

# APPROXIMATE NULLSPACE ITERATIONS FOR KKT SYSTEMS IN MODEL BASED OPTIMIZATION

KAZUFUMI ITO\*, KARL KUNISCH†, VOLKER SCHULZ‡, AND ILIA GHERMAN§

**Abstract.** The aim of the paper is to provide a theoretical basis for approximate reduced SQP methods. In contrast to inexact reduced SQP methods, the forward and the adjoint problem accuracies are not increased when zooming in to the solution of an optimization problem. Only linear-quadratic problems are treated, where approximate reduced SQP methods can be viewed as null-space iterations for KKT systems. Theoretical convergence results are given. Numerical examples illustrate the results and show that convergence also holds in cases when the assumptions guaranteeing convergence are not satisfied.

**Key words.** KKT systems, approximate reduced SQP methods, iterative solvers.

**AMS subject classifications.** 65F10, 65K05, 90C20, 93C20.

**1. Introduction.** We consider optimization problems of the form

$$\min_{x,p} f(x,p) \tag{1.1}$$

$$\text{s.t. } c(x,p) = 0 \tag{1.2}$$

where  $x \in \mathbb{R}^{n_x}$ ,  $p \in \mathbb{R}^{n_p}$  are the variable vectors of the optimization problem,  $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_p} \rightarrow \mathbb{R}$  is the objective of the problem, and  $c : \mathbb{R}^{n_x} \times \mathbb{R}^{n_p} \rightarrow \mathbb{R}^{n_c}$  the equality constraint. The Karush-Kuhn-Tucker necessary optimality conditions for this problem can be expressed as

$$\nabla_x L(x,p,\lambda) = 0 \tag{1.3}$$

$$\nabla_p L(x,p,\lambda) = 0 \tag{1.4}$$

$$c(x,p) = 0, \tag{1.5}$$

where  $L(x,p,\lambda) = f(x,p) + \lambda^\top c(x,p)$  defines the Lagrangian, and  $\lambda \in \mathbb{R}^{n_c}$  denotes the adjoint variable. The distinction of two types of variables,  $x$  on the one hand, and  $p$  on the other hand, is typical for model based optimization problems, where we assume that the constraint  $c(x,p) = 0$  is a mathematical description for a certain process to be driven into an optimal state defined by the objective  $f$  by means of proper choice of  $p$ . We call  $x$  the state vector and assume that the model is always solvable with respect to  $x$ , i.e.,  $\partial c / \partial x$  is always invertible.

In particular for large scale systems, reduced SQP methods [Sch97, Hei96, BNS95, KS92] are used successfully as a highly efficient solution approach. These reduced SQP techniques require frequent solutions of linear systems with  $\partial c / \partial x$  or  $(\partial c / \partial x)^\top$  as system matrix. Of course, this is usually not performed exactly, but only approximately. In [HV01] inexact reduced SQP techniques are analysed, where the solution accuracy of these systems is increased with the closeness to the optimal solution. In [Sch97] a reformulation of the reduced SQP method is presented for approximate solutions of these systems, which does not require that the solution accuracy is increased but delivers nevertheless the optimal solution, if convergence is achieved. To the best knowledge of the authors, the convergence itself is not yet guaranteed theoretically.

In the present paper, we simplify the situation to linear quadratic optimization problems

---

\* Department of Mathematics North Carolina State University, Raleigh, North Carolina, USA(kito@math.ncsu.edu)

†Institut für Mathematik, Karl-Franzens-Universität Graz, Heinrichstr. 36, A-8010 Graz, Austria (karl.kunisch@uni-graz.at).

‡Department of Mathematics, University of Trier, Universitätsring 15, 54286 Trier, Germany (volker.schulz@uni-trier.de).

§Department of Mathematics, University of Trier, Universitätsring 15, 54286 Trier, Germany (ilia.gherman@uni-trier.de).

(QP) of the form

$$\min_{x,p} \frac{1}{2} x^\top H_x x + \frac{1}{2} p^\top H_p p + f_x^\top x + f_p^\top p \quad (1.6)$$

$$\text{s.t. } C_x x + C_p p + c = 0. \quad (1.7)$$

The general case including mixing terms is considered in section 4. The QP may stand alone or it may arise as a QP-subproblem within an SQP method for a nonlinear optimization problem as studied above, so that equation (1.7) can be thought of as the linearization of a nonlinear model which is to be optimized in the sense of (1.6). The matrices  $H_x \in \mathbb{R}^{n_x \times n_x}$  and  $H_p \in \mathbb{R}^{n_p \times n_p}$  are supposed to be symmetric and the (stiffness) matrix  $C_x \in \mathbb{R}^{n_x \times n_x}$  is supposed to be invertible. The matrix  $C_p \in \mathbb{R}^{n_x \times n_p}$  determines the influence of the parameter vector  $p$  on the system. The dimensions of the respective vectors are  $f_x \in \mathbb{R}^{n_x}$ ,  $f_p \in \mathbb{R}^{n_p}$  and  $c \in \mathbb{R}^{n_x}$ .

In order to have a well posed problem, we assume that the reduced Hessian,

$$S = H_p + C_p^\top C_x^{-1} H_x C_x^{-1} C_p \quad (1.8)$$

is positive definite (or coercive, if we think of it as an operator in a function space). The QP (1.6, 1.7) is equivalent to the system of linear equations

$$\begin{bmatrix} H_x & 0 & C_x^\top \\ 0 & H_p & C_p^\top \\ C_x & C_p & 0 \end{bmatrix} \begin{pmatrix} x \\ p \\ \lambda \end{pmatrix} = \begin{pmatrix} -f_x \\ -f_p \\ -c \end{pmatrix} \quad (1.9)$$

Here, we should note that the reduced Hessian  $S$  can also be interpreted as the Schur complement of the KKT matrix in (1.9) with respect to the variables  $(x, \lambda)$ .

Model based optimization problems usually start from an already established solution technique of the model equation  $C_x x + C_p p + c = 0$ . That means that there is a-priori knowledge, e.g., in the form of an approximation  $A$  for  $C_x$ , which is easily invertible and can be used for an iterative solution of the state equation (1.7). For that, we assume

$$\rho(I - A^{-1}C_x) < 1$$

where  $\rho$  denotes the spectral radius. And we assume to have some approximation  $B$  of the reduced Hessian  $S$ , as well.

