UCLA
COMPUTATIONAL AND APPLIED MATHEMATICS

Multilevel Subspace Correction for Large-Scale Optimization Problems (Ph.D. Thesis)

Ilya A. Sharapov

July 1997

CAM Report 97-31

Department of Mathematics
University of California, Los Angeles
Los Angeles, CA. 90095-1555

UNIVERSITY OF CALIFORNIA

Los Angeles

Multilevel Subspace Correction for
Large-Scale Optimization Problems

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Mathematics

by

Ilya A. Sharapov

1997

The dissertation of Ilya A. Sharapov is approved.

Bjorn Engquist

Stanley Osher

Olivier Ledoit

Tony F. Chan, Committee Chair

University of California, Los Angeles

1997

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to express my gratitude to my thesis advisor, Professor Tony Chan, for his patience, excellent guidance, and encouragement throughout the completion of my thesis. My sincere thanks also go to my committee members, Professors Bjorn Engquist, Stanley Osher, and Olivier Ledoit, for their involvement in this project. I would also like to thank the faculty, staff and students of the Department of Mathematics for their help and support. In particular, I am thankful to Susie Go and Ludmil Zikatanov for helping me with the numerical experiments that involve unstructured grids.

This work was partially supported by the NSF under contract ASC 92-01266 and ONR under contract ONR-N00014-92-J-1890.

VITA

January 18, 1968	Born in Moscow, Soviet Union
1991	M.S. (Diploma), Applied Mathematics and Information Technology Moscow Institute of Physics and Technology, Russia
1992-1997	Research Assistant and Teaching Assistant Department of Mathematics University of California, Los Angeles
1993	M.A., Mathematics University of California, Los Angeles
1995,1996	Summer Intern The MacNeal-Schwendler Corporation, Los Angeles, CA

PUBLICATIONS

T. Chan and I. Sharapov. Subspace correction methods for eigenvalue problems. In *Domain Decomposition Methods in Scientific and Engineering Computing: proceedings of the Ninth International Conference on Domain Decomposition.* Wiley and Sons, 1996.

A. Knyazev and I. Sharapov. Variational Raleigh quotient iteration methods for a symmetric eigenvalue problem. *East-West Journal of Numerical Mathematics*, pages 121–128, 1993.

L. Komzsik, P. Poschmann, and I. Sharapov. A preconditioning technique for indefinite linear systems. To appear in *Finite Elements in Analysis and Design*, 1997.

ABSTRACT OF THE DISSERTATION

Multilevel Subspace Correction for
Large-Scale Optimization Problems

by

Ilya A. Sharapov
Doctor of Philosophy in Mathematics
University of California, Los Angeles, 1997
Professor Tony F. Chan, Chair

In this work we study the application of domain decomposition and multigrid techniques to optimization. We illustrate the resulting algorithms by applying them to optimization problems derived from discretizations of partial differential equations, as well as to purely algebraic optimization problems arising in mathematical finance. For the analysis of the presented algorithms we utilize the subspace correction framework (cf. Xu, 1992).

We discuss the cases of convex non-smooth and smooth non-convex optimization, as well as constrained optimization, and present the convergence analysis for the multiplicative Schwarz algorithms for these problems. For PDE-based optimization problems we also discuss the effect of coarse grid correction and analyze the convergence rate of the corresponding multiplicative and additive Schwarz methods.

We consider the application of the multiplicative subspace correction method to the variational formulation of the elliptic eigenvalue problem and show that, as in the linear case, if the coarse grid correction is used, the convergence rate is independent of both the number of subdomains and the meshsize. We discuss the generalization of this method for simultaneous computation of several eigenfunctions and its applications to the problem of partitioning a graph based on spectral bisection.

In the final chapter we consider the application of the subspace correction methods to some algebraic optimization problems arising in mathematical finance. We restrict our attention to the minimization of the Frobenius distance used in covariance matrix estimation, the factor analysis problem, and the gain-loss optimization problem. Numerical results illustrating the convergence behavior of the subspace correction methods applied to these problems are presented.

CHAPTER 1

Introduction

In this introduction we provide the motivation for applying the subspace correction techniques to optimization problems. We also give a brief survey of recent literature on subspace correction and domain decomposition methods.

1.1 Background and Motivation

Domain decomposition and multigrid methods are increasingly popular techniques for handling problems arising form partial differential equations. These iterative methods can be applied directly or in combination with other methods such as the conjugate gradient method.

The idea behind the domain decomposition method is to represent the spatial domain of the problem as a union of several subdomains. The subdomains are usually chosen in such a way that the restricted subproblems are easy to solve. This approach allows to make use of parallel machine architecture and can exploit certain geometrical properties of the original problem.

A two-level extension of the domain decomposition method uses a coarser discretization of the problem along with local subdomains. This approach dramatically improves the convergence properties of the method since the use of a coarse component allows the global propagation of data. The multigrid method exploits different levels of coarsening of the problem and can be viewed as a recursive application of a two-level domain decomposition method.

Both domain decomposition and the multigrid methods can be analyzed using the subspace correction framework (cf. Xu [30]). To illustrate this approach we consider a second order self-adjoint coercive elliptic problem

$$\begin{aligned} Lu \equiv \textstyle\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(a\frac{\partial u}{\partial x_j}\right) = f \qquad \text{in } \Omega \subset R^d \\ u = 0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{1.1}$$

which after discretization with n degrees of freedom yields a system

$$Ax = b, \tag{1.2}$$

where $x \in V = R^n$ and A is a sparse symmetric matrix.

Let $\Omega = \cup_{i=1}^{J}\Omega_i$ be a given partition of the domain of the problem and let I_i be the set of indices of the nodes in the interior of subdomains Ω_i. This partition induces decomposition

$$V_1 + V_2 + \cdots + V_J = V,$$

where the subspaces V_i are defined as

$$V_i = \{x \in R^n \mid x_{(k)} = 0 \text{ if } k \notin I_i\}. \tag{1.3}$$

In the case of two-level or multigrid methods the set of subspaces can be appended by one or more subspaces that contain functions corresponding to coarse grids.

For each subspace V_i of dimension $n_i = card(I_i)$ we define the corresponding $n \times n_i$ prolongation matrix P_i which in the simplest case of subspaces given by (1.3) is defined as

$$(P_i x_i)_k = \begin{cases} (x_i)_k & \text{for } k \in I_i \\ 0 & \text{for } k \in I - I_i, \quad \text{where } I = \cup_{k=1}^{J} I_k. \end{cases}$$

Given an approximation x to the exact solution

$$x^* = A^{-1} b$$

of (1.2) we can construct a new update using a correction from i-th subspace by

$$x' = x + P_i d_i \qquad d_i \in V_i. \tag{1.4}$$

The errors corresponding to the iterates x and x'

$$\begin{aligned} e &\equiv x - x^* \\ e' &\equiv x' - x^* = e + P_i d_i \end{aligned}$$

satisfy

$$\begin{aligned} \|e'\|_A^2 &= d_i^T P_i^T A P_i d_i + 2 d_i^T P_i^T A e + e^T A e \\ &= d_i^T P_i^T A P_i d_i + 2 d_i^T P_i^T (Ax - b) + e^T A e \end{aligned}$$

and therefore the maximal reduction in the A-norm of the error is achieved when d_i in (1.4) is

$$d_i = (P_i^T A P_i)^{-1} P_i^T (b - Ax). \tag{1.5}$$

If we introduce the subspace restrictions of the matrix A

$$A_i = P_i^T A P_i$$

we get the iteration

$$x' = x + P_i A_i^{-1} P_i^T (b - Ax), \tag{1.6}$$

which minimizes the A-norm of the error e' over the i-th subdomain. The corresponding equation is

$$e' = e - P_i A_i^{-1} P_i^T A e$$

or

$$e' = (I - T_i)e,$$

where T_i is given by

$$T_i = P_i A_i^{-1} P_i^T A. \tag{1.7}$$

Applying the corrections (1.6) from all the subspaces sequentially or in parallel results in well-known block Gauss-Seidel or Jacobi methods respectively. The corresponding equations for the error are

$$e^{n+1} = \prod_{i=1}^{J} (I - T_i) e^n$$

and

$$e^{n+1} = (I - \sum_{i=1}^{J} T_i) e^n.$$

Since the exact solution of (1.2) is given by $x^* = A^{-1}b$, the minimization of the A-norm of the error

$$\|e(x)\|_A^2 = x^T A x - 2x^T b + b^T A^{-1} b$$

is equivalent to minimizing the functional

$$f(x) = x^T A x - 2x^T b. \tag{1.8}$$

Therefore, given x, the optimal step (1.4) of the form (1.5) can be also obtained by performing a subspace search

$$f(x') = f(x + P_i d_i) = \min_{d \in V_i} f(x + P_i d).$$

This subspace correction step for a general function $f(x)$, not necessarily of the quadratic form (1.8), can be performed sequentially or in parallel with the corrections taken from subspaces V_i. The resulting methods are generalizations of the block Gauss-Seidel and Jacobi methods for optimization problems

$$\min_x f(x). \tag{1.9}$$

In this work we will consider these generalizations and their applications.

1.2 Recent Literature

1.2.1 Domain Decomposition Methods

The convergence properties of the subspace correction methods applied to (1.2) were analyzed by Bramble, Pasciak, Wang, and Xu [6]. They showed that the error reduction operator for the multiplicative Schwarz algorithm

$$E = \prod_{i=1}^{J} (I - T_i)$$

with T_i given by (1.7) satisfies

$$\|Ev\|^2 \leq \left(1 - O\left(\frac{d^2}{n^2}\right)\right) \|v\|^2 \qquad \text{for all } v \in V \qquad ,$$

where d is a characteristic size of subdomains and n is the maximal number of intersecting subdomains.

Xu [30] proved that if one of the subspaces contains the functions corresponding to a quasi-uniform triangulation of Ω of size d the estimate becomes

$$\|Ev\|^2 \leq \gamma \|v\|^2 \qquad \text{for all } v \in V$$

and the parameter $\gamma < 1$ in the above expression is independent of the parameters of discretization. The key role in the proof this estimate is played by the two assumptions which in our notation can be written as

Assumption 1.1 *For any $x \in V$ there exists a decomposition $x = \sum_{i=1}^{J} x_i$ such that $x_i \in V_i$ and*

$$\sum_{i=1}^{J} \|x_i\|_{A_i}^2 \leq C_1^2 \|x\|_A^2 \tag{1.10}$$

and

Assumption 1.2

$$\sum_{i=1}^{J} \sum_{j=1}^{J} (P_i x_i, P_j y_i)_A \leq C_2 \left(\sum_{i=1}^{J} \|x_i\|_{A_i}^2\right)^{\frac{1}{2}} \left(\sum_{j=1}^{J} \|y_j\|_{A_j}^2\right)^{\frac{1}{2}}$$

for any choice of $x_i \in V_i$ and $y_j \in V_j$.

