \geq \lambda_2 \geq .. \geq \lambda_n$ be the (ordered) eigenvalues of the correlation matrix R. Then, the minimum of $d(Q)$ for all projections Q on subspaces of dimension k is

$$\sum_{j=k+1}^{n} \lambda_j.$$

[1]

[1] The formula in [Hall et al. (2002)] has a misprint (?)

Best approximating subspace (1)

Let X be the $n \times N$ matrix whose columns are $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$. Let $\{\mathbf{v_1}, ..., \mathbf{v_n}\}$ be orthonormal eigenvectors of R and denote by V_k the matrix whose columns are the first k eigenvectors.

Then D:=diagonal$(\lambda_1, ..., \lambda_n) = V_n^{-1} R V_n = V_n^T X X^T V_n$; the projection Q on the subspace S_k spanned by $\{\mathbf{v_1}, ..., \mathbf{v_k}\}$ is given by $V_k V_k^T$ and $V_k^T \mathbf{x}^{(i)}$ gives the coordinates of $Q\left(\mathbf{x}^{(i)}\right)$ with respect to the selected basis on S_k.

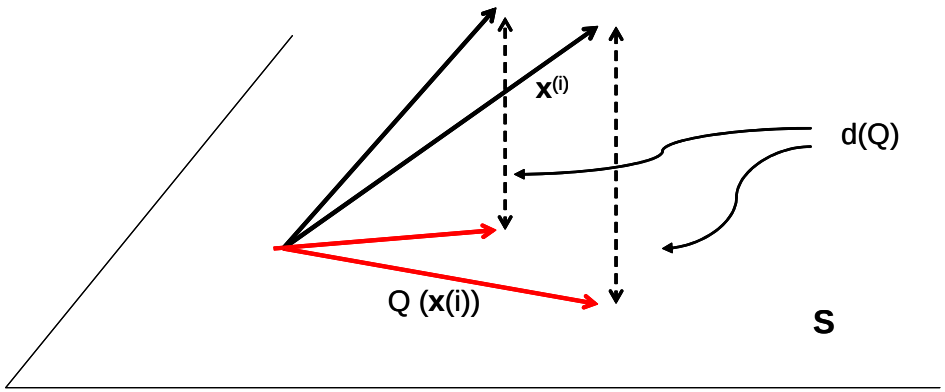
Denote by D_k the matrix constructed by the k-th principal minor of D. Thus $D_k = V_k^T X X^T V_k$, that is, the product of the matrix whose columns are $Q\left(\mathbf{x}^{(i)}\right)$ by its transpose. Then

Best approximating subspace (2)

$$\sum_{i=1}^{N} || Q(\mathbf{x}^{(i)}) ||^2 = \sum_{i=1}^{N} \left(\sum_{j=1}^{k} \left\{ \left(Q(\mathbf{x}^{(i)}) \right)_j \right\}^2 \right) =$$
$$\sum_{j=1}^{k} \left(\sum_{i=1}^{N} \left\{ \left(Q(\mathbf{x}^{(i)}) \right)_j \right\}^2 \right) = \sum_{j=1}^{k} (D_k)_{jj} = \sum_{j=1}^{k} \lambda_j .$$

and

$$d(Q) = \sum_{i=1}^{N} || \mathbf{x}^{(i)} ||^2 -$$
$$\sum_{i=1}^{N} || \left(Q(\mathbf{x}^{(i)}) \right)_j ||^2 =$$
$$\sum_{j=1}^{n} \lambda_j - \sum_{j=1}^{k} \lambda_j =$$
$$\sum_{j=k+1}^{n} \lambda_j$$



which is the minimum of $d(Q)$ over all rank k projections.

Affine correction

In order to obtain the optimal affine subspace, one should proceed with the covariance matrix C instead of R:

$$C_{rs} := \sum_{i=1}^{N} \left(x_r^{(i)} - \overline{x}_r \right) \cdot \left(x_s^{(i)} - -\overline{x}_s \right)$$

Finally, one should decide the rank of the projection subspace. The goal is to choose k such that the quotient of the sum of selected C-eigenvalues with respect to the total sum,

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{n} \lambda_j}$$

(that is, the proportion of the "energy" in the subspace) is close to one, yet k is sufficiently small.

Galerkin projection

We use Galerkin projection to construct lower-order models for a given dynamical system

$$\dot{\mathbf{x}}(t) = f\left(\mathbf{x}(t)\right).$$

First of all we should subtract the mean: $\widetilde{\mathbf{x}}(t) := \mathbf{x}(t) - \overline{\mathbf{x}}$. Thus,

$$\dot{\widetilde{\mathbf{x}}}(t) = f\left(\widetilde{\mathbf{x}}(t) + \overline{\mathbf{x}}\right)$$

Then, if $\mathbf{y}$ are the coordinates in the subspace S, (that is, $\mathbf{y}(t) = V_k^T \widetilde{\mathbf{x}}(t)$), the reduced-order approximation is given by

$$\dot{\mathbf{y}}(t) = V_k^T \dot{\widetilde{\mathbf{x}}}(t) = V_k^T f\left(V\mathbf{y}(t) + \overline{\mathbf{x}}\right).$$

One can expect such a procedure to work well within some region of state space, where the data should be collected. The principal advantage of this method is that it requires only matrix computations, despite its application to nonlinear systems.

