Systems & Control Letters 67 (2014) 55–64

Contents lists available at ScienceDirect

Systems & Control Letters

journal homepage: www.elsevier.com/locate/sysconle

On optimality of approximate low rank solutions of large-scale matrix equations

Peter Benner [a,b], Tobias Breiten [a,c,*]

[a] Computational Methods in Systems and Control Theory, Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg, Germany

[b] Mathematik in Industrie und Technik, Fakultät für Mathematik, TU Chemnitz, 09107 Chemnitz, Germany

[c] Institute for Mathematics and Scientific Computing, Karl-Franzens-Universität Graz, Heinrichstr. 36, 8010 Graz, Austria

ARTICLE INFO

Article history:
Received 22 November 2012
Received in revised form
10 February 2014
Accepted 10 February 2014
Available online 17 March 2014

Keywords:
Matrix equations
Low rank approximations
Rational Krylov subspaces
$\mathcal{H}_2$-model reduction

ABSTRACT

In this paper, we discuss some optimality results for the approximation of large-scale matrix equations. In particular, this includes the special case of Lyapunov and Sylvester equations, respectively. We show a relation between the iterative rational Krylov algorithm and a Riemannian optimization method which recently has been shown to locally minimize a certain energy norm of the underlying Lyapunov operator. Moreover, we extend the results for a more general setting leading to a slight modification of IRKA. By means of some numerical test examples, we show the efficiency of the proposed methods.

1. Introduction

In this paper, we discuss optimality results for low rank approximations of solutions $\mathbf{X} \in \mathbb{R}^{n\times q}$ of large-scale matrix equations of the form

$$\mathbf{AXM} + \mathbf{EXH} + \mathbf{BC} = \mathbf{0}, \tag{1}$$

where $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{n\times n}$, $\mathbf{M}, \mathbf{H} \in \mathbb{R}^{q\times q}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$ and $\mathbf{C} \in \mathbb{R}^{m\times q}$. As is well-known (see [1]), these types of equations play an important role in analyzing the structure of dynamical systems. For example, in case of $\mathbf{H} = \mathbf{A}^T$, $\mathbf{M} = \mathbf{E}^T$ and $\mathbf{C} = \mathbf{B}^T$, the resulting Lyapunov equation characterizes the stability properties of the associated dynamical system

$$\mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{Ax}(t) + \mathbf{Bu}(t), \qquad \mathbf{y}(t) = \mathbf{Cx}(t). \tag{2}$$

Throughout the paper we assume the matrix pencils $(\mathbf{A}, \mathbf{E})$ and $(\mathbf{H}, \mathbf{M})$ to be stable, i.e., all eigenvalues are included in the open left complex plane. While for small-to-medium scale equations the exact solution can be computed explicitly by means of the Bartels–Stewart algorithm (see [2]) or Hammarlings method (see [3]), this will no longer be possible for $n \sim 10^5$–10^6 since the complexity scales with $\mathcal{O}(n^3)$, making even the storage of the dense solution matrix $\mathbf{X}$ hardly possible. However, in a variety of interesting real-life applications, we have rank $(\mathbf{BC}) \ll n, m$ often inducing a fast singular value decay of the solution matrix $\mathbf{X}$. For a detailed discussion on this topic, we refer to, e.g., [4–7]. The fact that $\mathbf{X}$ can be well approximated by low rank matrices, i.e., $\mathbf{X} \approx \mathbf{UV}^T$ with $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n\times\hat{n}}$, $\hat{n} \ll n$, has motivated the development of several numerically efficient low rank methods for solving (1). Here, the most prominent ones are the alternating direction implicit (ADI) iteration as well as the Krylov-plus-inverted-Krylov method (KPIK), see, e.g., [8–10]. Recently, for the case of the Lyapunov equation, a rather different approach has been proposed in [11] which relies on Riemannian optimization and minimizes a certain energy norm associated with the underlying Lyapunov operator.

The main result of our paper is to connect the latter method with the iterative rational Krylov algorithm (IRKA), a well-established approach in the context of model order reduction of dynamical systems of the form (2), see [12]. We note that all our results are primarily of theoretical interest and we hope that they will lead to new insight into and better understanding of the relations between the most popular methods for solving

* Corresponding author at: Institute for Mathematics and Scientific Computing, Karl-Franzens-Universität Graz, Heinrichstr. 36, 8010 Graz, Austria. Tel.: +43 3163808546.