It is illustrative for the subsequent discussion to have a look at an SQP step for problem (1.6, 1.7). We need the following definitions:

$$T := \begin{bmatrix} C_x^{-1} C_p \\ I \end{bmatrix}, \quad H := \begin{bmatrix} H_x & 0 \\ 0 & H_p \end{bmatrix}, \quad C := [ C_x \quad C_p ]$$

Thus, we obtain another representation of the reduced Hessian

$$S = T^\top H T$$

which can be considered the null-space Schur complement of the system matrix (KKT matrix) in (1.9). We collect  $x, p$  in the vector  $y := \begin{pmatrix} x \\ p \end{pmatrix}$  and also  $f := \begin{pmatrix} f_x \\ f_p \end{pmatrix}$

The solution of the QP, which is exactly one SQP step formulated as in [KS93], can be written as

$$y = -TS^{-1}T^\top f + TS^{-1}T^\top \begin{pmatrix} H_x C_x^{-1} c \\ 0 \end{pmatrix} - \begin{pmatrix} C_x^{-1} c \\ 0 \end{pmatrix}$$

$$\lambda = -C_x^{-1} (H_x x + f_x)$$

In contrast a reduced SQP step omits terms containing  $H_x$  and consequently it does not give the solution after one step. The iterations are of the form

$$y^{k+1} = y^k + \Delta y$$

$$\lambda^{k+1} = \lambda^k + \Delta \lambda$$

with

$$\begin{bmatrix} 0 & 0 & C_x^\top \\ 0 & S & C_p^\top \\ C_x & C_p & 0 \end{bmatrix} \begin{pmatrix} \Delta x \\ \Delta p \\ \Delta \lambda \end{pmatrix} = - \begin{bmatrix} H_x & 0 & C_x^\top \\ 0 & H_p & C_p^\top \\ C_x & C_p & 0 \end{bmatrix} \begin{pmatrix} x^k \\ p^k \\ \lambda^k \end{pmatrix} - \begin{pmatrix} f_x \\ f_p \\ c \end{pmatrix}$$

In the present paper, we substitute  $C_x$  in  $T$  by an approximation  $A_a$  and the exact reduced Hessian  $S$  by an approximation  $B$  and investigate convergence conditions for the resulting iteration. This leads to the following linear iteration:

$$\begin{pmatrix} x^{k+1} \\ p^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} x^k \\ p^k \\ \lambda^k \end{pmatrix} - \begin{bmatrix} 0 & 0 & A_a \\ 0 & B & C_p^\top \\ A_f & C_p & 0 \end{bmatrix}^{-1} \left( \begin{bmatrix} H_x & 0 & C_x^\top \\ 0 & H_p & C_p^\top \\ C_x & C_p & 0 \end{bmatrix} \begin{pmatrix} x^k \\ p^k \\ \lambda^k \end{pmatrix} + \begin{pmatrix} f_x \\ f_p \\ c \end{pmatrix} \right) \quad (1.10)$$

where we use a slightly more general formulation so that  $A_f$  is some approximation to  $C_x$  and  $A_a$  is some approximation to  $C_x^\top$ .

In practical examples [GS05] the following fact has been observed which seems surprising at the first glance: The method works better, if  $B$  is an approximation of the reduced Hessian consistent with the choice of  $A_f, A_a$  as approximations to  $C_x, C_x^\top$  in the form

$$S_A = H_p + C_p^\top A_a^{-1} H_x A_f^{-1} C_p \quad (1.11)$$

rather than the exact reduced Hessian  $S$  from (1.8). We will give an explanation for this observation. One should note that this is in line with similar studies for variational saddle point problems as in [BWY90]. However, the convergence theory there cannot be carried over to the iteration (1.10).

Since the iteration concept is based on an approximate nullspace representation in  $T$  ( $C_x$  is substituted by  $A_f$  and  $C_x^\top$  is substituted by  $A_a$ ) we call it an approximate nullspace iterative technique. The resulting method resembles preconditioning approaches in [BS01], where theoretical results are derived for the case that  $C_x, C_x^\top$  are not substituted by approximations.

The iterations considered in this paper are related to the so-called piggy-back iterations in [?, ?]. This will be described in more detail in section 4. - In contrast to [?], we do not consider preconditioners to be used within some Krylov method, but rather complete linear iterations.

In the next section 2, we will give the main theoretical results of this paper. Section 3 is devoted to the application of these results in the context of a generic optimal control problem often found in literature. Section 4 generalizes the framework to QP with cross-terms. Numerical experiments supporting the theory of section 2 are given in section 5.

**2. Convergence results.** In this section, we will show that the above iteration (1.10) is convergent and we will also give criteria for the convergence. The whole convergence theory of this section is based on a perturbation analysis. First we show finite step convergence for a reduced-type exact solver.

LEMMA 2.1. *If  $A_f, A_a$  and the Schur complement  $S_A$  are invertible, then the iteration*

$$\begin{pmatrix} x^{k+1} \\ p^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} x^k \\ p^k \\ \lambda^k \end{pmatrix} - \begin{bmatrix} 0 & 0 & A_a \\ 0 & S_A & C_p^\top \\ A_f & C_p & 0 \end{bmatrix}^{-1} \left( \begin{bmatrix} H_x & 0 & A_a \\ 0 & H_p & C_p^\top \\ A_f & C_p & 0 \end{bmatrix} \begin{pmatrix} x^k \\ p^k \\ \lambda^k \end{pmatrix} + \begin{pmatrix} f_x \\ f_p \\ c \end{pmatrix} \right)$$

*converges after three steps to the exact solution of the (perturbed) problem*

$$\begin{bmatrix} H_x & 0 & A_a \\ 0 & H_p & C_p^\top \\ A_f & C_p & 0 \end{bmatrix} \begin{pmatrix} x \\ p \\ \lambda \end{pmatrix} + \begin{pmatrix} f_x \\ f_p \\ c \end{pmatrix} = 0.$$

**Proof.** One could prove this result by proceeding in parallel to reduced SQP methods, i.e. to perform actually three iterations of a reduced SQP method applied to the QP starting from zero and observe that the outcome is actually the exact solution. However, this will

not lead to a proof strategy consistent with the further development. Rather we consider the iteration matrix of the iteration above and show that it is nilpotent. First we give the exact inverse in block form

$$\begin{bmatrix} 0 & 0 & A_a \\ 0 & S_A & C_p^\top \\ A_f & C_p & 0 \end{bmatrix}^{-1} = \begin{bmatrix} A_f^{-1}C_pS_A^{-1}C_p^\top A_a^{-1} & -A_f^{-1}C_pS_A^{-1} & A_f^{-1} \\ -S_A^{-1}C_p^\top A_a^{-1} & S_A^{-1} & 0 \\ A_a^{-1} & 0 & 0 \end{bmatrix}.$$

Now we compute explicitly the iteration matrix

$$\begin{aligned} M &= I - \begin{bmatrix} 0 & 0 & A_a \\ 0 & S_A & C_p^\top \\ A_f & C_p & 0 \end{bmatrix}^{-1} \begin{bmatrix} H_x & 0 & A_a \\ 0 & H_p & C_p^\top \\ A_f & C_p & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & A_a \\ 0 & S_A & C_p^\top \\ A_f & C_p & 0 \end{bmatrix}^{-1} \left( \begin{bmatrix} 0 & 0 & A_a \\ 0 & S_A & C_p^\top \\ A_f & C_p & 0 \end{bmatrix} - \begin{bmatrix} H_x & 0 & A_a \\ 0 & H_p & C_p^\top \\ A_f & C_p & 0 \end{bmatrix} \right) \\ &= \begin{bmatrix} A_f^{-1}C_pS_A^{-1}C_p^\top A_a^{-1} & -A_f^{-1}C_pS_A^{-1} & A_f^{-1} \\ -S_A^{-1}C_p^\top A_a^{-1} & S_A^{-1} & 0 \\ A_a^{-1} & 0 & 0 \end{bmatrix} \begin{bmatrix} -H_x & 0 & 0 \\ 0 & S_A - H_p & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} -A_f^{-1}C_pS_A^{-1}C_p^\top A_a^{-1}H_x & -A_f^{-1}C_pS_A^{-1}(S_A - H_p) & 0 \\ S_A^{-1}C_p^\top A_a^{-1}H_x & S_A^{-1}(S_A - H_p) & 0 \\ -A_a^{-1}H_x & 0 & 0 \end{bmatrix}. \end{aligned}$$