Introduction (1)

Let us consider a nonlinear controlled dynamical system

$$\dot{\mathbf{x}}(t) = f\big(\mathbf{x}(t),\, \mathbf{u}(t)\big)$$
$$\mathbf{z}(t) = h\big(\mathbf{x}(t)\big)$$

where $\mathbf{x}(t)$ are the states, $\mathbf{u}(t)$ the input(s) and $\mathbf{z}(t)$ the output(s).

In [L 02], authors define a new resolution technique for such systems which rely on classical model reduction, but introduces a balancing algorithm in order to deal with nonlinearities.

Definition

Balancing means to apply a kind of linear transformations to two different matrices (here, the gramians) to obtain in booth cases the same diagonal matrix.

Introduction (2)

This technique includes several steps:

- To evaluate the (discrete) empirical gramians
- To balance booth empirical gramians and to evaluate the squared eigenvalues of the common diagonal matrix (Hankel singular values)
- According to the magnitudes of the eigenvalues, to choose the rang of the projection subspace.
- To solve the reduced model obtained by Galerkin projection onto a suitable subspace

Definition of the empirical gramians (1)

Empirical gramians are a sort of "covariance matrix", where covariance stands for the relation between steady state and perturbated solutions of the dynamical system.

The construction of empirical gramians depends on some parameters:

- n, the number of states; and p, the number of inputs;
- $\mathcal{T}^r = \{T_1, ..., T_r\}$, a set of orthogonal $n \times n$ matrices that will span the perturbation directions;
- $\mathcal{M} = \{c_1, ..., c_s\}$, a set of s positive constants (the different sizes of the perturbations); and
- $\mathcal{E}^p$, the set of standard unit vectors in $\mathbb{R}^p$.

Definition of the empirical gramians (2)

Let $x^{ilm}(t)$ be te state corresponding to the impulsive input $u(t) = c_m T_l e_i \delta(t)$.

Recall the definition of the temporal mean of any function $g(t)$:

$$\overline{g}(t) := \lim_{T\to\infty} \frac{1}{T} \int_0^T g(t)\,\mathrm{d}t.$$

From a theoretical point of view [2], empirical gramians have the following definition:

[2] that is, to ensure that in the linear case classical gramians are obtained

Definition of the empirical gramians (3)

Definition

Controllability empirical grammian

$$G_c := \sum_{i=1}^{p} \sum_{l=1}^{r} \sum_{m=1}^{s} \frac{1}{rsc_m^2} \int_0^\infty \left(x^{ilm}(t) - \overline{x}^{ilm}\right) \left(x^{ilm}(t) - \overline{x}^{ilm}\right)^T$$

Observability empirical grammian

$$G_o := \sum_{l=1}^{r} \sum_{m=1}^{s} \frac{1}{rsc_m^2} \int_0^\infty T_l \Psi^{lm} T_l^T \, dt,$$

where Ψ^{lm} is the $n \times n$ matrix given by

$$\left(\Psi^{lm}\right)_{ij} = \left(z^{ilm}(t) - \overline{z}^{ilm}\right)^T \left(z^{jlm}(t) - \overline{z}^{jlm}\right)$$

Definition of the empirical gramians (4)

For practical applications [3], one should use a finite sum of a sampled trajectory instead an infinite integral. Also, temporal means are replaced by steady states. The later implies that **one should know an input reference $\mathbf{u}_{ss}$**, (probably suggested by physical meaning of the underlying problem).

Then, the steady state $\mathbf{x}_{ss}$ is obtained from $f\big(\mathbf{x}_{ss}(t),\, \mathbf{u}_{ss}(t)\big) = 0$ and the corresponding output is denoted by $\mathbf{z}_{ss}$.

