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 problems for parabolic partial differential equations (PDEs) on infinite time horizons requires the solution of large scale algebraic Riccati equations (ARE). The Newton-ADI iteration is an efficient numerical method for this task. It includes the solution of a Lyapunov equation by the alternating directions implicit (ADI) algorithm in each iteration step. On finite time intervals the solution of a large scale differential Riccati equation is required. This can be solved by a backward differentiation formula (BDF) method, which needs to solve an ARE in each time 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 complete complex spectrum is crucial for computing the optimal solution, this is infeasible for the large scale systems arising from finite element discretization of PDEs. Therefore several alternatives for computing suboptimal parameters are discussed and compared for numerical examples.

Key words. ARE, DRE, Newton-ADI, shift parameters, Lyapunov equations, 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 [14, 24, 25, 9, 10]. 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 [27, 24, 25]. Gibson [17] and Banks/Kunisch [3] present an approximation technique 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, 37].

In Section 1.1 we briefly summarize the basic results for the LQR control of parabolic PDEs. 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 by the ADI algorithm in every iteration step. Furthermore we introduce the rational minimax problem related to the parameter selection problem there, which is the main topic of this paper. We give a brief summary of Wachspress's results and a heuristic choice of parameters described in [33], as well as a Leja point approach [38, 39] 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

*Fakultät für Mathematik, Technische Universität Chemnitz, D-09107 Chemnitz, benner@mathematik.tu-chemnitz.de

†Departamento de Matemática, Escuela Politécnica Nacional, Quito - Ecuador, hmena@server.epn.edu.ec

‡Fakultät für Mathematik, Technische Universität Chemnitz, D-09107 Chemnitz, jens.saak@mathematik.tu-chemnitz.de

even in case of very large systems. The forth section will show the efficiency of our method compared to the Wachspress parameters for test examples, where the complete spectrum can still be computed numerically and thus Wachspress's method can be used to compute the optimal parameters. We close this article with some conclusions in Section 5.

1.1. Problem background. Consider nonlinear parabolic diffusion-convection and diffusion-reaction systems of the form

$$\frac{\partial \mathbf{x}}{\partial t} + \nabla \cdot (\mathbf{c}(\mathbf{x}) - \mathbf{k}(\nabla \mathbf{x})) + \mathbf{q}(\mathbf{x}) = \mathcal{B}\mathbf{u}(t), \qquad t \in [0, T_f], \tag{1.1}$$

in $\Omega \subset \mathbb{R}^d, d = 1, 2, 3$, with appropriate initial and boundary conditions. The equation can be split into the convective term $\mathbf{c}$, the diffusive part $\mathbf{k}$ and the uncontrolled reaction given by $\mathbf{q}$. The state $\mathbf{x}$ of the system depends on $\xi \in \Omega$ and the time $t \in [0, T_f]$ and is denoted by $\mathbf{x}(\xi, t)$.

Notation. Note that we use **bold** letters for the infinite-dimensional setting and regular letters for the discretized case. We also write $\mathbf{x}(t) \in \mathcal{X}$ in the abstract setting, while $\mathbf{x}(\xi, t)$ is used if concrete problems (PDEs) are considered.

We consider applications where the control $\mathbf{u}(t)$ is assumed to depend only on the time $t \in [0, T_f]$ while the linear operator $\mathcal{B}$ may depend on $\xi \in \Omega$. Let $\hat{J}(\mathbf{x}, \mathbf{u})$ be a given performance index, then the control problem is given as:

$$\min_{\mathbf{u}} \hat{J}(\mathbf{x}, \mathbf{u}) \quad \text{subject to (1.1).} \tag{1.2}$$

If (1.1) is in fact linear, then a variational formulation leads to an abstract Cauchy problem for a linear evolution equation of the form

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \qquad \mathbf{x}(0) = \mathbf{x}_0 \in \mathcal{X}, \tag{1.3}$$

for linear operators

$$\begin{aligned} &\mathbf{A} : \mathrm{dom}(\mathbf{A}) \subset \mathcal{X} \to \mathcal{X}, \\ &\mathbf{B} : \mathcal{U} \to \mathcal{X}, \\ &\mathbf{C} : \mathcal{X} \to \mathcal{Y}, \end{aligned} \tag{1.4}$$

where the state space $\mathcal{X}$, the observation space $\mathcal{Y}$, and the control space $\mathcal{U}$ are assumed to be separable Hilbert spaces. Additionally, $\mathcal{U}$ is assumed to be finite dimensional, i.e. there are only a finite number of independent control inputs to (1.1). Here $\mathbf{C}$ maps the states of the system to its outputs, such that

$$\mathbf{y} = \mathbf{C}\mathbf{x}. \tag{1.5}$$

If (1.1) is nonlinear, model predictive control technics can be applied [5, 21, 22]. There the equation is linearized at certain working points or around reference trajectories and linear problems for equations as in (1.3) have to be solved on subintervals of $[0, T_f]$.

