**Electronic Transactions on Numerical Analysis.**
**Volume 29, pp. 136-149, 2008.**

**ISSN 1068-9613.**

**ETNA**
**Kent State University**
**http://etna.math.kent.edu**

# ON THE PARAMETER SELECTION PROBLEM IN THE NEWTON-ADI ITERATION FOR LARGE-SCALE RICCATI EQUATIONS*

PETER BENNER†, HERMANN MENA‡, AND JENS SAAK§

**Abstract.** The numerical treatment of linear-quadratic regulator (LQR) problems for parabolic partial differential equations (PDEs) on infinite-time horizons requires the solution of large-scale algebraic Riccati equations (AREs). The Newton-ADI iteration is an efficient numerical method for this task. It includes the solution of a Lyapunov equation by the alternating direction implicit (ADI) algorithm at each iteration step. Here, we study the selection of shift parameters for the ADI method. This leads to a rational min-max problem which has been considered by many authors. Since knowledge about the exact shape of the complex spectrum is crucial for computing the optimal solution, this is often infeasible for the large-scale systems arising from finite element discretization of PDEs. Therefore, several methods for computing suboptimal parameters are discussed and compared on numerical examples.

**Key words.** algebraic Riccati equation, Newton-ADI, shift parameters, Lyapunov equation, rational min-max problem, Zolotarev problem

**AMS subject classifications.** 15A24, 30E10, 65B99

## 1. Introduction.

Optimal control problems governed by partial differential equations are a topic of current research. Many control, stabilization, and parameter identification problems can be reduced to the linear-quadratic regulator (LQR) problem, see [10, 13, 21, 22]. Particularly, LQR problems for parabolic systems have been studied in detail in the past 30 years, and several results concerning existence theory and numerical approximation can be found, e.g., in [21, 22, 24] and the references therein. Gibson [16] and Banks and Kunisch [3] present approximation techniques to reduce the inherently infinite-dimensional problem of the distributed regulator problem for parabolic PDEs to (large) finite-dimensional analogues.

The solution of these finite-dimensional problems can be reduced to the solution of a matrix Riccati equation. In the finite-time horizon case, this is a first-order differential equation and in the infinite–time horizon case an algebraic one, see, e.g., [4, 31].

In Section 1.1, we state the Riccati equations of interest and introduce the matrices and basic notations used in the remainder. Then, we review the Newton-ADI iteration for the solution of large-scale matrix Riccati equations in Section 1.2, showing how this involves the solution of a Lyapunov equation with specially structured matrices by the alternating direction implicit (ADI) algorithm in every iteration step. Furthermore, we introduce the rational min-max problem related to the parameter selection problem there, which is the main topic of this paper. We give a brief summary of Wachspress' results and a heuristic choice of parameters described in [28], as well as a Leja point approach [32, 33] in Section 2. In Section 3, we show how the first two of these methods can be combined to have a parameter computation which can be applied efficiently even in case of very large systems. Section 4 shows the efficiency of our method compared to the Wachspress parameters for test examples, where the complete

*Received October 31, 2006. Accepted for publication March 7, 2008. Published online on June 26, 2008. Recommended by A. Frommer. This work was supported by the DFG project "Numerische Lösung von Optimalsteuerungsproblemen für instationäre Diffusions-Konvektions- und Diffusions-Reaktionsgleichungen", grant BE3715/1-1 and DAAD program "Acciones Integradas Hispano-Alemanas", grant D/05/25675.

†Fakultät für Mathematik, Technische Universität Chemnitz, 09107 Chemnitz, Germany (benner@mathematik.tu-chemnitz.de).

‡Departamento de Matemática, Escuela Politécnica Nacional, Quito, Ecuador (hmena@server.epn.edu.ec).

§Fakultät für Mathematik, Technische Universität Chemnitz, 09107 Chemnitz, Germany (jens.saak@mathematik.tu-chemnitz.de).

**ETNA**
Kent State University
http://etna.math.kent.edu



spectrum can still be computed numerically and thus Wachspress' method can be used to compute the optimal parameters. Finally, we state some conclusions in Section 5.

**1.1. Notation and background.** In this paper, we concentrate on solving large sparse matrix Riccati equations arising in the optimal control of semidiscretized PDEs (see, e.g., [6, 9]). Depending on whether the control problems are formulated on infinite- or finite-time horizons, these Riccati equations are

$$0 = \Re_h(\mathbf{X}) = \mathbf{C}^T\tilde{\mathbf{Q}}\mathbf{C} + \mathbf{A}^T\mathbf{X} + \mathbf{X}\mathbf{A} - \mathbf{X}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{X} \tag{1.1}$$

or

$$\dot{\mathbf{X}} = -\Re_h(\mathbf{X}) = -\mathbf{C}^T\tilde{\mathbf{Q}}\mathbf{C} - \mathbf{A}^T\mathbf{X} - \mathbf{X}\mathbf{A} + \mathbf{X}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{X}, \tag{1.2}$$

respectively. Typically, the coefficient matrices of these Riccati equations have a given structure (e.g., sparse, symmetric, or low rank). Efficient numerical methods for-large scale problems have to exploit this structure. The main focus of our research is how this can be achieved within an ADI parameter selection procedure.

The algebraic Riccati equation (ARE) is a nonlinear system of equations, so it is natural to apply Newton's method to find its solutions. This approach has been investigated; details and further references can be found in [4, 14, 20, 26, 29, 30]. Differential Riccati equations can efficiently be solved by BDF methods known from ordinary differential equations [8, 12, 15]. This involves solving algebraic equations of type (1.1) in each time step. Thus, an improvement in the solution of AREs will lead to substantial improvement in solving (1.2).

**1.2. Newton-ADI iteration.** Observing that the (Frechét) derivative of $\Re_h$ at $\mathbf{P}$ is given by the Lyapunov operator

$$\Re_h'|_P : \mathbf{X} \mapsto \left(\mathbf{A} - \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}\right)^T\mathbf{X} + \mathbf{X}\left(\mathbf{A} - \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}\right),$$

Newton's method for AREs can be written as

$$\begin{aligned} \mathbf{N}_\ell &:= \left(\Re_h'|_{\mathbf{P}_\ell}\right)^{-1}\Re_h(\mathbf{P}_\ell), \\ \mathbf{X}_{\ell+1} &:= \mathbf{X}_\ell + \mathbf{N}_\ell. \end{aligned}$$

Then, one step of the Newton iteration for a given starting matrix can be implemented as shown in Algorithm 1.1.