Discrete empirical gramians are calculated as follows: Let Δt be the sampling interval and let q be the number of points in the sample (usually, $q \sim 2000$). **One should to be aware that the system will have reached equilibrium for some time $t_f < q \cdot \Delta t$.**

[3] see [H 02]

Definition of the empirical gramians (5)

Put

$$x_k^{ilm} := x^{ilm}(k\,\Delta t), \quad z_k^{ilm} := z^{ilm}(k\,\Delta t)$$
$$\left(\Psi_k^{lm}\right)_{ij} := \left(z_k^{ilm} - z_{ss}^{ilm}\right)^T \left(z_k^{jlm} - z_{ss}^{jlm}\right)$$

Definition

Discrete controllability empirical grammian

$$W_c := \sum_{i=1}^{p}\sum_{l=1}^{r}\sum_{m=1}^{s}\frac{1}{rsc_m^2}\sum_{k=0}^{q}\left(x_k^{ilm} - x_{ss}^{ilm}\right)\left(x_k^{ilm} - x_{ss}^{ilm}\right)^T \Delta t$$

Discrete observability empirical grammian

$$W_o := \sum_{l=1}^{r}\sum_{m=1}^{s}\frac{1}{rsc_m^2}\sum_{k=0}^{q} T_l \Psi_k^{lm} T_l^T \Delta t.$$

Balancing

Definition

To balance the model, one should find a linear transformation L such that $L\,W_c\,L^T = \left(L^{-1}\right)^T\,W_o\,\left(L^{-1}\right) = \Sigma$, where Σ is a diagonal matrix.

Well know numerical techniques (Schur singular values decomposition) are applied to find the change of coordinates L. If gramians do not are full rank, only a decomposition of the following kind is obtained:

$$LW_cL^T = \begin{pmatrix} \Sigma_1 & 0 & 0 & 0 \\ 0 & Id & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \left(L^{-1}\right)^T W_o \left(L^{-1}\right) = \begin{pmatrix} \Sigma_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \Sigma_2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and one proceeds with Σ_1 instead of Σ.

Hankel singular values

Definition

Hankel singular values are the eigenvalues of $G_o\,G_c$.

Hankel singular values indicate the importance of the corresponding states in the transfer of energy from past inputs to future outputs. Then, the procedure will be to truncate states corresponding to small Hankel singular values, i.e., states that contribute very little to the input-output system behavior.

To find which state correspond to some Hankel singular values one should obtain the balanced realization of the system, because Hankel singular values are exactly its squared eigenvalues.

In the linear case, changes of coordinates don't affect Hankel singular values (even though the gramians are not invariant). For nonlinear systems, this result no longer holds.

Projection

The reduced model is then obtained in the usual way, but now the projection matrix V_k^T is a product of the transformation matrix L that balances gramians and a projection matrix $P = [Id_k \ 0]$ where k is the number of preserved states. The reduced system reads

$$\dot{\mathbf{y}}(t) = PL\, f\left(L^{-1}P^T\, \mathbf{y}(t),\, \mathbf{u}(t)\right)$$

$$\mathbf{z}(t) = h\left(L^{-1}P^T\, \mathbf{y}(t)\right)$$

Modified algorithm

In [H 02] an improvement is proposed for the case where steady states are different from zero. The modified system includes new equations and substitutes the eliminated states by its equilibrium value:

Let $\mathbf{y}_1$ be the states in the reduced model and let $\mathbf{y}_2$ be the states truncated by the projection. Put $\widehat{\mathbf{y}} := (\mathbf{y}_1, \mathbf{y}_2)^T$ and solve

$$\dot{\mathbf{y}}_1(t) = PL\, f\left(L^{-1}\,\widehat{\mathbf{y}}(t),\, \mathbf{u}(t)\right) \quad [4]$$

$$\mathbf{y}_2(t) = (\mathbf{y}_2)_{ss}$$

$$\mathbf{z}(t) = h\left(L^{-1}\,\widehat{\mathbf{y}}(t)\right)$$

[4] Notice that, instead a true projection, truncated states have been replaced with their steady state value

Preliminary remarks

J. Hahn and co-workers have write some MATLAB algorithms to solve nonlinear controlled systems using the procedure presented in [L 02] and developed for computational purposes in [H 02]. All these m-files are collected in the URL

http://cheweb.tamu.edu/orgs/groups/Hahn/Model_Reduction

Some demo-examples and a short user-guide are also available.

Prior to describe the routines, we shall consider two aspects:

- The necessary changes when system is not control-affine.
- The convenience of scaling in the control-affine case.

Non control-affine systems

Following [H 02], empirical gramians are suitable only for control-affine systems, because general nonlinear operations are not defined for impulses: remember the formulæ of gramians, where the inputs include a Dirac's delta.

Otherwise, one should take $u(t) = c_m T_l e_i v(t)$ where $v(t)$ is a series of steps. Yet $W_c(t)$ and $W_o(t)$ do not correspond to empirical gramians but a so-called controllability and observability covariance matrices.

For the rest, the method follows as usual. Only minor changes in the syntaxis of routines are needed.

Scaling systems (1)

The set of routines in Hahn's web is complete only for the "scaled" case. Otherwise, routines for model reduction and comparison should be adapted. The reason to impose scaled form seems rather eccentric and remains unknown to us[5].

Definition

Scaling simply means to apply a linear transformation such that all non-null coordinates of reference input and steady states became 1.