In many applications in engineering the performance index $\hat{J}(\mathbf{x}, \mathbf{u})$ is given in quadratic form. We assume (1.3) to have a unique solution for each input $\mathbf{u}$ so that $\mathbf{x} = \mathbf{x}(\mathbf{u})$. Thus we can write the cost functional as $J(\mathbf{u}) := \hat{J}(\mathbf{x}(\mathbf{u}), \mathbf{u})$. Then

$$J(\mathbf{u}) = \frac{1}{2} \int_0^{T_f} \langle \mathbf{x}, \mathbf{Q}\mathbf{x} \rangle_{\mathcal{X}} + \langle \mathbf{u}, \mathbf{R}\mathbf{u} \rangle_{\mathcal{U}} \, dt + \langle \mathbf{x}_{T_f}, \mathbf{G}\mathbf{x}_{T_f} \rangle_{\mathcal{X}}, \tag{1.6}$$

where $\mathbf{Q}$, $\mathbf{G}$ are selfadjoint operators on the state space $\mathcal{X}$, $\mathbf{R}$ is a selfadjoint operator on the control space $\mathcal{U}$ and $\mathbf{x}_{T_f}$ denotes $\mathbf{x}(.,T_f)$. To guarantee unique solvability of the control problem $\mathbf{R}$ is assumed positive definite. Since often only a few measurements of the state are available as the outputs of the system, the operator $\mathbf{Q} := \mathbf{C}^*\tilde{\mathbf{Q}}\mathbf{C}$ (here and in the following * denotes the Hilbert space adjoint) generally is only positive semidefinite as well as $\mathbf{G}$. In many applications one simply has $\tilde{\mathbf{Q}} = \mathbf{I}$ (e.g., in the examples in Section 4).

If the standard assumptions that

- $\mathbf{A}$ is the infinitesimal generator of a $\mathcal{C}^0$-semigroup $\mathbf{T}(t)$,
- $\mathbf{B}, \mathbf{C}$ are linear bounded operators and
- for every initial value there exists an admissible control $\mathbf{u} \in L^2(0,\infty;\mathcal{U})$

hold then the solution of the abstract LQR problem can be obtained analogously to the finite-dimensional case (see [44, 13, 17]). We then have to consider the operator Riccati equations

$$0 = \Re(\mathbf{X}) := \mathbf{C}^*\mathbf{Q}\mathbf{C} + \mathbf{A}^*\mathbf{X} + \mathbf{X}\mathbf{A} - \mathbf{X}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^*\mathbf{X} \tag{1.7}$$

and

$$\dot{\mathbf{X}} = -\Re(\mathbf{X}) \tag{1.8}$$

depending on whether $T_f < \infty$ (1.8) or not (1.7). If $T_f = \infty$ then $\mathbf{G} = 0$ and the linear operator $\mathbf{X}$ is the solution of (1.7), i.e. $\mathbf{X} : \operatorname{dom}\mathbf{A} \to \operatorname{dom}\mathbf{A}^*$ and $\langle \hat{\mathbf{x}}, \Re(\mathbf{X})\mathbf{x}\rangle = 0$ for all $\mathbf{x}, \hat{\mathbf{x}} \in \operatorname{dom}(\mathbf{A})$. The optimal control is then given as the *feedback control*

$$\mathbf{u}_*(t) = -\mathbf{R}^{-1}\mathbf{B}^*\mathbf{X}_\infty\mathbf{x}_*(t), \tag{1.9}$$

which has the form of a regulator or closed-loop control. Here, $\mathbf{X}_\infty$ is the minimal nonnegative self-adjoint solution of (1.7), $\mathbf{x}_*(t) = \mathbf{S}(t)\mathbf{x}_0(t)$, and $\mathbf{S}(t)$ is the $\mathcal{C}^0$-semigroup generated by $\mathbf{A} - \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^*\mathbf{X}_\infty$. In problems where $T_f < \infty$, the optimal control is defined similarly to (1.9) but then $\mathbf{X}_\infty$ represents the unique nonnegative solution of the differential Riccati equation (1.8) with initial condition $\mathbf{X}_{T_f} = \mathbf{G}$ and therefore depends on time, i.e., it has to be replaced by $\mathbf{X}_\infty(t)$ in (1.9). Most of the required conditions, particularly the restrictive assumption that $\mathbf{B}$ is bounded, can be weakened [24, 25, 35].

In order to solve the infinite-dimensional LQR problem numerically we use a Galerkin projection of the variational formulation of the PDE (1.1) onto a finite-dimensional space $\mathcal{X}_h$ spanned by a finite set of basis functions (e.g., finite element ansatz functions).

If we now choose the space of test functions as the space generated by finite element (fem) ansatz functions for a finite element semidiscretization in space, then the operators above have matrix representations in the fem basis. So we have to solve the discrete problem

$$\min_{u\in L^2(0,T_f;\mathcal{U})} \frac{1}{2}\int_0^{T_f} \langle x, Qx\rangle_{\mathcal{X}_h} + \langle u, \mathbf{R}u\rangle_{\mathcal{U}}\, dt + \langle x_{T_f}, Gx_{T_f}\rangle_{\mathcal{X}_h}, \tag{1.10}$$

with respect to

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ x(.,0) &= I_h\mathbf{x}_0, \\ y &= Cx. \end{aligned} \tag{1.11}$$

Here I_h is the interpolation operator from the space discretization method (here fem). Approximation results in terms of approximation of the Riccati solution operator $\mathbf{X}$ and the solution semigroup $\mathbf{S}(t)$ for the closed loop system, validating this technique have been considered e.g. in [25, 3, 8, 20, 30, 31]. Note that the control space is considered finite-dimensional and therefore does not change under spatial semi-discretization, i.e., we can directly apply the control computed for the discretized systems (1.11) to the infinite-dimensional system (1.3), although it might be suboptimal there. The estimation of the sub-optimality of that approach will be considered elsewhere.

