SIAM J. OPTIM.
Vol. 12, No. 1, pp. 188–208

MINIMIZING A QUADRATIC OVER A SPHERE*

WILLIAM W. HAGER†

Abstract. A new method, the sequential subspace method (SSM), is developed for the problem of minimizing a quadratic over a sphere. In our scheme, the quadratic is minimized over a subspace which is adjusted in successive iterations to ensure convergence to an optimum. When a sequential quadratic programming iterate is included in the subspace, convergence is locally quadratic. Numerical comparisons with other recent methods are given.

Key words. trust region subproblem, large-scale optimization, sparse optimization, quadratic optimization, quadratic programming, minimal residual, preconditioning, Krylov space, Arnoldi orthogonalization, symmetric successive overrelaxation, Gauss–Seidel

AMS subject classifications. 90C20, 65F10, 65Y20

PII. S1052623499356071

1. Introduction.

In this paper we consider the following problem of minimizing a quadratic over a sphere:

$$\text{minimize } \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} - 2\mathbf{b}^{\mathsf{T}}\mathbf{x} \quad \text{subject to } \|\mathbf{x}\| \le r, \tag{1.1}$$

where $\mathbf{A}$ is a symmetric $n \times n$ matrix, $\mathbf{b} \in \mathbf{R}^n$, T denotes transpose, and $\|\cdot\|$ is the Euclidean norm. This minimization problem is often called the trust region subproblem since it must be solved in each step of a trust region algorithm [1, 2, 3, 15, 19]. Problems of this form arise in many other applications including regularization methods for ill-posed problems [14, 26] and graph partitioning problems [10].

Although the solution to (1.1) can be expressed in terms of a diagonalization of $\mathbf{A}$, this representation is practical only when n is small. In this paper, we focus on the large-scale case. One approach to the large-scale case, developed by Golub and von Matt in [5] (also see [4]), is to (partially) tridiagonalize $\mathbf{A}$ using the Lanczos process and then solve tridiagonal problems to obtain an approximate solution to (1.1). For further developments of this approach, including preconditioning and a Fortran 90 implementation HSL_VF05 in the Harwell subroutine library, see Gould et al. [7]. For the method developed in this paper, we use an approach in the spirit of the Golub/von Matt/Gould et al. scheme to obtain a starting guess.

Parametric eigenvalue approaches to the sphere constrained problem (1.1) are developed by Sorensen [24] and by Rendl and Wolkowicz [20]. The relationship between these two approaches is discussed in detail in [20]. Roughly, Sorensen's approach involves constructing an approximation to the solution of (1.1) from the solution to a related eigenvalue problem. Since this approximation may not satisfy the bound on the norm of the solution, a series of eigenvalue problems are solved, and in the limit, the bound on the norm of the solution is fulfilled. In the approach of Rendl and Wolkowicz, the same eigenvalue problem is solved in each iteration; however, the bound on the norm of the solution is satisfied by maximizing a related dual function. The eigenvalue problems arising in either approach can be solved using Arnoldi

*Received by the editors May 10, 1999; accepted for publication (in revised form) November 7, 2000; published electronically July 2, 2001. This work was supported by the National Science Foundation.

http://www.siam.org/journals/siopt/12-1/35607.html

†Department of Mathematics, University of Florida, Gainesville, FL 32611 (hager@math.ufl.edu, http://www.math.ufl.edu/∼hager).

techniques such as those developed in [13]. In the "hard case" (see [16]), where $\mathbf{b}$ is orthogonal to the eigenvectors associated with the smallest eigenvalue of $\mathbf{A}$, Sorensen's approach needs to be modified. An efficient algorithm for the hard case is developed by Rojas in her thesis [21]. She also uses this algorithm to solve some difficult ill-posed problems of Hansen [11, 12]. The approach of Rendl and Wolkowicz does not need modification in the hard case; however, the convergence of algorithms for the eigenvalue problem may be slower when the computed eigenvalue is not simple.

The approach in this paper, which we call the sequential subspace method (SSM), involves solving (1.1) with the additional constraint that $\mathbf{x}$ is contained in a subspace. We show that convergence is locally quadratic (locally cubic when $\mathbf{b} = \mathbf{0}$) if the subspace contains the iterate generated by one step of the sequential quadratic programming (SQP) algorithm applied to (1.1). The convergence is quadratic even when the original problem is degenerate with multiple solutions and with a singular Jacobian for the first-order optimality system. Descent of the cost at a nonoptimal point can be ensured by including in the subspace either the cost gradient or an eigenvector associated with the smallest eigenvalue of $\mathbf{A}$. We observe in numerical experiments that appropriate small dimensional subspaces are generated by preconditioned Krylov space and minimum residual techniques. Comparisons with the algorithms of Sorensen [24], Rendl and Wolkowicz [20], and Gould et al. [7] are given in section 5.

A solution of the problem

$$\text{minimize } \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} \quad \text{subject to } \|\mathbf{x}\| = r \tag{1.2}$$

is any eigenvector associated with the smallest eigenvalue of $\mathbf{A}$. In comparing the SSM approach to algorithms for solving the eigenproblem, it follows from the discussion of Sleijpen and Van der Vorst in [22] that an SQP iterate for (1.2) is closely connected to the Rayleigh quotient iteration [18, p. 70], which is cubically convergent [18, p. 73]. In [22] approximate solutions to the SQP system are used to build up subspaces containing the approximation to the eigenvector. In this paper, we solve the SQP system relatively precisely, and we form a small dimensional subspace containing the SQP iterate. After computing the new approximation in the subspace, the previous information is discarded; hence, the computer memory requirements are relatively small.

2. Complete diagonalization. If there exists a solution $\mathbf{y}$ of (1.1) with $\|\mathbf{y}\| < r$, then $\mathbf{A}$ is positive semidefinite and $\mathbf{y}$ is the global minimizer of the quadratic $\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} - 2\mathbf{b}^{\mathsf{T}}\mathbf{x}$. Thus, when a minimizer of (1.1) lies in the interior of the constraining sphere, the constraint can be ignored and the optimization problem can be approached using techniques for unconstrained optimization. Consequently, we restrict our attention to the following equality constrained problem:

$$\text{minimize } \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} - 2\mathbf{b}^{\mathsf{T}}\mathbf{x} \quad \text{subject to } \|\mathbf{x}\| = r. \tag{2.1}$$

The solutions to (2.1) are characterized by the following result (see [23, Lemmas 2.4 and 2.8]).

LEMMA 2.1. *The vector* $\mathbf{x}$ *is a solution of* (2.1) *if and only if* $\|\mathbf{x}\| = r$ *and there exists* μ *such that* $\mathbf{A} + \mu\mathbf{I}$ *is positive semidefinite and* $(\mathbf{A} + \mu\mathbf{I})\mathbf{x} = \mathbf{b}$.

The solution to (2.1) can be expressed in terms of the eigenpairs of $\mathbf{A}$. Let $\mathbf{A} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^{\mathsf{T}}$ be a diagonalization of $\mathbf{A}$, where $\mathbf{\Lambda}$ is a diagonal matrix with diagonal elements $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ and $\mathbf{\Phi}$ is the matrix whose columns $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_n$ are orthonormal eigenvectors of $\mathbf{A}$. Defining $\beta_i = \mathbf{b}^{\mathsf{T}}\boldsymbol{\phi}_i$, $\mathcal{E}_1 = \{i : \lambda_i = \lambda_1\}$, and $\mathcal{E}_+ = \{i : \lambda_i > \lambda_1\}$, Lemma 2.1 yields the following.

LEMMA 2.2. *The vector* $\boldsymbol{\phi} = \sum_{i=1}^{n} c_i \boldsymbol{\phi}_i$ *is a solution of* (2.1) *if and only if* $\mathbf{c}$ *is chosen in the following way:*

(a) Degenerate case: *If* $\beta_i = 0$ *for all* $i \in \mathcal{E}_1$ *and*

$$\sum_{i \in \mathcal{E}_+} \frac{\beta_i^2}{(\lambda_i - \lambda_1)^2} \le r^2, \tag{2.2}$$

then $\mu = -\lambda_1$ *in Lemma* 2.1 *and* $c_i = \beta_i/(\lambda_i - \lambda_1)$ *for* $i \in \mathcal{E}_+$*; the* c_i *for* $i \in \mathcal{E}_1$ *are arbitrary scalars satisfying the condition*

$$\sum_{i \in \mathcal{E}_1} c_i^2 = r^2 - \sum_{i \in \mathcal{E}_+} \frac{\beta_i^2}{(\lambda_i - \lambda_1)^2}.$$

(b) Nondegenerate case: *If* (a) *does not hold, then* $c_i = \beta_i/(\lambda_i + \mu)$, $1 \le i \le n$, *where* $\mu > -\lambda_1$ *is chosen so that*

$$\sum_{i=1}^{n} \frac{\beta_i^2}{(\lambda_i + \mu)^2} = r^2. \tag{2.3}$$

Proof. Simply check that the sufficient optimality conditions of Lemma 2.1 are satisfied. The degenerate case, where the Jacobian of the first-order optimality system may be singular, coincides with the "hard case" of Moré and Sorensen [16], where $\mathbf{b}$ is orthogonal to the eigenspace associated with the smallest eigenvalue of $\mathbf{A}$ and the multiplier μ is equal to $-\lambda_1$. In the nondegenerate case, the multiplier μ is chosen so that $\mathbf{A} + \mu\mathbf{I}$ is positive definite and the solution $\mathbf{x} = \mathbf{x}(\mu)$ to $(\mathbf{A} + \mu\mathbf{I})\mathbf{x} = \mathbf{b}$ satisfies the constraint $\mathbf{x}^{\mathsf{T}}\mathbf{x} = r^2$. □

In the nondegenerate case, (2.3) leads to upper and lower bounds for the multiplier μ. Since $\lambda_i + \mu \ge \lambda_1 + \mu > 0$, $1 \le i \le n$, we have

$$r^2 = \sum_{i=1}^{n} \frac{\beta_i^2}{(\lambda_i + \mu)^2} \le \sum_{i=1}^{n} \frac{\beta_i^2}{(\lambda_1 + \mu)^2} = \frac{\|\mathbf{b}\|^2}{(\lambda_1 + \mu)^2}.$$

Since $\lambda_1 + \mu > 0$, it follows that

$$\mu \le \frac{\|\mathbf{b}\|}{r} - \lambda_1 := \mu_u. \tag{2.4}$$

To obtain a lower bound, observe that

$$r^2 = \sum_{i=1}^{n} \frac{\beta_i^2}{(\lambda_i + \mu)^2} \ge \frac{1}{(\lambda_1 + \mu)^2} \sum_{i \in \mathcal{E}_1} \beta_i^2,$$

which yields the relation

$$\mu \ge -\lambda_1 + \frac{1}{r} \left(\sum_{i \in \mathcal{E}_1} \beta_i^2 \right)^{1/2} := \mu_l. \tag{2.5}$$

Utilizing the upper and lower bounds μ_u and μ_l and the strict convexity of the left side of (2.3) on the interval $(\mu_l, \mu_u]$, it is easy to devise efficient algorithms to compute a solution μ of (2.3).

3. Incomplete diagonalization; local convergence. At iteration k in the SSM for (2.1), we impose the additional constraint that $\mathbf{x}$ lies in a subspace $\mathcal{S}_k$ of $\mathbf{R}^n$. Hence, the new iterate $\mathbf{x}_{k+1}$ is a solution of the problem

$$\text{minimize } \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} - 2\mathbf{b}^\mathsf{T}\mathbf{x} \quad \text{subject to } \|\mathbf{x}\| = r, \quad \mathbf{x} \in \mathcal{S}_k. \tag{3.1}$$

We show that the convergence is locally quadratic, even when the original problem (2.1) is degenerate, if we include an SQP iterate associated with $\mathbf{x}_k$ in $\mathcal{S}_k$.