A comprehensive survey of the domain decomposition methods and the convergence estimates is given by Chan and Mathew in [8].

1.2.2 Subspace Correction for Optimization Problems and Nonlinear Problems

The multiplicative Schwarz algorithm applied to optimization problems was proposed by Lions [18]. He proved the convergence of the method for the case of a smooth convex objective function and two subspaces.

This result was extended by Tai and Espedal [29] who generalized the theory for subspace correction methods for linear problems ([30] and [6]) and applied it to a more general class of optimization problems (1.9). They consider twice continuously differentiable convex objective functions $f(x)$ satisfying generalizations of Assumptions 1.1, 1.2 and

Assumption 1.3 *There exist constants $K > 0$ and $L < \infty$ such that*

$$\begin{aligned} (f'(x) - f'(y), x - y) &\geq K\|x - y\|^2 \\ \|f'(x) - f'(y)\| &\leq L\|x - y\| \end{aligned}$$

for any x and y.

The convergence rate of Schwarz algorithms is estimated in terms of the constants in Assumptions 1.1-1.3.

Subspace correction technique for constrained nonlinear optimization was analyzed by Gelman and Mandel [13]. For the problem

$$\min_{x \in C} f(x) \tag{1.11}$$

they use subspace corrections to generate a sequence that stays within the feasible set C and monotonically decreases the objective function $f(x)$. It is shown that all the limit points of the generated sequence are in the solution set of (1.11). The recursive application of this techniques results in a multilevel algorithm.

Application of the multilevel and multigrid techniques was considered in the book by McCormick [24] who mostly focuses on linear problems but also discusses multilevel techniques for solving the eigenvalue problem and the variational form of the Riccati equation. In his book on multigrid methods Hackbusch [15] also discusses nonlinear applications including applications to the eigenvalue problem.

Several domain decomposition-based methods for the eigenvalue problem were proposed by Lui [19], in particular the method based on a nonoverlapping partitioning with the interface problem solved either using a discrete analogue of a Steklov-Poincare operator or using Schur complement-based techniques. The former approach is similar to the component mode synthesis method (cf. Bourquin and Hennezel [5]) which is an approximation rather than iterative technique for solving eigenproblems also used by Farhat and Géradin [12]. Stathopoulos, Saad and Fischer [28] considered iterations based on Schur complement of the block corresponding to the interface variables.

The subspace correction methods for the eigenvalue problem were described by Kaschiev [16] and Maliassov [21]. We will discuss this algorithms in more details in Chapter 3.

CHAPTER 2

Subspace Correction Framework for Optimization Problems

In this chapter we analyze the multiplicative and additive Schwarz methods applied to optimization problems. We present the convergence results for convex unconstrained and constrained problems as well as for smooth but not convex problems. Comparing the general case with the quadratic optimization we discuss the effect of coarse grid correction for the multiplicative and additive algorithms when applied to the optimization problems arising from discretization of PDE-based functionals. Finally, we exploit the recursion to formulate a more computationally efficient multilevel algorithm.

2.1 Multiplicative Schwarz Method

2.1.1 Basic Method

First we consider a general unconstrained nonlinear optimization problem

$$\min f(x), \qquad x \in R^n. \tag{2.1}$$

Here $f(x)$ can either be a purely algebraic function or can be a discrete representation of an integral-differential functional over some spatial domain in R^d. We restrict our consideration to the latter case in the next section of this chapter and in Chapter 3 while applications for purely algebraic problem are discussed in Chapter 4.

To apply the subspace correction methods to (2.1) we introduce J (possibly overlapping) subspaces $V_i \in R^n$, that span the entire space

$$span\{V_i\}_{i=1}^J = V = R^n.$$

In an important special case of this decomposition the subspaces V_i correspond to sets of indices

$$I_i \subset \{1, 2, \dots n\}, \qquad i = 1, \dots, J$$

such that

$$\cup_{i=1}^J I_i = \{1, 2, \dots n\}$$

and

$$V_i = \{x \in R^n \mid x_{(k)} = 0 \text{ if } k \notin I_i\}.$$

We can formulate a standard multiplicative Schwarz algorithm for solving (2.1) (see for example [18]).

```
Alg. 2.1 - Basic Multiplicative Schwarz Algorithm
    Starting with $x^0$ for $k = 0$ until convergence
        for $i = 1 : J$
            find $x^{k+i/J}$ such that
                $f(x^{k+i/J}) = \min_{d_i^k \in V_i} f(x^{k+(i-1)/J} + d_i^k)$        (2.2)
        end
        $x^{k+1} = x^{k+J/J}$
    end
end
```

In the next two sections we consider the convergence properties of this algorithm applied to convex problems as well as to the problems with smooth locally convex objective functions $f(x)$.

2.1.2 Convex Problems

In this section we restrict our attention to the objective functions $f(x)$ that satisfy

Assumption 2.1 *Function $f(x)$ in (2.1) is strictly convex.*

To ensure the existence of the solution we also require

Assumption 2.2 *Function $f(x)$ in (2.1) is coercive: the set $\{x \mid f(x) \leq h\}$ is bounded for any h.*

Assumptions 2.1 and 2.2 guarantee that the original problem (2.1) has a unique solution. In this section we do not assume smoothness of $f(x)$ which makes the convergence analysis applicable to a more general class of problems than the class of smooth convex problems considered by Lions [18] and Tai, Espedal [29].

Theorem 2.1 *Under Assumptions 2.1 and 2.2 the Basic Multiplicative Schwarz Algorithm applied to problem (2.1) generates a convergent sequence of iterates $\{x^k\}$ and the limit $\tilde{x}$ satisfies*

$$f(\tilde{x}) = \min_{d_i \in V_i} f(\tilde{x} + d_i), \qquad i = 1 \dots J. \tag{2.3}$$

Proof. Since Assumption 2.2 implies that all the iterates $x^{k+i/J}$ are contained in the compact set $\{x \mid f(x) \leq f(x^0)\}$ it suffices to show that $x^{k+i/J}$ cannot have more than one limit point. The contrary would imply that there exists a sequence k_j and an index i' such that

$$\tilde{x_1} = \lim_{k_j \to \infty} x^{k_j + (i'-1)/J} \neq \lim_{k_j \to \infty} x^{k_j + i'/J} = \tilde{x_2}. \tag{2.4}$$

Since $f(x^{k+i/J})$ is non-increasing we have

$$f(\tilde{x_1}) = f(\tilde{x_2}) \leq f(x^{k+i/J}) \tag{2.5}$$

for any (k, i) and because $f(x)$ is strictly convex the midpoint $\frac{\tilde{x_1}+\tilde{x_2}}{2}$ satisfies

$$f\left(\frac{\tilde{x_1}+\tilde{x_2}}{2}\right) < f(\tilde{x_1}).$$

Convexity of f also implies its continuity and therefore for some $\epsilon > 0$ we have

$$f\left(\frac{\tilde{x_1}+\tilde{x_2}}{2} + r\right) < f(\tilde{x_1}) \tag{2.6}$$

for any r, such that $\|r\| < \epsilon$. Here $\| \cdot \|$ denotes the Euclidean norm on R^n. Condition (2.4) guarantees that there exists j' such that

$$\|x^{k_{j'}+(i'-1)/J} - \tilde{x_1}\| < \frac{\epsilon}{2} \qquad \text{and} \qquad \|x^{k_{j'}+i'/J} - \tilde{x_2}\| < \frac{\epsilon}{2}, \tag{2.7}$$

but since the corresponding increment satisfies

$$d'_i = x^{k_{j'}+i'/J} - x^{k_{j'}+(i'-1)/J} \in V_i$$

and because of (2.2), (2.6) and (2.7) we have

$$f(x^{k_{j'}+i'/J}) \leq f\left(x^{k_{j'}+(i'-1)/J} + \frac{d'_i}{2}\right) < f(\tilde{x_1}),$$

which contradicts (2.5).

This contradiction shows that there could be only one limit point of $x^{k+i/J}$ and therefore the iterates converge to some $\tilde{x} \in V$. To show that $\tilde{x}$ satisfies (2.3) we assume that for some i there exists $d_i \in V_i$ such that

$$f(\tilde{x} + d_i) < f(\tilde{x}).$$

From the continuity of $f(x)$ it follows that for some $\epsilon > 0$

$$f(\tilde{x} + r + d_i) < f(\tilde{x})$$

for any r satisfying $\|r\| < \epsilon$. Therefore, if

$$\|x^{k+(i-1)/J} - \tilde{x}\| < \epsilon$$

we have

$$f(x^{k+i/J}) < f(\tilde{x})$$

and since $f(x^{k+i/J})$ is decreasing

$$\lim_{k\to\infty} x^k \neq \tilde{x}. \qquad \Box$$

Corollary 2.1 *If in addition to Assumptions 2.1 and 2.2 the objective function $f(x)$ is smooth then the Basic Multiplicative Schwarz Algorithm converges to the minimum of $f(x)$.*

Proof. If $f(x)$ is smooth condition (2.3) implies that the gradient of f satisfies

$$P_i^T \nabla f(\tilde{x}) = 0 \qquad i = 1, \ldots, J,$$

where the P_i^T are the projection operators corresponding to subspaces V_i. Since $spanV_i = V = R^n$ we have $\nabla f(\tilde{x}) = 0$ which for convex $f(x)$ means that $\tilde{x}$ is the minimum of f. $\Box$

2.1.3 Nonconvex Problems

Now we release the convexity assumption and replace it with the assumptions on smoothness of $f(x)$ and its convexity near its local minima.

Assumption 2.3 *Function $f(x)$ in (2.1) is continuously differentiable.*

We can formulate the following result for the case of two subspaces.

Theorem 2.2 *In case of two subdomains ($J = 2$) and under Assumptions 2.2 and 2.3 every convergent subsequence of $\{x^k\}$ generated by Algorithm 2.1 converges to a critical point of $f(x)$.*

Proof. We will show that for any $\tilde{x}$ which is not a critical point of $f(x)$ there is a neighborhood containing that point with the property that once the iterates get in that neighborhood the next iterate will drive the objective below $f(\tilde{x})$. That will prove that the iterates cannot accumulate near $\tilde{x}$.