---

ALGORITHM 1.1
*Newton's method for AREs*

**Require:** $\mathbf{P}_l$, such that $\mathbf{A}_l$ is stable
1: $\mathbf{A}_\ell \leftarrow \mathbf{A} - \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_\ell$
2: Solve the Lyapunov equation $\mathbf{A}_\ell^T\mathbf{N}_\ell + \mathbf{N}_\ell\mathbf{A}_\ell = -\Re_h(\mathbf{P}_\ell)$
3: $\mathbf{P}_{\ell+1} \leftarrow \mathbf{P}_\ell + \mathbf{N}_\ell$

---

Newton's iteration for AREs can be reformulated as a one-step iteration rewriting it such that the next iteration is computed directly from the Lyapunov equation in Step 2 of Algorithm 1.1,

$$\begin{aligned} (\mathbf{A} - \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_\ell)^T\mathbf{P}_{\ell+1} + \mathbf{P}_{\ell+1}(\mathbf{A} - \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_\ell) \\ = -\mathbf{C}^T\tilde{\mathbf{Q}}\mathbf{C} - \mathbf{P}_\ell\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_\ell =: -\mathbf{W}_\ell\mathbf{W}_\ell^T. \end{aligned} \tag{1.3}$$

So we have to solve a Lyapunov equation

$$\mathbf{F}^T\mathbf{X} + \mathbf{X}\mathbf{F} = -\mathbf{W}\mathbf{W}^T \tag{1.4}$$

with stable $\mathbf{F}$ in each Newton step. Equation (1.4) will be solved using the ADI iteration, which can be written as [36]

$$\begin{aligned} (\mathbf{F}^T + p_j\mathbf{I})\mathbf{X}_{j-1/2} &= -\mathbf{W}\mathbf{W}^T - \mathbf{X}_{j-1}(\mathbf{F} - p_j\mathbf{I}), \\ (\mathbf{F}^T + p_j\mathbf{I})\mathbf{X}_j^T &= -\mathbf{W}\mathbf{W}^T - \mathbf{X}_{j-1/2}(\mathbf{F} - p_j\mathbf{I}). \end{aligned} \tag{1.5}$$

Note that from (1.3) we see that $\mathbf{F}$ in (1.4) and (1.5) can be represented as the sum of a sparse matrix ($\mathbf{A}$) and a low-rank perturbation ($-\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_\ell$). This allows us to exploit the Sherman-Morrison-Woodbury formula (see, e.g., [17]) in the solver for (1.5). Therefore, we consider the problem sparse if $\mathbf{F}$ has this specific structure. Note that for problems from a finite element discretization, $\mathbf{A} = -\mathbf{M}^{-1}\mathbf{S}$ with sparse mass matrix $\mathbf{M}$ and sparse stiffness matrix $\mathbf{S}$. Despite the fact that, in this case, $\mathbf{A}$ will in general be dense, the problem can still be considered sparse as all linear algebra operations required involve only sparse matrix multiplication, and sparse system solves as $\mathbf{A}$ never needs to be formed to implement the ADI method; see [5].

If the shift parameters $p_j$ are chosen appropriately, then $\lim_{j\to\infty}\mathbf{X}_j = \mathbf{X}$ with a superlinear convergence rate. In order to make this iteration work for large-scale problems, we apply the low-rank Newton-ADI method presented in [7, 28] (based upon the iterative technique by Wachspress [36]) to the AREs.

Practical experience shows that it is crucial to have good shift parameters to get fast convergence in the ADI process. If the parameters are real[1], the error in iterate $j$ is given by $\mathbf{e}_j = \mathbf{R}_j\mathbf{e}_{j-1}$, where

$$\mathbf{R}_j := (\mathbf{F} + p_j\mathbf{I})^{-1}(\mathbf{F}^T - p_j\mathbf{I})(\mathbf{F}^T + p_j\mathbf{I})^{-1}(\mathbf{F} - p_j\mathbf{I})$$

and $\mathbf{e}_0 := \mathbf{X}_0 - \mathbf{X}$. Thus, the error after $J$ iterations satisfies $\mathbf{e}_J = \mathbf{G}_J\mathbf{e}_0$, where $\mathbf{G}_J := \prod_{j=1}^J \mathbf{R}_j$. Unrolling matrices into vectors in (1.4), one observes that with the Kronecker products $\mathbf{I}\otimes\mathbf{F}^T$ and $\mathbf{F}^T\otimes\mathbf{I}$ also the factors in $\mathbf{R}_j$ commute and $\|\mathbf{G}_J\|_2 = \rho(\mathbf{G}_J)$. Therefore,

$$\|\mathbf{e}_J\|_2 \le \rho(\mathbf{G}_J)\,\|\mathbf{e}_0\|_2, \quad \rho(\mathbf{G}_J) = \ell(\mathbf{p})^2, \tag{1.6}$$

where $\mathbf{p} = \{p_1, p_2, \ldots, p_J\}$ and

$$\ell(\mathbf{p}) = \max_{\lambda\in\sigma(\mathbf{F})}\left|\prod_{j=1}^J \frac{(p_j - \lambda)}{(p_j + \lambda)}\right|.$$

By this, the ADI parameters are chosen in order to minimize $\rho(\mathbf{G}_J)$, which leads to the rational min-max problem

$$\min_{\{p_j\in\mathbb{R}:j=1,\ldots,J\}} \ell(\mathbf{p}) \tag{1.7}$$

for the shift parameters $p_j$; see, e.g., [37]. This minimization problem is also known as the rational Zolotarev problem since, in the real case, i.e., $\sigma(\mathbf{F}) \subset \mathbb{R}$, it is equivalent to the third of four approximation problems solved by Zolotarev in the 19th century; see [23]. For a complete historical overview; see [35].

[1] This is the desired case for efficiency reasons and can be assured in many applications to optimal control problems for diffusion-reaction-convection equations.

**2. Review of existing parameter selection methods.** Many procedures for constructing optimal or suboptimal shift parameters have been proposed in the literature [19, 27, 33, 37]. Most of the approaches cover the spectrum of $\mathbf{F}$ by a domain $\Omega \subset \mathbb{C}_-$ and solve (1.7) with respect to $\Omega$ instead of $\sigma(\mathbf{F})$. In general, one must choose among the various approaches to find effective ADI iteration parameters for specific problems. One could even consider sophisticated algorithms like the one proposed by Istace and Thiran [19] in which the authors use numerical techniques for nonlinear optimization problems to determine optimal parameters. However, it is important to make sure that the time spent in computing parameters does not outweigh the convergence improvement derived therefrom.

Wachspress [37] computes the optimum parameters when the spectrum of the matrix $\mathbf{F}$ is real or, in the complex case, if the spectrum of $\mathbf{F}$ can be embedded in an elliptic function region (a precise definition will be given in Section 2.2), which often occurs in practice. These parameters may be chosen real, even if the spectrum is complex, as long as the imaginary parts of the eigenvalues are *small* compared to their real parts; see [25, 37] for details. The method applied by Wachspress in the complex case is similar to the technique of embedding the spectrum into an ellipse and then using Chebyshev polynomials. In case that the spectrum is not well represented by the elliptic functions region, a more general development by Starke [33] describes how generalized Leja points yield asymptotically optimal iteration parameters. Finally, an inexpensive heuristic procedure for determining ADI shift parameters, which often works well in practice, was proposed by Penzl [27]. We summarize next these approaches.