If $\mathbf{V}$ is an $n \times l$ matrix with orthonormal columns that span $\mathcal{S}_k$, then (3.1) is equivalent to the problem

$$\text{minimize } \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} - 2\mathbf{b}^\mathsf{T}\mathbf{x} \quad \text{subject to } \|\mathbf{x}\| = r, \quad \mathbf{x} = \mathbf{V}\mathbf{y}. \tag{3.2}$$

After substituting for $\mathbf{x}$, (3.2) reduces to the following problem in $\mathbf{R}^l$:

$$\text{minimize } \mathbf{y}^\mathsf{T}\mathbf{B}\mathbf{y} - 2\mathbf{c}^\mathsf{T}\mathbf{y} \quad \text{subject to } \|\mathbf{y}\| = r, \tag{3.3}$$

where $\mathbf{B} = \mathbf{V}^\mathsf{T}\mathbf{A}\mathbf{V}$ and $\mathbf{c} = \mathbf{V}^\mathsf{T}\mathbf{b}$. If l is small, then (3.3) can be solved by complete diagonalization as in section 2 or, if $\mathbf{B}$ has a sparse factorization, then (3.3) can be solved quickly using the Newton approach developed in [16].

In theory, a tridiagonal $\mathbf{B}$ is generated using the Lanczos process [6]. In particular, if $\mathbf{v}_1$ is a unit vector and $\mathbf{v}_i$ is the ith column of $\mathbf{V}$, then the Lanczos process can be expressed as follows.

ALGORITHM 1 (LANCZOS).
$u_0 = 0$
for $j = 1 : l - 1$
  $\mathbf{s} \leftarrow \mathbf{A}\mathbf{v}_j$
  $d_j \leftarrow \mathbf{s}^\mathsf{T}\mathbf{v}_j$
  $\mathbf{s} \leftarrow \mathbf{s} - d_j\mathbf{v}_j - u_{j-1}\mathbf{v}_{j-1}$
  $u_j \leftarrow \|\mathbf{s}\|$
  $\mathbf{v}_{j+1} \leftarrow \mathbf{s}/u_j$
end
END ALGORITHM 1

Here $\mathbf{d}$ is the diagonal and $\mathbf{u}$ is the superdiagonal of the tridiagonal matrix $\mathbf{B}$. If $u_j = 0$ for some j, then the Lanczos process is terminated and the column spaces of $\mathbf{V}$ and $\mathbf{AV}$ coincide.

It is well known that the columns of $\mathbf{V}$ generated by this process may deviate significantly from orthogonality due to the propagation of rounding errors. When this happens, (3.2) is no longer equivalent to (3.3). Nonetheless, Gould et al. observe in [7] that the solution to (3.3) often provides a good approximation to the solution of (3.2) despite the loss of orthogonality. The Lanczos process can be repaired, in order to restore orthogonality, by using a Householder process to generate the columns of $\mathbf{V}$. This process, however, requires products between a vector and each of the previously computed columns of $\mathbf{V}$. Thus, the overhead needed to maintain orthogonality grows as nl^2 in the number of flops and as nl in storage. This overhead can be significant when n or l is large. On the other hand, to compute a high accuracy solution, we need to maintain orthogonality in order to obtain an equivalent problem (3.3). This leads us to focus on approaches that involve subspaces where l is much smaller than n. In particular, for an implementation (Algorithm 4) of the SSM proposed later, l is either 4 or 5.

Since SQP techniques often converge rapidly, with a good starting guess, we always include the SQP approximation in the subspace $\mathcal{S}_k$. The SQP method is equivalent to Newton's method applied to the nonlinear system

$$(\mathbf{A}+\mu\mathbf{I})\mathbf{x}-\mathbf{b}=\mathbf{0}, \quad \frac{1}{2}\mathbf{x}^{\mathsf{T}}\mathbf{x}-\frac{1}{2}r^2=0. \tag{3.4}$$

If $\mathbf{x}_k$ is the current iterate, which we assume satisfies the constraint $\|\mathbf{x}\|=r$, and μ_k is the current approximation to the multiplier associated with the constraint, then the Newton iterate can be expressed in the following way: $\mathbf{x}_{\rm SQP}=\mathbf{z}+\mathbf{x}_k$ and $\mu_{\rm SQP}=\mu_k+\nu$, where $\mathbf{z}$ and ν are solutions of the linear system

$$(\mathbf{A}+\mu_k\mathbf{I})\mathbf{z}+\mathbf{x}_k\nu=\mathbf{b}-(\mathbf{A}+\mu_k\mathbf{I})\mathbf{x}_k, \tag{3.5}$$

$$\mathbf{x}_k^{\mathsf{T}}\mathbf{z}=0. \tag{3.6}$$

When the coefficient matrix in (3.5)–(3.6) is singular, we let $(\mathbf{z},\nu)$ be the minimum residual/minimum norm solution; that is, $(\mathbf{z},\nu)$ is obtained (in theory) by multiplying the right side by the pseudoinverse of the coefficient matrix (see [8]).

A solution $\mathbf{x}_{k+1}$ to the subspace problem (3.1) is an approximation to the solution of (2.1). To obtain an estimate for the multiplier of Lemma 2.1, we minimize the Euclidean norm of the residual $\mathbf{b}-\mathbf{A}\mathbf{x}_{k+1}-\mu\mathbf{x}_{k+1}$ over the scalar μ. This works out to give

$$\mu_{k+1}=\rho(\mathbf{x}_{k+1}), \quad \text{where } \rho(\mathbf{x})=\frac{(\mathbf{b}-\mathbf{A}\mathbf{x})^{\mathsf{T}}\mathbf{x}}{\|\mathbf{x}\|^2}. \tag{3.7}$$

This is the standard least squares approximation to the solution of the overdetermined linear system $\mu\mathbf{x}_{k+1}=\mathbf{b}-\mathbf{A}\mathbf{x}_{k+1}$.

We now examine the local convergence of a solution $\mathbf{x}_{k+1}$ of (3.1) and the multiplier estimate (3.7) under the assumption that $\mathcal{S}_k$ contains $\mathbf{x}_{\rm SQP}=\mathbf{z}+\mathbf{x}_k$, where $(\mathbf{z},\nu)$ is a solution to (3.5). Let $\mathcal{S}^*$ denote the set of minimizers of (2.1), and let μ^* be the multiplier given by Lemma 2.1. In the nondegenerate setting, where $\mathbf{A}+\mu^*\mathbf{I}$ is positive definite, we show that the iteration is locally, quadratically convergent to the unique solution of (2.1). In the degenerate case $\mu^*=-\lambda_1$, where $\mathcal{S}^*$ has more than one element, we obtain local quadratic convergence to $\mathcal{S}^*$, where distance is measured in the usual way:

$$\operatorname{dist}(\mathbf{x},\mathcal{S}^*)=\inf\{\|\mathbf{x}-\mathbf{x}^*\| : \mathbf{x}^*\in\mathcal{S}^*\}.$$

In the nondegenerate-degenerate case, where $\mu^*=-\lambda_1$ but $\mathcal{S}^*$ contains a single element, we obtain local quadratic convergence for a "safe-guarded" choice of μ_k. Our convergence result in the special nondegenerate-degenerate case is given later in Lemma 3.4, while our local convergence result in either the nondegenerate case or the degenerate case with multiple solutions is as follows.

THEOREM 3.1. *Let μ^* be the multiplier of Lemma* 2.1 *associated with the set of solutions $\mathcal{S}^*$ of* (2.1)*, and suppose that either $\mathbf{A}+\mu^*\mathbf{I}$ is positive definite or $\mu^*=-\lambda_1$ with* (2.2) *a strict inequality. Then there exist positive constants η and C with the property that for any $(\mathbf{x}_k,\mu_k)$ such that*

$$|\mu_k-\mu^*|+\operatorname{dist}(\mathbf{x}_k,\mathcal{S}^*)\le\eta, \quad \|\mathbf{x}_k\|=r,$$

and for any subspace $\mathcal{S}_k$ that contains the SQP *iterate $\mathbf{x}_{\rm SQP}$ associated with* (3.5)–(3.6)*, any solution $\mathbf{x}_{k+1}$ of* (3.1) *and associated multiplier μ_{k+1} given by* (3.7) *satisfy*

the estimate

$$\operatorname{dist}(\mathbf{x}_{k+1}, \mathcal{S}^*) + |\mu_{k+1} - \mu_*| \le C\big(\operatorname{dist}(\mathbf{x}_k, \mathcal{S}^*)^2 + |\mu_k - \mu^*|^2\big).$$

The eigenvalue problem (1.2), corresponding to $\mathbf{b} = \mathbf{0}$, is always degenerate (with multiple solution) and the error has the special form

$$\operatorname{dist}(\mathbf{x}_{k+1}, \mathcal{S}^*) \le C|\mu_k + \lambda_1|\operatorname{dist}(\mathbf{x}_k, \mathcal{S}^*).$$

When the multiplier is estimated using (3.7), it can be shown, when $\mathbf{b} = \mathbf{0}$, that the error in the multiplier is bounded by a constant times the error in the solution vector squared (see the remark at the end of section 3.1). It follows that for some constant C,

$$\operatorname{dist}(\mathbf{x}_{k+1}, \mathcal{S}^*) \le C\operatorname{dist}(\mathbf{x}_k, \mathcal{S}^*)^3 \quad \text{and} \quad |\mu_{k+1} + \lambda_1| \le C\operatorname{dist}(\mathbf{x}_k, \mathcal{S}^*)^6, \tag{3.8}$$

which is the same as the convergence result for the Rayleigh quotient iteration.

3.1. Nondegenerate problems. We begin the derivation of Theorem 3.1 with the nondegenerate case.

LEMMA 3.2. *If* (2.1) *has a solution* $\mathbf{x}^*$ *and an associated multiplier* μ^* *with* $\mu^* > -\lambda_1$, *then there exist a neighborhood* $\mathcal{N}$ *of* $(\mathbf{x}^*, \mu^*)$ *and a constant* C *with the property that for any* $(\mathbf{x}_k, \mu_k) \in \mathcal{N}$ *with* $\|\mathbf{x}_k\| = r$, *and for any subspace* $\mathcal{S}_k$ *that contains the* SQP *iterate* $\mathbf{x}_{\mathrm{SQP}}$ *associated with* (3.5)–(3.6), *any solution* $\mathbf{x}_{k+1}$ *of* (3.1) *and associated multiplier* μ_{k+1} *given by* (3.7) *satisfy the estimate*

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| + |\mu_{k+1} - \mu^*| \le C(\|\mathbf{x}_k - \mathbf{x}^*\|^2 + |\mu_k - \mu^*|^2).$$

Proof. Since $\mu^* > -\lambda_1$, the matrix $\mathbf{A} + \mu^*\mathbf{I}$ is positive definite, and the Jacobian of the nonlinear system (3.4) is nonsingular at $(\mathbf{x}^*, \mu^*)$. By the standard convergence theorem for Newton's method applied to a smooth system of equations, there exist a neighborhood $\mathcal{N}$ of $(\mathbf{x}_k, \mu_k)$ and a constant c such that

$$\|\mathbf{x}_{\mathrm{SQP}} - \mathbf{x}^*\| + |\mu_{\mathrm{SQP}} - \mu^*| \le c(\|\mathbf{x}_k - \mathbf{x}^*\|^2 + |\mu_k - \mu^*|^2)$$

whenever $(\mathbf{x}_k, \mu_k) \in \mathcal{N}$.