When studying  $M^2$ , we have to keep in mind the definition (1.11) for  $S_A$ . We investigate each block of the  $3 \times 3$ -block matrix  $M^2$ , which is not obviously zero, separately.

$$\begin{aligned} (M^2)_{(1,1)} &= A_f^{-1}C_pS_A^{-1}C_p^\top A_a^{-1}H_xA_f^{-1}C_pS_A^{-1}C_p^\top A_a^{-1}H_x - A_f^{-1}C_pS_A^{-1}(S_A - H_p)S_A^{-1}C_p^\top A_a^{-1}H_x \\ &= A_f^{-1}C_pS_A^{-1} \underbrace{(C_p^\top A_a^{-1}H_xA_f^{-1}C_p - S_A + H_p)}_0 S_A^{-1}C_p^\top A_a^{-1}H_x = 0 \end{aligned}$$

$$\begin{aligned} (M^2)_{(1,2)} &= A_f^{-1}C_pS_A^{-1}C_p^\top A_a^{-1}H_xA_f^{-1}C_pS_A^{-1}(S_A - H_p) - A_f^{-1}C_pS_A^{-1}(S_A - H_p)S_A^{-1}(S_A - H_p) \\ &= A_f^{-1}C_pS_A^{-1} \underbrace{(C_p^\top A_a^{-1}H_xA_f^{-1}C_p - S_A + H_p)}_0 S_A^{-1}(S_A - H_p) = 0 \end{aligned}$$

$$\begin{aligned} (M^2)_{(2,1)} &= -S_A^{-1}C_p^\top A_a^{-1}H_xA_f^{-1}C_pS_A^{-1}C_p^\top A_a^{-1}H_x + S_A^{-1}(S_A - H_p)S_A^{-1}C_p^\top A_a^{-1}H_x \\ &= -S_A^{-1} \underbrace{(C_p^\top A_a^{-1}H_xA_f^{-1}C_p - S_A + H_p)}_0 S_A^{-1}C_p^\top A_a^{-1}H_x = 0 \end{aligned}$$

$$\begin{aligned} (M^2)_{(2,2)} &= -S_A^{-1}C_p^\top A_a^{-1}H_xA_f^{-1}C_pS_A^{-1}(S_A - H_p) + S_A^{-1}(S_A - H_p)S_A^{-1}(S_A - H_p) \\ &= -S_A^{-1} \underbrace{(C_p^\top A_a^{-1}H_xA_f^{-1}C_p - S_A + H_p)}_0 S_A^{-1}(S_A - H_p) = 0 \end{aligned}$$

$$(M^2)_{(3,1)} = A_a^{-1}H_xA_f^{-1}C_pS_A^{-1}C_p^\top A_a^{-1}H_x \neq 0 \quad (\text{in general})$$

$$(M^2)_{(3,2)} = A_a^{-1}H_xA_f^{-1}C_pS_A^{-1}(S_A - H_p) \neq 0 \quad (\text{in general})$$

Therefore,  $M^2$  is of the form

$$M^2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ * & * & 0 \end{bmatrix}$$

and obviously  $M^3 = M \cdot M^2 = 0$ .  $\square$

Henceforth we use the following abbreviations:

$$R := \begin{bmatrix} 0 & 0 & A_a \\ 0 & B & C_p^\top \\ A_f & C_p & 0 \end{bmatrix} \quad K := \begin{bmatrix} H_x & 0 & C_x^\top \\ 0 & H_p & C_p^\top \\ C_x & C_p & 0 \end{bmatrix}$$

$$\tilde{K} := \begin{bmatrix} H_x & 0 & A_a \\ 0 & \tilde{H}_p & C_p^\top \\ A_f & C_p & 0 \end{bmatrix} \quad \tilde{H}_p := H_p - S_A + B.$$

Note that the matrix  $B$  is the exact Schur complement of the matrix  $\tilde{K}$  which arises by eliminating the first and third variables, because

$$\tilde{H}_p + C_p^\top A_a^{-1} H_x A_f^{-1} C_p = H_p - S_A + B + C_p^\top A_a^{-1} H_x A_f^{-1} C_p = B.$$

Therefore, we can conclude with lemma 2.1 that

$$(I - R^{-1} \tilde{K})^3 = 0.$$

The iteration matrix for the iteration (1.10) is given by

$$I - R^{-1} K = I - R^{-1} \tilde{K} + R^{-1} (\tilde{K} - K) =: M + N, \quad (2.1)$$

where  $M$  is a nilpotent matrix of nilpotency degree 3 and  $N$  can be considered as a perturbation of  $M$ . For any norm  $\|\cdot\|$ , this yields

$$\|(M + N)^3\| \leq \left[ \|M^2 + MN + NM + N^2\| + \|M^2 + NM\| + \|M\|^2 \right] \|N\|.$$

The following lemma is based on a perturbation argument and estimates the influence of the perturbation induced by  $K$ .

LEMMA 2.2. *Define*

$$\theta := \|M^2 + MN + NM + N^2\| + \|M^2 + NM\| + \|M\|^2, \quad r := \|N\| = \|R^{-1} (\tilde{K} - K)\|$$

If  $\theta \cdot r < 1$ , then the iteration (1.10) converges with the upper bound for the convergence rate

$$\kappa := \sqrt[3]{\theta \cdot r} < 1.$$

**Proof.** By Gelfand's theorem, we obtain

$$\begin{aligned} \rho(I - R^{-1} K) &= \sqrt[3]{\rho((I - R^{-1} K)^3)} = \sqrt[3]{\lim_{n \rightarrow \infty} \|(I - R^{-1} K)^{3n}\|^{1/n}} = \sqrt[3]{\lim_{n \rightarrow \infty} \|(M + N)^{3n}\|^{1/n}} \\ &\leq \sqrt[3]{\theta \cdot r} < 1 \end{aligned}$$

$\square$

Henceforth we need to specify a norm which we fix as the  $\ell_2$ - norm.