Let $||\nabla f(\tilde{x})|| \neq 0$ and let $\nabla_i f(\tilde{x})$ be the orthogonal projection of $\nabla f(\tilde{x})$ on V_i, with $i = 1, 2$. Here and below $||\cdot||$ denotes the Euclidean norm. We assume that i is chosen in such a way that $\nabla_i f(\tilde{x})$ is the larger projection (of the two) and therefore

$$||\nabla f(\tilde{x})|| \leq \sqrt{2}||\nabla_i f(\tilde{x})||.$$

Since $\nabla_i f(\tilde{x}) \neq \overline{0}$ and because of continuity of $\nabla_i f(x)$ there exists $\delta = \delta(\tilde{x}) > 0$ such that

$$||\nabla f(x)|| < 2||\nabla_i f(\tilde{x})|| \tag{2.8}$$

and

$$\frac{\nabla_i f(x)^T \nabla_i f(\tilde{x})}{||\nabla_i f(\tilde{x})||^2} > \frac{1}{2}. \tag{2.9}$$

for all $x \in R^n$ such that $\|x - \tilde{x}\| \leq \delta$.

Let x be such that $\|x - \tilde{x}\| \leq \frac{\delta}{6}$, then using (2.8) we have

$$\begin{aligned} f(x) - f(\tilde{x}) &= \int_{\tilde{x}}^{x} \nabla f^T \overline{dl} \\ &\leq 2\|\nabla_i f(\tilde{x})\|\frac{\delta}{6} \\ &= \frac{\delta}{3}\|\nabla_i f(\tilde{x})\|. \end{aligned} \tag{2.10}$$

If we choose a search vector

$$s = -\frac{\nabla_i f(\tilde{x})}{\|\nabla_i f(\tilde{x})\|}$$

then the point

$$x' = x + \frac{5\delta}{6}s$$

satisfies $\|x' - \tilde{x}\| \leq \delta$ and because of (2.9) we have

$$\begin{aligned} f(x) - f(x') &= -\int_0^{\frac{5\delta}{6}} \nabla_i f(x + ts)^T s dt \\ &= \int_0^{\frac{5\delta}{6}} \nabla_i f(x + ts)^T \frac{\nabla_i f(\tilde{x})}{\|\nabla_i f(\tilde{x})\|} dt \\ &\geq \frac{1}{2}\int_0^{\frac{5\delta}{6}} \|\nabla_i f(\tilde{x})\| dt \\ &> \frac{\delta}{3}\|\nabla_i f(\tilde{x})\|. \end{aligned} \tag{2.11}$$

Comparing (2.10) and (2.11) we see that if $\|\nabla f(\tilde{x})\| \neq 0$ and δ defined by (2.8) and (2.9) for any x in the $\frac{\delta_i}{6}$-neighborhood of $\tilde{x}$ there is a point

$$x' \in x + V_i \tag{2.12}$$

such that $f(x') < f(\tilde{x})$. In (2.12) $i = 1$ or $i = 2$ depending on the choice we made in the beginning of the proof. If $i = 1$ then for x^n that satisfies $\|x^n - \tilde{x}\| \leq \frac{\delta}{6}$ we have

$$f(x^{n+\frac{1}{2}}) < f(\tilde{x})$$

and since the sequence $f(x^k)$ is not increasing we have $f(x^k) < f(\tilde{x})$ for $k > n$ and $\tilde{x}$ is not a limit point of x^k. On the other hand if in (2.12) $i = 2$ and $\|x^n - \tilde{x}\| \leq \frac{\delta}{6}$ we have

$$f(x^n) < f(\tilde{x})$$

because otherwise x^n does not minimize $f(x)$ over $x^{n-\frac{1}{2}} + V_2$. The same argument applies and we conclude that $\tilde{x}$ cannot be a limit point of x^k. □

2.2 Coarse Grid Correction for Schwarz Methods

In this section we assume that $f(x)$ in (2.1) comes from a discretization of an integral-differential functional over some spatial region $\Omega \in R^d$ and $x \in V = R^n$ is an n-vector in a finite element space corresponding to this discretization. To apply the domain decomposition technique to this problem we can represent Ω as a union of J overlapping subdomains

$$\Omega = \cup_{i=1}^{J} \Omega_i$$

and consider subspaces $\{V_i\}_{i=1}^{J}$ that contain discretized functions whose support is contained in corresponding subdomains.

We can modify the algorithm presented in the previous section by adding a coarse grid correction step after a loop over subdomains is completed. By doing so in the case of a linear problem with sufficient subdomain overlap the convergence rate becomes independent of both meshsize and the number of subdomains [6], [30].

Let the space of coarse grid functions be V_c then a modified algorithm becomes:

Alg. 2.2 - ***Multiplicative Schwarz Algorithm with Coarse Grid Correction***

```
  Starting with $x^0$ for $k = 0$ until convergence
    for $i = 1 : J$
      find $x^{k+i/J}$ such that
            $f(x^{k+i/J}) = \min_{d_i^k \in V_i} f(x^{k+(i-1)/J} + d_i^k)$
    end
    find $x_c^k$ such that
            $f(x_c^k) = \min_{d_c^k \in V_c} f(x^{k+(J-1)/J} + d_c^k)$
    $x^{k+1} = x_c^k$
  end
end
```

We will present the numerical examples that illustrate the effect of adding the coarse grid correction in Chapter 3.

2.2.1 Asymptotic Convergence Rate

For the local convergence rate analysis we can use the theory developed for the minimization of quadratic functional

$$f(x) = x^T Ax + 2b^T x \tag{2.13}$$

with symmetric positive definite A.

It is known [6] that the iterates x^k generated by Algorithm 2.2 applied to (2.13) converge to the exact solution $x^* = -A^{-1}b$ and satisfy

$$\frac{\|x^{k+1} - x^*\|_A^2}{\|x^k - x^*\|_A^2} \leq \delta < 1, \tag{2.14}$$

where δ is independent of the discretization parameters. Since $f(x)$ is given by (2.13) this condition can be also written as

$$\frac{f(x^{k+1}) - f(x^*)}{f(x^k) - f(x^*)} \leq \delta < 1. \tag{2.15}$$

Now we consider the general minimization problem (2.1) and assume that the iterates x^k converge to the solution x^*. In case of smooth and strictly locally convex $f(x)$ we have

$$\nabla f(x^*) = 0 \text{ and } H = \nabla^2 f(x^*) > 0$$

and the local representation of f

$$\begin{aligned} f(x^* + \epsilon) &= f(x^*) + \nabla f(x^*)\epsilon + \epsilon^T H\epsilon + o(\|\epsilon\|^2) \\ &= f(u^*) + \epsilon^T H\epsilon + o(\|\epsilon\|^2) \end{aligned}$$

shows that, up to the higher order terms, the problem can be viewed as a quadratic minimization and the estimate (2.15) holds asymptotically.

The global convergence rate analysis for the minimization problem in the general setting (2.1) is complicated because the local behavior of the function is not related to the global distribution of its critical points. We can, however, give the convergence rate estimate in a special case when the objective function $f(x)$ in (2.1) is a perturbation of a quadratic function.

Theorem 2.3 *Let the objective function $f(x)$ in (2.1) satisfy*

$$0 < A \leq \nabla^2 f(x) \leq B \tag{2.16}$$

for some symmetric positive definite matrices A and B and let x^* be the solution of (2.1) then the reduction of $f(x)$ during one subspace iteration of Algorithm 2.2 satisfies

$$\frac{f(x^{k+i/J}) - f(x^*)}{f(x^{k+(i-1)/J}) - f(x^*)} \leq \delta_i(A, x^{k+(i-1)/J}) \times \rho(A^{-1}B),$$

where $\delta_i(A, x^{k+(i-1)/J})$ is the corresponding reduction for the quadratic objective function with Hessian A in one subiteration over V_i initialized at $x^{k+(i-1)/J}$ and $\rho(\cdot)$ denotes the spectral radius of a matrix.

Proof. Without the loss of generality we assume that the objective function takes its minimum at the origin

$$\min_x f(x) = f(\overline{0})$$

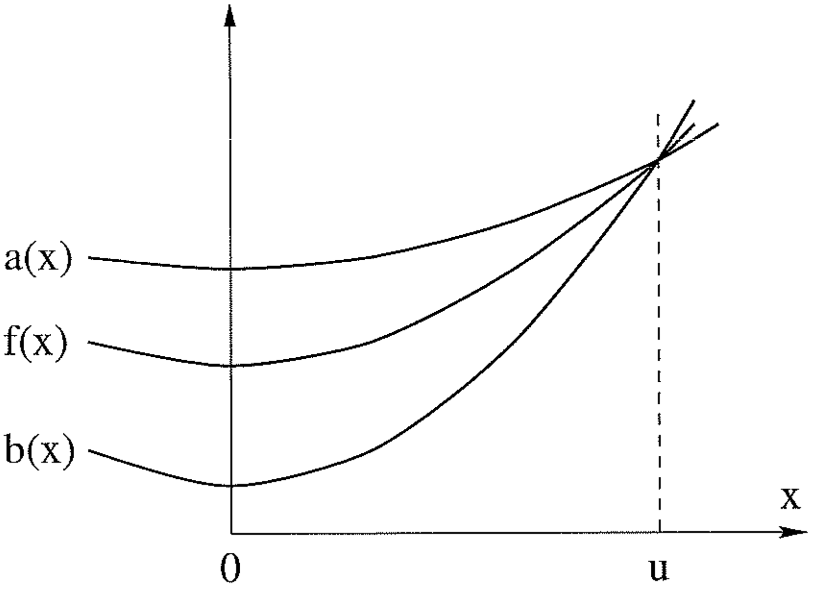


Figure 2.1: Theorem 2.3.

and for the sake of compactness we denote

$$u = x^{k+(i-1)/J}$$

$$u' = x^{k+i/J}.$$

We consider two quadratic functionals on V

$$a(x) = x^T Ax - u^T Au + f(u),$$

$$b(x) = x^T Bx - u^T Bu + f(u),$$

for which we have

$$a(u) = b(u) = f(u)$$

and the minimum of $a(x)$ and $b(x)$ is attained at $x = \overline{0}$ (see Figure 2.1). Besides, condition (2.16) implies

$$b(\overline{0}) \leq f(\overline{0}) \leq a(\overline{0})$$

and

$$f(x) - f(\overline{0}) \leq b(x) - b(\overline{0})$$

for any $x \in R^n$.