Let α and β be positive scalars chosen so that

$$\alpha\|\mathbf{x}\|^2 \le \mathbf{x}^{\mathsf{T}}(\mathbf{A} + \mu^*\mathbf{I})\mathbf{x} \le \beta\|\mathbf{x}\|^2 \tag{3.9}$$

for all $\mathbf{x} \in \mathbf{R}^n$, let f be the cost function in (2.1), $f(\mathbf{x}) = \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} - 2\mathbf{b}^{\mathsf{T}}\mathbf{x}$, and let $\mathcal{L}$ be the Lagrangian defined by

$$\mathcal{L}(\mathbf{x}, \mu) = f(\mathbf{x}) + \mu(\mathbf{x}^{\mathsf{T}}\mathbf{x} - r^2).$$

A Taylor expansion around $\mathbf{x}^*$ yields the relation

$$f(\mathbf{x}) = \mathcal{L}(\mathbf{x}, \mu^*) = f(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^{\mathsf{T}}(\mathbf{A} + \mu^*\mathbf{I})(\mathbf{x} - \mathbf{x}^*)$$

for any $\mathbf{x} \in \mathcal{B}_r = \{\mathbf{x} \in \mathbf{R}^n : \|\mathbf{x}\| = r\}$. Combining this with (3.9) gives

$$\alpha\|\mathbf{x} - \mathbf{x}^*\|^2 \le f(\mathbf{x}) - f(\mathbf{x}^*) \le \beta\|\mathbf{x} - \mathbf{x}^*\|^2 \tag{3.10}$$

for any $\mathbf{x} \in \mathcal{B}_r$.

If $\mathbf{p}$ is the projection of $\mathbf{x}_{\rm SQP}$ onto $\mathcal{B}_r$, then

$$\|\mathbf{x}_{\rm SQP} - \mathbf{p}\| \le \|\mathbf{x}_{\rm SQP} - \mathbf{x}^*\| \le c(\|\mathbf{x}_k - \mathbf{x}^*\|^2 + |\mu_k - \mu^*|^2). \tag{3.11}$$

Hence, we have

$$\|\mathbf{p} - \mathbf{x}^*\| \le \|\mathbf{p} - \mathbf{x}_{\rm SQP}\| + \|\mathbf{x}_{\rm SQP} - \mathbf{x}^*\| \le 2c(\|\mathbf{x}_k - \mathbf{x}^*\|^2 + |\mu_k - \mu^*|^2).$$

Since $\mathbf{p} = \gamma \mathbf{x}_{\rm SQP}$ for some γ, it follows that $\mathbf{p} \in \mathcal{S}_k$ and $f(\mathbf{x}_{k+1}) \le f(\mathbf{p})$. Combining this inequality with (3.10) and (3.11) gives

$$\begin{aligned}\alpha\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 &\le f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)\\ &\le f(\mathbf{p}) - f(\mathbf{x}^*)\\ &\le \beta\|\mathbf{p} - \mathbf{x}^*\|^2\\ &\le 4c^2\beta(\|\mathbf{x}_k - \mathbf{x}^*\|^2 + |\mu_k - \mu^*|^2)^2,\end{aligned}$$

which implies that

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \le 2c\sqrt{\beta/\alpha}(\|\mathbf{x}_k - \mathbf{x}^*\|^2 + |\mu_k - \mu^*|^2). \tag{3.12}$$

Since $\mathbf{b} = (\mathbf{A} + \mu^*\mathbf{I})\mathbf{x}^*$, we have, for any $\mathbf{x} \in \mathcal{B}_r$,

$$\begin{aligned}r^2\rho(\mathbf{x}) &= (\mathbf{b} - \mathbf{A}\mathbf{x})^{\mathsf T}\mathbf{x} = ((\mathbf{A} + \mu^*\mathbf{I})\mathbf{x}^* - \mathbf{A}\mathbf{x})^{\mathsf T}\mathbf{x}\\ &= \mathbf{x}^{\mathsf T}(\mathbf{A} + \mu^*\mathbf{I})(\mathbf{x}^* - \mathbf{x}) + \mu^* r^2.\end{aligned} \tag{3.13}$$

Making this substitution gives

$$|\mu_{k+1} - \mu^*| = |\rho(\mathbf{x}_{k+1}) - \mu^*| \le \frac{\lambda_n + \mu^*}{r}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|. \tag{3.14}$$

Combining (3.14) with (3.12), the proof is complete. ∎

Remark. For the eigenvalue problem (1.2), we have $\mathbf{x}^* = r\boldsymbol{\phi}_1$, $\mu^* = -\lambda_1$, and $\boldsymbol{\phi}_1^{\mathsf T}(\mathbf{A} - \lambda_1\mathbf{I}) = \mathbf{0}$. In this case, (3.13) yields

$$r^2\rho(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^*)^{\mathsf T}(\mathbf{A} - \lambda_1\mathbf{I})(\mathbf{x}^* - \mathbf{x}) - \lambda_1 r^2,$$

and (3.14) becomes

$$|\mu_{k+1} - \mu^*| \le \frac{\lambda_n - \lambda_1}{r^2}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2.$$

3.2. Degenerate problems. Now consider local convergence in the degenerate case where $\mu^* = -\lambda_1$. Referring to Lemma 2.2, the degenerate case can happen only when $\beta_i = \mathbf{b}^{\mathsf T}\boldsymbol{\phi}_i = 0$ for all $i \in \mathcal{E}_1$. Any solution to (2.1) in the degenerate case can be expressed as $\mathbf{x}^* = \boldsymbol{\Phi}_1 + \boldsymbol{\Phi}_+$, where

$$\boldsymbol{\Phi}_+ = \sum_{i \in \mathcal{E}_+} c_i\boldsymbol{\phi}_i, \quad c_i = \beta_i/(\lambda_i - \lambda_1), \tag{3.15}$$

and $\boldsymbol{\Phi}_1$ is any linear combination of the vectors $\boldsymbol{\phi}_i$, $i \in \mathcal{E}_1$, satisfying the relation

$$\|\boldsymbol{\Phi}_1\|^2 + \|\boldsymbol{\Phi}_+\|^2 = r^2.$$

Initially, we suppose that $\|\mathbf{\Phi}_1\| = \delta > 0$, in which case the projection of $\mathcal{S}^*$ on the eigenspace associated with $\mathcal{E}_1$ contains a sphere of radius δ. Our convergence result is the following.

LEMMA 3.3. *Suppose that the multiplier μ^* of Lemma 2.1 associated with the set of solutions $\mathcal{S}^*$ of (2.1) is given by $\mu^* = -\lambda_1$ and that $\|\mathbf{\Phi}_1\| = \delta > 0$, where $\mathbf{\Phi}_1$ is the component of an element of $\mathcal{S}^*$ in the eigenspace associated with $\mathcal{E}_1$. Then there exist positive constants η and C with the property that for any $(\mathbf{x}_k, \mu_k)$ such that*

$$|\mu_k + \lambda_1| + \operatorname{dist}(\mathbf{x}_k, \mathcal{S}^*) \leq \eta, \quad \|\mathbf{x}_k\| = r,$$

and for any subspace $\mathcal{S}_k$ that contains the SQP *iterate $\mathbf{x}_{\mathrm{SQP}}$ associated with (3.5)–(3.6), any solution $\mathbf{x}_{k+1}$ of (3.1) and associated multiplier μ_{k+1} given by (3.7) satisfy the estimate*

$$\operatorname{dist}(\mathbf{x}_{k+1}, \mathcal{S}^*) + |\mu_{k+1} + \lambda_1| \leq C\big(\operatorname{dist}(\mathbf{x}_k, \mathcal{S}^*)^2 + |\mu_k + \lambda_1|^2\big).$$

Proof. Initially, let us assume that μ_k is near $-\lambda_1$, but $\mu_k \neq -\lambda_1$. In this case, the linear system (3.5)–(3.6) is nonsingular, and there exists a unique solution $(\mathbf{z}, \nu)$. We expand $\mathbf{z}$ and $\mathbf{x}_k$ in terms of the eigenvectors of $\mathbf{A}$ writing $\mathbf{z} = \sum_{i=1}^n \zeta_i \boldsymbol{\phi}_i$ and $\mathbf{x}_k = \sum_{i=1}^n \chi_i \boldsymbol{\phi}_i$. Utilizing (3.5), we obtain

$$\zeta_i = \frac{-\chi_i \nu}{\lambda_i + \mu_k} + \frac{\beta_i - (\lambda_i + \mu_k)\chi_i}{\lambda_i + \mu_k}. \tag{3.16}$$

Substituting this in (3.6) gives

$$\nu = \frac{\sum_{i=1}^n \chi_i(\beta_i - (\lambda_i + \mu_k)\chi_i)/(\lambda_i + \mu_k)}{\sum_{i=1}^n \chi_i^2/(\lambda_i + \mu_k)}. \tag{3.17}$$

Let us define $\mathbf{R} = \mathbf{b} - (\mathbf{A} + \mu_k \mathbf{I})\mathbf{x}_k$ and $\rho_i = \mathbf{R}^{\mathsf{T}} \boldsymbol{\phi}_i$. For $i \in \mathcal{E}_1$, $\beta_i = 0$ and

$$\nu = \frac{-(\lambda_1 + \mu_k)\left(\sum_{i \in \mathcal{E}_1} \chi_i^2 + \sum_{i \in \mathcal{E}_+} \frac{\chi_i \rho_i}{\lambda_i + \mu_k}\right)}{\sum_{i \in \mathcal{E}_1} \chi_i^2 + (\lambda_1 + \mu_k) \sum_{i \in \mathcal{E}_+} \frac{\chi_i^2}{\lambda_i + \mu_k}}. \tag{3.18}$$

If $\mathbf{x}^* \in \mathcal{S}^*$, then since $\mathbf{b} = (\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{x}^*$ and $\|\mathbf{x}_k\| = r$, we have

$$\begin{aligned} \|\mathbf{R}\| = \|\mathbf{b} - (\mathbf{A} + \mu_k \mathbf{I})\mathbf{x}_k\| &\leq r|\lambda_1 + \mu_k| + \|\mathbf{A} - \lambda_1 \mathbf{I}\|\|\mathbf{x}_k - \mathbf{x}^*\| \\ &\leq \max\{r, \|\mathbf{A} - \lambda_1 \mathbf{I}\|\}(|\lambda_1 + \mu_k| + \|\mathbf{x}_k - \mathbf{x}^*\|). \end{aligned} \tag{3.19}$$

Let ϵ_k be the error at step k defined by

$$\epsilon_k = |\lambda_1 + \mu_k| + \operatorname{dist}(\mathbf{x}_k, \mathcal{S}^*).$$

By (3.19), we have $\|\mathbf{R}\| = O(\epsilon_k)$, while (3.18) gives

$$\nu = -(\lambda_1 + \mu_k)(1 + O(\epsilon_k)) \tag{3.20}$$

$$= -(\lambda_1 + \mu_k) + O(\epsilon_k^2) \tag{3.21}$$

since $\sum_{i \in \mathcal{E}_1} \chi_i^2$ is near $\delta^2 > 0$ when $\mathbf{x}_k$ is near $\mathcal{S}^*$. From (3.16), we have

$$\zeta_i + \chi_i = \frac{\beta_i - \chi_i \nu}{\lambda_i + \mu_k}. \tag{3.22}$$

Since $\beta_i = 0$ and $\lambda_i = \lambda_1$ for $i \in \mathcal{E}_1$, (3.20) and (3.22) give

$$\zeta_i + \chi_i = \chi_i + O(\epsilon_k) \quad \text{for } i \in \mathcal{E}_1. \tag{3.23}$$

Let $\mathbf{x}^*$ be the closest element of $\mathcal{S}^*$ to $\mathbf{x}_k$ and define $\chi_i^* = \boldsymbol{\phi}^{\mathsf{T}}\mathbf{x}^*$. Then we have

$$|\chi_i - \chi_i^*| = |(\mathbf{x}_k - \mathbf{x}^*)^{\mathsf{T}}\boldsymbol{\phi}_i| \le \|\mathbf{x}_k - \mathbf{x}^*\| \le \epsilon_k. \tag{3.24}$$

By (3.23) the $\boldsymbol{\phi}_i$ component of $\mathbf{x}_{\text{SQP}} = \mathbf{z} + \mathbf{x}_k$ for $i \in \mathcal{E}_1$ is in error by $O(\epsilon_k)$ since χ_i, the $\boldsymbol{\phi}_i$ component of $\mathbf{x}_k$, is in error by $O(\epsilon_k)$ by (3.24).