**2.1. Leja points.** Gonchar [18] characterizes the general min-max problem and shows how asymptotically optimal parameters can be obtained with generalized Leja or Fejér points. Starke [32] applies this theory to the ADI min-max problem (1.7). The generalized Leja points are defined as follows. Given $\mathcal{E}, \mathcal{F} \subset \mathbb{C}$ containing the spectra of $\mathbf{I} \otimes \mathbf{F}^T$ and $\mathbf{F}^T \otimes \mathbf{I}$, as well as arbitrary points $\varphi_1, \ldots, \varphi_j \in \mathcal{E}$ and $\psi_1, \ldots, \psi_j \in \mathcal{F}$, then for $j = 1, 2, \ldots,$ the new points $\varphi_{j+1} \in \mathcal{E}$ and $\psi_{j+1} \in \mathcal{F}$ are chosen recursively in such a way that, with

$$r_j(z) = \prod_{i=1}^{j} \frac{z - \varphi_j}{z - \psi_j},$$

the two conditions $\max_{x \in \mathcal{E}} |r_j(z)| = |r_j(\varphi_{j+1})|$ and $\max_{x \in \mathcal{F}} |r_j(z)| = |r_j(\psi_{j+1})|$ are fulfilled. Bagby [2] shows that the rational functions $r_j$ obtained by this procedure are asymptotically minimal for the rational Zolotarev problem.

The generalized Leja points can be determined numerically for a large class of boundary curves $\partial\mathcal{E}$ and $\partial\mathcal{F}$. On the other hand, Wachspress [37] notes that in many situations when the optimal parameter choice leads to relatively few iterations to attain the prescribed accuracy based on (1.6), choosing Leja points instead of the Wachspress parameters may lead to poor convergence. Moreover, the computation of Leja points is quite time-consuming when their number becomes large.

**2.2. Optimal parameters.** In this section, we summarize the parameter selection procedure given in [37].

Define the spectral bounds $a$, $b$ and a sector angle $\alpha$ for the matrix $\mathbf{F}$ as

$$a = \min_i \left( \operatorname{Re}\{\lambda_i\} \right), \quad b = \max_i \left( \operatorname{Re}\{\lambda_i\} \right), \quad \alpha = \tan^{-1} \max_i \left| \frac{\operatorname{Im}\{\lambda_i\}}{\operatorname{Re}\{\lambda_i\}} \right|, \tag{2.1}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $-\mathbf{F}$. It is assumed that the spectrum of $-\mathbf{F}$ lies inside the elliptic functions region determined by $a$, $b$, $\alpha$, as defined in [37]: let

$$\cos^2\beta = \frac{2}{1+\frac{1}{2}\left(\frac{a}{b}+\frac{b}{a}\right)}, \quad m = \frac{2\cos^2\alpha}{\cos^2\beta} - 1. \tag{2.2}$$

If $\alpha < \beta$, then $m \geq 1$ and the parameters are real. We define

$$k_1 = \frac{1}{m+\sqrt{m^2-1}}, \quad k = \sqrt{1-{k_1}^2}, \tag{2.3}$$

and the elliptic integrals $K$ and $v$ via

$$L[\psi, k] = \int_0^{\psi} \frac{dx}{\sqrt{1-k^2\sin^2 x}}, \tag{2.4}$$

as

$$K = K(k) = L\left[\frac{\pi}{2}, k\right], \quad v = L\left[\sin^{-1}\sqrt{\frac{a}{bk_1}}, k_1\right], \tag{2.5}$$

where $L$ is the incomplete elliptic integral of the first kind, $k$ is its modulus, and $\psi$ is its amplitude.

With this, we can give a precise definition of the region containing $\sigma(-\mathbf{F})$. This region is tangent to the ray $\sqrt{k_1}e^{\imath Kr}$ from the origin at angle $\alpha$, yielding $r = \alpha/K$ [37, Section 4.3].

DEFINITION 2.1 [37]. *The elliptic function region corresponding to* $\mathbf{F}$ *is defined as*

$$D(r) = \left\{p = \mathrm{dn}(zK, k) \mid z = x + \imath y,\ 0 \leq x \leq 1 \text{ and } |y| \leq r\right\} \subset \mathbb{C},$$

*where $r$ is defined above, $k, K = K(k)$ are as in (2.3) and (2.5), respectively, and* $\mathrm{dn}$ *is the well-known Jacobi elliptic function [1, Chapter 16].*

The number of the ADI iterations required to achieve $\ell(\mathbf{p})^2 \leq \epsilon$ is $J = \lceil \frac{K}{2v\pi} \log \frac{4}{\epsilon} \rceil$, and the ADI parameters are given by

$$p_j = -\sqrt{\frac{ab}{k_1}}\,\mathrm{dn}\left[\frac{(2j-1)K}{2J}, k\right], \quad j = 1, 2, \ldots, J, \tag{2.6}$$

with the Jacobi elliptic function $\mathrm{dn}(u, k)$ as in Definition 2.1.

If $m < 1$, the parameters are complex. We define the dual elliptic spectrum,

$$a' = \tan\left(\frac{\pi}{4} - \frac{\alpha}{2}\right), \quad b' = \frac{1}{a'}, \quad \alpha' = \beta.$$

Substituting $a'$ in (2.2), we find that

$$\beta' = \alpha, \quad m' = \frac{2\cos^2\beta}{\cos^2\alpha} - 1.$$

By construction, $m'$ must now be greater than 1. Therefore, we may compute the optimum real parameters $p'_j$ for the dual problem. The corresponding complex parameters for the actual spectrum can then be computed from

$$\cos\alpha_j = \frac{2}{p'_j + \frac{1}{p'_j}}, \tag{2.7}$$

yielding

$$p_{2j-1} = \sqrt{ab}\,e^{\imath\alpha_j}, \quad p_{2j} = \sqrt{ab}\,e^{-\imath\alpha_j}, \quad j = 1, 2, \ldots, \left\lceil\frac{1+J}{2}\right\rceil. \tag{2.8}$$

**ETNA**
**Kent State University**
**http://etna.math.kent.edu**



**2.3. Heuristic parameters.** The bounds needed to compute optimal parameters are too expensive to be computed exactly in case of large-scale systems because they need the knowledge of the whole spectrum of $\mathbf{F}$. In fact, this computation would be more expensive than the application of the ADI method itself.