1.2. Newton-ADI iteration. In this note we will concentrate on the step of solving the large sparse matrix Riccati equations

$$0 = \Re_h(X) = C^T\tilde{Q}C + A^TX + XA - XB\mathbf{R}^{-1}B^TX, \tag{1.12}$$

or

$$\dot{X} = -\Re_h(X) = -C^T\tilde{Q}C - A^TX - XA + XB\mathbf{R}^{-1}B^TX, \tag{1.13}$$

respectively. The initial value for the latter initial value problem is $X(T_f) = G$. Such initial value problems can efficiently be solved by BDF methods known from ordinary differential equations [7, 16, 12]. This involves solving algebraic equations of type (1.12) in each time step. 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 [36, 23, 29, 34, 4, 15].

Observing that the (Frechét) derivative of $\Re_h$ at P is given by the Lyapunov operator

$$\Re_h'|_P: \quad Q \mapsto (A_h - B_hR^{-1}B_h^TP)^TQ + Q(A_h - B_hR^{-1}B_h^TP),$$

Newton's method for AREs can be written as

$$\begin{aligned} N_\ell &:= \left(\Re_h'|_{P_\ell}\right)^{-1}\Re_h(P_\ell), \\ X_{\ell+1} &:= X_\ell + N_\ell. \end{aligned}$$

Then one step of the Newton iteration for a given starting matrix can be implemented as follows:

Algorithm 1.1 Newton's method for AREs

Require: P_l, such that A_l is stable

1: $A_\ell \leftarrow A_h - B_hR^{-1}B_h^TP_\ell$
2: Solve the Lyapunov equation $A_\ell^TN_\ell + N_\ell A_\ell = -\Re_h(P_\ell)$
3: $P_{\ell+1} \leftarrow P_\ell + N_\ell$

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

$$\begin{aligned} (A_h - B_hR^{-1}B_h^TP_\ell)^TP_{\ell+1} + P_{\ell+1}(A_h - B_hR^{-1}B_h^TP_\ell) &= \\ -C_h^T\tilde{Q}_hC_h - P_\ell B_hR^{-1}B_h^TP_\ell &=: -W_\ell W_\ell^T. \end{aligned}$$

So we have to solve a Lyapunov equation

$$F^T X + XF = -WW^T \tag{1.14}$$

with stable F in each Newton step. (1.14) will be solved using the *alternating direction implicit*(ADI) iteration, which can be written as [42]

$$\begin{aligned} (F^T + p_j I)Q_{(j-1)/2} &= -WW^T - Q_{j-1}(F - p_j I), \\ (F^T + \overline{p_j} I)Q_j^T &= -WW^T - Q_{(j-1)/2}(F - \overline{p_j} I), \end{aligned} \tag{1.15}$$

where $\overline{p}$ denotes the complex conjugate of $p \in \mathbb{C}$. If the shift parameters p_j are chosen appropriately, then $\lim_{j\to\infty} Q_j = Q$ 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 [6, 33] (based upon the iterative technique by Wachspress [42]) to the AREs.

Practical experience shows that it is crucial to have good shift parameters to get fast convergence in the ADI process. The error in iterate j is given by $e_j = R_j e_{j-1}$, where

$$R_j := (F + p_j I)^{-1}(F^T - p_j I)(F^T + p_j I)^{-1}(F - p_j I).$$

Thus the error after J iterations satisfies

$$e_J = G_J e_0, \quad G_J := \prod_{j=1}^{J} R_j,$$

due to the fact that G_J is symmetric,

$$||e_J|| \leq \rho(G_J)||e_0||, \quad \rho(G_J) = k(\mathbf{p})^2,$$

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

$$k(\mathbf{p}) = \max_{\lambda\in\sigma(F)} \left| \prod_{j=1}^{J} \frac{(p_j - \lambda)}{(p_j + \lambda)} \right|. \tag{1.16}$$

By this the ADI parameters are chosen in order to minimizes $\rho(G_J)$ which leads to the rational minimax problem

$$\min_{\{p_j\in\mathbb{R}: j=1,\ldots,J\}} k(\mathbf{p}) \tag{1.17}$$

for the shift parameters p_j, see e.g. [43]. This minimization problem is also known as the rational Zolotarev problem since, in the real case, i.e $\sigma(F) \subset \mathbb{R}$, it is equivalent to the third of four approximation problems solved by Zolotarev in the 19th century, see [26]. For a complete historical overview see [41].

2. Review of existing parameter selection methods. Many procedures for constructing optimal or suboptimal shift parameters have been proposed in the literature [19, 32, 39, 43]. Most of the approaches cover the spectrum of F by a domain $\Omega \subset \mathbb{C}_-$ and solve (1.17) with respect to Ω instead of $\sigma(F)$. 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 take care that the time spent in computing parameters does not outweigh the convergence improvement derived therefrom.