Let u'_a be a solution of the subspace problem with respect to the functional $a(x)$:

$$a(u'_a) = \min_{d_i \in V_i} a(u + d_i).$$

We have

$$\begin{aligned}
\frac{f(u') - f(\overline{0})}{f(u) - f(\overline{0})} &\leq \frac{f(u'_a) - f(\overline{0})}{f(u) - f(\overline{0})} \\
&\leq \frac{b(u'_a) - b(\overline{0})}{a(u) - a(\overline{0})} \\
&= \frac{a(u'_a) - a(\overline{0})}{a(u) - a(\overline{0})} \times \frac{b(u'_a) - b(\overline{0})}{a(u'_a) - a(\overline{0})} \\
&= \delta_i(A, u) \times \frac{{u'_a}^T B u'_a}{{u'_a}^T A u'_a} \\
&\leq \delta_i(A, u) \times \rho(A^{-1}B). \qquad \square
\end{aligned}$$

This theorem shows that if matrices A and B in (2.16) are close enough then the reduction of the general objective function (2.1) produced by Algorithm 2.2 can be arbitrarily close to the corresponding reduction for the quadratic objective (2.15).

2.2.2 Convergence Rate for Additive Schwarz Methods

In this section we consider the additive Schwarz method for solving 2.1.

Alg. 2.3 - ***Additive Schwarz Algorithm***

```
    Choose initial approximation $x^0$ and relaxation parameters

        $\alpha_i > 0, \quad \text{such that} \quad \sum_{i=1}^{J} \alpha_i = 1 \qquad (2.17)$

    for $k = 0$ until convergence
      for $i = 1 : J$
        in parallel find $d_i^k \in V_i$ such that

          $f(x^k + d_i^k) = \min_{d_i \in V_i} f(x^k + d_i) \qquad (2.18)$

      end
      $x^{k+1} = x^k + \sum_{i=1}^{J} \alpha_i d_i^k$
    end
end
```

Remark 2.1 *The additive Schwarz method was considered by Tai and Espedal [29] with*

$$\alpha_i > 0 \qquad \text{and} \qquad \sum_{i=1}^{J} \alpha_i \leq 1. \tag{2.19}$$

In the above algorithm the condition $\sum_{i=1}^{J} \alpha_i = 1$ is not restrictive because if the parameters α_i satisfy (2.19) we can append the set of V_i by a trivial subspace $V_{J+1} = \overline{0}$ and choose the corresponding $\alpha_{J+1} = 1 - \sum_{i=1}^{J} \alpha_i$.

We assume that the global error in the objective function can be bounded in terms of local gains over the subspaces. Later in this section we will give an equivalent form of this assumption that can be viewed as an analog of Assumption 1.1 (cf. [30]) for non-quadratic minimization.

Assumption 2.4 *There is a constant C such that for any $x \in V$ there exist such vectors $d_i \in V_i$ that the following estimate holds*

$$f(x) - f(x^*) \leq C \sum_{i=1}^{J} \alpha_i (f(x) - f(x + d_i)), \tag{2.20}$$

where x^* is the minimizer of $f(x)$.

Theorem 2.4 *Under assumptions 2.1 and 2.4 the additive Schwarz algorithm converges and the reduction of the objective function can be characterized as*

$$\frac{f(x^{k+1}) - f(x^*)}{f(x^k) - f(x^*)} \leq 1 - \frac{1}{C},$$

where C is given by (2.20).

Proof. The vectors d_i^k defined by (2.18) provide the optimal reduction in f and therefore

$$f(x^k + d_i^k) \leq f(x^k + d_i)$$

for any $d_i \in V_i$. This and (2.20) imply

$$f(x^k) - f(x^*) \leq C \sum_{i=1}^{J} \alpha_i (f(x^k) - f(x + d_i^k)). \tag{2.21}$$

From the convexity of $f(x)$ and the conditions on the relaxation parameters (2.17) it follows that

$$\begin{aligned} f(x^{k+1}) &= f\left(x^k + \sum_{i=1}^{J} \alpha_i d_i^k\right) \\ &= f\left(\sum_{i=1}^{J} \alpha_i (x^k + d_i^k)\right) \\ &\leq \sum_{i=1}^{J} \alpha_i (f(x^k + d_i^k)) \end{aligned}$$

and form (2.21) we get

$$f(x^k) - f(x^*) \leq C(f(x^k) - f(x^{k+1}))$$

or

$$C(f(x^{k+1}) - f(x^*)) \leq (C-1)(f(x^k) - f(x^*))$$

and

$$\frac{f(x^{k+1}) - f(x^*)}{f(x^k) - f(x^*)} \leq 1 - \frac{1}{C}. \qquad \square$$

As we pointed out Assumption 2.4 can be put in a different form. Using condition (2.17) we can perform the following chain of transformations starting with (2.20)

$$\begin{aligned}
f(x) - f(x^*) &\leq Cf(x) - C\sum_{i=1}^{J} \alpha_i f(x+d_i) \\
C\sum_{i=1}^{J} \alpha_i f(x+d_i) &\leq (C-1)f(x) + f(x^*) \\
C\left(\sum_{i=1}^{J} \alpha_i f(x+d_i) - f(x^*)\right) &\leq (C-1)\left(f(x) - f(x^*)\right) \\
\sum_{i=1}^{J} \alpha_i \left(f(x+d_i) - f(x^*)\right) &\leq \frac{C-1}{C}\left(f(x) - f(x^*)\right)
\end{aligned}$$

and therefore Assumption 2.4 is equivalent to

Assumption 2.5 *There is a constant $C > 0$ such that for any $x \in V$ there we can find $d_i \in V_i$ such that*

$$\sum_{i=1}^{J} \alpha_i \left(f(x+d_i) - f(x^*)\right) \leq \left(1 - \frac{1}{C}\right)\left(f(x) - f(x^*)\right). \tag{2.22}$$

We can compare this assumption with Assumption 1.1 we mentioned in Introduction. In the case when $f(x)$ is quadratic (2.13) the term $(f(x) - f(x^*))$ in the right hand side of the above inequality is equal to $\|x - x^*\|_A^2$ and therefore coincides with the right hand side term of (1.10) applied to the error vector $x - x^*$. The left hand side of (2.22) generalizes the decomposition of the error vector weighted with the relaxation parameters α_i.

Remark 2.2 *An important distinction between our approach and the theory of Tai and Espedal [29] is that in this section we assume the convexity of $f(x)$ in (2.1) but not its smoothness.*

2.3 Constrained Optimization

In this section we allow constraints in the problem (2.1):

$$\min_{x \in C} f(x), \tag{2.23}$$

where $C \subseteq R^n$ is a closed convex feasible set and the function $f(x)$ satisfies the following assumption on the level surfaces of $f(x)$.

Assumption 2.6 *The sets*

$$f_h = \{x \mid f(x) \leq h\} \tag{2.24}$$

are strictly convex for any h.

We will discuss examples of problems satisfying the assumptions of this section in Chapter 4.

Remark 2.3 *Assumption 2.6 is weaker than the requirement for f to be convex, for example $f(x) = \|x\|^{\frac{1}{2}}$ is not convex in R^n but satisfies the above assumption.*

Also, as we did in the first section of this chapter, to ensure that (2.23) has a solution we require, that f is coercive on C:

Assumption 2.7 *The sets*

$$f_h \cap C = \{x \mid x \in C \ , \ f(x) \leq h\} \tag{2.25}$$

are bounded for all h.

The following result is straightforward.

Lemma 2.1 *Under Assumptions 2.6 and 2.7 the minimization problem (2.23) has a unique solution.*

Proof. First we chose h for which

$$C_h = f_h \cap C$$

is nonempty. We can replace (2.23) with an equivalent restricted problem

$$\min_{x \in C_h} f(x). \tag{2.26}$$

As an intersection of two closed sets the set C_h is closed. Besides, the assumption that $f(x)$ is coercive implies that C_h is bounded and therefore, since $C_h \in R^n$, it is compact. That and continuity of $f(x)$ imply that the restricted problem (2.26) has a solution. To show uniqueness we assume that x_1 and x_2 are solutions of (2.23):

$$f(x_1) = f(x_2) = \min_{x \in C} f(x) = H. \tag{2.27}$$

Convexity of C implies that the convex combination $\frac{1}{2}(x_1 + x_2)$ is in C, but since the set f_H is strictly convex we have

$$f\left(\frac{x_1 + x_2}{2}\right) < H,$$

which violates (2.27). □

We can adjust the multiplicative algorithm for the constrained problem (2.23).

```
Alg. 2.4 - Multiplicative Method for Constrained Optimization
    Starting with $x^0 \in C$ for $k = 0$ until convergence
        for $i = 1 : J$
            find $x^{k+i/J} = x^{k+(i-1)/J} + d_i^k \subset C$ such that
                $f(x^{k+i/J}) = \min_{\substack{d_i \in V_i \\ x^{k+(i-1)/J} + d_i \subset C}} f(x^{k+(i-1)/J} + d_i)$    (2.28)
        end
    end
end
```

Remark 2.4 *Since the restricted problem (2.28) satisfies the assumptions we made for the original problem (2.23) its solution is well-defined (Lemma 2.1).*

We can formulate the theorem that under the assumptions of the above lemma the iterates generated by Algorithm 2.4 converge. Its proof is similar to the proof of Theorem 2.1.

Theorem 2.5 *Under the assumptions of Lemma 2.1 Algorithm 2.4 applied to the problem (2.23) produces a convergent sequence of iterates and the limit*

$$\lim_{k \to \infty} x^k = \tilde{x} \in C$$

satisfies

$$f(\tilde{x}) = \min_{\substack{d_i \in V_i \\ \tilde{x}+d_i \subset C}} f(\tilde{x} + d_i), \qquad i = 1 \dots J. \tag{2.29}$$

Proof. Since all the iterates $x^{k+i/J}$ are contained in the compact set

$$\{x \mid f(x) \leq f(x^0)\} \cap C$$

they have a limit point. To show that this limit point is unique we assume the contrary. That would imply that there is a sequence of indices k_j and an integer $i' \geq 0$ such that

$$\tilde{x_1} = \lim_{k_j \to \infty} x^{k_j + (i'-1)/J} \neq \lim_{k_j \to \infty} x^{k_j + i'/J} = \tilde{x_2}. \tag{2.30}$$

We have

$$\tilde{x_2} - \tilde{x_1} = \lim_{k_j \to \infty} (x^{k_j + i'/J} - x^{k_j + (i'-1)/J}) \in V_i. \tag{2.31}$$

Since the sequence $f(x^{k+i/J})$ is non-increasing we must have

$$f(\tilde{x_1}) = f(\tilde{x_2}) = h$$

and since by assumption the set $\{x \mid f(x) \leq h\}$ is strictly convex the midpoint satisfies

$$f\left(\frac{\tilde{x_1} + \tilde{x_2}}{2}\right) < h$$

and, because f is continuous, there exists an $\epsilon > 0$ such that

$$f\left(\frac{\tilde{x_1} + \tilde{x_2}}{2} + r\right) < h \tag{2.32}$$

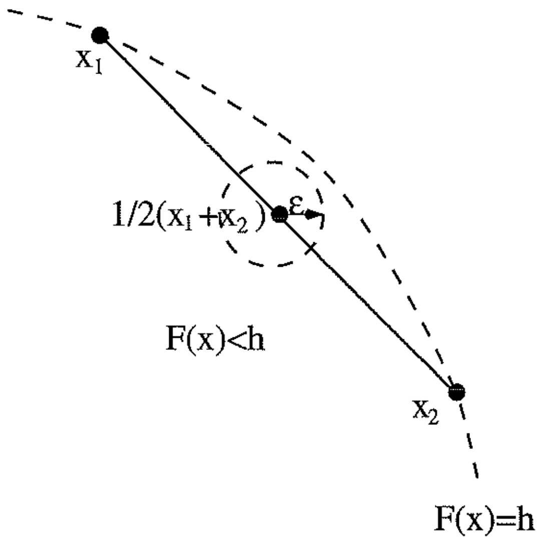


Figure 2.2: Theorem 2.5.

for any $\|r\| < \epsilon$ (see figure).