An alternative was proposed by Penzl in [27]. He presents a heuristic procedure which determines suboptimal parameters based on the idea of replacing $\sigma(\mathbf{F})$ by an approximation $\mathcal{R}$ of the spectrum in (1.7). Specifically, $\sigma(\mathbf{F})$ is approximated using the Ritz values computed by the Arnoldi process (or any other large-scale eigensolver). Due to the fact that the Ritz values tend to be located near the largest magnitude eigenvalues, the inverses of the Ritz values related to $\mathbf{F}^{-1}$ are also computed to get an approximation of the smallest magnitude eigenvalues of $\mathbf{F}$ yielding a better approximation of $\sigma(\mathbf{F})$. The suboptimal parameters $\mathcal{P} = \{p_1, \ldots, p_k\}$ are chosen among the elements of this approximation because the function

$$s_{\mathcal{P}}(t) = \frac{\left|(t-p_1)\cdots(t-p_k)\right|}{\left|(t+p_1)\cdots(t+p_k)\right|}$$

becomes small over $\sigma(\mathbf{F})$ if there is one of the shifts $p_j$ in the neighborhood of each eigenvalue. The procedure determines the parameters as follows. First, the element $p_j \in \mathcal{R}$ which minimizes the function $s_{\{p_j\}}$ over $\mathcal{R}$ is chosen. The set $\mathcal{P}$ is initialized by either $\{p_j\}$ or the pair of complex conjugates $\{p_j, \bar{p}_j\}$. Now $\mathcal{P}$ is successively enlarged by the elements or pairs of elements of $\mathcal{R}$, for which the maximum of the current $s_{\mathcal{P}}$ is attained. Doing this, the elements of $\mathcal{R}$ giving the largest contributions to the value of $s_{\mathcal{P}}$ are successively canceled out. Therefore, the resulting $s_{\mathcal{P}}$ is nonzero only in the elements of $\mathcal{R}$ where its value is comparably small anyway. In this sense, (1.7) is solved heuristically.

**2.4. Discussion.** In the considered applications from PDE constraint control, we are mainly concerned with problems where the diffusive part dominates the convection terms. Thus, the resulting operator has a spectrum with only moderately large imaginary parts compared to the real parts. Only for this kind of problems, Newton-ADI appears to be a suitable method as for convection-dominated problems, the low-rank property of the solution which makes the approach feasible for large-scale problems will in general not hold. Hence, we will assume that the spectrum of $\mathbf{A}$ is contained in a sector with moderate opening angle in the left half-plane. Note that from numerical experiments it seems that this property is inherited by the $\mathbf{A}_\ell$ in the Newton iteration despite the fact that they will in general be nonsymmetric even if $\mathbf{A}$ is symmetric negative definite. In this situation, the Wachspress approach should always be applicable and lead to real shift parameters in many cases. In problems, where the reactive and convective terms are absent, i.e., we are considering a plain heat equation and therefore the spectrum is part of the real axis, the Wachspress parameters are proven to be optimal. The heuristics proposed by Penzl then require considerably more expensive computations, and Starke notes in [32] that the generalized Leja approach will not be competitive here since it is only asymptotically optimal. For the complex spectra case, common strategies to determine the generalized Leja points generalize the idea of enclosing the spectrum by a polygonal domain, where the starting roots are placed in the corners. So one needs quite exact information about the shape of the spectrum there. In practice, this computation will be too expensive unless one knows some a priori information about the spectrum.

**3. Suboptimal parameter computation.** In this section, we discuss our new contribution to the parameter selection problem. The idea is to avoid the problems of the methods reviewed in the previous section and on the other hand combine their advantages.

Since the important information that we need to know for the Wachspress approach is the outer shape of the spectrum of the matrix $\mathbf{F}$, we will describe an algorithm approximating the outer spectrum. With this approximation the input parameters $a$, $b$, $\alpha$ for the Wachspress method are determined and the optimal parameters for the approximated spectrum are computed. Obviously, these parameters have to be considered suboptimal for the original problem, but if we can approximate the outer spectrum using a few Ritz values only, we end up with a method giving nearly optimal parameters at a drastically reduced computational cost. Algorithm 3.1 is based on these ideas.

ALGORITHM 3.1
*Approximate optimal ADI parameter computation*

**Require:** $\mathbf{F}$ Hurwitz stable

**if** $\sigma(\mathbf{F}) \subset \mathbb{R}$ **then**
  Compute the spectral bounds and set $a = \min \sigma(-\mathbf{F})$ and $b = \max \sigma(-\mathbf{F})$,
  $k_1 = \frac{a}{b}$ , $k = \sqrt{1 - k_1^2}$,
  $K = L(\frac{\pi}{2}, k)$ , $v = L(\frac{\pi}{2}, k_1)$.
  Compute $J$ and the parameters according to (2.6).
**else**
  Compute $\tilde{a} = \min \operatorname{Re}(\sigma(-\mathbf{F}))$, $\tilde{b} = \max \operatorname{Re}(\sigma(-\mathbf{F}))$ and $c = \frac{\tilde{a}+\tilde{b}}{2}$.
  Compute $l$ largest magnitude eigenvalues $\hat{\lambda}_i$ for the shifted matrix $-\mathbf{F} + cI$ by an Arnoldi process or alike.
  Shift these eigenvalues back, i.e., set $\tilde{\lambda}_i = \hat{\lambda}_i + c$.
  Compute $a$, $b$, and $\alpha$ from the $\tilde{\lambda}_i$ as in (2.1).
  **if** $m \geq 1$ in (2.2) **then**
    Compute the parameters by (2.2)–(2.6).
  **else** {The ADI parameters are complex in this case}
    Compute the dual variables.
    Compute the parameters for the dual variables by (2.2)–(2.6).
    Use (2.7) and (2.8) to get the complex shifts.
  **end if**
**end if**

In the following, we discuss the main computational steps in Algorithm 3.1.

*Real spectra.* In the case where the spectrum is real, we can simply compute the upper and lower bounds of the spectrum by the Arnoldi (or, if $\mathbf{F} = \mathbf{F}^T$, the Lanczos) process and enter the Wachspress computation with these values for $a$ and $b$, and set $\alpha = 0$, i.e., we only have to compute two complete elliptic integrals by an arithmetic geometric mean process. This is very cheap since it is a quadratically converging scalar computation (see below). Note that particularly in the symmetric case leading naturally to a real spectrum, applying the Lanczos process to $\mathbf{F}$ with its simultaneous convergence to the eigenvalues of the smallest and largest magnitude [17, Section 9.1], no eigenvalue computation using $\mathbf{F}^{-1}$ is necessary. In any case, as the accurate computation of $a$, $b$ usually requires only few Arnoldi or Lanczos steps, the parameter calculation will usually be significantly more efficient than Penzl's heuristic which requires many Ritz values of $\mathbf{F}$ and $\mathbf{F}^{-1}$.

*Complex spectra.* For complex spectra, we introduce an additional shifting step to be able to apply the Arnoldi process more efficiently. Since we are dealing with stable systems[2], we compute the largest and smallest magnitude eigenvalues and use the arithmetic mean of their real parts as a horizontal shift such that the spectrum is centered about the origin. Now

[2]Note that the Newton-ADI-iteration assumes that we know a stabilizing initial feedback, or the system is stable itself.

ETNA
Kent State University
http://etna.math.kent.edu



Arnoldi's method is applied to the shifted spectrum to compute a number of largest magnitude eigenvalues. These will now automatically include the smallest magnitude eigenvalues of the original system after shifting back. So we can avoid extensive application of the Arnoldi method to the inverse of $\mathbf{F}$. We only need it to get a rough approximation of the smallest magnitude eigenvalue to determine $\tilde{a}$ and $\tilde{b}$ for the shifting step.