Lemma 2.2 implies that $\beta_i = \chi_i^*(\lambda_i - \lambda_1)$ for $i \in \mathcal{E}_+$. Combining this with (3.21) and (3.22) gives

$$\begin{aligned}
\zeta_i + \chi_i &= \frac{\beta_i - \chi_i\nu}{\lambda_i + \mu_k} = \frac{\beta_i + \chi_i(\lambda_1 + \mu_k)}{\lambda_i + \mu_k} + O(\epsilon_k^2) \\
&= \frac{\chi_i^*(\lambda_i - \lambda_1) + \chi_i(\lambda_1 + \mu_k)}{\lambda_i + \mu_k} + O(\epsilon_k^2) \\
&= \frac{\chi_i^*(\lambda_i - \lambda_1) + \chi_i^*(\lambda_1 + \mu_k)}{\lambda_i + \mu_k} + O(\epsilon_k^2) = \chi_i^* + O(\epsilon_k^2). \qquad (3.25)
\end{aligned}$$

Hence, for $i \in \mathcal{E}_+$ the $\boldsymbol{\phi}_i$ component of $\mathbf{x}_{\text{SQP}}$ is in error by $O(\epsilon_k^2)$.

Let $\|\cdot\|_+$ be the seminorm associated with projection into the eigenspace associated with $\mathcal{E}_+$:

$$\|\mathbf{x}\|_+^2 = \sum_{i \in \mathcal{E}_+} (\mathbf{x}^{\mathsf{T}}\boldsymbol{\phi}_i)^2.$$

Then we have

$$(\lambda_+ - \lambda_1)\|\mathbf{x}\|_+^2 \le \mathbf{x}^{\mathsf{T}}(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{x} \le (\lambda_n - \lambda_1)\|\mathbf{x}\|_+^2 \tag{3.26}$$

for all $\mathbf{x} \in \mathbf{R}^n$, where $\lambda_+ = \min\{\lambda_i : \lambda_i > \lambda_1, 1 \le i \le n\}$. Proceeding as we did earlier, but replacing norms with seminorms,

$$\begin{aligned}
\alpha\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_+^2 &\le f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \\
&\le f(\mathbf{p}) - f(\mathbf{x}^*) \\
&\le \beta\|\mathbf{p} - \mathbf{x}^*\|_+^2, \qquad (3.27)
\end{aligned}$$

where $\mathbf{p}$ is the projection of $\mathbf{x}_{\text{SQP}}$ onto the ball $\mathcal{B}_r$, and $\mathbf{p} = \gamma\mathbf{x}_{\text{SQP}}$ for some $\gamma \ge 0$. Since $\|\mathbf{z}\| = O(\epsilon_k)$ by (3.23) and (3.25), and $\mathbf{z}$ is perpendicular to $\mathbf{x}_k$ by (3.6), we have

$$\|\mathbf{x}_{\text{SQP}}\|^2 = \|\mathbf{z} + \mathbf{x}_k\|^2 = \|\mathbf{z}\|^2 + \|\mathbf{x}_k\|^2 = r^2 + O(\epsilon_k^2).$$

This implies that $\|\mathbf{x}_{\text{SQP}}\| = r + O(\epsilon_k^2)$, and $\gamma = 1 + O(\epsilon_k^2)$. For $i \in \mathcal{E}_+$,

$$\mathbf{p}^{\mathsf{T}}\boldsymbol{\phi}_i = \gamma\mathbf{x}_{\text{SQP}}^{\mathsf{T}}\boldsymbol{\phi}_i = (1 + O(\epsilon_k^2))(\zeta_i + \chi_i) = (1 + O(\epsilon_k^2))(\chi_i^* + O(\epsilon_k^2)) = \chi_i^* + O(\epsilon_k^2).$$

Consequently, $\|\mathbf{p} - \mathbf{x}^*\|_+ = O(\epsilon_k^2)$, which combines with (3.27) to give

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_+ = O(\epsilon_k^2). \tag{3.28}$$

By the triangle inequality,

$$\|\mathbf{x}^*\|_+ - O(\epsilon_k^2) \le \|\mathbf{x}_{k+1}\|_+ \le \|\mathbf{x}^*\|_+ + O(\epsilon_k^2).$$

Let $\|\cdot\|_1$ be the seminorm defined by

$$\|\mathbf{x}\|_1^2 = \sum_{i\in\mathcal{E}_1} (\mathbf{x}^\mathsf{T}\boldsymbol{\phi}_i)^2, \tag{3.29}$$

and recall that $\|\mathbf{x}^*\|_1 = \delta$ for any $\mathbf{x}^* \in \mathcal{S}^*$. By the Pythagorean theorem and the fact that $\mathbf{x}_{k+1}$ has length r, we have

$$\|\mathbf{x}_{k+1}\|_1^2 = r^2 - \|\mathbf{x}_{k+1}\|_+^2 = r^2 - \|\mathbf{x}^*\|_+^2 + O(\epsilon_k^2) = \|\mathbf{x}^*\|_1^2 + O(\epsilon_k^2) = \delta^2 + O(\epsilon_k^2),$$

which implies that

$$\|\mathbf{x}_{k+1}\|_1 = \delta + O(\epsilon_k^2). \tag{3.30}$$

The distance from $\mathbf{x}_{k+1}$ to $\mathcal{S}^*$ is given by

$$\operatorname{dist}(\mathbf{x}_{k+1}, \mathcal{S}^*)^2 = \|\mathbf{x}_{k+1} - \mathbf{x}^*\|_+^2 + (\delta - \|\mathbf{x}_{k+1}\|_1)^2, \tag{3.31}$$

where $\mathbf{x}^*$ is any element of $\mathcal{S}^*$. Relations (3.28)–(3.31) yield $\operatorname{dist}(\mathbf{x}_{k+1}, \mathcal{S}^*) = O(\epsilon_k^2)$, while (3.14) gives $|\mu_{k+1} - \mu^*| = O(\epsilon_k^2)$. Combining these estimates, we have $\epsilon_{k+1} = O(\epsilon_k^2)$.

This analysis was given under the assumption that $\mu_k \neq -\lambda_1$. In the special case $\mu_k = -\lambda_1$, we now show how the analysis should be modified. With the change of variables $\mathbf{z} = \sum_{i=1}^n \zeta_i \boldsymbol{\phi}_i$ and the substitution $\mathbf{x}_k = \sum_{i=1}^n \chi_i \boldsymbol{\phi}_i$, the SQP system (3.5)–(3.6) is equivalent, by orthogonal transformation, to

$$\left[\begin{array}{c|c} \mathbf{D} & \boldsymbol{\chi} \\ \boldsymbol{\chi}^\mathsf{T} & \mathbf{0} \end{array}\right] \left[\begin{array}{c} \boldsymbol{\zeta} \\ \nu \end{array}\right] = \left[\begin{array}{c} \boldsymbol{\beta} - \mathbf{D}\boldsymbol{\chi} \\ 0 \end{array}\right], \tag{3.32}$$

where $\mathbf{D}$ is a diagonal matrix with diagonal elements $d_{ii} = \lambda_i - \lambda_1$. If $\mathcal{E}_1$ has s elements, then the first s diagonal elements of $\mathbf{D}$ and the first s components of $\boldsymbol{\beta} - \mathbf{D}\boldsymbol{\chi}$ vanish. Hence, the first s equations in (3.32) imply that $\nu = 0$. The next $n - s$ equations give

$$\zeta_i = -\chi_i + \beta_i/(\lambda_i - \lambda_1) = -\chi_i + \chi_i^*, \quad i \in \mathcal{E}_+, \tag{3.33}$$

while the last equation in (3.32) gives

$$\sum_{i\in\mathcal{E}_1} \chi_i\zeta_i = -\sum_{i\in\mathcal{E}_+} \chi_i\zeta_i.$$

The minimum norm solution to this last equation is

$$\zeta_i = -\left(\frac{\sum_{i\in\mathcal{E}_+} \zeta_i\chi_i}{\sum_{i\in\mathcal{E}_1} \chi_i^2}\right)\chi_i \quad \text{for } i \in \mathcal{E}_1. \tag{3.34}$$

By (3.33), $\zeta_i + \chi_i = \chi_i^*$ and $|\zeta_i| \le \epsilon_k$ for $i \in \mathcal{E}_+$. By (3.34), $|\zeta_i| = O(\epsilon_k)$ for $i \in \mathcal{E}_1$. Combining these bounds, we have $\|\mathbf{z}\| = O(\epsilon_k)$. With these relations, all the analysis from (3.26) onward can be applied, leading us to the estimate $\epsilon_{k+1} = O(\epsilon_k^2)$. ∎

Lemmas 3.2 and 3.3 yield Theorem 3.1.

3.3. Nondegenerate-degenerate problems. Finally, let us consider the nondegenerate-degenerate case, where $\mu^* = -\lambda_1$, $\mathbf{x}^* = \mathbf{\Phi}_1 + \mathbf{\Phi}_+$, and the $\mathbf{\Phi}_1$ component of $\mathbf{x}^*$ in the eigenspace associated with the smallest eigenvalue of $\mathbf{A}$ vanishes. Our convergence result is the following.