THEOREM 2.3. *We define the numerical spectral norms*

$$r_A^f := \|I - A_f^{-1} C_x\|_2, \quad r_A^a := \|I - A_a^{-1} C_x^\top\|_2, \quad r_S := \|I - B^{-1} S_A\|_2$$

$$\text{If } \max\{r_A^f, r_A^a, r_S\} < 1/\tilde{\theta}, \quad (2.2)$$

$$\text{with } \tilde{\theta} := \theta \cdot \varphi, \quad \text{and } \varphi := \left\| \begin{bmatrix} A_f^{-1} C_p B^{-1} C_p^\top & -A_f^{-1} C_p & I \\ -B^{-1} C_p^\top & I & 0 \\ I & 0 & 0 \end{bmatrix} \right\|_2,$$

then the iteration (1.10) converges with an upper bound for the convergence rate given by

$$\kappa = \sqrt[3]{\theta \cdot r} < 1.$$

**Proof.** We observe that

$$R^{-1}(\tilde{K} - K) = \begin{bmatrix} A_f^{-1}C_pB^{-1}C_p^\top A_a^{-1} & -A_f^{-1}C_pB^{-1} & A_f^{-1} \\ -B^{-1}C_p^\top A_a^{-1} & B^{-1} & 0 \\ A_a^{-1} & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & A_a - C_x^\top \\ 0 & B - S_A & 0 \\ A_f - C_x & 0 & 0 \end{bmatrix} = \begin{bmatrix} A_f^{-1}C_pB^{-1}C_p^\top & -A_f^{-1}C_p & I \\ -B^{-1}C_p^\top & I & 0 \\ I & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & I - A_a^{-1}C_x^\top \\ 0 & I - B^{-1}S_A & 0 \\ I - A_f^{-1}C_x & 0 & 0 \end{bmatrix}$$

where

$$\left\| \begin{bmatrix} 0 & 0 & I - A_a^{-1}C_x^\top \\ 0 & I - B^{-1}S_A & 0 \\ I - A_f^{-1}C_x & 0 & 0 \end{bmatrix} \right\|_2 = \max\{r_A^f, r_A^a, r_S\}$$

Application of lemma 2.2 gives

$$\begin{aligned} \rho(I - R^{-1}K) &\leq \sqrt[3]{\|(I - R^{-1}K)^3\|_2} \leq \underbrace{\sqrt[3]{\theta \cdot \|R^{-1}(\tilde{K} - K)^3\|_2}}_{\kappa} \\ &\leq \sqrt[3]{\theta \cdot \varphi \cdot \max\{r_A^f, r_A^a, r_S\}} < 1 \end{aligned}$$

□

REMARK 1. *Since  $M$  and  $N$  defined in (2.1) depend on  $A_f$  and  $A_a$  and therefore also  $\tilde{\theta}$ , we need to verify that the condition (2.2) can be satisfied for some choice of  $A_f$  and  $A_a$ . In particular, we observe that  $\tilde{\theta}$  stays bounded from above, when  $A_f \rightarrow C_x$  and  $A_a \rightarrow C_x^\top$ , because then*

$$\begin{aligned} N &\rightarrow 0 \\ M &\rightarrow I - \begin{bmatrix} 0 & 0 & C_x^\top \\ 0 & S & C_p^\top \\ C_x & C_p & 0 \end{bmatrix}^{-1} \begin{bmatrix} H_x & 0 & C_x^\top \\ 0 & H_p & C_p^\top \\ C_x & C_p & 0 \end{bmatrix} =: M_{\text{lim}} \\ \Rightarrow \theta &\rightarrow 2\|M_{\text{lim}}^2\| + \|M_{\text{lim}}\|^2 =: \theta_{\text{lim}} \\ \varphi &\rightarrow \left\| \begin{bmatrix} C_x^{-1}C_pS^{-1}C_p^\top & -C_x^{-1}C_p & I \\ -S^{-1}C_p^\top & I & 0 \\ I & 0 & 0 \end{bmatrix} \right\|_2 =: \varphi_{\text{lim}} \\ \Rightarrow \tilde{\theta} &\rightarrow \theta_{\text{lim}} \cdot \varphi_{\text{lim}} < \infty. \end{aligned}$$

Theorem 2.3 shows that the convergence behaviour of the approximate nullspace iteration is limited both by the approximation quality in the forward and in the adjoint system and by the approximation quality of the consistent Schur complement in a worst case fashion. Often, one might choose  $A_a = A_f^\top =: A$ . In this situation, we can give a refined version of the theorem above, if additionally  $C_x^\top = C_x$  and one chooses the spectral norm

$$\left\| \begin{pmatrix} x \\ p \\ \lambda \end{pmatrix} \right\|_R := \left\| \begin{pmatrix} A^{1/2}x A^{-1/2} \\ B^{1/2}p B^{-1/2} \\ A^{1/2}\lambda A^{-1/2} \end{pmatrix} \right\|_2$$

COROLLARY 2.4. *Choose  $A_a = A_f^\top = A$  and  $\|\cdot\| := \|\cdot\|_R$ , and define*

$$\rho_A := \rho(I - A^{-1}C_x), \quad \rho_S := \rho(I - B^{-1}S_A).$$

Then if  $\max\{\rho_A^f, \rho_S\} < 1/\bar{\theta}$ , with

$$\bar{\theta} := \theta \cdot \bar{\varphi}, \quad \text{and } \bar{\varphi} := \rho \left( \begin{bmatrix} A^{-1/2}C_p B^{-1}C_p^\top A^{-1/2} & -A^{-1/2}C_p B^{-1/2} & I \\ -B^{-1/2}C_p^\top A^{-1/2} & I & 0 \\ I & 0 & 0 \end{bmatrix} \right),$$

then the iteration (1.10) converges with the upper bound for the convergence rate

$$\kappa = \sqrt[3]{\bar{\theta} \cdot r} < 1.$$

**Proof.** We give more refined representations of the factors

$$\begin{aligned} & \left\| \begin{bmatrix} A^{-1}C_p B^{-1}C_p^\top & -A^{-1}C_p & I \\ -B^{-1}C_p^\top & I & 0 \\ I & 0 & 0 \end{bmatrix} \right\|_R = \\ & \left\| \begin{bmatrix} A^{-1/2}C_p B^{-1}C_p^\top A^{-1/2} & -A^{-1/2}C_p B^{-1/2} & I \\ -B^{-1/2}C_p^\top A^{-1/2} & I & 0 \\ I & 0 & 0 \end{bmatrix} \right\|_2 = \\ & \rho \left( \begin{bmatrix} A^{-1/2}C_p B^{-1}C_p^\top A^{-1/2} & -A^{-1/2}C_p B^{-1/2} & I \\ -B^{-1/2}C_p^\top A^{-1/2} & I & 0 \\ I & 0 & 0 \end{bmatrix} \right) \end{aligned}$$

and

$$\begin{aligned} & \left\| \begin{bmatrix} 0 & 0 & I - A^{-1}C_x^\top \\ 0 & I - B^{-1}S_A & 0 \\ I - A^{-1}C_x & 0 & 0 \end{bmatrix} \right\|_R = \\ & = \left\| \begin{bmatrix} 0 & 0 & I - A^{-1/2}C_x A^{-1/2} \\ 0 & I - B^{-1/2}S_A B^{-1/2} & 0 \\ I - A^{-1/2}C_x A^{-1/2} & 0 & 0 \end{bmatrix} \right\|_2 \\ & = \rho \left( \begin{bmatrix} 0 & 0 & I - A^{-1/2}C_x A^{-1/2} \\ 0 & I - B^{-1/2}S_A B^{-1/2} & 0 \\ I - A^{-1/2}C_x A^{-1/2} & 0 & 0 \end{bmatrix} \right) = \max\{\rho_A^f, \rho_S\}. \end{aligned}$$