E-mail addresses: benner@mpi-magdeburg.mpg.de (P. Benner), tobias.breiten@uni-graz.at, breiten@mpi-magdeburg.mpg.de (T. Breiten).

http://dx.doi.org/10.1016/j.sysconle.2014.02.005
0167-6911/

large-scale Lyapunov and Sylvester equations. In a certain sense, the approximations we propose are optimal and we provide numerical algorithms which may converge to these optimal approximations. Nevertheless, since all of them are of iterative nature, they strongly depend on the convergence rate and the cost per iteration. In most cases they will not be competitive in terms of numerical efficiency when compared to methods like, e.g., the ADI method and rational Krylov, but we expect our findings to be helpful in further optimizing these methods and possibly arriving at a hybrid method aggregating their best features.

The structure of this paper now is as follows. We start with a brief review of $\mathcal{H}_2$-optimal model reduction of linear systems and subsequently we show that for symmetric state-space systems, constructing locally $\mathcal{H}_2$-optimal models is equivalent to minimizing an energy norm naturally arising for the Lyapunov equation corresponding to the system. For unsymmetric systems, we propose a method that for a given reduced rank $\hat{n}$ aims at locally minimizing the residual of the Lyapunov equation of size $n \times n$. We briefly discuss the generalization for the case of the Sylvester equation. By means of some numerical examples, we test the new methods and underscore our theoretical results.

Throughout the paper, all matrices and vectors are denoted in bold letters and $\mathbf{I}$ always denotes the identity matrix whose dimension should be clear from the context. With $\otimes$ we denote the Kronecker product of two matrices and vec $(\cdot)$ will be understood as the operation of stacking the columns of a matrix into a long vector. For a matrix $\mathbf{V} \in \mathbb{R}^{n\times\hat{n}}$, we denote $\mathcal{V} = \mathrm{Ran}(\mathbf{V})$. Further, tr $(\cdot)$ denotes the trace of a matrix, i.e., the sum of its diagonal entries while $\hat{\mathbf{A}} \in \mathbb{R}^{\hat{n}\times\hat{n}}$ always denotes that the matrix is related to a reduced-order model, meaning that $\hat{n} \ll n$. With $\mathbf{A} \succ 0$ or $\mathbf{A} \prec 0$ we denote a symmetric positive or negative definite matrix, respectively. For matrices $\mathbf{A} \prec 0, \mathbf{H} \prec 0, \mathbf{E} \succ 0$ and $\mathbf{M} \succ 0$, we associate norms induced by $\mathcal{L} = -\mathbf{E}\otimes\mathbf{A} - \mathbf{A}\otimes\mathbf{E}$ and $\mathcal{L}_S = -\mathbf{M}\otimes\mathbf{A} - \mathbf{H}\otimes\mathbf{E}$ as follows:

$$\|\cdot\|_{\mathcal{L}} = \sqrt{\langle\cdot,\cdot\rangle} \quad \text{with } \langle\mathbf{U},\mathbf{V}\rangle_{\mathcal{L}} = \langle \mathrm{vec}\,(\mathbf{U}),\, \mathcal{L}\,\mathrm{vec}\,(\mathbf{V})\rangle, \tag{3}$$

$$\|\cdot\|_{\mathcal{L}_S} = \sqrt{\langle\cdot,\cdot\rangle} \quad \text{with } \langle\mathbf{U},\mathbf{W}\rangle_{\mathcal{L}_S} = \langle \mathrm{vec}\,(\mathbf{U}),\, \mathcal{L}_S\mathrm{vec}\,(\mathbf{W})\rangle, \tag{4}$$

for matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n\times n}$ and $\mathbf{W} \in \mathbb{R}^{q\times q}$.

2. $\mathcal{H}_2$-Optimal model reduction

In this section, we briefly introduce the topic of model order reduction of a dynamical system. Basically, the goal is to replace a system of the form (2) by another system with the same structure but much fewer states, i.e.,

$$\hat{\mathbf{E}}\dot{\hat{\mathbf{x}}}(t) = \hat{\mathbf{A}}\hat{\mathbf{x}}(t) + \hat{\mathbf{B}}\mathbf{u}(t), \qquad \hat{\mathbf{y}}(t) = \hat{\mathbf{C}}\hat{\mathbf{x}}(t), \tag{5}$$