Finally, since $\tilde{x_1}$ is a limit point of $\{x^{k_j+(i'-1)/J}\}$, there should be some j' such that

$$\|f(x^{k_{j'}+(i'-1)/J}) - \tilde{x_1}\| = \|r\| < \epsilon$$

and therefore because of (2.31) and (2.32)

$$f(x^{k_{j'}+i'/J}) \leq f\left(x^{k_{j'}+(i'-1)/J} + \frac{\tilde{x_2} - \tilde{x_1}}{2}\right) = f\left(\frac{\tilde{x_1} + \tilde{x_2}}{2} + r\right) < h$$

which violates (2.30) since the sequence $\{f(x^{k_j+i/J})\}$ is non-increasing. Therefore x^k has a unique limit point $\tilde{x}$ which is its limit.

To prove (2.29) we again assume the contrary which in this case implies that there exists $d_i \in V_i$ for some i such that $\tilde{x} + d_i \in C$ and

$$f(\tilde{x} + d_i) < f(\tilde{x}).$$

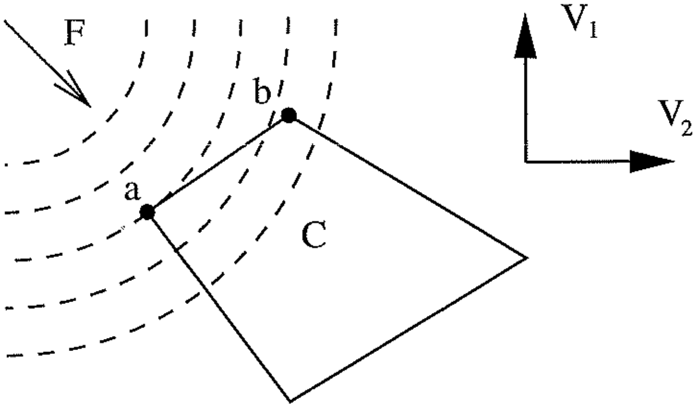


Figure 2.3: Possible convergence to a non-solution.

Since C is convex and f is continuous there exists $\tilde{d}_i \in V_i$ and $\epsilon > 0$ such that

$$f(\tilde{x} + \tilde{d}_i + r) < f(\tilde{x}) \tag{2.33}$$

and

$$\tilde{x} + \tilde{d}_i + r \in C$$

for any $\|r\| < \epsilon$. If $\tilde{x}$ is the limit of $\{x^{k+i/J}\}$ we have for some j

$$\|x^{k_j+(i-1)/J} - \tilde{x}\| < \epsilon$$

and therefore, because of (2.33)

$$f(x^{k_j+i/J}) \leq f(x^{k_j+(i-1)/J} + \tilde{d}_i) < f(\tilde{x}),$$

which is impossible since $f(x^{k+i/J})$ is non-increasing and its limit is $f(\tilde{x})$. □

We should point out here that condition (2.29) doesn't imply that $\tilde{x}$ is the solution of the problem (2.23). In Figure 2.3 the circles show the level curves for a function $F : R^2 \to R$ whose minimum over the region C is attained at the point a. However, with the choice of two subspaces shown in the figure point b satisfies the condition (2.29) as well as any other point on the line segment connecting a and b. In the last section of Chapter 4 we will observe a similar situation in the numerical examples for the l_1-approximation problem.

2.4 Multilevel Method

Each subspace step of the algorithms presented in this chapter results in a minimization problem with fewer unknowns than there are in the original problem (2.1). If this reduction in the number of unknowns results in a significant reduction in computational work we can make the multiplicative algorithms more efficient if we apply them recursively. In order to do that we have to represent each subdomain and the coarse grid as a union of overlapping subdomains and add a coarse grid for each of them. To apply an l-level method we should perform l nested steps of the recursion. This recursion applied to the multiplicative method with coarse grid correction can be viewed as a multilevel method and the iterations performed on each level as the smoothing of the solution. The recursion can be stopped once the dimensions of the subspace problems reach some small enough fixed size C_c.

We present the resulting method with m smoothing steps.

```
Alg. 2.5 - Multilevel method
    Starting with $x^0$ for $k = 0, 1, \dots$
        for $i = 1 : J$
            if the size of subproblem is smaller than $C_c$ solve or
            otherwise apply $m$ steps of Alg. 2.5 to find $x^{k+i/J}$ such that
                $f(x^{k+i/J}) = \min_{d_i^k \in V_i} f(x^{k+(i-1)/J} + d_i^k)$
        end
        if the size of the coarse subproblem is smaller than $C_c$ solve or
        otherwise apply $m$ steps of Alg. 2.5 to find $x_c^k$ such that
            $f(x_c^k) = \min_{d_c^k \in V_c} f(u^{k+(J-1)/J} + d_c)$
        $x^{k+1} = x_c^k$
    end
end
```

If we choose m to be large, the subspace problems will be solved with high accuracy and the algorithm above becomes equivalent to Algorithm 2.2. On the other hand we can make m relatively small. In this case we don't solve the subproblems exactly but perform few smoothening steps reducing $f(x)$.

CHAPTER 3

Applications to PDE-Based Optimization Problems

In this chapter we apply the subspace correction methods to the variational formulation of a symmetric PDE-based eigenvalue problem and present theoretical and numerical results that support the conclusions of Chapter 2. We generalize this methods to accommodate the problem of finding several lowest eigenmodes also presented in a variational form. We also apply the domain decomposition methods for eigenvalue problem to the graph partitioning using spectral bisection.

An important feature of the problems considered in this chapter is that the application of subspace correction methods results in subspace problems that are of the same type as the original problem. This feature allows the recursive application resulting in multilevel methods.

3.1 Domain Decomposition Methods for Elliptic Eigenvalue Problems

In this section we analyze the application of the multiplicative Schwarz methods to the eigenvalue problem. The idea to use the coordinate relaxation applied directly to a matrix eigenvalue problem goes back to the book by Fadeev and Fadeeva [11] (1963) where they applied a technique similar to the Gauss-Seidel method for minimizing the Rayleigh Quotient. This idea was generalized by Kaschiev [16] and Maliassov [21] who considered the subspace correction method for eigenvalue problems. We will extend the results of [16] and [21] for the multiplicative Schwarz method by considering a two-level scheme. We prove the convergence of the algorithm and present the convergence rate analysis. We also describe the recursive implementation of the method which results in a multilevel algorithm. Finally, we discuss an alternative variational formulation of the eigenvalue problem which is mathematically equivalent to the minimization of the Rayleigh Quotient but is more suitable for theoretical considerations.

3.1.1 Subspace Correction for Eigenvalue Problems

Let us consider the problem of finding the smallest eigenvalue λ and the corresponding eigenvector u of

$$Lu \equiv -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i} a_{i,j} \frac{\partial u}{\partial x_j} + p(\overline{x})u = \lambda u \tag{3.1}$$

$$x \in \Omega \ , \qquad u \mid_{\partial\Omega} = 0 \qquad a_{i,j} > 0,$$

where Ω is a bounded region in R^2 and $a_{i,j}(\overline{x}) = a_{j,i}(\overline{x})$, $p(\overline{x}) \geq 0$ are piecewise smooth real functions.

To discretize the problem, we can perform a triangulation of Ω with triangles of quasi-uniform size h and use the standard finite element approach to represent

(3.1) as

$$Ax = \lambda Mx, \tag{3.2}$$

where x is the unknown n-vector and $A = A^T > 0$, $M = M^T > 0$ are the stiffness and the mass matrices respectively that satisfy

Assumption 3.1 *The discretized stiffness matrix A is an M-matrix and the mass matrix M is nonnegative componentwise.*

We will also need an assumption of the irreducibility of matrices A and M which follows from

Assumption 3.2 *The triangulation of Ω that gives rise to matrices A and M forms a connected graph in R^2.*

We have

Lemma 3.1 *Under Assumptions 3.1 and 3.2 the smallest eigenvalue of (3.2) λ_1 is simple and the associated eigenvector x_1 can be taken componentwise positive.*

Proof. We can rewrite (3.2) as

$$A^{-1}Mx = \lambda^{-1}x. \tag{3.3}$$

From Assumption 3.1 it follows that all the entries of A^{-1} and M are nonnegative and Assumption 3.2 implies that these matrices are irreducible. Therefore the product $A^{-1}M$ is nonnegative and irreducible. The Perron-Frobenius theorem (see for example [25]) guarantees that (3.3) has a simple largest eigenvalue ρ_{max} and the corresponding eigenvector x_1 can be chosen componentwise positive. The original eigenvalue problem has the smallest eigenvalue $\lambda_1 = \rho_{max}^{-1}$ with the same eigenvector x_1 □.

This lemma provides a discrete version of the corresponding result for the continuous problem (3.1) [14].

We will compute the smallest eigenvalue of (3.2) minimizing the Rayleigh quotient

$$\lambda_1 = \min_{x \neq 0} \rho(x) = \min_{x \neq 0} \frac{x^T A x}{x^T M x}. \tag{3.4}$$

To apply the domain decomposition technique to this problem we represent Ω as a union of overlapping subdomains: $\Omega = \cup_{i=1}^{J} \Omega_i$. Let $\{V_i\}_{i=1}^{J}$ be the finite element subspaces corresponding to this partition and let P_i^T denote the orthogonal projection into the subspace V_i, its transpose P_i is the prolongation operator from V_i to $V = R^n$. We also introduce the M-norm of a vector $\|x\|_M = (x^T M x)^{1/2}$.