The number of eigenvalues we compute can be seen as a tuning parameter here. The more eigenvalues we compute, the better the approximation of the shape of the spectrum is and the closer we get to the exact $a$, $b$, and $\alpha$, but obviously the computation becomes more and more expensive. Especially, the dimension of the Krylov subspaces is increasing with the number of parameters requested and with it the memory consumption in the Arnoldi process. But in cases where the spectrum is filling a rectangle or an egg-like shape, a few eigenvalues are sufficient here; compare Section 4.1.

A drawback of this method can be that in case of small (compared to the real parts) imaginary parts of the eigenvalues, one may need a large number of eigenvalue approximations to find the ones with large imaginary parts, which are crucial to determine $\alpha$ accurately. On the other hand, in that case the spectrum is *almost* real, and therefore it will be sufficient to compute the parameters for the approximate real spectrum in most applications.

*Computation of the elliptic integrals.* The new as well as the Wachspress parameter algorithms require the computation of certain elliptic integrals presented in (2.4). These are equivalent to the integral

$$L[\psi, k] = \int_0^{\psi} \frac{dx}{\sqrt{(1-k^2)\sin^2 x + \cos^2 x}} = \int_0^{\psi} \frac{dx}{\sqrt{(k_1^2)\sin^2 x + \cos^2 x}}. \tag{3.1}$$

In the case of real spectra, $\psi = \frac{\pi}{2}$ and $L[\frac{\pi}{2}, k]$ is a complete elliptic integral of the form

$$I(a,b) = \int_0^{\frac{\pi}{2}} \frac{dx}{\sqrt{a^2\cos^2 x + b^2 \sin^2 x}}$$

and $I(a,b) = \pi/2M(a,b)$, where $M(a,b)$ is the arithmetic geometric mean of $a$ and $b$. The proof for the quadratic convergence of the arithmetic geometric mean process is given in many textbooks; see, e.g., [34].

For incomplete elliptic integrals, i.e., the case $\psi < \pi/2$, an additional Landen's transformation has to be performed. Here, first the arithmetic geometric mean is computed as above, then a descending Landen's transformation is applied (see [1, Chapter 17]), which comes in at the cost of a number of scalar tangent computations equal to the number of iteration steps taken in the arithmetic geometric mean process above.

The value of the elliptic function $\mathrm{dn}$ from equation (2.6) is also computed by an arithmetic geometric mean process; see [1, Chapter 16].

To summarize the advantages of the proposed method, we can say the following.

(i) We compute real shift parameters even in many cases of complex spectra, where the heuristic method would compute complex ones. This results in a significantly cheaper ADI iteration considering memory consumption and computational effort, since complex computations are avoided.

(ii) We have to compute less Ritz values compared to the heuristic method, reducing the time spent in the computational overhead for the acceleration of the ADI method. In particular, the number of applications of $\mathbf{F}^{-1}$ in the eigenvalue computations is drastically reduced or even avoided completely.

(iii) We compute a good approximation of the Wachspress parameters at a drastically reduced computational cost compared to their exact computation.

**4. Numerical results.** For the numerical tests, we used the LyaPack[3] software package [28]. A test program similar to `demo_r1` from the LyaPack examples is used for the computation, where the ADI parameter selection is switched between the methods described in the previous sections. We are here concentrating on the case where the ADI shift parameters can be chosen real.

**4.1. FDM semidiscretized diffusion-convection-reaction equation.** Here, we consider the finite difference semidiscretized partial differential equation

$$\frac{\partial x}{\partial t} - \Delta x - \begin{bmatrix} 20 \\ 0 \end{bmatrix} \cdot \nabla x + 180x = f(\xi)u(t), \tag{4.1}$$

where $x$ is a function of time $t$, vertical position $\xi_1$ and horizontal position $\xi_2$ on the square with opposite corners $(0,0)$ and $(1,1)$. The example is taken from the SLICOT collection of benchmark examples for model reduction of linear time-invariant dynamical systems; see [11, Section 2.7] for details. It is given in semidiscretized state space model representation:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \quad \mathbf{y} = \mathbf{C}\mathbf{x}. \tag{4.2}$$

The matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ for this system can be found on the SLICOT web site[4].

Figures 4.1(a) and 4.1(b) show the spectrum and sparsity pattern of the system matrix $\mathbf{A}$. The iteration history, i.e., the numbers of ADI steps in each step of Newton's method, is plotted in Figure 4.1(c). There, we can see that in fact the semioptimal parameters work exactly like the optimal ones by the Wachspress approach. This is what we would expect since the rectangular spectrum is an optimal case for our idea, because the parameters $a$, $b$, and $\alpha$ are exactly (up to the accuracy of Arnoldi's method) computed here. Note especially that for the heuristic parameters even more outer Newton iterations than for our parameters are required.

**4.2. FDM semidiscretized heat equation.** In this example, we tested the parameters for the finite difference semidiscretized heat equation on the unit square $(0,1) \times (0,1)$:

$$\frac{\partial x}{\partial t} - \Delta x = f(\xi)u(t). \tag{4.3}$$

The data is generated by the routines `fdm_2d_matrix` and `fdm_2d_vector` from the examples of the LyaPack package. Details on the generation of test problems can be found in the documentation of these routines (comments and MATLAB help). Since the differential operator is symmetric here, the matrix $A$ is symmetric and its spectrum is real in this case. Hence, $\alpha = 0$, and for the Wachspress parameters only the largest and smallest magnitude eigenvalues have to be found to determine $a$ and $b$. That means we only need to compute two Ritz values by the Arnoldi process (which here is in fact a Lanczos process because of symmetry) compared to about 30 (which seems to be an adequate number of shifts) for the heuristic approach. We used a test example with 400 unknowns here to still be able to compute the complete spectrum using `eig` for comparison.

In Figure 4.2, we plotted the sparsity pattern of $A$ and the iteration history for the solution of the corresponding ARE. We can see (Figure 4.2(b)) that iteration numbers only differ very slightly. Hence, we can choose quite independently which parameters to use. Since the Wachspress approach needs a good approximation of the smallest magnitude eigenvalue, it might be a good idea to choose the heuristic parameters here (even though they are much more expensive to compute) if the smallest magnitude eigenvalue is known to be close to the origin (e.g., in case of finite element discretizations with fine meshes).

[3] http://www.netlib.org/lyapack/ or http://www.tu-chemnitz.de/sfb393/lyapack/.
[4] http://www.slicot.org/index.php?site=benchmodred.