where $\hat{\mathbf{E}}, \hat{\mathbf{A}} \in \mathbb{R}^{\hat{n}\times\hat{n}}$, $\hat{\mathbf{B}} \in \mathbb{R}^{\hat{n}\times m}$, $\hat{\mathbf{C}} \in \mathbb{R}^{p\times\hat{n}}$ and $\hat{n} \ll n$. Here, the reduced-order matrices usually are obtained by a Petrov–Galerkin projection of the form $\mathcal{P} = \mathbf{V}\mathbf{W}^T, \mathbf{W}^T\mathbf{V} = \mathbf{I}$ meaning that $\hat{\mathbf{E}} = \mathbf{W}^T\mathbf{E}\mathbf{V}$, $\hat{\mathbf{A}} = \mathbf{W}^T\mathbf{A}\mathbf{V}$, $\hat{\mathbf{C}} = \mathbf{C}\mathbf{V}, \hat{\mathbf{B}} = \mathbf{W}^T\mathbf{B}$. We abbreviate the above structure by denoting $\hat{\boldsymbol{\Sigma}} = (\hat{\mathbf{E}}; \hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}})$. Since (5) should approximate (2) in some sense, one usually demands $\mathbf{y} \approx \hat{\mathbf{y}}$. For linear systems, the deviation from a reduced system to the original one can for instance be measured in terms of the $\mathcal{H}_2$-norm

$$\|\boldsymbol{\Sigma}\|_{\mathcal{H}_2} := \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{tr}\left(\mathbf{H}(-i\omega)\mathbf{H}(i\omega)^T\right)\mathrm{d}\omega\right)^{\frac{1}{2}},$$

where $\mathbf{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$ is called the transfer function of $\boldsymbol{\Sigma}$. Note that due to [1,12] we have

$$\|\boldsymbol{\Sigma}\|_{\mathcal{H}_2}^2 = \mathrm{tr}\left(\mathbf{C}\mathbf{P}\mathbf{C}^T\right) = \mathrm{tr}\left(\mathbf{B}^T\mathbf{Q}\mathbf{B}\right),$$

where

$$\mathbf{A}\mathbf{P}\mathbf{E}^T + \mathbf{E}\mathbf{P}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = \mathbf{0}, \qquad \mathbf{A}^T\mathbf{Q}\mathbf{E} + \mathbf{E}^T\mathbf{Q}\mathbf{A} + \mathbf{C}^T\mathbf{C} = \mathbf{0}.$$

Similarly, for two systems $\boldsymbol{\Sigma}_1 = (\mathbf{E}_1; \mathbf{A}_1, \mathbf{B}_1, \mathbf{C}_1)$ and $\boldsymbol{\Sigma}_2 = (\mathbf{E}_2; \mathbf{A}_2, \mathbf{B}_2, \mathbf{C}_2)$ the $\mathcal{H}_2$-inner product can be derived as

$$\langle\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2\rangle_{\mathcal{H}_2} = \mathbf{C}_1\mathbf{Z}\mathbf{C}_2^T, \tag{6}$$

where $\mathbf{A}_1\mathbf{Z}\mathbf{E}_2^T + \mathbf{E}_1\mathbf{Z}\mathbf{A}_2^T + \mathbf{B}_1\mathbf{B}_2^T = \mathbf{0}$. For measuring the quality of a reduced-order model with respect to the $\mathcal{H}_2$-norm, one defines the error system as

$$\mathbf{E}_e = \begin{bmatrix}\mathbf{E} & \mathbf{0}\\ \mathbf{0} & \hat{\mathbf{E}}\end{bmatrix}, \qquad \mathbf{A}_e = \begin{bmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \hat{\mathbf{A}}\end{bmatrix}, \qquad \mathbf{B}_e = \begin{bmatrix}\mathbf{B}\\ \hat{\mathbf{B}}\end{bmatrix}, \qquad \mathbf{C}_e = \begin{bmatrix}\mathbf{C} & -\hat{\mathbf{C}}\end{bmatrix}. \tag{7}$$