LEMMA 3.4. *If* (2.1) *has a solution* $\mathbf{x}^* = \mathbf{\Phi}_+$, *where* $\mathbf{\Phi}_+$ *is given by* (3.15), *then there exist a neighborhood* $\mathcal{N}$ *of* $(\mathbf{x}^*, -\lambda_1)$ *and a constant* C *with the property that for any* $(\mathbf{x}_k, \mu_k) \in \mathcal{N}$ *with*

$$\mu_k \geq -\lambda_1 + \|\mathbf{b} - (\mathbf{A} + \rho(\mathbf{x}_k)\mathbf{I})\mathbf{x}_k\|, \quad \|\mathbf{x}_k\| = r, \tag{3.35}$$

and for any subspace $\mathcal{S}_k$ *that contains the* SQP *iterate* $\mathbf{x}_{\rm SQP}$ *associated with* (3.5)–(3.6), *the solution* $\mathbf{x}_{k+1}$ *of* (3.1) *and associated multiplier* μ_{k+1} *given by* (3.7) *satisfy the estimate*

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| + |\mu_{k+1} - \mu^*| \leq C(\|\mathbf{x}_k - \mathbf{x}^*\|^2 + |\mu_k - \mu^*|^2).$$

In the case that $\mu_k = -\lambda_1 + \|\mathbf{b} - (\mathbf{A} + \rho(\mathbf{x}_k)\mathbf{I})\mathbf{x}_k\|$, C *can be chosen so that*

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| + |\mu_{k+1} - \mu^*| \leq C\|\mathbf{x}_k - \mathbf{x}^*\|^2.$$

Proof. Focusing on the numerator in (3.17), and substituting $\beta_i = (\lambda_i - \lambda_1)\chi_i^*$, we have

$$\begin{aligned}
&\sum_{i=1}^n \frac{\chi_i(\beta_i - (\lambda_i + \mu_k)\chi_i)}{\lambda_i + \mu_k} \\
&= \sum_{i=1}^n \frac{\chi_i\big((\lambda_i - \lambda_1)(\chi_i^* - \chi_i + \chi_i) - (\lambda_i + \mu_k)\chi_i\big)}{\lambda_i + \mu_k} \\
&= -(\lambda_1 + \mu_k)\sum_{i=1}^n \frac{\chi_i^2}{\lambda_i + \mu_k} + \sum_{i\in\mathcal{E}_+} \frac{\chi_i(\lambda_i - \lambda_1)(\chi_i^* - \chi_i)}{\lambda_i + \mu_k} \\
&= -(\lambda_1 + \mu_k)\sum_{i=1}^n \frac{\chi_i^2}{\lambda_i + \mu_k} + \sum_{i\in\mathcal{E}_+} \chi_i(\chi_i^* - \chi_i) - (\lambda_1 + \mu_k)\sum_{i\in\mathcal{E}_+} \frac{\chi_i(\chi_i^* - \chi_i)}{\lambda_i + \mu_k}.
\end{aligned}$$

With this substitution for the numerator of ν in (3.17), we obtain

$$\nu = -(\lambda_1 + \mu_k) + \frac{\sum_{i\in\mathcal{E}_+} \chi_i(\chi_i^* - \chi_i)}{\sum_{i=1}^n \chi_i^2/(\lambda_i + \mu_k)} - \frac{(\lambda_1 + \mu_k)\sum_{i\in\mathcal{E}_+} \frac{\chi_i(\chi_i^* - \chi_i)}{\lambda_i + \mu_k}}{\sum_{i=1}^n \chi_i^2/(\lambda_i + \mu_k)}. \tag{3.36}$$

Since $\mu_k > -\lambda_1$, the denominator terms in (3.36) have the lower bound

$$\sum_{i=1}^n \frac{\chi_i^2}{\lambda_i + \mu_k} \geq \sum_{i=1}^n \frac{\chi_i^2}{\lambda_n + \mu_k} = \frac{r^2}{\lambda_n + \mu_k}. \tag{3.37}$$

Another lower bound is gotten by neglecting terms corresponding to indices $i \in \mathcal{E}_+$:

$$\sum_{i=1}^n \frac{\chi_i^2}{\lambda_i + \mu_k} \geq \sum_{i\in\mathcal{E}_1} \frac{\chi_i^2}{\lambda_i + \mu_k} = \frac{\|\mathbf{x}_k\|_1^2}{\lambda_1 + \mu_k}, \tag{3.38}$$

where the seminorm $\|\cdot\|_1$ is defined in (3.29). Combining (3.36)–(3.38) yields

$$\nu = -(\lambda_1 + \mu_k)(1 + O(\|\mathbf{x}_k - \mathbf{x}^*\|_+)) + \frac{O(\|\mathbf{x}^* - \mathbf{x}_k\|_+)}{\max\{1, \|\mathbf{x}_k\|_1^2/(\lambda_1 + \mu_k)\}}. \tag{3.39}$$

Returning to our previous analysis of the degenerate case, it follows from (3.22) and (3.39) that for $i \in \mathcal{E}_1$, we have

$$\begin{aligned} \zeta_i + \chi_i &= \frac{\beta_i - \chi_i \nu}{\lambda_i + \mu_k} = \frac{-\chi_i \nu}{\lambda_1 + \mu_k} \\ &= \chi_i + O(\epsilon_k) \left(1 + \frac{\|\mathbf{x}_k - \mathbf{x}^*\|_+}{\max\{\lambda_1 + \mu_k, \|\mathbf{x}_k\|_1^2\}} \right). \end{aligned} \tag{3.40}$$

Here we exploit the fact that for $i \in \mathcal{E}_1$, $|\chi_i| \leq \|\mathbf{x}_k\|_1 \leq \epsilon_k$. In order to analyze (3.40), we consider two separate cases: (i) $\|\mathbf{x}_k\|_1^2 \geq \sigma \|\mathbf{x}_k - \mathbf{x}^*\|_+$ and (ii) $\|\mathbf{x}_k\|_1^2 < \sigma \|\mathbf{x}_k - \mathbf{x}^*\|_+$, where σ is any fixed constant satisfying

$$0 < \sigma < \frac{r(\lambda_+ - \lambda_1)}{\lambda_n - \lambda_1}, \quad \lambda_+ = \min\{\lambda_i : \lambda_i > \lambda_1,\ 1 \leq i \leq n\}. \tag{3.41}$$

In case (i),

$$\frac{\|\mathbf{x}_k - \mathbf{x}^*\|_+}{\max\{\lambda_1 + \mu_k, \|\mathbf{x}_k\|_1^2\}} \leq \frac{\|\mathbf{x}_k - \mathbf{x}^*\|_+}{\max\{\lambda_1 + \mu_k, \sigma \|\mathbf{x}_k - \mathbf{x}^*\|_+\}} \leq \frac{1}{\sigma}. \tag{3.42}$$

We now derive a similar bound for the left side of (3.42) in case (ii). In this case, it follows from (3.35) that

$$\frac{\|\mathbf{x}_k - \mathbf{x}^*\|_+}{\max\{\lambda_1 + \mu_k, \|\mathbf{x}_k\|_1^2\}} \leq \frac{\|\mathbf{x}^* - \mathbf{x}_k\|_+}{\|\mathbf{b} - (\mathbf{A} + \rho(\mathbf{x}_k)\mathbf{I})\mathbf{x}_k\|}.$$

Since $\mathbf{b} = (\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{x}^*$, we have

$$\begin{aligned} \mathbf{b} - (\mathbf{A} + \rho(\mathbf{x})\mathbf{I})\mathbf{x} &= (\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{x}^* - (\mathbf{A} + \rho(\mathbf{x})\mathbf{I})\mathbf{x} \\ &= (\mathbf{A} - \lambda_1 \mathbf{I})(\mathbf{x}^* - \mathbf{x}) - (\rho(\mathbf{x}) + \lambda_1)\mathbf{x} \\ &= (\mathbf{A} - \lambda_1 \mathbf{I})(\mathbf{x}^* - \mathbf{x})_+ - (\rho(\mathbf{x}) + \lambda_1)\mathbf{x} \end{aligned}$$

for any $\mathbf{x} \in \mathbf{R}^n$, where a + subscript on a vector is used to denote its projection on the eigenspace associated with $\mathcal{E}_+$. After substituting for ρ using (3.13), we obtain

$$\mathbf{b} - (\mathbf{A} + \rho(\mathbf{x})\mathbf{I})\mathbf{x} = (\mathbf{A} - \lambda_1 \mathbf{I})(\mathbf{x}^* - \mathbf{x}) - r^{-2}\big(\mathbf{x}^{\mathsf{T}}(\mathbf{A} - \lambda_1 \mathbf{I})(\mathbf{x}^* - \mathbf{x})_+\big)\mathbf{x} \tag{3.43}$$

for any $\mathbf{x} \in \mathcal{B}_r$. Assuming $\mathbf{x}_k \neq \mathbf{x}^*$, it follows that

$$\frac{\|\mathbf{x}^* - \mathbf{x}_k\|_+}{\|\mathbf{b} - (\mathbf{A} + \rho(\mathbf{x}_k)\mathbf{I})\mathbf{x}_k\|} = \frac{1}{\|(\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{y} - r^{-2}\big(\mathbf{x}_k^{\mathsf{T}}(\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{y}\big)\mathbf{x}_k\|}, \tag{3.44}$$

where $\mathbf{y} = (\mathbf{x}^* - \mathbf{x}_k)_+ / \|\mathbf{x}^* - \mathbf{x}_k\|_+$ is a unit vector (note that when $\|\mathbf{x}_k\| = r$, $\|\mathbf{x}^* - \mathbf{x}_k\|_+ = 0$ if and only if $\mathbf{x}_k = \mathbf{x}^*$ since $\|\mathbf{x}^*\|_1 = 0$).

We will establish a uniform bound for the expression (3.44) when $\mathbf{x}_k$ is near $\mathbf{x}^*$, $\|\mathbf{x}_k\|_1^2 \leq \sigma \|\mathbf{x}_k - \mathbf{x}^*\|_+$, and $\|\mathbf{x}_k\| = r$. To facilitate this analysis, we first consider whether the equation

$$(\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{y} = r^{-2}\big(\mathbf{y}^{\mathsf{T}}(\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{x}^*\big)\mathbf{x}^* \tag{3.45}$$

has a solution of the form $\mathbf{y} = (\mathbf{x}^* - \mathbf{x})_+ / \|\mathbf{x}^* - \mathbf{x}\|_+$ with $\mathbf{x}$ near $\mathbf{x}^*$, $\|\mathbf{x}\| = r$, and $\|\mathbf{x}\|_1^2 \leq \sigma \|\mathbf{x} - \mathbf{x}^*\|_+$. Since $\|\mathbf{y}\| = 1$ for $\mathbf{y}$ of this form, the Schwarz inequality gives

$$|\mathbf{y}^{\mathsf{T}}(\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{x}^*| \leq (\lambda_n - \lambda_1)\|\mathbf{y}\|\|\mathbf{x}^*\| = r(\lambda_n - \lambda_1). \tag{3.46}$$

Since the unit vector $\mathbf{y}$ is orthogonal to the eigenspace associated with λ_1,

$$\mathbf{y}^{\mathsf{T}}(\mathbf{A}-\lambda_1\mathbf{I})\mathbf{y} \geq \lambda_+ - \lambda_1. \tag{3.47}$$

Multiplying (3.45) by $\mathbf{y}^{\mathsf{T}}$ and using both (3.46) and (3.47) gives

$$|\mathbf{y}^{\mathsf{T}}\mathbf{x}^*| \geq \frac{r(\lambda_+ - \lambda_1)}{\lambda_n - \lambda_1} > \sigma. \tag{3.48}$$

For any $\mathbf{x} \in \mathcal{B}_r$, we have

$$\begin{aligned} r^2 = \|\mathbf{x}\|^2 &= r^2 + 2(\mathbf{x}-\mathbf{x}^*)^{\mathsf{T}}\mathbf{x}^* + \|\mathbf{x}-\mathbf{x}^*\|^2 \\ &= r^2 - 2\|\mathbf{x}^*-\mathbf{x}\|_+\mathbf{y}^{\mathsf{T}}\mathbf{x}^* + \|\mathbf{x}-\mathbf{x}^*\|^2, \end{aligned}$$

which implies that

$$\begin{aligned} \mathbf{y}^{\mathsf{T}}\mathbf{x}^* &= \frac{\|\mathbf{x}-\mathbf{x}^*\|^2}{2\|\mathbf{x}-\mathbf{x}^*\|_+} = \frac{\|\mathbf{x}-\mathbf{x}^*\|_+^2 + \|\mathbf{x}-\mathbf{x}^*\|_1^2}{2\|\mathbf{x}-\mathbf{x}^*\|_+} \\ &= \frac{1}{2}\left(\|\mathbf{x}-\mathbf{x}^*\|_+ + \frac{\|\mathbf{x}\|_1^2}{\|\mathbf{x}-\mathbf{x}^*\|_+}\right) \end{aligned} \tag{3.49}$$

since $\|\mathbf{x}-\mathbf{x}^*\|_1 = \|\mathbf{x}\|_1$. If $\|\mathbf{x}\|_1^2 \leq \sigma\|\mathbf{x}-\mathbf{x}^*\|_+$, then (3.49) yields the relation

$$0 \leq \mathbf{y}^{\mathsf{T}}\mathbf{x}^* \leq \frac{1}{2}(\|\mathbf{x}-\mathbf{x}^*\|_+ + \sigma). \tag{3.50}$$

Referring to (3.48), we have a contradiction when $\|\mathbf{x}-\mathbf{x}^*\|_+ \leq \sigma$.