Wachspress et al. [43] compute the optimum parameters when the spectrum of the matrix F is real or, in the complex case, if the spectrum of F can be embedded in an elliptic functions region, 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 [28, 43] 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 use Chebyshev polynomials. In case that the spectrum is not well represented by the elliptic functions region a more general development by Starke [39] 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 [32]. We will summarize these approaches here.

2.1. Leja Points. Gonchar [18] characterizes the general minimax problem and shows how asymptotically optimal parameters can be obtained with generalized Leja or Fejér points. Starke [38] applies this theory to the ADI minimax problem (1.17). The generalized Leja points are defined as follows. Given $\varphi \in E$ and $\psi \in F$ arbitrarily, E, F subsets of $\mathbb{C}$, for $j = 1, 2, \ldots$, the new points $\varphi_j \in E$ and $\psi_j \in F$ are chosen recursively in such a way that, with

$$r_j(z) = \prod_{i=1}^{j} \frac{z - \varphi_j}{z - \psi_j} \tag{2.1}$$

the two conditions

$$\begin{aligned} &\max_{x \in E} |r_j(z)| = |r_j(\varphi_{j+1})|, \\ &\max_{x \in F} |r_j(z)| = |r_j(\psi_{j+1})| \end{aligned} \tag{2.2}$$

are fullfilled. Bagby [2] shows that the rational functions r_j obtained by this procedure are asymptotically minimal for the rational Zolotarev problem. Starke considers a general ADI iteration, so for ADI applied to the Lyapunov equation (1.15) the generalized Leja points will be defined as follows:

Given $p_0 \in E$, E is a complex subset such that $\sigma(F) \subset E$, for $j = 1, 2, \ldots$, the new points $p_j \in E$ are chosen recursively in such a way that, with

$$r_j(z) = \prod_{i=1}^{j} \frac{z - p_j}{z + p_j} \tag{2.3}$$

the condition

$$\max_{x \in E} |r_j(z)| = |r_j(p_{j+1})|, \tag{2.4}$$

holds. The generalized Leja points can be determined numerically for a large class of boundary curves ∂E. When relatively few iterations are needed to attain the prescribed accuracy, the Leja points may be poor. Moreover their computation can be quite time consuming when the number of Leja points generated is large, since the computation gets more and more expensive the more prior Leja points are already calculated.

2.2. Optimal parameters. We will briefly summarize the parameter selection procedure given in [43] in this section.

Define the spectral bounds a, b and a sector angle α for the matrix F as

$$a = \min_i(Re\{\lambda_i\}), \quad b = \max_i(Re\{\lambda_i\}), \quad \alpha = \tan^{-1}\max_i\left|\frac{Im\{\lambda_i\}}{Re\{\lambda_i\}}\right|, \tag{2.5}$$

where $\lambda_1, \ldots, \lambda_n$ are eigenvalues of $-F$. It is assumed that the spectrum of $-F$ lies inside the elliptic functions region determined by a, b, α, as defined in [43]. Let

$$\cos^2\beta = \frac{2}{1+\frac{1}{2}\left(\frac{a}{b}+\frac{b}{a}\right)}, \quad m = \frac{2\cos^2\alpha}{\cos^2\beta} - 1. \tag{2.6}$$

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

$$k_1 = \frac{1}{m+\sqrt{m^2-1}}, \quad k = \sqrt{1-{k_1}^2}. \tag{2.7}$$

Define the elliptic integrals K and v via

$$F[\psi, k] = \int_0^{\psi} \frac{dx}{\sqrt{1-k^2\sin^2 x}}, \tag{2.8}$$

as

$$K = K(k) = F\left[\frac{\pi}{2}, k\right], \quad v = F\left[\sin^{-1}\sqrt{\frac{a}{bk_1}}, k_1\right], \tag{2.9}$$

where F is the incomplete elliptic integral of the first kind, k is its modulus and ψ is its amplitude.

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

$$p_j = -\sqrt{\frac{ab}{k_1}}\,\mathrm{dn}\left[\frac{(2j-1)K}{2J}, k\right], \quad j = 1, 2, \ldots, J, \tag{2.10}$$

where $\mathrm{dn}(u,k)$ is the elliptic function (see [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 (2.6), 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:

$$\cos\alpha_j = \frac{2}{p'_j + \frac{1}{p'_j}}, \tag{2.11}$$

for $j = 1, 2, \ldots, \lceil\frac{1+J}{2}\rceil$

$$p_{2j-1} = \sqrt{ab}\exp[\imath\alpha_j], \quad p_{2j} = \sqrt{ab}\exp[-\imath\alpha_j] \tag{2.12}$$

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 F. In fact, this computation would be more expensive than the application of the ADI method itself.

An alternative was proposed by Penzl in [32]. He presents a heuristic procedure which determines suboptimal parameters based on the idea of replacing $\sigma(F)$ by an approximation $\mathcal{R}$ of the spectrum in (1.17). Specifically, $\sigma(F)$ 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 F^{-1} are also computed to get an approximation of the smallest magnitude eigenvalues of F yielding a better approximation of $\sigma(F)$. The suboptimal parameters $\mathcal{P} = \{p_1, \ldots, p_k\}$ are chosen among the elements of this approximation because the function

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

becomes small over $\sigma(F)$ 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 (1.17) is solved heuristicly.