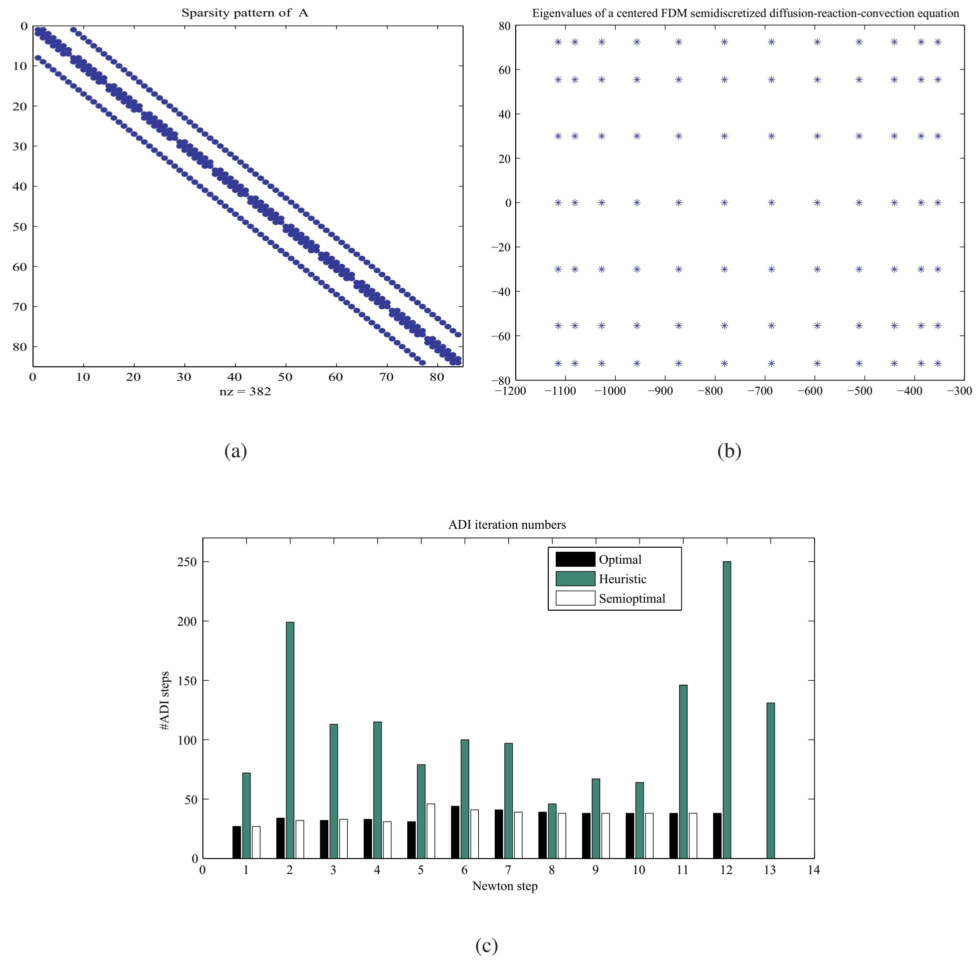



FIG. 4.1. *(a) Sparsity pattern of the FDM semidiscretized operator for equation* (4.1) *and (b) its spectrum. (c) Iteration history for the Newton ADI method applied to* (4.1).

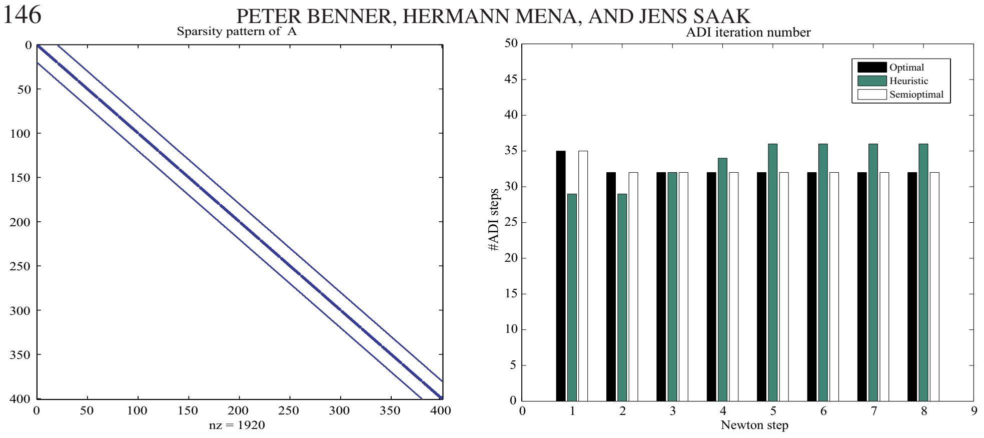
**4.3. FEM semidiscretized convection-diffusion equation.** The last example is a system appearing in the optimal heating/cooling of a fluid flow in a tube. An application is the temperature regulation of certain reagent inflows in chemical reactors. The model equations are

$$\begin{aligned}
\frac{\partial x}{\partial t} - \kappa\Delta x + v \cdot \nabla x &= 0 && \text{in } \Omega,\\
x &= x_0 && \text{on } \Gamma_{\text{in}},\\
\frac{\partial x}{\partial n} &= \sigma(u - x) && \text{on } \Gamma_{\text{heat1}} \cup \Gamma_{\text{heat2}},\\
\frac{\partial x}{\partial n} &= 0 && \text{on } \Gamma_{\text{out}}.
\end{aligned} \tag{4.4}$$

Here, $\Omega$ is the rectangular domain shown in Figure 4.3(a). The inflow $\Gamma_{\text{in}}$ is at the left part of the boundary and the outflow $\Gamma_{\text{out}}$ the right one. The control is applied via the upper and lower

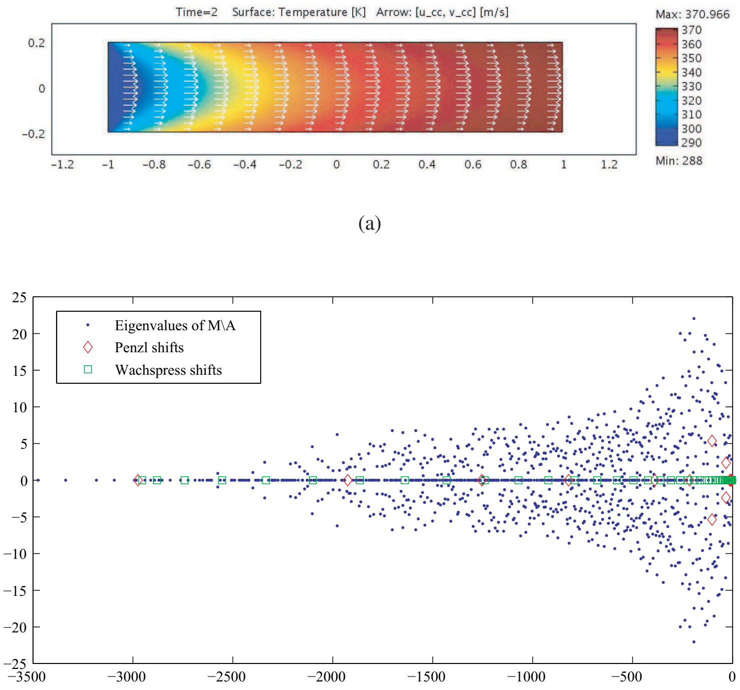
ETNA
Kent State University
http://etna.math.kent.edu






(a) (b)