In summary, (3.45) has no solution over the set $\mathcal{Y}$ consisting of those $\mathbf{y}$ that satisfy the conditions $\mathbf{y} = (\mathbf{x}^*-\mathbf{x})_+/\|\mathbf{x}^*-\mathbf{x}\|_+$, $\mathbf{x} \neq \mathbf{x}^*$, $\|\mathbf{x}\|_1^2 \leq \sigma\|\mathbf{x}-\mathbf{x}^*\|_+$, $\|\mathbf{x}-\mathbf{x}^*\|_+ \leq \sigma$, and $\|\mathbf{x}\| = r$. If $\mathbf{y}$ lies in the closure of $\mathcal{Y}$, then by (3.50), $\mathbf{y}^{\mathsf{T}}\mathbf{x}^* \leq \sigma$; since any solution of (3.45) satisfies (3.48), $\mathbf{y}$ cannot be a solution of (3.45). Since (3.45) has no solution over the closure of $\mathcal{Y}$, the following constant δ is strictly positive:

$$\delta = \min_{\mathbf{y}\in\mathcal{Y}} \|(\mathbf{A}-\lambda_1\mathbf{I})\mathbf{y} - r^{-2}\big(\mathbf{y}^{\mathsf{T}}(\mathbf{A}-\lambda_1\mathbf{I})\mathbf{x}^*\big)\mathbf{x}^*\|.$$

Since

$$\lim_{\mathbf{x}_k\to\mathbf{x}^*} \min_{\mathbf{y}\in\mathcal{Y}} \|(\mathbf{A}-\lambda_1\mathbf{I})\mathbf{y} - r^{-2}\big(\mathbf{y}^{\mathsf{T}}(\mathbf{A}-\lambda_1\mathbf{I})\mathbf{x}_k\big)\mathbf{x}_k\| = \delta,$$

(3.44) is bounded uniformly over all $\mathbf{x}_k$ near $\mathbf{x}^*$ with $\|\mathbf{x}_k\| = r$ and $\|\mathbf{x}_k\|_1^2 < \sigma\|\mathbf{x}_k - \mathbf{x}^*\|_+$. Thus in either case (i) or (ii), the left side of (3.42) is bounded and, by (3.40), we have

$$\zeta_i + \chi_i = \chi_i + O(\epsilon_k) \quad \text{for } i \in \mathcal{E}_1,$$

which is the same as relation (3.23) in the degenerate case.

To establish the analogue of (3.25) for indices $i \in \mathcal{E}_+$, we need a different bound for the next to last term in (3.36). From the identity $\sum_{i=1}^n \chi_i^2 = \sum_{i=1}^n \chi_i^{*2} = r^2$, we obtain

$$\sum_{i=1}^n (\chi_i^* + \chi_i)(\chi_i^* - \chi_i) = 0. \tag{3.51}$$

Hence, we have

$$-\sum_{i=1}^{n}\chi_i(\chi_i^*-\chi_i) = -\sum_{i=1}^{n}\chi_i(\chi_i^*-\chi_i) + \frac{1}{2}(\chi_i^*+\chi_i)(\chi_i^*-\chi_i)$$

$$= \frac{1}{2}\sum_{i=1}^{n}(\chi_i^*-\chi_i)^2. \tag{3.52}$$

Since $\chi_i^* = 0$ for $i \in \mathcal{E}_1$, (3.52) implies that

$$-\sum_{i\in\mathcal{E}_+}\chi_i(\chi_i^*-\chi_i) = \frac{1}{2}\sum_{i\in\mathcal{E}_+}(\chi_i^*-\chi_i)^2 - \frac{1}{2}\sum_{i\in\mathcal{E}_1}\chi_i^2$$

$$= \frac{1}{2}\sum_{i\in\mathcal{E}_+}(\chi_i^*-\chi_i)^2 - \frac{1}{2}\sum_{i\in\mathcal{E}_1}(\chi_i^*-\chi_i)^2.$$

It follows that

$$\left|\sum_{i\in\mathcal{E}_+}\chi_i(\chi_i^*-\chi_i)\right| \le \|\mathbf{x}^*-\mathbf{x}_k\|^2.$$

This estimate, along with the lower bound (3.37) for the denominator in (3.36), yields the relation

$$\nu = -(\lambda_1+\mu_k) + O(\epsilon_k^2).$$

The remainder of the analysis is identical to that given for the degenerate case (Lemma 3.3), starting with (3.25). Since $\mathcal{S}^* = \{\mathbf{x}^*\}$, it follows from the analysis of Lemma 3.3 that

$$\|\mathbf{x}_{k+1}-\mathbf{x}^*\| + |\mu_{k+1}+\lambda_1| \le C(\|\mathbf{x}_k-\mathbf{x}^*\|^2 + |\mu_k+\lambda_1|^2). \tag{3.53}$$

In the special case $\mu_k = -\lambda_1 + \|\mathbf{b}-(\mathbf{A}+\rho(\mathbf{x}_k)\mathbf{I})\mathbf{x}_k\|$, (3.43) gives

$$|\mu_k+\lambda_1| = \|\mathbf{b}-(\mathbf{A}+\rho(\mathbf{x}_k)\mathbf{I})\mathbf{x}_k\| = O(\|\mathbf{x}_k-\mathbf{x}^*\|).$$

Hence, the $|\mu_k+\lambda_1|^2$ term in (3.53) can be absorbed in the $\|\mathbf{x}_k-\mathbf{x}^*\|^2$ term. This completes the proof. ∎

4. Implementation. In our experimentation with the SSM, we put the following four vectors in $\mathcal{S}_k$ in each iteration: $\mathbf{x}_{\mathrm{SQP}}$, $\mathbf{x}_k$, $\mathbf{b}-\mathbf{A}\mathbf{x}_k$, and an estimate for an eigenvector of $\mathbf{A}$ associated with the smallest eigenvalue. By including $\mathbf{x}_k$ in $\mathcal{S}_k$, the value of the cost function can only decrease in consecutive iterations. The multiple $\mathbf{b}-\mathbf{A}\mathbf{x}_k$ of the cost function gradient ensures descent if the current iterate does not satisfy the first-order optimality conditions. The eigenvector associated with the smallest eigenvalue will dislodge the iterates from a nonoptimal stationary point. We also use this vector in a "safe-guard" strategy designed to keep $\mathbf{A}+\mu_k\mathbf{I}$ positive definite.

4.1. The SQP system. Now consider the SQP system (3.5)–(3.6). According to (3.6), $\mathbf{z}$ is orthogonal to the prior iterate $\mathbf{x}_k$. Let $\mathbf{P}$ be the matrix that projects a vector into the space perpendicular to $\mathbf{x}_k$:

$$\mathbf{P} = \mathbf{I} - \frac{\mathbf{x}_k\mathbf{x}_k^{\mathsf{T}}}{\mathbf{x}_k^{\mathsf{T}}\mathbf{x}_k}.$$

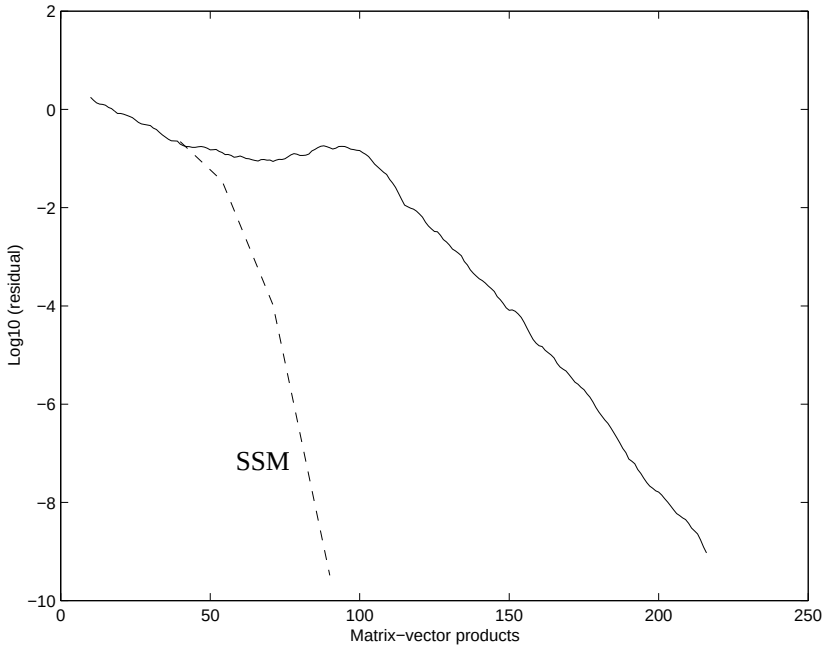

FIG. 4.1. *Convergence of the tridiagonalization approach* (*solid*) *and SSM* (*dashed*) *for the second test problem from* [24].

Multiplying (3.5) by $\mathbf{P}$ yields

$$\mathbf{P}(\mathbf{A}+\mu_k\mathbf{I})\mathbf{z} = \mathbf{P}(\mathbf{b}-\mathbf{A}\mathbf{x}_k).$$

Since $\mathbf{P}\mathbf{z} = \mathbf{z}$, according to (3.6), we have

$$\mathbf{P}(\mathbf{A}+\mu_k\mathbf{I})\mathbf{P}\mathbf{z} = \mathbf{P}(\mathbf{b}-\mathbf{A}\mathbf{x}_k). \tag{4.1}$$

We have found that preconditioned Krylov space methods, such as the Gauss–Seidel scheme in [9], converge very quickly when applied to (4.1). As a small illustration, let us consider the second test problem from [24] with $r = 100$ and $\mathbf{A} = \mathbf{Q}\boldsymbol{\Delta}\mathbf{Q}$, where $\boldsymbol{\Delta}$ is a 1000×1000 diagonal matrix with diagonal elements selected randomly from a uniform distribution on $(-.5, .5)$ and $\mathbf{Q} = \mathbf{I} - 2\mathbf{q}\mathbf{q}^{\mathsf{T}}$, where $\mathbf{q}$ is obtained by first generating random numbers on $(-.5, .5)$ and then scaling the resulting vector to have unit length. The vector $\mathbf{b}$ is generated in the same way as $\mathbf{q}$. The solid curve in Figure 4.1 gives the convergence when a Lanczos type process (Algorithm 1, with starting vector $\mathbf{v}_1 = \mathbf{P}\mathbf{b}$) is used to generate the matrix $\mathbf{V}$ used in (3.2). The Lanczos process was modified to ensure orthogonality of the columns of $\mathbf{V}$. For each value of l in Algorithm 1, we solve the $l \times l$ tridiagonal problem (3.3) to obtain an approximate solution $\mathbf{x}$ and associated multiplier $\mu = \rho(\mathbf{x})$ for the original problem (2.1). In the solid curve of Figure 4.1, we plot the base 10 logarithm of the norm of the residual $\|\mathbf{b} - (\mathbf{A} + \mu\mathbf{I})\mathbf{x}\|$. According to Lemma 2.1, the residual vanishes at an optimal solution.