2.4. Discussion. We are searching for a parameter set for the ADI method applied to a control problem, where in the PDE constraint (1.1) the diffusive part is dominating the reaction or convection terms, respectively. Thus the resulting operator has a spectrum with only moderately large imaginary components compared to the real parts. In these problems 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 is more expensive to compute there and Starke notes in [38], 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 would need to be able to compute the eigenvalues with largest imaginary parts already for a simple rectangular enclosure of the spectrum. Since this still doesn't work reliable, we decided to avoid the comparison with that approach in this publication, although it might proof useful in cases where the Wachspress parameters are no longer applicable or one knows some a-priori information on 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 F, we will describe an algorithm approximating the outer spectrum. With this approximation the input parameters a, b and α 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 at a similar cost to that of the heuristic parameter choice we end up with a method giving nearly optimal parameters at a drastically reduced computational cost compared to the optimal parameters.

Algorithm 3.1 approximate optimal ADI parameter computation

Require: F Hurwitz stable
1: **if** $\sigma(F) \subset \mathbb{R}$ **then**
2: Compute the spectral bounds and set $a = \min \sigma(-F)$ and $b = \max \sigma(-F)$,
3: $k_1 = \frac{a}{b}$, $k = \sqrt{1 - k_1^2}$,
4: $K = F(\frac{\pi}{2}, k)$, $v = F(\frac{\pi}{2}, k_1)$.
5: Compute J and the parameters according to (2.10).
6: **else**
7: Compute $\tilde{a} = \min \mathrm{Re}\,(\sigma(-F))$, $\tilde{b} = \max \mathrm{Re}\,(\sigma(-F))$ and $c = \frac{\tilde{a}+\tilde{b}}{2}$.
8: Compute l largest magnitude eigenvalues $\hat{\lambda}_i$ for the shifted matrix $-F + cI$ by an Arnoldi process or alike.
9: Shift these Eigenvalues back, i.e. $\tilde{\lambda}_i = \hat{\lambda}_i + c$.
10: Compute a, b and α from the $\tilde{\lambda}_i$ like in (2.5).
11: **if** $m \geq 1$ in (2.6) **then**
12: Compute the parameters by (2.6)–(2.10).
13: **else** {The ADI parameters are complex in this case}
14: Compute the dual variables.
15: Compute the parameters for the dual variables by (2.6)–(2.10).
16: Use (2.11) and (2.12) to get the complex shifts.
17: **end if**
18: **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 an Arnoldi 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).

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[1] we compute the largest magnitude and smallest magnitude eigenvalues and use the arithmetic mean of their real parts as a horizontal shift, such that the spectrum is centered around the origin. Now 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

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

to the inverse of 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 α, but obviously the computation becomes more and more expensive. Especially the dimension of the Krylov subspaces is rising 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 α 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 (2.8). These are equivalent to the integral

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

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

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

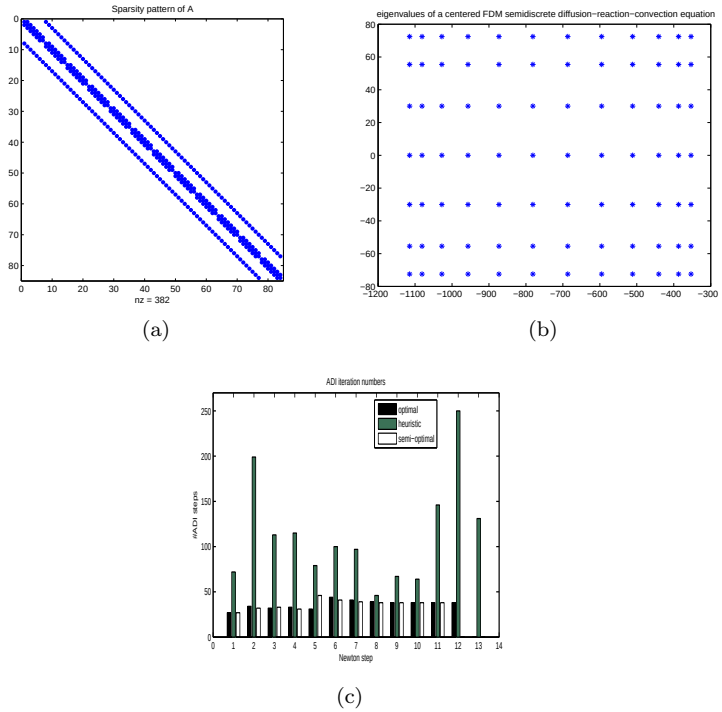
and $I(a,b) = \frac{\pi}{2M(a,b)}$, where $M(a,b)$ is the arithmetic geometric mean of a and b. The proof for the quadratic convergence of the arithmetic geometric mean process is given in many textbooks (e.g.,[40]).

For incomplete elliptic integrals, i.e., the case $\psi < \frac{\pi}{2}$, 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 dn from equation (2.10) is also computed by an arithmetic geometric mean process (see [1, Chapter 16]).

To summarize the advantages of the proposed method we can say:

- We compute real shift parameters even in case of many 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.
- 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.
- We compute a good approximation of the Wachspress parameters at a drastically reduced computational cost compared to their exact computation.