The $\mathcal{H}_2$-error of a reduced-order model $\hat{\boldsymbol{\Sigma}}$ thus is given as

$$\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_{\mathcal{H}_2}^2 = \mathrm{tr}\left(\mathbf{C}_e\mathbf{P}_e\mathbf{C}_e^T\right) = \mathrm{tr}\left(\mathbf{B}_e^T\mathbf{Q}_e\mathbf{B}_e\right), \tag{8}$$

where $\mathbf{P}_e$ and $\mathbf{Q}_e$ solve the Lyapunov equations of the error system, i.e.,

$$\begin{aligned}\mathbf{A}_e\mathbf{P}_e\mathbf{E}_e^T + \mathbf{E}_e\mathbf{P}_e\mathbf{A}_e^T + \mathbf{B}_e\mathbf{B}_e^T &= \mathbf{0},\\ \mathbf{A}_e^T\mathbf{Q}_e\mathbf{E}_e + \mathbf{E}_e^T\mathbf{Q}_e\mathbf{A}_e + \mathbf{C}_e^T\mathbf{C}_e &= \mathbf{0}.\end{aligned} \tag{9}$$

Interpreting $J := \|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_{\mathcal{H}_2}^2$ as a cost functional depending on the reduced-order system matrices $\hat{\mathbf{E}}$, $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$, we can derive first-order necessary conditions by setting $\nabla J_{\hat{\mathbf{E}}}$, $\nabla J_{\hat{\mathbf{A}}}$, $\nabla J_{\hat{\mathbf{B}}}$ and $\nabla J_{\hat{\mathbf{C}}}$ equal to zero:

$$\begin{aligned}\mathbf{Q}_{12}^T\mathbf{A}\mathbf{P}_{12} + \mathbf{Q}_{22}\hat{\mathbf{A}}\mathbf{P}_{22} &= \mathbf{0}, \qquad \mathbf{Q}_{12}^T\mathbf{E}\mathbf{P}_{12} + \mathbf{Q}_{22}\hat{\mathbf{E}}\mathbf{P}_{22} = \mathbf{0},\\ \mathbf{Q}_{12}^T\mathbf{B} + \mathbf{Q}_{22}\hat{\mathbf{B}} &= \mathbf{0}, \qquad \hat{\mathbf{C}}\mathbf{P}_{22} - \mathbf{C}\mathbf{P}_{12} = \mathbf{0},\end{aligned} \tag{10}$$

where we assumed that $\mathbf{P}_e$ and $\mathbf{Q}_e$ are partitioned as follows

$$\mathbf{P}_e = \begin{bmatrix}\mathbf{P}_{11} & \mathbf{P}_{12}\\ \mathbf{P}_{12}^T & \mathbf{P}_{22}\end{bmatrix}, \qquad \mathbf{Q}_e = \begin{bmatrix}\mathbf{Q}_{11} & \mathbf{Q}_{12}\\ \mathbf{Q}_{12}^T & \mathbf{Q}_{22}\end{bmatrix}.$$

The above conditions have been derived in [13] and usually are referred to as the Wilson conditions for $\mathcal{H}_2$-optimality.

Alternatively, one can characterize optimality in terms of bitangential Hermite-type interpolation conditions for the reduced transfer function. Assume that

$$\hat{\mathbf{H}}(s) = \hat{\mathbf{C}}(s\hat{\mathbf{E}} - \hat{\mathbf{A}})^{-1}\hat{\mathbf{B}} = \sum_{k=1}^{\hat{n}}\frac{\mathbf{c}_k\mathbf{b}_k^T}{s - \hat{\lambda}_k},$$

is the pole-residue representation of the reduced-order system. Similar to the analysis in the SISO case [14], one can show that a locally optimal reduced-order system with simple poles at $\{\hat{\lambda}_1, \ldots, \hat{\lambda}_{\hat{n}}\}$ fulfills:

$$\mathbf{H}(-\hat{\lambda}_k)\mathbf{b}_k = \hat{\mathbf{H}}(-\hat{\lambda}_k)\mathbf{b}_k, \tag{11a}$$

$$\mathbf{c}_k^T\mathbf{H}(-\hat{\lambda}_k) = \mathbf{c}_k^T\hat{\mathbf{H}}(-\hat{\lambda}_k), \tag{11b}$$

$$\mathbf{c}_k^T\mathbf{H}'(-\hat{\lambda}_k)\mathbf{b}_k = \mathbf{c}_k^T\hat{\mathbf{H}}'(-\hat{\lambda}_k)\mathbf{b}_k, \tag{11c}$$

for $k = 1, \ldots, \hat{n}$. An iterative algorithm (IRKA), see Algorithm 1, that upon convergence yields a locally $\mathcal{H}_2$-optimal reduced-order system has been introduced in [12]. Recently, in [15], the authors have shown that for symmetric state-space systems, IRKA in fact is a locally attractive fixed point iteration.