The dashed curve of Figure 4.1, based on the SSM approach, is obtained in the following way: Taking $l = 40$ in Algorithm 1, we generate a $\mathbf{V}$ with 40 orthonormal columns. Solving (3.3), we obtain a starting guess of $\mathbf{x}_0$. In iteration k of the SSM

phase, we start with the vector $\mathbf{v}_1 = \mathbf{P}(\mathbf{b}-\mathbf{A}\mathbf{x}_k)$ and we use the Gauss–Seidel/Krylov space approach of [9] to generate a matrix $\mathbf{V}$, with orthonormal columns, that approximately contains a solution of (4.1) in its range. Using the $\mathbf{V}$ generated in this way, we solve (3.2) to obtain the next iterate $\mathbf{x}_{k+1}$. The associated multiplier is estimated using (3.7). Each kink in the dashed curve of Figure 4.1 corresponds to the number of iterations needed to obtain an approximate solution of (4.1). In this example, roughly 15 multiplications by the elements of the matrix $\mathbf{A}$ are used to solve (4.1). The quadratic convergence of SSM is reflected in the rapid decay of the residual norm.

This approach for generating $\mathbf{V}$, using a nonsymmetric Gauss–Seidel matrix, Krylov spaces, and orthogonalization, can become expensive when n is really large since each of the columns of $\mathbf{V}$ should be stored in memory. Hence, in the remainder of this paper, we focus on low-storage symmetric techniques for solving (4.1), which we compare to other approaches.

We solve (4.1) using a preconditioned version of Paige and Saunders' MINRES algorithm [17]. More precisely, we use Algorithms 3 and 3a in [9] and three different choices for the symmetrizing preconditioner $\mathbf{W}$ in that paper: (i) $\mathbf{W} = \mathbf{I}$, corresponding to unconditioned iterations; (ii) $\mathbf{W} = \mathbf{D}^{1/2}$, where $\mathbf{D}$ is the diagonal matrix whose diagonal matches that of $\mathbf{C} = \mathbf{P}(\mathbf{A} + \mu_k\mathbf{I})\mathbf{P}$ (Jacobi symmetrization); (iii) $\mathbf{W} = \mathbf{D}^{-1/2}(\mathbf{L} + \mathbf{D})$, where $\mathbf{L}$ is the strictly lower triangular matrix whose lower triangle matches that of $\mathbf{C}$ (SSOR symmetrization). The implementations of SSM associated with the latter two preconditioners are denoted SSM_d and SSM_l, respectively.

Typically, the $\mathbf{L}$ matrix associated with $\mathbf{C} = \mathbf{P}(\mathbf{A}+\mu_k\mathbf{I})\mathbf{P}$ is dense, even when $\mathbf{A}$ is sparse, since $\mathbf{P}$ is often dense. Nonetheless, linear systems of the form $(\mathbf{L}+\mathbf{D})\mathbf{y} = \mathbf{g}$ can be solved in time proportional to the number of nonzero elements in the lower triangle of $\mathbf{A}$, due to the special structure of $\mathbf{C}$. In terms of the vectors $\mathbf{w}$, $\mathbf{q}$, and $\mathbf{p}$ defined by

$$\mathbf{w} = \mathbf{x}_k/\|\mathbf{x}_k\|, \quad \mathbf{q} = (\mathbf{A} + \mu_k\mathbf{I})\mathbf{w}, \quad \text{and} \quad \mathbf{p} = \mathbf{q} - (\mathbf{q}^{\mathsf{T}}\mathbf{w})\mathbf{w},$$

the diagonal $\mathbf{d}$ of $\mathbf{C}$ can be expressed

$$d_i = a_{ii} + \mu_k - (p_i + q_i)w_i,$$

while the off-diagonal elements of $\mathbf{C}$ are

$$c_{ij} = a_{ij} - w_i q_j - p_i w_j, \quad i \neq j.$$

Exploiting this structure, it can be shown that the solution to $(\mathbf{L} + \mathbf{D})\mathbf{y} = \mathbf{g}$ can be computed in the following way.

ALGORITHM 2 ($\mathbf{y} = (\mathbf{L} + \mathbf{D})^{-1}\mathbf{g}$, $\mathbf{L} + \mathbf{D} + \mathbf{L}^{\mathsf{T}} = \mathbf{P}(\mathbf{A} + \mu_k\mathbf{I})\mathbf{P}$, $\mathbf{P} = \mathbf{I} - \mathbf{w}\mathbf{w}^{\mathsf{T}}$).

$\mathbf{y} = \mathbf{g}$, $s = 0$, $t = 0$
for $i = 1 : n - 1$
  $y_i = (y_i + sw_i + tp_i)/d_i$
  $s = s + q_i y_i$
  $t = t + w_i y_i$
  $y_{i+1:n} = y_{i+1:n} - y_i a_{i+1:n,i}$
end
$y_n = (y_n + sw_n + tp_n)/d_n$
END ALGORITHM 2

The statement $y_{i+1:n} = y_{i+1:n} - y_i a_{i+1:n,i}$ of Algorithm 2 requires only the nonzero elements in column i of $\mathbf{A}$ beneath the diagonal. Hence, the number of floating point operations for Algorithm 2 is $O(n)$ plus the number of nonzero elements in the lower triangle of $\mathbf{A}$.

The analogous procedure for the transposed system is the following.

ALGORITHM 3 ($\mathbf{y} = (\mathbf{L}+\mathbf{D})^{-\mathsf{T}}\mathbf{g}$, $\mathbf{L}+\mathbf{D}+\mathbf{L}^{\mathsf{T}} = \mathbf{P}(\mathbf{A}+\mu_k\mathbf{I})\mathbf{P}$, $\mathbf{P} = \mathbf{I} - \mathbf{w}\mathbf{w}^{\mathsf{T}}$).

$\mathbf{y} = \mathbf{g}$, $s = 0$, $t = 0$
for $i = n : -1 : 2$
  $y_i = (y_i + sw_i + tq_i)/d_i$
  $s = s + p_i y_i$
  $t = t + w_i y_i$
  $y_{1:i-1} = y_{1:i-1} - y_i a_{1:i-1,i}$
end
$y_1 = (y_1 + sw_1 + tq_1)/d_1$

END ALGORITHM 3

4.2. Positive definiteness. In theory, the MINRES algorithm we use to solve (4.1) can be applied to any symmetric matrix. In practice, convergence can be extremely slow when $\mathbf{C}$ is indefinite. For this reason, we try to choose μ_k so that $\mathbf{A}+\mu_k\mathbf{I}$ is positive definite. If $\mathbf{e}$ is an eigenvector of the matrix $\mathbf{B}$ in (3.3) associated with the smallest eigenvalue σ, then the pair $(\mathbf{v}, \sigma)$, where $\mathbf{v} = \mathbf{V}\mathbf{e}/\|\mathbf{V}\mathbf{e}\|$, approximates an eigenpair of $\mathbf{A}$ corresponding to the smallest eigenvalue. The error in σ can be estimated in the following way: If σ is closer to λ_1 than the other eigenvalues of $\mathbf{A}$, then after substituting

$$\mathbf{v} = \sum_{i=1}^{n} \nu_i \boldsymbol{\phi}_i, \quad \nu_i = \mathbf{v}^{\mathsf{T}}\boldsymbol{\phi}_i,$$

in the residual $\mathbf{r} = \mathbf{A}\mathbf{v} - \sigma\mathbf{v}$, we have

$$\|\mathbf{r}\|^2 = \sum_{i=1}^{n} |\sigma - \lambda_i|^2 \nu_i^2 \geq \sum_{i=1}^{n} |\sigma - \lambda_1|^2 \nu_i^2 = |\sigma - \lambda_1|^2,$$

since $\sum_{i=1}^{n} \nu_i^2 = 1$. Thus $|\sigma - \lambda_1| \leq \|\mathbf{r}\|$, which implies that

$$\lambda_1 \geq \sigma - \|\mathbf{r}\|.$$

With this insight, we replace the least squares estimate (3.7) by the safe-guarded estimate

$$\mu_k = \max\{\|\mathbf{r}\| - \sigma, \rho(\mathbf{x}_k)\}. \tag{4.2}$$

This choice for μ_k helps to ensure that $\mathbf{A} + \mu_k\mathbf{I}$ is positive definite, often leading to faster convergence of iterative methods applied to (4.1).

When the approximate eigenpair $(\mathbf{v}, \sigma)$ is not very accurate, then the safe-guarded step (4.2) is a safe, but poor, approximation to μ^*. Hence, whenever $\mu_k = \|\mathbf{r}\| - \sigma$, we apply one iteration of SSM to the quadratic eigenvalue problem (1.2) in order to compute a more accurate eigenpair. Due to the third- and sixth-order estimates in (3.8), simply one iteration of SSM for the eigenproblem often yields a highly accurate eigenpair.

4.3. The algorithm. We now collect our observations and present the algorithm that was used to generate the numerical results of the next section. To simplify the presentation, we introduce the following subroutines:

- $\mathbf{V} = \text{Lanczos}(\mathbf{A}, \mathbf{v}_1, l)$: This routine applies Algorithm 1 to the matrix $\mathbf{A}$, starting from the vector $\mathbf{v}_1$, to generate a matrix $\mathbf{V}$ with columns $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_l$.
- $(\mathbf{x}, \mu, \mathbf{v}, \sigma) = \text{SSM}(\mathbf{A}, \mathbf{b}, \mathcal{S}_k)$: This routine solves the problem (3.1), generating a solution denoted $\mathbf{x}$ and an associated multiplier $\mu = \rho(\mathbf{x})$. If $\mathbf{V}$ is a matrix whose columns are an orthonormal basis for $\mathcal{S}_k$, then an estimate $(\mathbf{v}, \sigma)$ for the smallest eigenvalue of $\mathbf{A}$ and an associated eigenvector is obtained by computing the smallest eigenvalue σ and an associated eigenvector $\mathbf{e}$ for $\mathbf{B} = \mathbf{V}^\mathsf{T}\mathbf{A}\mathbf{V}$ and setting $\mathbf{v} = \mathbf{V}\mathbf{e}$.
- $\mathbf{z} = \text{SQP}(\mathbf{A}, \mu, \mathbf{b}, \mathbf{x})$: This routine computes a (minimum residual, minimum norm) solution $(\mathbf{z}, \nu)$ of the linear system

$$\left[\begin{array}{c|c} \mathbf{A}+\mu\mathbf{I} & \mathbf{x} \\ \mathbf{x}^\mathsf{T} & \mathbf{0} \end{array}\right]\left[\begin{array}{c} \mathbf{z} \\ \nu \end{array}\right] = \left[\begin{array}{c} \mathbf{b}-(\mathbf{A}+\mu\mathbf{I})\mathbf{x} \\ 0 \end{array}\right].$$

Our implementation of the sequential subspace method combines these three routines and the safe-guarded step (4.2).

ALGORITHM 4 (SAFE-GUARDED SSM WITH LANCZOS STARTUP).