(a) (b)

(c)

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. Numerical results. For the numerical tests we used the LyaPack[2] software package [33]. 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

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

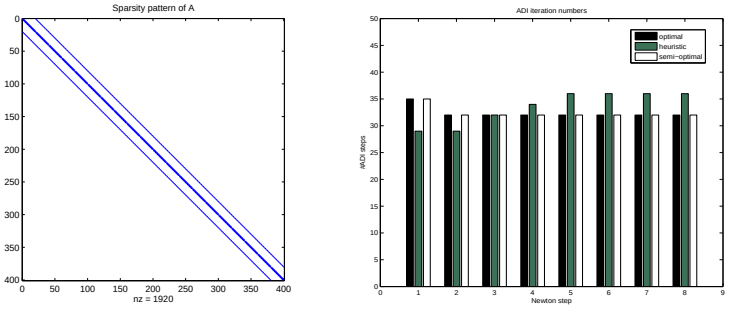
where $\mathbf{x}$ is a function of time t, vertical position ξ_1 and horizontal position ξ_2 on the square with opposite corners $(0,0)$ and $(1,1)$. The example is taken from the SLICOT collection of benchmark examples for model reduction of linear time-invariant dynamical systems (see [11, Section 2.7] for details). It is given in semidiscretized state space model representation:

$$\dot{x} = Ax + Bu, \quad y = Cx. \tag{4.2}$$

The matrices A, B, C for this system can be found on the NICONET web site[3].

[2]available from: `http://www.netlib.org/lyapack/` or `http://www.tu-chemnitz.de/sfb393/lyapack/`

[3]`http://www.icm.tu-bs.de/NICONET/benchmodred.html`



(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*

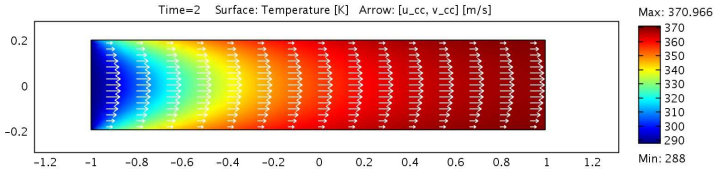
Figure 4.1 (a),(b) show the spectrum and sparsity pattern of the system matrix A. The iteration history, i.e., the numbers of ADI steps in each step of Newton's method are plotted in Figure 4.1 (c). There we can see that in fact the semi-optimal parameters work exactly like the optimal ones by the Wachspress approach. This is what we would expect since the rectangular spectrum is an optimal case for our idea, because the parameters a, b and α are exactly (to the accuracy of Arnoldi's method) met 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 $(0,1) \times (0,1)$.

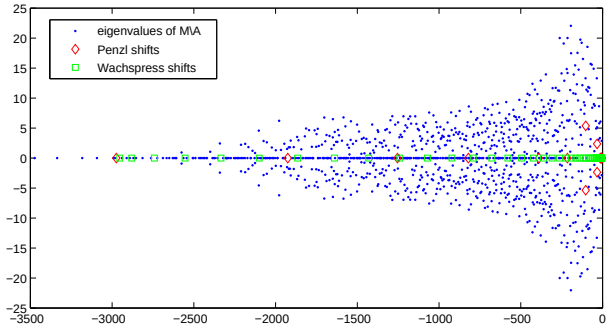
$$\frac{\partial \mathbf{x}}{\partial t} - \Delta \mathbf{x} = \mathbf{f}(\xi)\mathbf{u}(t). \tag{4.3}$$

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 magnitude 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 (which here is in fact a Lanczos process because of symmetry) process 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).



(a)



(b)

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

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:

$$\begin{aligned} \tfrac{\partial \mathbf{x}}{\partial t} - \alpha\Delta\mathbf{x} + \mathbf{v}\cdot\nabla\mathbf{x} &= 0 && \text{in } \Omega \\ \mathbf{x} &= \mathbf{x}_0 && \text{on } \Gamma_{in} \\ \tfrac{\partial \mathbf{x}}{\partial n} &= \sigma(\mathbf{u}-\mathbf{x}) && \text{on } \Gamma_{heat1}\cup\Gamma_{heat2} \\ \tfrac{\partial \mathbf{x}}{\partial n} &= 0 && \text{on } \Gamma_{out}. \end{aligned} \tag{4.4}$$

Here Ω is the rectangular domain shown in Figure 4.3 (a). The inflow Γ_{in} is at the left part of the boundary and the outflow Γ_{out} the right one. The control is applied via the upper and lower boundaries. We can restrict ourselves to this 2d-domain assuming rotational symmetry, i.e., non-turbulent diffusion dominated flows. The test matrices have been created using the COMSOL Multiphysics software and $\alpha = 0.06$, 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 model is of the form