FIG. 4.2. *(a) Sparsity pattern of the FDM semidiscretized operator for equation (4.3), and (b) iteration history for the Newton-ADI method.*




(a)

(b)

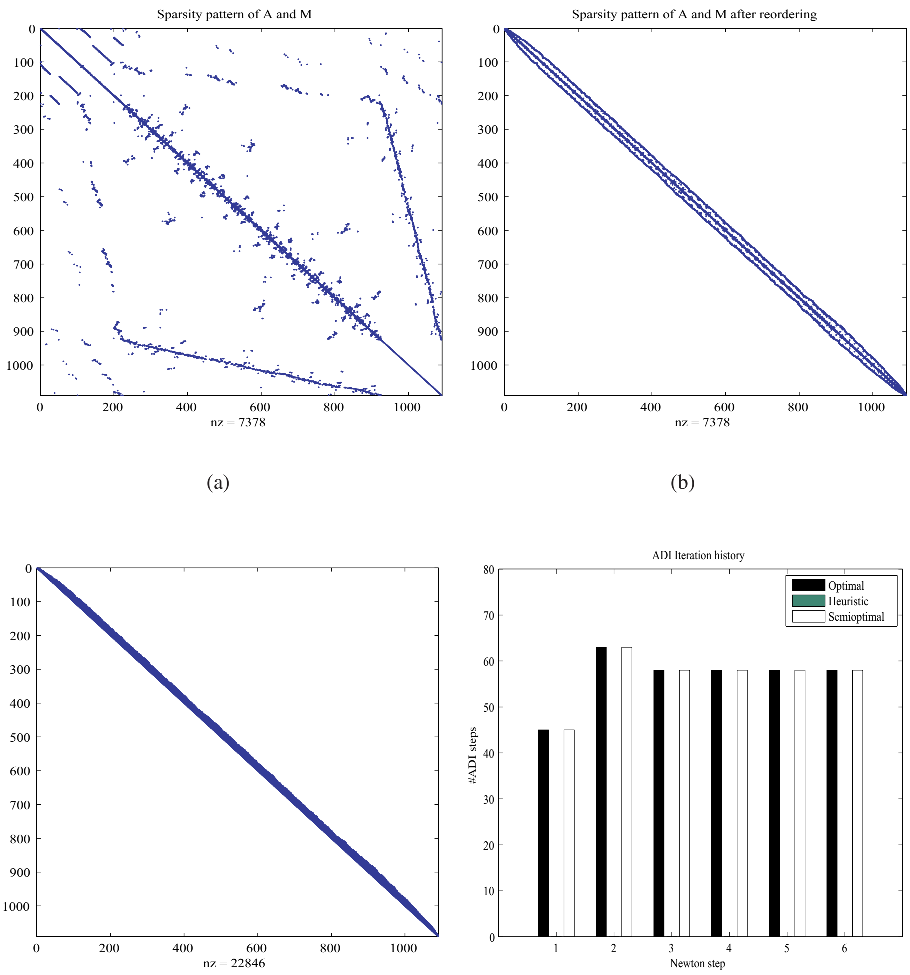
FIG. 4.3. *(a) A 2d cross-section of the liquid flow in a round tube. (b) Eigenvalue and shift parameter distributions.*

boundaries. We can restrict ourselves to this 2d domain assuming rotational symmetry, i.e., nonturbulent diffusion-dominated flows. The test matrices have been created using the COMSOL Multiphysics software and $\kappa = 0.06$, resulting in the eigenvalue and shift distributions shown in Figure 4.3(b).

Since a finite element discretization in space has been applied here, the semidiscrete

ETNA
Kent State University
http://etna.math.kent.edu






(a) (b) (c) (d)

FIG. 4.4. *(a) Sparsity pattern of $A$ and $M$ in* (4.5)*, (b) sparsity pattern of $A$ and $M$ in* (4.5) *after reordering for bandwidth reduction, (c) sparsity pattern of the Cholesky factor of reordered $M$, and (d) iteration history for the Newton-ADI method.*

model is of the form

$$\begin{aligned} \mathbf{M}\dot{\mathbf{x}} &= \tilde{\mathbf{A}}\mathbf{x} + \tilde{\mathbf{B}}\mathbf{u}, \\ \mathbf{y} &= \tilde{\mathbf{C}}\mathbf{x}. \end{aligned} \tag{4.5}$$

This is transformed into a standard system (4.2) using the sparse Cholesky decomposition $\mathbf{M} = \mathbf{M}_L\mathbf{M}_L^T$ (Note that $\mathbf{M}$ is symmetric positive definite.). Sparse reverse Cuthill-McKee ordering is used to reduce the fill in in the Cholesky factors; see Figure 4.4(a)-(c) for sparsity patterns and nonzero counts. Then defining $\tilde{\mathbf{x}} := \mathbf{M}_L^T\mathbf{x}$, $\mathbf{A} := \mathbf{M}_L^{-1}\tilde{\mathbf{A}}\mathbf{M}_L^{-T}$, $\mathbf{B} := \mathbf{M}_L^{-1}\tilde{\mathbf{B}}$, and $\mathbf{C} := \tilde{\mathbf{C}}\mathbf{M}_L^{-T}$ (without computing any of the inverses explicitly in the code), we end up with a standard system for $\tilde{\mathbf{x}}$ having the same inputs $\mathbf{u}$ as (4.5).

Figure 4.4(d) shows the iteration history for the Newton-ADI method with the suggested parameter choices. Note that the heuristic parameters do not appear in the results bar graphics there. This is due to the fact that the LyaPack software crashed while applying the complex shift computed by the heuristics. Numerical tests only using the real ones of the heuristic parameters lead to very poor convergence in the inner loop, which is generally stopped by the maximum iteration number stopping criterion. Thus, no convergence of the Newton iteration is obtained.

**5. Conclusions.** In this paper, we have reviewed existing methods for determining sets of ADI parameters, and based on this review we suggest a new procedure which combines the best features of two of those. For the real case, the parameters computed by the new method are optimal and in many complex cases their performance is quite satisfactory as one can see in the numerical examples. The computational cost depends only on that of the Arnoldi process for the matrix involved and on the computation of elliptic integrals. Since the latter is a quadratically converging scalar iteration, the Arnoldi process is the dominant computation here, which makes this method suitable for the large-scale systems arising from finite element discretization of PDEs. The main advantages of the new method are that it is cheaper to compute than the existing ones and that it avoids complex computations in the ADI iteration for many cases where the others would result in complex ADI iterations.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, eds., *Pocketbook of Mathematical Functions*. Abridged edition of *Handbook of Mathematical Functions*. Material selected by M. Danos and J. Rafelski, Verlag Harri Deutsch, Frankfurt am Main, 1984.

[2] T. BAGBY, *On interpolation by rational functions*, Duke Math. J., 36 (1969), pp. 95–104.

[3] H. T. BANKS AND K. KUNISCH, *The linear regulator problem for parabolic systems*, SIAM J. Control Optim., 22 (1984), pp. 684–698.