The multiplicative Schwarz algorithm for solving (3.4) analogous to Algorithm 2.1 of Chapter 2 was proposed Kaschiev [16] and Maliassov [21]. They showed that the subspace minimization (2.2) for problem (3.4)

$$\rho(x^{k+i/J}) = \min_{d_i \in V_i} \rho(x^{k+(i-1)/J} + P_i d_i) \tag{3.5}$$

results in an eigenvalue problem of size $(n_i+1) \times (n_i+1)$, where n_i is the dimension of the subspace V_i.

To see that we introduce the notation

$$\tilde{d}_i^k = \begin{pmatrix} d_i^k \\ \\ 1 \end{pmatrix} \tag{3.6}$$

$$P_i^k = \begin{pmatrix} P_i & x^{k+\frac{i-1}{J}} \end{pmatrix} \tag{3.7}$$

and represent the subsequent iterate

$$x^{k+\frac{i}{J}} = x^{k+\frac{i-1}{J}} + P_i d_i^k$$

as

$$x^{k+\frac{i}{J}} = P_i^k \tilde{d}_i^k.$$

We can rewrite (3.5) as

$$\begin{aligned} \min_{\tilde{d}_i^k} \rho(P_i^k \tilde{d}_i^k) &= \min_{\tilde{d}_i^k} \frac{(P_i^k \tilde{d}_i^k)^T A (P_i^k \tilde{d}_i^k)}{(P_i^k \tilde{d}_i^k)^T M (P_i^k \tilde{d}_i^k)} \\ &= \min_{\tilde{d}_i^k} \frac{\tilde{d}_i^{k^T} A_i^k \tilde{d}_i^k}{\tilde{d}_i^{k^T} M_i^k \tilde{d}_i^k}, \end{aligned} \tag{3.8}$$

where

$$\begin{aligned} A_i^k &= P_i^{k^T} A P_i^k \\ M_i^k &= P_i^{k^T} M P_i^k. \end{aligned} \tag{3.9}$$

The form of (3.8) suggests that the subspace problem (3.5) is the eigenvalue problem with matrices A_i^k and M_i^k

$$A_i^k \tilde{d}_i^k = \rho(x^{k+\frac{i}{J}}) M_i^k \tilde{d}_i^k. \tag{3.10}$$

The matrices A_i^k and M_i^k preserve the sparsity of the original matrices A and M except for the last row and column, therefore the minimization subproblem can be efficiently solved by standard methods such as inverse iteration.

Algorithm 2.1 of Chapter 2 applied to the eigenvalue problem (3.4) becomes

```
Alg. 3.1 - Multiplicative Algorithm for Eigenvalue Problem
    Starting with $x^0$ for $k = 0$ until convergence
        for $i = 1 : J$
            using $x^{k+(i-1)/J}$ construct matrices $A_i^k$ and $M_i^k$ and solve (3.10)
            for min. eigenvector $\tilde{d}_i^k = \begin{pmatrix} d_i^k \\ 1 \end{pmatrix}$, then update and normailize

                $x^{k+\frac{i}{J}} = x^{k+(i-1)/J} + P_i d_i^k$        (3.11)

                $x^{k+\frac{i}{J}} \leftarrow \frac{x^{k+\frac{i}{J}}}{||x^{k+\frac{i}{J}}||_M}$        (3.12)

        end
    end
end
```

The normalization step (3.12) can be performed either after each subspace iteration or after a loop over all the subspaces is completed. The convergence of this algorithm was proven in [16] and [21] with the assumption that the initial guess x^0 satisfies

$$\lambda_1 < \rho(x^0) < \lambda_2. \tag{3.13}$$

Lui [20] pointed out that the algorithm can break down in certain degenerate cases when it faces the problem of not being able to find the required correction from a current subspace (3.5). This happens when the solution to the eigenvalue problem (3.10) has a zero component corresponding to the previous iterate

$$A_i^k \begin{pmatrix} d_i^k \\ 0 \end{pmatrix} = \rho(x^{k+\frac{i}{J}}) M_i^k \begin{pmatrix} d_i^k \\ 0 \end{pmatrix} \tag{3.14}$$

and therefore cannot be scaled to match (3.6). Lui proved convergence to the smallest eigenvalue λ_1 for a modified algorithm in the case of two subdomains under condition (3.13).

Before we proceed we formulate a lemma that guarantees that once initialized with a componentwise positive vector the iterates produced by Algorithm 3.1 cannot have components of mixed signs.

Lemma 3.2 *Under Assumptions 3.1 and 3.2 if Algorithm 3.1 is initialized with a componentwise positive x^0 and if the subsequent iterates $x^{k+\frac{i}{J}}$ are defined, they are also componentwise positive.*

Proof. We will argue by contradiction. Suppose x is the first iterate of Algorithm 3.1 that has a non-positive component. First we assume that x does not have zero components and therefore can be represented as $x = \begin{pmatrix} x_p \\ -x_n \end{pmatrix}$, where components of x_p and x_n are positive. Let

$$A = \begin{pmatrix} A_{pp} & A_{pn} \\ A_{pn}^T & A_{nn} \end{pmatrix} \qquad M = \begin{pmatrix} M_{pp} & M_{pn} \\ M_{pn}^T & M_{nn} \end{pmatrix}$$

be the corresponding partitionings of A and M. For the vector $x_+ = \begin{pmatrix} x_p \\ x_n \end{pmatrix}$ we have

$$\begin{aligned} x_+^T A x_+ &= x_p^T A_{pp} x_p + 2x_p^T A_{pn} x_n + x_n^T A_{nn} x_n \\ &< x_p^T A_{pp} x_p - 2x_p^T A_{pn} x_n + x_n^T A_{nn} x_n \\ &= x^T A x. \end{aligned}$$

The inequality is strict because Assumption 3.2 implies that A and M are not block diagonal and therefore the block A_{pn} cannot be zero. Similarly

$$x_+^T M x_+ > x^T M x$$

and

$$\rho(x_+) < \rho(x).$$

Since we assumed that x is produced from a positive vector by changing some of its components we conclude that x_+ can be formed the same way and therefore x cannot correspond to the optimal choice of the subspace correction.

If x has zero components then for x_+ that is again formed of the absolute values of the components of x the above inequalities are not necessarily strict

$$\rho(x_+) \leq \rho(x).$$

In this case the representation of x_+ is

$$x_+ = \begin{pmatrix} x_p \\ x_0 \end{pmatrix}, \tag{3.15}$$

where $x_p > 0$ and x_0 is a zero block. Then the vector

$$\tilde{x}_+ = \begin{pmatrix} x_p \\ -A_{00}^{-1} A_{p0}^T x_p \end{pmatrix},$$

where A_{00} and A_{p0} are the blocks of A corresponding to the partition (3.15), is componentwise positive. We have

$$\begin{aligned} \tilde{x}_+^T A \tilde{x}_+ &= x_+^T A x_+ - x_p^T A_{p0} A_{00}^{-1} A_{p0}^T x_p \\ &< x_+^T A x_+ \\ &\leq x^T A x \end{aligned}$$

and

$$\begin{aligned} \tilde{x}_+^T M \tilde{x}_+ &> x_+^T M x_+ \\ &\geq x^T M x. \end{aligned}$$

For the same reasons as above the vector x cannot result from the optimal subspace correction of a positive vector because the vector $\tilde{x}_+$ can be produced by a correction from the same subspace and results in a greater reduction of ρ. □

The breakdown of the algorithm (3.14) occurs if the eigenvector corresponding to the minimum eigenvalue is contained in one of the subspaces V_i. To prevent this we can formulate a natural assumption that this eigenvector is not contained in any of the subspaces V_i, which means that the Rayleigh Quotient over each of the subspaces V_i should exceed the minimum eigenvalue λ_1.

Assumption 3.3 *For all subspaces V_i the constants*

$$\lambda_1^{(i)} \equiv \min_{d_i \in V_i} \rho(d_i) = \min_{d_i \in V_i} \frac{d_i^T P_i^T A P_i d_i}{d_i^T P_i^T M P_i d_i} \tag{3.16}$$

satisfy $\lambda_1^{(i)} > \lambda_1$.

Lemma 3.3 *Let Assumptions 3.1, 3.2 and 3.3 hold and let Algorithm 3.1 be initialized with a componentwise positive vector x^0. Then for all i and k the eigenvector $\tilde{d}_i^k$ corresponding to the lowest eigenvalue of (3.10) has a nonzero last component and therefore the update step (3.11) of Algorithm 3.1 is well-defined.*

Proof. Assuming the contrary we can rewrite (3.14) using (3.7) and (3.9) as

$$P_i^T A P_i d_i^k = \rho(P_i d_i^k) P_i^T M P_i d_i^k$$

$$x^{k+(i-1)/J^T}(A P_i d_i^k - \rho(P_i d_i^k) M P_i d_i^k) = 0. \tag{3.17}$$

From the first condition it follows that $(\rho(P_i d_i), d_i)$ is the lowest eigenpair of the problem with matrices $P_i^T A P_i$ and $P_i^T M P_i$. These matrices satisfy Assumption 3.1, and because of Lemma 3.1 the eigenvector d_i^k can be taken componentwise positive. The residual vector

$$r_i^k = A P_i d_i^k - \rho(P_i d_i^k) M P_i d_i^k$$

satisfies $P_i^{\perp} r_i^k \leq 0$ componentwise since, as an M-matrix, A has non-positive off-diagonal block $P_i^{\perp} A P_i$ and the corresponding block of M is nonnegative. Besides, $P_i r_i^k = 0$ and therefore, $r_i^k \leq 0$ componentwise. Since $x^{k+(i-1)/J}$ is componentwise positive (Lemma 3.2) from (3.17) it follows that $r_i^k = 0$ which in turn implies that $P_i d_i$ is an eigenvector of the unrestricted problem (3.2). Since $P_i d_i$ has all its components of the same sign we have $\rho(P_i d_i) = \lambda_1$ which violates Assumption 3.3. □

This lemma shows that the subspace correction update (3.11) is always well-defined. Now we can formulate the convergence result.

Theorem 3.1 *Let Assumptions 3.1, 3.2 and 3.3 hold then vectors x^k and the corresponding Rayleigh quotients $\rho(x^k)$ produced by Algorithm 3.1 converge to the lowest eigenpair of discretized problem (3.2) if x^0 is chosen componentwise positive.*

Proof. Since the sequence $\rho(x^{k+i/J})$ is non-increasing and bounded from below by λ_1 it converges. We will show that its limit $\tilde{\rho}$ is an eigenvalue of (3.2) and that the sequence x^k converges to the corresponding eigenvector $\tilde{x}$. Then using Lemmas 3.1 and 3.2 we will show that $\tilde{\rho} = \lambda_1$.