We conclude analogously to the proof of theorem 2.3

$$\begin{aligned} \rho(I - R^{-1}K) & \leq \sqrt[3]{\|(I - R^{-1}K)^3\|_R} \leq \underbrace{\sqrt[3]{\theta \cdot \|R^{-1}(\tilde{K} - K)^3\|_R}}_{\kappa} \\ & \leq \sqrt[3]{\bar{\theta} \cdot \bar{\varphi} \cdot \max\{\rho_A, \rho_S\}} < 1 \end{aligned}$$

□

In many cases a good choice for  $A$  approximating  $C_x$  will be available. However, the only part of  $S_A$  which is easily accessible is  $H_p$ . Therefore, a natural question arises about the usefulness of just using  $H_p$  as an approximation to  $S_A$ . For the analysis of this effect, one has to take into account more refined problem characteristics. This will be performed in the next section.

**3. Application to optimal control.** The iterative methods discussed above are of particular importance for the solution of optimal control problems. A generic version of them is the problem

$$\begin{aligned} & \min_{x,p} \frac{1}{2}(x - \hat{x}, x - \hat{x})_2 + \frac{\mu}{2}(p, p)_2 \\ \text{s.t.} & \quad Ly + \Pi p = 0 \end{aligned}$$

where  $L : H^2(\Omega) \cap H_0^1(\Omega) \rightarrow L_2(\Omega)$  is a linear mapping for functions defined on an open region  $\Omega$  and  $(\cdot, \cdot)_2$  is the scalar product in  $L_2$ . The operator  $\Pi$  is assumed to be bounded. Discretization, e.g., by finite differences with meshsize  $h$  gives

$$H_x = h^d \cdot M_1, \quad H_p = h^d \cdot M_1, \quad C_x = L_h = h^{-2} M_2, \quad C_p = \Pi_h$$

where  $d$  is the space dimension of  $\Omega$  and

$$\|M_1\| \leq m_1 < \infty, \quad \|M_2\| \leq m_2 < \infty, \quad \|\Pi_h\| \leq \pi < \infty$$

For iteration purposes, we may assume that the approximation  $A$  to  $L_h$  is of the form

$$A = h^{-2}W, \quad \text{with} \quad \frac{1}{w} \cdot I \leq W \leq w \cdot I$$

for some  $w \in \mathbb{R}, w \geq 0$ . Then, the "wrong" Schur complement takes the form

$$\begin{aligned} S_A &= \mu \cdot h^d \cdot M_1 + \Pi_h^\top \cdot h^2 \cdot W^{-\top} \cdot h^d \cdot M_1 \cdot h^2 \cdot W^{-1} \cdot \Pi_h \\ &= \mu \cdot h^d \cdot M_1 + h^{d+4} \cdot \Pi_h^\top W^{-\top} M_1 W^{-1} \Pi_h \end{aligned}$$

The  $H_p$ -part of the Schur complement is easily available. Therefore we consider the choice of  $B = H_p$  in the Schur complement part of the iterations. The resulting iteration matrix takes the form

$$I - B^{-1}S_A = I - H_p^{-1}S_A = -\frac{h^4}{\mu} \cdot \Pi_h^\top W^{-\top} M_1 W^{-1} \Pi_h$$

Now, we can make the following observations.

**COROLLARY 3.1.** *For discretized optimal control problems with the characteristics described above, we obtain a good convergence rate in the Schur complement for  $B = H_p$ , provided that  $\mu$  is large enough or the discretization  $h$  is fine enough.*

**Proof.** For the norm  $\|\cdot\| = (\cdot, \cdot)^{1/2}$  we see that

$$\begin{aligned} \rho_S &= \rho(I - H_p^{-1}S_A) \leq \left\| \frac{h^4}{\mu} \cdot \Pi_h^\top W^{-\top} M_1 W^{-1} \Pi_h \right\| \\ &\leq \frac{h^4}{\mu} \cdot \pi^2 \cdot w^2 \cdot m_1 \end{aligned}$$

Therefore,  $\rho_S < 1$ , if  $\mu$  large enough or  $h$  small enough.  $\square$

Now, we can easily achieve  $\rho_S < 1$ . The forward and adjoint system can also be assumed to be solvable with  $\rho_A < 1$ . If we take a close look at theorem 2.3 and corollary 2.4, we see that these properties for  $\rho_S, \rho_A$  are not enough to guarantee overall convergence. At least in corollary 2.4, we observe, that  $\bar{\varphi}$  is close to 1, if  $h$  is small enough or  $\mu$  large enough. However, the parameter  $\theta$  increases noticeably for decreasing  $h$ , when performing numerical experiments as below. Therefore, for small  $h$  the conditions for  $\rho_S, \rho_A$  become very restrictive in order to be able to apply the convergence theorems 2.3 and 2.4. But, the numerical results below show, that convergence is also achieved in cases, where theorem 2.3 and corollary 2.4 are not applicable.

**4. Generalizations.** For ease of presentation, the discussions in sections 1 and 2 have been limited to linear-quadratic problems of the type (1.6, 1.7). However, QP within an SQP method for the problem (1.1, 1.2) or derived from a Newton method for the necessary conditions (1.3-1.5) contain cross-terms in the form

$$\min_{x,p} \frac{1}{2} (x^\top H_x x + x^\top H_{xp} p + p^\top H_{px} x + p^\top H_p p) + f_x^\top x + f_p^\top p \quad (4.1)$$

$$\text{s.t.} \quad C_x x + C_p p + c = 0 \quad (4.2)$$

which is equivalent to the linear system

$$\begin{bmatrix} H_x & H_{xp} & C_x^\top \\ H_{px} & H_p & C_p^\top \\ C_x & C_p & 0 \end{bmatrix} \begin{pmatrix} x \\ p \\ \lambda \end{pmatrix} = \begin{pmatrix} -f_x \\ -f_p \\ -c \end{pmatrix}. \quad (4.3)$$

Here, we show that all results can be generalized to this case, as well, i.e. to the iteration  $(x^{k+1}, p^{k+1}, \lambda^{k+1}) = (x^k, p^k, \lambda^k) + (\Delta x^k, \Delta p^k, \Delta \lambda^k)$

$$\begin{pmatrix} \Delta x^k \\ \Delta p^k \\ \Delta \lambda^k \end{pmatrix} = - \begin{bmatrix} 0 & 0 & A_a \\ 0 & B & C_p^\top \\ A_f & C_p & 0 \end{bmatrix}^{-1} \left( \begin{bmatrix} H_x & H_{xp} & C_x^\top \\ H_{px} & H_p & C_p^\top \\ C_x & C_p & 0 \end{bmatrix} \begin{pmatrix} x^k \\ p^k \\ \lambda^k \end{pmatrix} + \begin{pmatrix} f_x \\ f_p \\ c \end{pmatrix} \right) \quad (4.4)$$

where  $B$  approximates the consistent Schur complement

$$S_A = H_p - H_{px} A_f^{-1} C_p - C_p^\top A_a^{-1} H_{xp} + C_p^\top A_a^{-1} H_x A_f^{-1} C_p \quad (4.5)$$