[4] P. BENNER, *Computational methods for linear-quadratic optimization*, Rend. Circ. Mat. Palermo (2) Suppl., 58 (1999), pp. 21–56. Extended version available as *Berichte aus der Technomathematik*, Report 98-04, Universität Bremen, August 1998, http://www.math.uni-bremen.de/zetem/berichte.html.

[5] ———, *Solving large-scale control problems*, IEEE Control Syst. Mag., 24 (2004), pp. 44–59.

[6] P. BENNER, S. GÖRNER, AND J. SAAK, *Numerical solution of optimal control problems for parabolic systems*, in Parallel Algorithms and Cluster Computing, vol. 52 of Lecture Notes in Computational Science and Engineering, Springer, Berlin, 2006, pp. 151–169.

[7] P. BENNER, J.-R. LI, AND T. PENZL, *Numerical solution of large Lyapunov equations, Riccati equations, and linear-quadratic control problems*. To appear in Numer. Linear Algebra Appl.

[8] P. BENNER AND H. MENA, *BDF methods for large-scale differential Riccati equations*, in Proceedings of Mathematical Theory of Network and Systems (MTNS 2004), B. De Moor, B. Motmans, J. Willems, P. Van Dooren, and V. Blondel, eds., 2004.

[9] P. BENNER AND J. SAAK, *Linear-quadratic regulator design for optimal cooling of steel profiles*, Tech. Report SFB393/05-05, Sonderforschungsbereich 393 *Parallele Numerische Simulation für Physik und Kontinuumsmechanik*, TU Chemnitz, Chemnitz, Germany, 2005. http://www.tu-chemnitz.de/sfb393/sfb05pr.html.

[10] A. BENSOUSSAN, G. DA PRATO, M. C. DELFOUR, AND S. K. MITTER, *Representation and Control of Infinite Dimensional Systems*, Systems & Control: Foundations & Applications, Birkhäuser Boston, Massachusetts, 2nd ed., 2007.

[11] Y. CHAHLAOUI AND P. VAN DOOREN, *A collection of benchmark examples for model reduction of linear time invariant dynamical systems*, SLICOT Working Note 2002-2, 2002. http://www.slicot.org.

[12] C. H. CHOI AND A. J. LAUB, *Efficient matrix-valued algorithms for solving stiff Riccati differential equations*, IEEE Trans. Automat. Control, 35 (1990), pp. 770–776.

[13] R. F. CURTAIN AND H. ZWART, *An Introduction to Infinite-Dimensional Linear Systems Theory*, vol. 21 of Texts in Applied Mathematics, Springer-Verlag, New York, 1995.

[14] B. N. DATTA, *Numerical Methods for Linear Control Systems*, Elsevier Academic Press, California, 2004.

[15] L. DIECI, *Numerical integration of the differential Riccati equation and some related issues*, SIAM J. Numer. Anal., 29 (1992), pp. 781–815.

[16] J. S. GIBSON, *The Riccati integral equations for optimal control problems on Hilbert spaces*, SIAM J. Control Optim., 17 (1979), pp. 537–565.

[17] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Maryland, 3rd ed., 1996.

[18] A. A. GONCHAR, *Zolotarev problems connected with rational functions*, Math. USSR Sb., 7 (1969), pp. 623–635.

[19] M.-P. ISTACE AND J.-P. THIRAN, *On the third and fourth Zolotarev problems in the complex plane*, SIAM J. Numer. Anal., 32 (1995), pp. 249–259.

[20] P. LANCASTER AND L. RODMAN, *Algebraic Riccati Equations*, Oxford Science Publications, Oxford University Press, New York, 1995.

[21] I. LASIECKA AND R. TRIGGIANI, *Differential and Algebraic Riccati Equations with Application to Boundary/Point Control Problems: Continuous Theory and Approximation Theory*, vol. 164 of Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, 1991.

[22] ———, *Control theory for partial differential equations: continuous and approximation theories. I*, vol. 74 of Encyclopedia of Mathematics and Its Applications, Cambridge University Press, Cambridge, 2000.

[23] V. I. LEBEDEV, *On a Zolotarev problem in the method of alternating directions*, U.S.S.R. Comput. Math. Math. Phys., 17 (1977), pp. 58–76.

[24] J.-L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, vol. 170 of Die Grundlehren der mathematischen Wissenschaften, Springer-Verlag, New York, 1971. Translated by S. K. Mitter.

[25] A. LU AND E. L. WACHSPRESS, *Solution of Lyapunov equations by alternating direction implicit iteration*, Comput. Math. Appl., 21 (1991), pp. 43–58.

[26] V. L. MEHRMANN, *The Autonomous Linear Quadratic Control Problem. Theory and Numerical Solution*, vol. 163 of Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, 1991.

[27] T. PENZL, *A cyclic low-rank Smith method for large sparse Lyapunov equations*, SIAM J. Sci. Comput., 21 (1999), pp. 1401–1418.

[28] ———, LYAPACK *Users Guide*, Tech. Report SFB393/00-33, Sonderforschungsbereich 393 *Numerische Simulation auf massiv parallelen Rechnern*, TU Chemnitz, Chemnitz, Germany, 2000. http://www.tu-chemnitz.de/sfb393/sfb00pr.html.

[29] P. H. PETKOV, N. D. CHRISTOV, AND M. M. KONSTANTINOV, *Computational Methods for Linear Control Systems*, Prentice-Hall, Hertfordshire, UK, 1991.

[30] V. SIMA, *Algorithms for Linear-Quadratic Optimization*, vol. 200 of Monographs and Textbooks in Pure and Applied Mathematics, Marcel Dekker Inc., New York, 1996.

[31] E. D. SONTAG, *Mathematical Control Theory. Deterministic Finite-Dimensional Systems*, vol. 6 of Texts in Applied Mathematics, Springer-Verlag, New York, 2nd ed., 1998.

[32] G. STARKE, *Rationale Minimierungsprobleme in der komplexen Ebene im Zusammenhang mit der Bestimmung optimaler ADI-Parameter*, PhD thesis, Fakultät für Mathematik, Universität Karlsruhe, December 1989.

[33] ———, *Optimal alternating direction implicit parameters for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1431–1445.

[34] U. STORCH AND H. WIEBE, *Lehrbuch der Mathematik. Band 1: Analysis einer Veränderlichen [Textbook of Mathematics. Vol. 1: Analysis of a Variable]*, Spektrum Akademischer Verlag, Heidelberg, 3rd ed., 2003.

[35] J. TODD, *Applications of transformation theory: a legacy from Zolotarev (1847–1878)*, in Approximation Theory and Spline Functions (St. John's, Nfld., 1983), vol. 136 of NATO Advanced Science Institutes Series C: Mathematical and Physical Sciences, Reidel, Dordrecht, 1984, pp. 207–245.

[36] E. L. WACHSPRESS, *Iterative solution of the Lyapunov matrix equation*, Appl. Math. Lett., 1 (1988), pp. 87–90.

[37] ———, *The ADI model problem*. Preprint, 1995.