First we point out that for any k and i we have

$$\rho(x^{k+\frac{i}{J}}) < \lambda_1^{(i)}, \tag{3.18}$$

where $\lambda_1^{(i)}$ is given by (3.16). To show that we notice that choosing $x = Pd_i$, where d_i is the optimal vector for (3.16) we have $\rho(x) = \lambda_1^{(i)}$ but, as it follows from

Lemma 3.3, the optimal $x^{k+\frac{i}{J}}$ is of the form $x^{k+\frac{i-1}{J}} + P_i d_i^k$ and since the lowest eigenvalue of the restricted problem is simple we have (3.18).

From (3.10) it follows that

$$P_i^T \left(A - \rho(x^{k+\frac{i}{J}})M\right) P_i d_i^k = P_i^T \left(\rho(x^{k+\frac{i}{J}})M - A\right) x^{k+\frac{i-1}{J}} \tag{3.19}$$

and

$$x^{k+\frac{i-1}{J}^T} \left(A - \rho(x^{k+\frac{i}{J}})M\right) P_i d_i^k = x^{k+\frac{i-1}{J}^T} M x^{k+\frac{i-1}{J}} \left(\rho(x^{k+\frac{i}{J}}) - \rho(x^{k+\frac{i-1}{J}})\right).$$

Combining the two equations above we get

$$d_i^{k^T} P_i^T \left(A - \rho(x^{k+\frac{i}{J}})M\right) P_i d_i^k = x^{k+\frac{i-1}{J}^T} M x^{k+\frac{i-1}{J}} \left(\rho(x^{k+\frac{i-1}{J}}) - \rho(x^{k+\frac{i}{J}})\right). \tag{3.20}$$

Since $\rho(x^{k+i/J})$ converges as $k \to \infty$ and because of normalization (3.12) the right hand side of the above expression converges to zero as $k \to \infty$ for every i therefore the left hand side converges to zero. Using (3.18) and (3.16) we can estimate the left hand side of (3.20)

$$\begin{aligned} d_i^{k^T} P_i^T (A - \rho(x^{k+\frac{i}{J}})M) P_i d_i^k &> \left(\lambda_1^{(i)} - \rho(x^{k+\frac{i}{J}})\right) d_i^{k^T} P_i^T M P_i d_i^k \\ &\geq \left(\lambda_1^{(i)} - \rho(x^{0+\frac{i}{J}})\right) d_i^{k^T} P_i^T M P_i d_i^k \\ &> \alpha d_i^{k^T} P_i^T M P_i d_i^k, \end{aligned}$$

where

$$\alpha = \min_{i=1}^{J} \left(\lambda_1^{(i)} - \rho(x^{0+\frac{i}{J}})\right) > 0$$

and since the left hand side of (3.20) converges to zero and M is positive definite we conclude that the vectors $P_i d_i^k$ converge to zero componentwise as $k \to \infty$ for every i and the iterates $x^{k+i/J}$ converge to an M-normalized limit $\tilde{x}$.

If we take the limit of (3.19) we get

$$P_i^T(\tilde{\rho}M - A)\tilde{x} = 0$$

and since the prolongation operators P_i span the entire space we have

$$(A - \tilde{\rho}M)\tilde{x} = 0$$

which shows that $(\tilde{\rho}, \tilde{x})$ is an eigenpair of (3.2). From Lemma 3.2 it follows that $\tilde{x}$ is componentwise positive and since the eigenvectors are M-orthogonal and the smallest eigenvalue is simple and the corresponding eigenvector is componentwise positive we conclude that $(\tilde{\rho}, \tilde{x})$ is the lowest eigenmode of (3.2). □

3.1.2 Coarse Grid Correction and Multilevel Method

We can modify Algorithm 3.1 to include the coarse grid correction after a loop over the subdomains is completed as we did for Algorithm 2.2. The effect of the this correction for a model problem of 2-D Laplacian in a unit square is shown in Figures 3.1 and 3.2. We can see that without the coarse grid the convergence rate is dependent on both the meshsize and the number of subdomains whereas after the coarse grid correction has been added the convergence rate becomes independent of both fine and coarse meshsizes h and H.

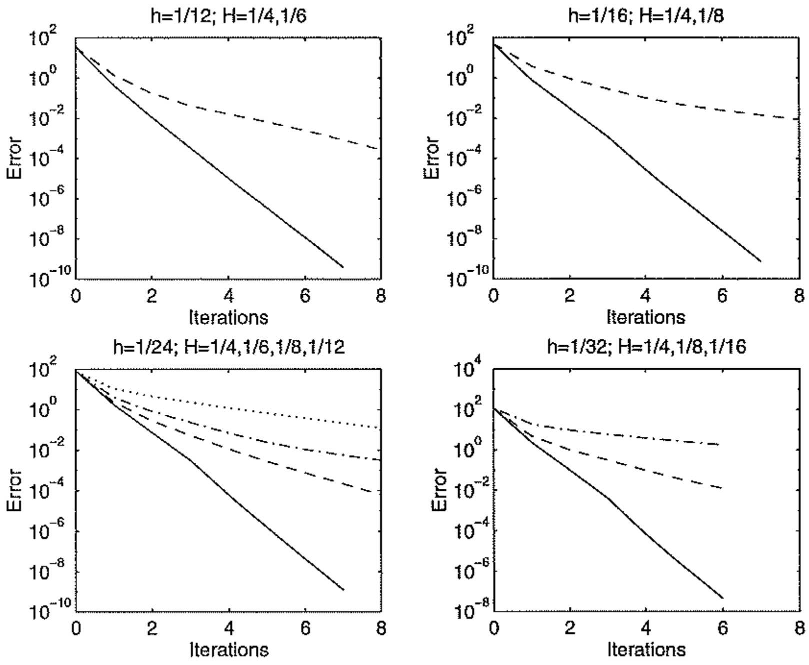


Figure 3.1: Algorithm 3.1 for 2D Laplacian without the coarse grid correction.

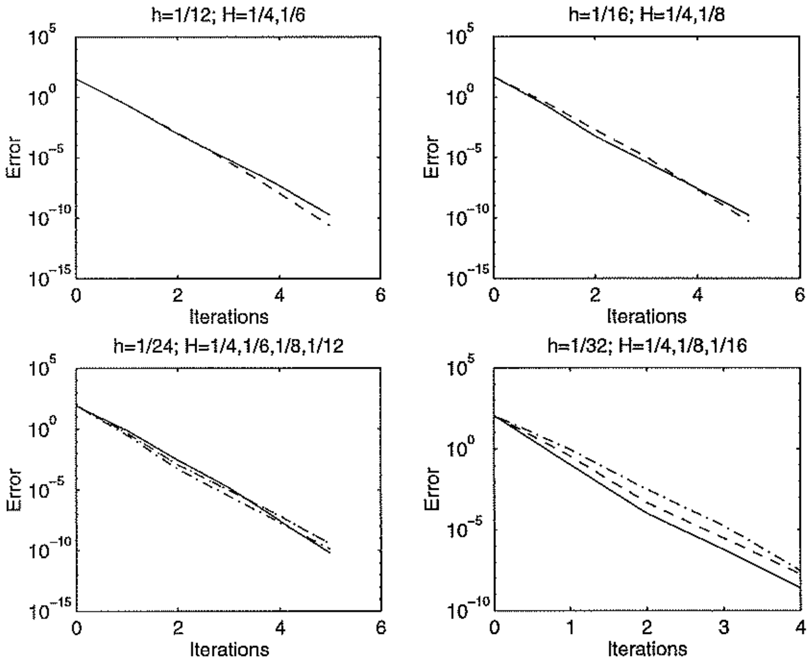


Figure 3.2: Algorithm 3.1 for 2D Laplacian with the coarse grid correction.

Since the subspace problem is of the same type as the original one i.e. a generalized eigenvalue problem, we can make the algorithm more efficient by applying it recursively. Instead of solving the eigenvalue subproblem over a subdomain by some other method we can apply several iterations of the same algorithm. Applying that recursion to the multiplicative method with coarse grid correction we can view the resulting scheme as a multilevel method and the iterations performed on each level as the smoothing of the solution. As we pointed out in section 2.4, the recursion can be stopped once the subproblems we are solving reached some small enough fixed size C_c and traditional inverse iteration can be applied.

We should remark that at each successive level the subproblem matrices get appended by a dense row and column. We may think of the corresponding component as of a fictitious gridpoint for a subdomain at the current level. At the coarsest level the subproblem matrices will be appended by rows and columns coming from the previous levels.

Though the presented algorithms are sequential we can add some degree of parallelism using multicoloring technique (see for example [8]).

3.1.3 Alternative Variational Characterization of the Smallest Eigenvalue

A different variational formulation for the symmetric positive definite eigenvalue problem (3.2) was described by Mathew and Reddy (94) [23]. They pointed out that the minimal eigenpair (x_1, λ_1) can be characterized as:

$$\sigma(x_1) = \min_x \sigma(x) \equiv \min_x (x^T Ax + \mu(1 - x^T Mx)^2) \tag{3.21}$$

with

$$\lambda_1 = 2\mu - \sqrt{4\mu^2 - 4\mu\sigma(x_1)}$$

and

$$\|x_1\|_M^2 = 1 - \frac{\lambda_1}{2\mu} \tag{3.22}$$

for any

$$\mu > \lambda_1/2. \tag{3.23}$$

Unlike the Rayleigh Quotient minimization, formulation (3.21) is unconstrained. The μ-term in $\sigma(x)$ serves as a barrier to pull x away from the trivial solution 0.