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

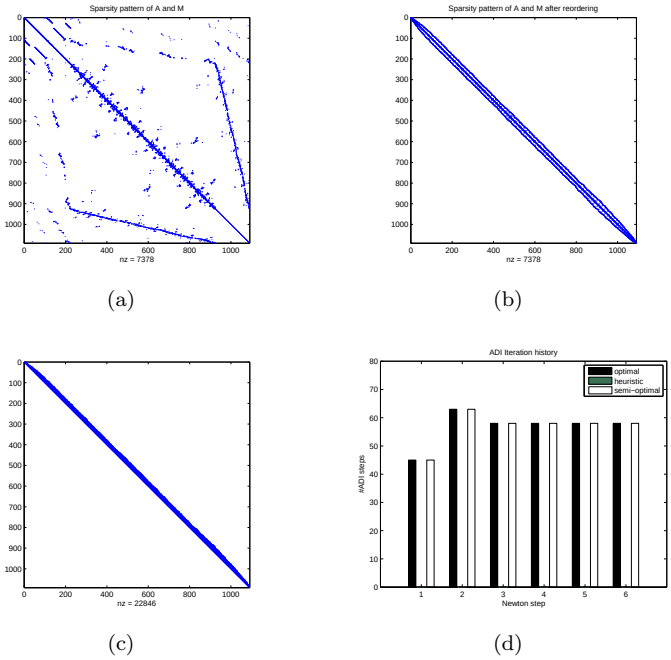


FIG. 4.4. *(a) sparsity pattern of A and M in (4.5) , (b) sparsity pattern of A and M in (4.5) after reordering for bandwidth reduction, (c) sparsity pattern of the Cholesky factor of reordered M and (d) Iteration history for the Newton ADI*

This is transformed into a standard system (4.2) by decomposing M into $M = M_L M_U$ where $M_L = M_U^T$ since M is symmetric. Then defining $\tilde{x} := M_U x$, $A := M_L^{-1} \tilde{A} M_U^{-1}$, $B := M_L^{-1}$ and $C := \tilde{C} M_U^{-1}$ (without computing any of the inverses explicitly in the code) we end up with a standard system for $\tilde{x}$ having the same inputs u as (4.5).

Note, that the heuristic parameters do not appear in the results bar graphics here. This is due to the fact, that the LyaPacksoftware crashed while applying the complex shift computed by the heuristics. Numerical tests where only the real ones of the heuristic parameters where used lead to very poor convergence in the inner loop, which is generally stopped by the maximum iteration number stopping criterion. This resulted in breaking the convergence in the outer Newton loop.

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 general their performance is quite satisfactory as one can see in the numerical examples. The computational cost depends only on the computation of an 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 iterations.

6. Acknowledgements. 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

REFERENCES

[1] M. Abramovitz and I. Stegun, eds., *Pocketbook of mathematical functions*, Verlag Harry Deutsch, 1984. Abridged edition of "Handbook of mathematical functions" (1964).
[2] T. Bagby, *On interpolation by rational functions*, Duke Math. J., 36 (1969), pp. 95–104.
[3] H. Banks and K. Kunisch, *The linear regulator problem for parabolic systems*, SIAM J. Cont. Optim., 22 (1984), pp. 684–698.
[4] P. Benner, *Computational methods for linear-quadratic optimization*, Supplemento ai Rendiconti del Circolo Matematico di Palermo, Serie II, No. 58 (1999), pp. 21–56.
[5] P. Benner, S. Görner, and J. Saak, *Numerical solution of optimal control problems for parabolicsystems*, in Parallel Algorithms and Cluster Computing. Implementations,Algorithms, and Applications, K. Hoffmann and A. Meyer, eds., vol. 52 of Lecture Notes in Computational Science and Engineering, Springer-Verlag, Berlin/Heidelberg, Germany, 2006.
[6] P. Benner, J.-R. Li, and T. Penzl, *Numerical solution of large Lyapunov equations, Riccati equations, and linear-quadratic control problems*, tech. rep.
[7] P. Benner and H. Mena, *BDF methods for large-scale differential Riccati equations*, in Proc. of Mathematical Theory of Network and Systems, MTNS 2004, B. D. Moor, B. Motmans, J. Willems, P. V. Dooren, and V. Blondel, eds., 2004.
[8] P. Benner and J. Saak, *Linear-quadratic regulator design for optimal cooling of steel profiles*, Tech. Rep. SFB393/05-05, Sonderforschungsbereich 393 *Parallele Numerische Simulation für Physik und Kontinuumsmechanik*, TU Chemnitz, D-09107 Chemnitz (Germany), 2005. Available from `http://www.tu-chemnitz.de/sfb393/sfb05pr.html`.
[9] A. Bensoussan, G. D. Prato, M. Delfour, and S. Mitter, *Representation and Control of Infinite Dimensional Systems, Volume I*, Systems & Control: Foundations & Applications, Birkäuser, Boston, Basel, Berlin, 1992.
[10] ———, *Representation and Control of Infinite Dimensional Systems, Volume II*, Systems & Control: Foundations & Applications, Birkäuser, Boston, Basel, Berlin, 1992.
[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, Feb. 2002. Available from `http://www.icm.tu-bs.de/NICONET/reports.html`.
[12] C. Choi and A. Laub, *Efficient matrix-valued algorithms for solving stiff Riccati differential equations*, IEEE Trans. Automat. Control, 35 (1990), pp. 770–776.
[13] R. Curtain and T. Pritchard, *Infinite Dimensional Linear System Theory*, vol. 8 of Lecture Notes in Control and Information Sciences, Springer-Verlag, New York, 1978.
[14] R. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, vol. 21 of Texts in Applied Mathematics, Springer-Verlag, New York, 1995.
[15] B. Datta, *Numerical Methods for Linear Control Systems*, Elsevier Academic Press, 2004.
[16] L. Dieci, *Numerical integration of the differential Riccati equation and some related issues*, SIAM J. Numer. Anal., 29 (1992), pp. 781–815.
[17] J. S. Gibson, *The Riccati integral equation for optimal control problems in Hilbert spaces*, SIAM J. Cont. Optim., 17 (1979), pp. 537–565.
[18] A. A. Gonchar, *Zolotarev problems connected with rational functions*, Math USSR Sbornik, 7 (1969), pp. 623–635.
[19] M. P. Istace and J. P. Thiran, *On the third and fourth Zoltarev problems in complex plane*, Math. Comp., (1993).
[20] K. Ito, *Finite-dimensional compensators for infinite-dimensional systems via galerkin-type approximation*, SIAContOpt, 28 (1990), pp. 1251–1269.
[21] K. Ito and K. Kunisch, *Receding horizon optimal control for infinite dimensional systems*, ESAIM: Control Optim. Calc. Var., 8 (2002), pp. 741–760.
[22] ———, *Receding horizon control with incomplete observations*, SIAM J. Control Optim., 45 (2006), pp. 207–225.
[23] P. Lancaster and L. Rodman, *The Algebraic Riccati Equation*, Oxford University Press, Oxford, 1995.
[24] I. Lasiecka and R. Triggiani, *Differential and Algebraic Riccati Equations with Application*