$it = \nu = \mu = 0,\ \mathbf{v} = \mathbf{x} = \mathbf{0},\ \mathbf{c} = \text{rand}(n,1) - .5$
$\mathbf{u} = \mathbf{c}/(100\|\mathbf{c}\|) + \mathbf{b}/(r\|\mathbf{A}\|)$
while ($\nu == \mu$ & $it = it + 1 \le \bar{it}$)
  $\mathbf{V} = \text{Lanczos}(\mathbf{A}, \mathbf{u}, l)$
  $(\mathbf{x}, \mu, \mathbf{v}, \sigma) = \text{SSM}(\mathbf{A}, \mathbf{b}, \text{span}(\mathbf{x}, \mathbf{v}, \mathbf{v}_1, \ldots, \mathbf{v}_l))$
  $\nu = \|(\mathbf{A} - \sigma\mathbf{I})\mathbf{v}\| - \sigma$
  if ($\nu > \mu$) $\mu = \nu$
  $\mathbf{u} = \mathbf{b} - (\mathbf{A} + \mu\mathbf{I})\mathbf{x}$
end
while ($\|\mathbf{b} - (\mathbf{A} + \mu\mathbf{I})\mathbf{x}\| > tol$)
  $\mathbf{z} = \text{SQP}(\mathbf{A}, \mu, \mathbf{b}, \mathbf{x})$
  $\mathcal{S} = \text{span}(\mathbf{x}, \mathbf{z}, \mathbf{v}, \mathbf{b} - \mathbf{A}\mathbf{x})$
  $(\mathbf{x}, \mu, \mathbf{v}, \sigma) = \text{SSM}(\mathbf{A}, \mathbf{b}, \mathcal{S})$
  $\nu = \|(\mathbf{A} - \sigma\mathbf{I})\mathbf{v}\| - \sigma$
  $\epsilon = \|\mathbf{b} - (\mathbf{A} + \mu\mathbf{I})\mathbf{x}\|$
  if ($\nu > \mu$ & $\nu + \sigma > \epsilon/r$)
    $\mathbf{z} = \text{SQP}(\mathbf{A}, \nu, \mathbf{0}, \mathbf{v})$
    $(\mathbf{x}, \mu, \mathbf{v}, \sigma) = \text{SSM}(\mathbf{A}, \mathbf{b}, \text{span}(\mathcal{S}, \mathbf{z}))$
    $\nu = \|(\mathbf{A} - \sigma\mathbf{I})\mathbf{v}\| - \sigma$
  end
  if ($\nu > \mu$) $\mu = \nu$
end

END ALGORITHM 4

For the computational results reported in the next section, we took $\bar{it} = 3$ and $l = \max\{10, .01n\}$. The "rand" function appearing at the start of Algorithm 4 generates a vector with components uniformly distributed on $[0, 1]$.

5. Computational results. In this section we compare the performance of SSM to the performance of the algorithms in [7, 20, 24], denoted GLRT, RW, and S, respectively, using the three test problems presented in [24]. The results that we

TABLE 5.1
Problem 1*, average number of matrix-vector products versus tolerance.*

Tolerance	S	RW	GLRT	SSM	SSM_d	SSM_l
10^{-4}	249.0 (04.2)	383.6 (3)	51.0	78.0	51.2	44.2
10^{-6}	824.0 (08.4)	460.7 (4)	65.7	107.1	65.5	54.3
10^{-8}	1633.4 (12.3)	465.7 (4)	86.7	124.3	86.7	70.7

TABLE 5.2
Problem 2*, average number of matrix-vector products versus radius.*

Radius	S	RW	GLRT	SSM	SSM_d	SSM_l
10	240 (08)	1437.9 (5.5)	27.0	88.3	42.3	54.1
100	579 (13)	2567.7 (7.8)	188.8	353.7	88.4	136.2

report for S were extracted from [24], while the results reported for GLRT and RW were obtained using codes provided by the authors. Each of these codes used different stopping criteria. GLRT stopped when $\|\mathbf{b}-(\mathbf{A}+\mu\mathbf{I})\mathbf{x}\|/\|\mathbf{b}\|$ was bounded by a given tolerance, while RW stopped when the gap between the value of the primal and dual problem, and hence the error in the primal cost function, was smaller than a given tolerance. In order to ensure that each code computed a solution with the same accuracy, we adjusted the error tolerance parameter of each code until the value of $\|\mathbf{b}-(\mathbf{A}+\mu\mathbf{I})\mathbf{x}\|$ for the computed solution was smaller than a given tolerance (specified below).

In the first test problem of [24], $\mathbf{A} = \mathbf{A}_0 - 5\mathbf{I}$, where $\mathbf{A}_0$ is the standard 2-D discrete Laplacian on the unit square based on a 5-point stencil with equally spaced mesh points. Taking $n = 32^2 = 1024$ and $r = 100$, a series of 20 problems was generated, where $\mathbf{b}$ was a vector with elements uniformly distributed on $[0, 1]$. Each of these problems was solved using three different tolerances, 10^{-4}, 10^{-6}, and 10^{-8}. In Table 5.1 we give the average number of matrix-vector products involving $\mathbf{A}$ for each algorithm. Each iteration of the preconditioned MINRES algorithm with lower triangular preconditioner involves roughly twice as many flops as an iteration of either the identity or the diagonal preconditioned schemes. Hence, in doing the bookkeeping, we charged for two matrix-vector products in each iteration of the triangular preconditioned scheme. As seen in Table 5.1, SSM_l converges more than twice as fast as the identity and diagonal preconditioned schemes and, overall, SSM_l uses the smallest number of matrix-vector products for this test problem. Since the parametric eigenvalue algorithms S and RW compute an extreme eigenvalue for a series of matrices, we also list in parentheses in Table 5.1 the number of these eigenproblems that are solved. Hence, RW is very economical in terms of the number of these eigenproblems that are solved.

The second suite of test problems in [24] utilizes the matrix described earlier in section 3. In these problems, the radius of the sphere is varied and the number of matrix-vector products is tabulated. For radii of one or smaller, solutions can be computed extremely quickly, so we focused on $r = 10$ and $r = 100$ and an error tolerance of 10^{-7}. In Table 5.2 we see that for $r = 100$, SSM_d had the fewest matrix-vector products, while GLRT had the fewest for $r = 10$.

The final problem of [24] again employed the discrete Laplacian matrix, but with $n = 16^2$ and $r = 100$. The vector $\mathbf{b}$ was designed to make the problem degenerate; first a random $\mathbf{b}$ was generated, then its $\boldsymbol{\phi}_1$ component was removed. Table 5.3 gives the results for the various algorithms.

TABLE 5.3
Problem 3, average number of matrix-vector products.

S	RW	GLRT	SSM	SSM_d	SSM_l
291 (11)	441.0 (5.4)	134.0*	179.3	179.2	161.5

We placed an asterisk by the result in Table 5.3 for GLRT since this routine reduced the error to 10^{-4}, not the 10^{-7} tolerance used by the other routines. Among the routines that achieved the error tolerance, SSM_l performed the best relative to the number of matrix-vector products. Note that the number of matrix-vector products given in Table 5.3 for S was taken from [24] while Rojas, in her recent thesis [21], developed a more efficient implementation of Sorensen's approach for degenerate problems.

In summary, a Lanczos type process seems to be very effective when the problem is very nondegenerate ($\mu^* >> -\lambda_1$). As the problem becomes more degenerate, preconditioned schemes such as SSM_d or SSM_l appear more effective. The number of times that RW computes an extreme eigenpair is often around 5. For the numerical experiments reported in this paper, Matlab's eigs routine was used to compute this extreme eigenpair. If this routine for computing an extreme eigenpair could be sped up, possibly using the Jacobi type methods of Sleijpen and Van der Vorst [22] or the truncated RQ iteration of Sorensen and Yang [25], the number of matrix-vector operations used in the parametric eigenvalue approach would be reduced.

Acknowledgments. The author gratefully acknowledges the comments and suggestions of the referees. He also thanks Henry Wolkowicz for pointing out the related paper [7] and for his comments and suggestions, and the authors of [24] for providing access to their codes.

REFERENCES

[1] R. H. BYRD, R. B. SCHNABEL, AND G. A. SCHULTZ, *A trust region algorithm for nonlinearly constrained optimization*, SIAM J. Numer. Anal., 24 (1987), pp. 1152–1170.
[2] M. R. CELIS, J. E. DENNIS, AND R. A. TAPIA, *A trust region strategy for nonlinear equality constrained optimization*, in Numerical Optimization 1984, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., SIAM, Philadelphia, PA, 1985, pp. 71–82.
[3] M. EL-ALEM, *A global convergence theory for the Celis–Dennis–Tapia trust-region algorithm for constrained optimization*, SIAM J. Numer. Anal., 28 (1991), pp. 266–290.
[4] W. GANDER, *Least squares with a quadratic constraint*, Numer. Math., 36 (1981), pp. 291–307.
[5] G. H. GOLUB AND U. VON MATT, *Quadratically constrained least squares and quadratic problems*, Numer. Math., 59 (1991), pp. 561–580.
[6] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1989.
[7] N. I. M. GOULD, S. LUCIDI, M. ROMA, AND P. L. TOINT, *Solving the trust-region subproblem using the Lanczos method*, SIAM J. Optim., 9 (1999), pp. 504–525.
[8] W. W. HAGER, *Applied Numerical Linear Algebra*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
[9] W. W. HAGER, *Iterative methods for nearly singular linear systems*, SIAM J. Sci. Comput., 22 (2000), pp. 747–766.
[10] W. W. HAGER AND Y. KRYLYUK, *Graph partitioning and continuous quadratic programming*, SIAM J. Discrete Math., 12 (1999), pp. 500–523.
[11] P. C. HANSEN, *Regularization tools: A MATLAB package for analysis and solution of discrete ill-posed problems*, Numer. Algorithms, 6 (1994), pp. 1–35.
[12] P. C. HANSEN, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM, Philadelphia, PA, 1998.
[13] R. B. LEHOUCQ, D. C. SORENSEN, AND C. YANG, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicity Restarted Arnoldi Methods*, SIAM, Philadelphia, PA, 1998.

[14] W. Menke, *Geophysical Data Analysis: Discrete Inverse Theory*, Academic Press, San Diego, CA, 1989.
[15] J. J. Moré, *Recent developments in algorithms and software for trust region methods*, in Mathematical Programming: State of the Art, A. Bachem, M. Grotschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 258–287.
[16] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Stat. Comput., 4 (1983), pp. 553–572.
[17] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[18] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[19] M. J. D. Powell and Y. Yuan, *A trust region algorithm for equality constrained optimization*, Math. Programming, 49 (1991), pp. 189–211.
[20] R. Rendl and H. Wolkowicz, *A semidefinite framework for trust region subproblems with applications to large scale minimization*, Math. Programming, 77 (1997), pp. 273–299.
[21] M. Rojas, *A Large-Scale Trust-Region Approach to the Regularization of Discrete Ill-Posed Problems*, Ph.D. thesis, Computational and Applied Mathematics, Rice University, Houston, TX, 1998.
[22] G. L. G. Sleijpen and H. A. Van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.
[23] D. C. Sorensen, *Newton's method with a model trust region modification*, SIAM J. Numer. Anal., 19 (1982), pp. 409–426.
[24] D. C. Sorensen, *Minimization of a large-scale quadratic function subject to a spherical constraint*, SIAM J. Optim., 7 (1997), pp. 141–161.
[25] D. C. Sorensen and C. Yang, *A truncated RQ iteration for large scale eigenvalue calculations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 1045–1073.
[26] A. Tarantola, *Inverse Problem Theory*, Elsevier, Amsterdam, The Netherlands, 1987.