**THEOREM 4.1.** *The results of theorem 2.3 and corollary 2.4 are also valid for the case of non-zero cross-terms, i.e. problem (4.1, 4.2) and iteration (4.4).*

**Proof.** First, we have to generalize lemma 2.1. I.e., we have to show that the matrix

$$\begin{aligned} M &= I - \begin{bmatrix} 0 & 0 & A_a \\ 0 & S_A & C_p^\top \\ A_f & C_p & 0 \end{bmatrix}^{-1} \begin{bmatrix} H_x & H_{xp} & A_a \\ H_{px} & H_p & C_p^\top \\ A_f & C_p & 0 \end{bmatrix} \\ &= \begin{bmatrix} M_{11} & M_{12} & 0 \\ (S_A^{-1} C_p^\top A_a^{-1} H_x - S_A^{-1} H_{px}) & (S_A^{-1} C_p^\top A_a^{-1} H_{xp} + S_A^{-1} (S_A - H_p)) & 0 \\ -A_a^{-1} H_x & -A_a^{-1} H_{xp} & 0 \end{bmatrix}, \end{aligned}$$

where

$$\begin{aligned} M_{11} &= -A_f^{-1} C_p S_A^{-1} C_p^\top A_a^{-1} H_x + A_f^{-1} C_p S_A^{-1} H_{px} \\ M_{12} &= -A_f^{-1} C_p S_A^{-1} C_p^\top A_a^{-1} H_{xp} - A_f^{-1} C_p S_A^{-1} (S_A - H_p) \end{aligned}$$

is nilpotent of degree 3. We proceed analogously to the proof of lemma 2.1. In order to simplify the notation, we define the formal expression

$$\mathcal{Z} := \left[ C_p^\top A_a^{-1} H_x A_f^{-1} C_p - H_{px} A_f^{-1} C_p - C_p^\top A_a^{-1} H_{xp} - S_A + H_p \right].$$

Of course,  $\mathcal{Z} = 0$ , but we have to keep in mind the formal expression above in order to be able to understand the subsequent arguments. We compute

$$\begin{aligned} (M^2)_{(1,1)} &= A_f^{-1} C_p S_A^{-1} \mathcal{Z} S_A^{-1} C_p^\top A_a^{-1} H_x - A_f^{-1} C_p S_A^{-1} \mathcal{Z} S_A^{-1} H_{px} = 0 \\ (M^2)_{(1,2)} &= A_f^{-1} C_p S_A^{-1} \mathcal{Z} S_A^{-1} C_p^\top A_a^{-1} H_{xp} + A_f^{-1} C_p S_A^{-1} \mathcal{Z} S_A^{-1} (S_A - H_p) = 0 \\ (M^2)_{(2,1)} &= S_A^{-1} \mathcal{Z} S_A^{-1} H_{xp} - S_A^{-1} \mathcal{Z} S_A^{-1} C_p A_a^{-1} H_x = 0 \\ (M^2)_{(2,2)} &= -S_A^{-1} \mathcal{Z} S_A^{-1} C_p A_a^{-1} H_{xp} - S_A^{-1} \mathcal{Z} S_A^{-1} (S_A - H_p) = 0 \end{aligned}$$

Therefore,  $M$  is again nilpotent of degree 3. The remainder of the discussion after lemma 2.1 need not be changed at all, if we use the appropriate definitions

$$K := \begin{bmatrix} H_x & H_{xp} & C_x^\top \\ H_{px} & H_p & C_p^\top \\ C_x & C_p & 0 \end{bmatrix} \quad \tilde{K} := \begin{bmatrix} H_x & H_{xp} & A_a \\ H_{px} & \tilde{H}_p & C_p^\top \\ A_f & C_p & 0 \end{bmatrix}$$

and again  $\tilde{H}_p := H_p - S_A + B$ .  $\square$

For comparison with other iterations and for generalization to nonlinear problems, it is useful to write the iteration (4.4) by use of the notation (1.3-1.5). We define the Lagrangian

$$L(x, p, \lambda) := \frac{1}{2} (x^\top H_x x + x^\top H_{xp} p + p^\top H_{px} x + p^\top H_p p) + f_x^\top x + f_p^\top p + \lambda^\top (C_x x + C_p p + c)$$

Then, the necessary conditions are

$$\begin{aligned} \nabla_x L(x, p, \lambda) &= H_x x + H_{xp} p + C_x^\top \lambda + f_x = 0 \\ \nabla_p L(x, p, \lambda) &= H_{px} x + H_p p + C_p^\top \lambda + f_p = 0 \\ \nabla_\lambda L(x, p, \lambda) &= C_x x + C_p p + c = 0 \end{aligned}$$

Iteration (4.4) can now be rewritten in a more compact form as

$$\begin{aligned}\lambda^{k+1} &= \lambda^k - A_a^{-1} \nabla_x L(x^k, p^k, \lambda^k) \\ p^{k+1} &= p^k - B^{-1} \nabla_p L(x^k, p^k, \lambda^{k+1}) \\ x^{k+1} &= x^k - A_f^{-1} \nabla_\lambda L(x^k, p^{k+1}, \lambda^{k+1})\end{aligned}$$

The difference to so-called piggy-back-iterations as in [?, ?] lies in the early usage of information as soon as it is available. Because, piggy-back iterations are derived from automatic differentiation, where functions and derivatives are evaluated simultaneously, the piggy-iteration for problem (4.1, 4.2) is written in this notation as

$$\begin{aligned}\lambda^{k+1} &= \lambda^k - A_a^{-1} \nabla_x L(x^k, p^k, \lambda^k) \\ p^{k+1} &= p^k - B^{-1} \nabla_p L(x^k, p^k, \lambda^k) \\ x^{k+1} &= x^k - A_f^{-1} \nabla_\lambda L(x^k, p^k, \lambda^k)\end{aligned}$$

or in matrix notation as

$$\begin{pmatrix} \Delta x^k \\ \Delta p^k \\ \Delta \lambda^k \end{pmatrix} = - \begin{bmatrix} 0 & 0 & A_a \\ 0 & B & 0 \\ A_f & 0 & 0 \end{bmatrix}^{-1} \left( \begin{bmatrix} H_x & H_{xp} & C_x^\top \\ H_{px} & H_p & C_p^\top \\ C_x & C_p & 0 \end{bmatrix} \begin{pmatrix} x^k \\ p^k \\ \lambda^k \end{pmatrix} + \begin{pmatrix} f_x \\ f_p \\ c \end{pmatrix} \right) \quad (4.6)$$

where the appropriate definition of the design preconditioner  $B$  is discussed in [?, ?].