to Boundary/Point Control Problems: Continuous Theory and Approximation Theory, no. 164 in Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, 1991.

[25] ———, *Control Theory for Partial Differential Equations: Continuous and Approximation Theories I. Abstract Parabolic Systems*, Cambridge University Press, Cambridge, UK, 2000.

[26] V. I. Lebedev, *On a Zolotarev problem in the method of alternating directions*, USSR Comput. Math. and Math. Phys., 17 (1977), pp. 58–76.

[27] J. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, FRG, 1971.

[28] A. Lu and E. Wachspress, *Solution of Lyapunov equations by alternating direction implicit iteration.*, Comput. Math. Appl., 21 (1991), pp. 43–58.

[29] V. Mehrmann, *The Autonomous Linear Quadratic Control Problem, Theory and Numerical Solution*, no. 163 in Lecture Notes in Control and Information Sciences, Springer-Verlag, Heidelberg, July 1991.

[30] K. Morris, *Convergence of controllers designed using state-space methods.*, IEEE Trans. Automat. Control, 39 (1994), pp. 2100–2104.

[31] ———, *Design of finite-dimensional controllers for infinite-dimensional systems by approximation*, J. Math. Systems, Estim. and Control, 4 (1994), pp. 1–30.

[32] T. Penzl, *A cyclic low rank Smith method for large sparse Lyapunov equations*, SIAM J. Sci. Comput., 21 (2000), pp. 1401–1418.

[33] ———, Lyapack *Users Guide*, Tech. Rep. SFB393/00-33, Sonderforschungsbereich 393 *Numerische Simulation auf massiv parallelen Rechnern*, TU Chemnitz, 09107 Chemnitz, Germany, 2000. Available from `http://www.tu-chemnitz.de/sfb393/sfb00pr.html`.

[34] P. Petkov, N. Christov, and M. Konstantinov, *Computational Methods for Linear Control Systems*, Prentice-Hall, Hertfordshire, UK, 1991.

[35] A. Pritchard and D. Salamon, *The linear quadratic control problem for infinite dimensional systems with unbounded input and output operators.*, SIAM J. Control Optimization, 25 (1987), pp. 121–144.

[36] V. Sima, *Algorithms for Linear-Quadratic Optimization*, vol. 200 of Pure and Applied Mathematics, Marcel Dekker, Inc., New York, NY, 1996.

[37] E. D. Sontag, *Mathematical control theory. Deterministic finite dimensional systems.*, no. 6. in Texts in Applied Mathematics, Springer-Verlag, New York, NY, 2nd ed., 1998.

[38] 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.

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

[40] U. Storch and H. Wiebe, *Textbook of mathematics. Vol. 1: Analysis of one variable. (Lehrbuch der Mathematik. Band 1: Analysis einer Veränderlichen.)*, Spektrum Akademischer Verlag, Heidelberg, 3 ed., 2003. (German).

[41] J. Todd, *Applications of transformation theory: A legacy from Zolotarev (1847-1878)*, in Approximation Theory and Spline Functions, S. P. S. et al., ed., no. C 136 in NATO ASI Ser., Dordrecht-Boston-Lancaster, 1984, D. Reidel Publishing Co., pp. 207–245. Proc. NATO Adv. Study Inst., St. John's/Newfoundland 1983.

[42] E. Wachspress, *Iterative solution of the Lyapunov matrix equation*, Appl. Math. Letters, 107 (1988), pp. 87–90.

[43] ———, *The ADI model problem*, 1995. Available from the author.

[44] J. Zabczyk, *Remarks on the algebraic Riccati equation*, Appl. Math. Optim., 2 (1976), pp. 251–258.