[5] That the condition should hold only for non-null steady states is only a personal assumption

Scaling systems (2)

More precisely, put

$$T_u := \text{diagonal}\left((u_{ss})_1, \ ..., \ (u_{ss})_p \right), \quad [6]$$
$$T_x := \text{diagonal}\left((x_{ss})_1, \ ..., \ (x_{ss})_n \right)$$

then, the scaled system is given by

$$\dot{\widetilde{\mathbf{x}}}(t) = (T_x)^{-1} \ f\left(T_x \, \widetilde{\mathbf{x}}(t), \, T_u \, \widetilde{\mathbf{u}}(t)\right)$$
$$\mathbf{z}(t) = h\left(T_x \, \widetilde{\mathbf{x}}(t)\right)$$

[6] (if some term vanishes, it should be replaced by 1)

Description of routines (1)

- ctrl_gram_cov*(a,[b],[c],d,e)*

  Parameters:

  1. a refers to a MATLAB m-file with the scaled dynamical system (Syntaxis: Function dx=a(t,x))
  2. *[b]* = [start time, end time, sampling time interval]
  3. *[c]* = [number of inputs, number of states, number of outputs, 2, sample length]
  4. *d* = the set of constants $\mathcal{M}$
  5. *e* = 0, for gramian; $\neq 0$, for covariance matrix

  Answer: W_c, the empirical controllability gramian (or covariance matrix)

This routine should be modified if some steady state vanishes.

Description of routines (2)

- obsv_gram_cov*(a,[b],[c],d,[f],g,e)*

  Parameters:

  1. *a* to *e* like the preceding case
  2. *[f]* = the index of the states that are the output of the system (The routine assumes that output are some states)
  3. g = the steady state x_{ss}

  Answer: W_o, the empirical observability gramian (or covariance matrix)

This routine should be modified if some steady state vanishes.

Remark: Notice that gramians are evaluated assuming $\mathcal{T}^r = \{\mathrm{Id}\}$ (see [L 02]).

Description of routines (3)

- bal_realization*(h,i,j)*

  Parameters:

  1. $h = W_c$
  2. $i = W_o$
  3. $j =$ number of states

  Answer:

  1. Trans = the transformation matrix for balancing and invTrans, its inverse
  2. $Wc,\ Wo =$ balanced controllability and observability gramians
  3. *svd_Wc, svd_Wo* = the Hankel singular values (as eigenvalues[7] of either Wc or Wo)

[7] In fact, the routine should be modified, because Hankel s.v. really are the squared eigenvalues

Description of routines (4)

Running the m-file nonlinear_test2 is a shortcut for all 3 previous routines. File should include the steady state x_{ss} and the reference input u_{ss}. As a result, it saves the output of bal_realization in a file.

Next, one can run the m-file nonlinear_test2_comparison that gives a figure showing (one selected) output booth for the full-order and the truncated system.

This file should also include the value of x_{ss} and u_{ss}, together with *red_n,* the number of remaining states in the reduced system.

Remark This file calls two ode function m-files, corresponding to full and reduced-order system.

Description of routines (5)

- Remarks on the ode m-file for the scaled system

  To introduce the previously explained scaling change of coordinates, some changes should be introduced on the m-file corresponding to the dynamical system:

  1. Before to write system equations, one should apply matrices T_x and T_u :

     $$u = diag(uss)*ud; \quad x = diag(xss)*x;$$

  2. and file should end with the inverse transformation

     $$dx = diag(1./xss)*dx;$$

Of course, one should be aware of the remark in footnote 6.

Description of routines (6)

- Remarks on the ode m-file for the reduced (scaled) system

  To embody balancing transform L and model reduction, the m-file for the scaled system should include some modifications:

  1. States are transformed now according to T_x, and L

     *x=diag(xss)*invTrans*x;*

  2. Inverse transformation and reduction enter at the end of the file [8]:

     *dx(1:red_n,1) = Trans(1:red_n,:)*dx;*

     (inverse transform for the conserved states)

     dx(red_n+1:n,1) = zeros(n - red_n,1);

     (steady state holds in truncated states)

[8] It seems to me that this procedure neither reduce CPU time nor takes advantage of the reduction in the number of equations

References

K 96 Kamen, W.: Fundamentals of linear time-varying systems, in *The Control Handbook*, (Levine, W.S. editor), CRC Press, 1996.

L 02 Lall, S., Marsden, J.E., Glavaški, S.: A subspace approach to balanced truncation for model reduction of nonlinear control systems. *Int. J. Robust Nonlinear Control* 2002; **12**: 519-535.

H 02 Hahn, J.,Edgar, T. F.: Balancing approach to minimal mealization and model reduction of stable nonlinear systems. *Ind. Eng. Chem. Res.* 2002; **41**: 2204-2212.