**5. Numerical experiments.** Here, we illustrate the theoretical results from the previous section by application to a common model problem, which serves as a standard test problem in PDE constrained optimization. For a given open computational region  $\Omega$ , we investigate the problem

$$\begin{aligned} \min_{x,p} \quad & \frac{1}{2} \int_{\Omega} (x(\xi) - \hat{x}(\xi))^2 d\xi + \frac{\mu}{2} \int_{\Omega} p(\xi)^2 d\xi \\ \text{s.t.} \quad & -\Delta x(\xi) = p(\xi), \quad \forall \xi \in \Omega \\ & x(\xi) = 0 \quad \forall \xi \in \partial\Omega \end{aligned}$$

The variables  $x$  and  $p$  are functions defined on the domain  $\Omega$  and  $\Delta$  denotes the Laplacian operator. The aim of the problem is to track a given function  $\hat{x}$  with the solution of a differential equation. Since we only aim at illustrating certain numerical effects and do not pretend to attack any real life problem in this paper (this has been done and will be done again by the authors in different publications), we even downsize the problem to 1D, i.e.,  $\Omega = [0, 1]$ . Then, this model problem simplifies to

$$\begin{aligned} \min_{x,p} \quad & \frac{1}{2} \int_0^1 (x(\xi) - \hat{x}(\xi))^2 d\xi + \frac{\mu}{2} \int_0^1 p(\xi)^2 d\xi \\ \text{s.t.} \quad & -x''(\xi) = p(\xi), \quad \forall \xi \in [0, 1] \\ & x(0) = 0 = x(1) \end{aligned}$$

This problem is discretized by finite differences on an equidistant mesh with meshsize  $h = 1/(N-1)$ ,  $N \in \mathbb{N}$ , i.e.

$$\begin{aligned} x_\ell &:= x(\ell \cdot h), \quad \ell = 0, \dots, N \\ p_\ell &:= p(\ell \cdot h), \quad \ell = 0, \dots, N \\ -x''(\ell \cdot h) &\approx \frac{1}{h^2} (-x_{\ell-1} + 2x_\ell - x_{\ell+1}), \quad \ell = 1, \dots, N-1 \\ \int_0^1 (x(\xi) - \hat{x}(\xi))^2 d\xi &\approx h \sum_{\ell=1}^{N-1} (x_\ell - \hat{x}(\ell \cdot h))^2 \\ \int_0^1 p(\xi)^2 d\xi &\approx h \sum_{\ell=1}^{N-1} p_\ell^2 \end{aligned}$$



$i$	$j$	$\rho_A$	$\rho_S$	$\rho_{It}$	$\theta$	$r$	$\bar{\theta}$	$\kappa$
0	0	0.9877	0.0000	0.9916	1.3122	1.1898	1.9251	1.1601
0	1	0.9877	0.0000	0.9916	1.3122	1.1898	1.9250	1.1601
0	5	0.9877	0.0000	0.9916	1.3122	1.1898	1.9250	1.1601
0	$S_A$	0.9877	0.0000	0.9916	1.3122	1.1898	1.9250	1.1601
0	$S_E$	0.9877	0.0934	0.9916	1.3122	1.1898	1.9250	1.1601
5	0	0.9284	0.0005	0.9284	2.9489	2.0941	7.9272	1.8346
5	1	0.9284	0.0000	0.9284	2.9485	2.0934	7.9238	1.8344
5	5	0.9284	0.0000	0.9284	2.9485	2.0934	7.9238	1.8344
5	$S_A$	0.9284	0.0000	0.9284	2.9485	2.0934	7.9238	1.8344
5	$S_E$	0.9284	0.0930	0.9284	2.8647	1.9838	7.3266	1.7845
150	0	0.1540	0.0738	0.1549	0.7571	2.7484	13.8025	1.2767
150	1	0.1540	0.0054	0.1801	0.9085	2.5358	15.4144	1.3207
150	5	0.1540	0.0000	0.1780	0.8989	2.5491	15.3299	1.3184
150	$S_A$	0.1540	0.0000	0.1780	0.8989	2.5491	15.3299	1.3184
150	$S_E$	0.1540	0.0266	0.1885	0.9447	2.4861	15.7093	1.3292
200	0	0.0829	0.0867	0.0835	0.2859	1.6347	5.6241	0.7761
200	1	0.0829	0.0075	0.1274	0.4010	1.4559	7.2422	0.8358
200	5	0.0829	0.0000	0.1244	0.3919	1.4662	7.1273	0.8313
200	$S_A$	0.0829	0.0000	0.1244	0.3919	1.4662	7.1273	0.8313
200	$S_E$	0.0829	0.0149	0.1302	0.4099	1.4465	7.3513	0.8401
250	0	0.0446	0.0941	0.0451	0.2232	0.9860	4.5628	0.6037
250	1	0.0446	0.0089	0.0935	0.2447	0.8096	4.5570	0.5829
250	5	0.0446	0.0000	0.0908	0.2379	0.8157	4.4681	0.5789
250	$S_A$	0.0446	0.0000	0.0908	0.2379	0.8157	4.4681	0.5789
250	$S_E$	0.0446	0.0082	0.0933	0.2441	0.8100	4.5499	0.5826
$C_x$	0	0.0000	0.1031	0.1031	0.2987	0.4758	6.3784	0.5219
$C_x$	1	0.0000	0.0106	0.0106	0.1835	0.0466	3.5354	0.2045
$C_x$	5	0.0000	0.0000	0.0000	0.1829	0.0000	3.5583	0.0099
$C_x$	$S_A$	0.0000	0.0000	0.0000	0.1829	0.0000	3.5583	0.0000
$C_x$	$S_E$	0.0000	0.0000	0.0000	0.1829	0.0000	3.5583	0.0000

TABLE 5.1  
Convergence results for  $N = 21$  and  $\mu = 0.1$

case  $i = 0, j = 0$   
case  $i = 250, j = 1$

FIG. 5.1. Convergence history and eigenvalues for  $h = \frac{1}{20}$  and  $\mu = 0.1$

problem is solved exactly). The column  $j$  gives the analogous information for  $B_j^{-1}$ . Furthermore,  $\rho_A$  denotes the spectral radius of the matrix  $I - A_i^{-1}C_x$ . Analogously,  $\rho_S$  denotes the spectral radius of the iteration matrix of the design equation and  $\rho_{It}$  denotes the spectral radius of the overall iteration matrix  $(I - R^{-1}K)$ . For the definitions of  $\theta$ ,  $r$ ,  $\bar{\theta}$  and  $\kappa$  see above and use the  $\|\cdot\|_R$ -norm. ??? (with the  $\|\cdot\|_R$ -norm for  $r$  and  $\theta$ ) ???

We observe that the estimation for  $\kappa$  in corollary 2.4 is conservative. ??? In fact, we have convergence in all cases of table 5.1 since  $\rho_{It} < 1$  but the convergence condition  $\max\{\rho_A, \rho_S\} < 1/\bar{\theta}$  formulated in corollary 2.4 is only satisfied from the row ( $i = 250, j = 1$ ) on downward. Figure 5.1 gives the convergence history and and the eigenvalues of the iteration matrix of the KKT system for two characteristic cases.

???

From table 5.1 we conclude that the choice  $B = S$  is inferior to other choices, since the spectral radius of the KKT iteration matrix is larger than in cases, where  $B$  is chosen closely to  $S_A$  (cf., the cases  $i = 5, 100, 250$ ).

In this setting, we can also check that it is not enough to achieve convergence in the forward, the adjoint and the design system in order to obtain convergence for the overall

FIG. 5.2. *Solution (State) for formulation (1)*

$i$	$j$	$\rho_A$	$\rho_S$	$\rho_{It}$	# It.(0)	# It. (1)
0	0	0.9995	0.0000	1.0011	—	—
0	1	0.9995	0.0000	1.0011	—	—
0	3	0.9995	0.0000	1.0011	—	—
0	$S_A$	0.9995	0.0000	1.0011	—	—
0	$S_E$	0.9995	0.9113	1.0011	—	—
1	0	0.9990	0.0000	0.9990	18771	18770
1	1	0.9990	0.0000	0.9990	18771	18770
1	3	0.9990	0.0000	0.9990	18771	18770
1	$S_A$	0.9990	0.0000	0.9990	18771	18770
1	$S_E$	0.9990	0.9112	0.9991	19701	19701
3	0	0.9980	0.0000	0.9980	9880	9880
3	1	0.9980	0.0000	0.9980	9880	9880
3	3	0.9980	0.0000	0.9980	9880	9880
3	$S_A$	0.9980	0.0000	0.9980	9880	9880
3	$S_E$	0.9980	0.9112	0.9982	10819	10819
5	0	0.9970	0.0001	0.9970	6590	6590
5	1	0.9970	0.0000	0.9970	6590	6590
5	3	0.9970	0.0000	0.9970	6590	6590
5	$S_A$	0.9970	0.0000	0.9970	6590	6590
5	$S_E$	0.9970	0.9112	0.9975	7865	7865

TABLE 5.2  
Convergence results for  $N = 101$  and  $\mu = 0.001$

KKT iteration. In the tests above ( $N = 21$ ), we have achieved convergence in all cases. ??? was ist hier gemeint: to obtain ""theoretically quaranteed"" convergence ???

Table 5.2 gives the results for  $N = 101 \Rightarrow h = 0.01$  and  $\mu = 0.001$ . The columns  $i, j, \rho_A, \rho_S, \rho_{It}$  are defined as above. In addition, we give the numbers for iterations in the columns [# It(0)], [# It(1)]until

$$\left\| \begin{pmatrix} x^{k+1} - x^k \\ p^{k+1} - p^k \\ \lambda^{k+1} - \lambda^k \end{pmatrix} \right\| = \left\| \begin{pmatrix} x^k \\ p^k \\ \lambda^k \end{pmatrix} \right\| \leq 10^{-6}$$

Here distinguish two problem formulations:

- (0)  $\bar{x}(\xi) = 0, \quad \forall \xi \in [0, 1]$
- (1)  $\bar{x}(\xi) = \begin{cases} 0.8 - \xi, & 0 \leq \xi \leq 0.4 \\ -2.6 + 2 \cdot \xi, & 0.4 < \xi \leq 1 \end{cases}$

The start of the iterations is  $x^\top = p^\top = \lambda^\top = (1, \dots, 1)^\top$ . The solution for formulation (0) is constant zero. The solution for formulation (1) is plotted in figure 5.2.

We observe that the KKT iteration does not converge for the choice  $A = D$  ( $i = 0$ ), although the iterations in the components (forward, adjoint, design) are convergent. This is in line with theorems 2.3 and 2.4, which state that each spectral radius of the components has to be below a certain limit, which may be significantly below 1, in order to guarantee the convergence of the overall KKT iteration.

Furthermore, we observe the effect of choosing  $B = S$ : The convergence deteriorates significantly, i.e., the choice  $B = S_A$  (resp.  $B \approx S_A$ ) is significantly better than  $B \approx S_E$ . ???

Finally, we give the history of the residuals and the respective eigenvalue distribution ??? weg ???for the cases  $i = 3$  and  $B = S_A$  resp.  $B = S_E$  in figure 5.3.

**6. Conclusions.** ??? stark gekuerzt

The aim of this paper is to investigate defect correcting iterations of the type (1.10) for

case  $i = 3$ ,  $B = S_A$  ■  
 case  $i = 3$ ,  $B = S_E$  ■

FIG. 5.3. *Convergence history and eigenvalues for  $N = 101$*

the solution of linear-quadratic optimization problems. Theoretical foundations for practically well-established iterations are given and the following facts are established:

- Iteration (1.10) is convergent if  $(A_f, A_a, B)$  are close enough to  $(C_x, C_x^\top, S_A)$ .
- *What matrix should be chosen as the matrix  $B$ , i.e., as a preconditioner in the Schur complement system?*

We have found theoretical justification for the statement that  $B$  should be chosen close to  $S_A$ , rather than  $S$  which would be the canonical first guess.

- *Are there examples which satisfy the restrictive assumptions of the convergence theorems?*

In the numerical results section, examples are given which show that the convergence theorems do not talk about the empty set.

Although presenting theoretical investigations into iterations of type (1.10) and giving some positive theoretical results, we observe, that there is still some theoretical gap, insofar, as there are many more cases, where the iteration converges numerically than where the assumptions of the convergence theorems are satisfied.

**Acknowledgements.** The third author wishes to thank the university of Graz for providing support for his stay at the university of Graz, during which most of the ideas in this paper have been developed.

#### REFERENCES

- [BNS95] L.T. Biegler, J. Nocedal, and C. Schmid, *A reduced Hessian method for large-scale constrained optimization*, SIAM Journal on Optimization **5** (1995), no. 2, 314–347.
- [BS01] A. Battermann and E. Sachs, *Block preconditioners for KKT systems in PDE-governed optimal control problems*, Fast solution of discretized optimization problems (K.-H. Hoffmann, R.H.W. Hoppe, and V. Schulz, eds.), ISNM, no. 138, Birkhaeuser, 2001, pp. 1–18.
- [BWY90] R. Bank, B. Welfert, and H. Yserentant, *A class of iterative methods for solving saddle point problems*, Numer. Math. **56** (1990), 645–666.
- [GS05] I. Gherman and V. Schulz, *Preconditioning of one-shot pseudo-timestepping methods for shape optimization*, PAMM **5** (2005), no. 1, 741–741.
- [Hei96] M. Heinkenschloss, *Projected sequential quadratic programming methods*, SIAM Journal on Optimization **6** (1996), 373–417.
- [HV01] M. Heinkenschloss and L. Vicente, *Analysis of inexact trust-region sqp algorithms*, SIAM Journal on Optimization **12** (2001), 283–302.
- [KS92] K. Kunisch and E. Sachs, *Reduced sqp methods for parameter identification problems*, SIAM Journal on Numerical Analysis **29** (1992), 1793–1820.
- [KS93] F.-S. Kupfer and E.W. Sachs, *Reduced SQP methods for nonlinear heat conduction control problems*, International Series of Numerical Mathematics **111** (1993), 145–160.
- [Sch97] V.H. Schulz, *Solving discretized optimization problems by partially reduced SQP methods*, Comput. Vis. Sci. **1** (1997), no. 2, 83–96.