Using

$$P_i^k = \begin{pmatrix} P_i & x^{k+(i-1)/J} \end{pmatrix}$$

$$\tilde{d}_i^k = \begin{pmatrix} d_i^k \\ \alpha_i^k \end{pmatrix},$$

we can write the minimizaition step (2.2) of the multiplicative algorithm as

$$\begin{aligned} \sigma(x^{k+i/J}) &= \min_{d_i \in V_i, \alpha} J(P_i^k \tilde{d}_i) \\ &= \min_{\tilde{d}_i} \left[(P_i^k \tilde{d}_i)^T A (P_i^k \tilde{d}_i) + \mu(1 - (P_i^k \tilde{d}_i)^T M (P_i^k \tilde{d}_i))^2 \right] \\ &= \min_{\tilde{d}_i} \left[\tilde{d}_i^T A_i^k \tilde{d}_i + \mu(1 - \tilde{d}_i^T M_i^k \tilde{d}_i)^2 \right] \end{aligned}$$

with

$$A_i^k = P_i^{k^T} A P_i^k$$

$$M_i^k = P_i^{k^T} M P_i^k.$$

We can see that the subspace problem is again of the same type as the original problem (3.21). To make it the eigenvalue problem with matrices (A_i^k, M_i^k) we should make a restriction on μ similar to (3.23) replacing λ_1 by the current value of the Rayleigh quotient $\rho(x^{k+i/J})$. Since the sequence of the Rayleigh Quotients is non-increasing we can choose μ satisfying

$$\mu > \rho(x_0)/2 = \rho_0/2. \tag{3.24}$$

Similarly condition (3.22) becomes

$$\begin{aligned} \|x^{k+i/J}\|_M^2 &= 1 - \tfrac{\rho(x^{k+i/J})}{2\mu} \\ &\geq 1 - \tfrac{\rho(x_0)}{2\mu}. \end{aligned} \tag{3.25}$$

Remark 3.1 *For any choice of μ satisfying (3.24), one subspace correction step for formulations (3.4) and (3.21) results in the same reduced generalized eigenvalue problem with matrices (3.9). The application of the multiplicative Schwarz algorithm to both formulations results in the same approximations to the lowest eigenvalue and the approximations to the eigenvector are the same up to normalization.*

The following lemma estimates the Hessian matrix of $\sigma(x)$

$$\begin{aligned} H(x) &\equiv \nabla^2 \sigma(x) \\ &= 2(A - 2\mu M) + 8\mu M(xx^T)M + 4\mu(x^T M x)M. \end{aligned} \tag{3.26}$$

in terms of A near the solution of (3.21). This estimate will allow us to use the theory of Multiplicative Schwarz methods for minimization problems developed by Tai and Espedal [29].

Lemma 3.4 *Let*

$$\rho(x) \leq \rho_0 < \frac{\lambda_1 + \lambda_2}{2} \tag{3.27}$$

and

$$1 - \frac{\rho_0}{2\mu} \leq \|x\|_M^2 \leq 1. \tag{3.28}$$

Then if $\mu = \frac{3}{4}\rho_0$ the Hessian (3.26) satisfies

$$c_1 A < H(x) < c_2 A, \tag{3.29}$$

where the constants c_1 and c_2 can be chosen independently of the discretization parameters.

Remark 3.2 *Conditions (3.27) and (3.28) provide the directional and the radial restrictions on x. The equivalence of the minimization of $\sigma(x)$ and $\rho(x)$ (Remark 3.1) implies that that if condition (3.27) is satisfied for the initial approximation x_0 it is satisfied for all the subsequent iterates x^k. The radial estimate (3.28) is not restrictive because even if it is not satisfied by x_0 condition (3.25) enforces it for all the subsequent iterates.*

Proof. Given x satisfying (3.27) and (3.28) we can represent any $y \in R^n$ as

$$y = y_x + y_\perp,$$

where y_x is the M-orthogonal projection of y on $span(x)$.

Condition (3.27) implies that there is $\alpha \in (0, \frac{1}{2})$ such that

$$\rho(x) < \rho_0 = (1-\alpha)\lambda_1 + \alpha\lambda_2 \tag{3.30}$$

and since $y_\perp$ is M-orthogonal to x we have

$$\rho(y_\perp) \geq \alpha\lambda_1 + (1-\alpha)\lambda_2. \tag{3.31}$$

Using the lower bound for $\|x\|_M^2$ from (3.28) we get

$$\begin{aligned}
y_\perp^T H y_\perp &\geq y_\perp^T \left(2(A - 2\mu M) + 8\mu M(xx^T)M + 4\mu(1 - \frac{\rho_0}{2\mu})M \right) y_\perp \\
&= y_\perp^T \left(2(A - \rho_0 M) + 8\mu M(xx^T)M \right) y_\perp \\
&\geq 2(y_\perp^T A y_\perp - \rho_0 y_\perp^T M y_\perp) \\
&= 2\left(1 - \frac{\rho_0}{\rho(y_\perp)}\right) y_\perp^T A y_\perp \\
&\geq 2\frac{(1-2\alpha)(\lambda_2 - \lambda_1)}{\lambda_1 + \lambda_2} y_\perp^T A y_\perp,
\end{aligned} \tag{3.32}$$

where the last inequality follows from (3.30) and (3.31). We also have

$$y_\perp^T H y_x = 2 y_\perp^T A y_x \tag{3.33}$$

and with our chose of $\mu = \frac{3}{4}\rho_0$

$$\begin{aligned}
y_x^T H y_x &\geq 2(y_x^T A y_x - \rho_0 y_x^T M y_x) + 8\mu(x^T M x)(y_x^T M y_x) \\
&\geq 2(y_x^T A y_x - \rho_0 y_x^T M y_x) + 8\mu \left(1 - \frac{\rho_0}{2\mu}\right)(y_x^T M y_x) \\
&= 2(y_x^T A y_x + (4\mu - 3\rho_0) y_x^T M y_x) \\
&= 2 y_x^T A y_x.
\end{aligned} \tag{3.34}$$

Since for $\alpha \in (0, \frac{1}{2})$

$$2\frac{(1-2\alpha)(\lambda_2-\lambda_1)}{\lambda_1+\lambda_2} < 2$$

from (3.32), (3.33) and (3.34) it follows that

$$y^T Hy \geq c_1 y^T Ay$$

for any c_1 satisfying

$$c_1 \leq 2\frac{(1-2\alpha)(\lambda_2-\lambda_1)}{\lambda_1+\lambda_2}. \tag{3.35}$$

To find an upper bound for H we use the upper bound for $\|x\|_M^2$ in (3.28).

$$\begin{aligned} y^T Hy &= 2y^T Ay - 4\mu y^T My + 8\mu(y^T Mx)^2 + 4\mu(y^T My)(x^T Mx) \\ &\leq 2y^T Ay + 8\mu(y^T My)(x^T Mx) \\ &\leq 2\left(1+\frac{4\mu}{\rho(y)}\right) y^T Ay \\ &= 2\left(1+\frac{3\rho_0}{\rho(y)}\right) y^T Ay \\ &\leq 2\frac{(4-3\alpha)\lambda_1+\alpha\lambda_2}{\lambda_1} y^T Ay \end{aligned}$$

and if c_2 satisfies

$$c_2 \geq 2\frac{(4-3\alpha)\lambda_1+\alpha\lambda_2}{\lambda_1} \tag{3.36}$$

we have the upper inequality in (3.29). The values of the parameters c_1 and c_2 satisfying (3.35) and (3.36) can be chosen independently of the discretization parameters of the problem. □

Remark 3.3 *The specific choice for $\mu = \frac{3}{4}\rho_0$ was made for the analysis purpose only. The algorithm results in the same iterates and therefore provides the same convergence rate for all values of μ satisfying (3.24).*

Condition (3.29) allows us to use the theory of Multiplicative Schwarz methods for minimization problems (Tai, Espedal, 96) [29]. The convergence rate parameter δ (2.14) for Algorithm 2.2 can bounded in terms of c_1 and c_2 and therefore is independent of the meshsize and the number of subdomains. Besides the equivalence of the formulations (3.4) and (3.21) stated in Remark 3.1 gives the following

Theorem 3.2 *The iterates produced by Algorithm 3.1 with coarse grid correction applied to (3.4) and initialized with x_0 such that $\rho(x_0) \leq \frac{\lambda_1+\lambda_2}{2}$ produces iterates that satisfy*

$$\frac{\|x_1 - x^{k+1}\|_A^2}{\|x_1 - x^k\|_A^2} \leq \delta < 1,$$

where (x_1, λ_1) is the minimum eigenpair of (3.2) with $\|x_1\|_M = 1$ and the value of δ is independent of the meshsize h and the number of subdomains J.

This theorem also provides an alternative proof of the convergence of Algorithm 3.1 in case of

$$\rho(x_0) \leq \frac{\lambda_1 + \lambda_2}{2}.$$

3.2 Simultaneous Computation of Several Eigenvalues

3.2.1 Generalization of the Rayleigh Quotient Minimization

In this section we consider the problem of finding m lowest eigenmodes of (3.1) which in the discretized form corresponds to finding m lowest eigenvalues and the corresponding eigenvectors of (3.2). This problem can be formulated as

$$AX = MX\Lambda,$$

where Λ is a diagonal matrix of m lowest eigenvalues of (3.2) and X is an $n \times m$ matrix whose columns are the corresponding eigenvectors.

In order to generalize the definition of the Rayleigh Quotient to the case of several eigenvalues we consider a matrix

$$R(X) \equiv (X^T MX)^{-1}(X^T AX) \tag{3.37}$$

and use its trace as a generalized Rayleigh Quotient of the m-dimensional subspace spanned by columns of X

$$\rho_{(m)}(X) \equiv tr(R(X)) = tr((X^T MX)^{-1}(X^T AX)). \tag{3.38}$$

The subscript (m) is introduced to distinguish the generalized Rayleigh Quotient from the standard one in case of $m = 1$.

The following lemma shows that the $\rho_{(m)}(X)$ defined by (3.38) is invariant with respect to linear transformations and therefore characterizes the column space of X.

Lemma 3.5 *For any $n \times m$ matrix X of full column rank and any nonsingular $m \times m$ matrix U we have*

$$\rho_{(m)}(XU) = \rho_{(m)}(X).$$

Proof. We have

$$\begin{aligned} R(XU) &= ((XU)^T MXU)^{-1}((XU)^T AXU) \\ &= U^{-1}(X^T MX)^{-1}U^{-T}U^T(X^T AX)U \\ &= U^{-1}R(X)U, \end{aligned}$$

and therefore, since matrices $R(XU)$ and $R(X)$ are similar, they have the same spectrum and trace. □

We should also point out that the order of multiplication in $R(X)$ (3.37) which is somewhat arbitrary does not affect $\rho_k(X)$.

Now we are ready to give the problem of finding m lowest eigenpairs $(\lambda_l, e_l)_{l=1\ldots m}$ of (3.2) a minimization formulation.

Theorem 3.3 *If X^* is a solution of*

$$\rho_{(m)}(X^*) = \min_X \rho_{(m)}(X) \tag{3.39}$$

then

$$\rho_{(m)}(X^*) = \sum_{l=1}^{m} \lambda_l$$

and the columns of X form a basis for the space spanned by m lowest eigenvectors of (3.2).

Proof. Since we can transform (3.2) to the equivalent problem with identity mass matrix

$$M^{-\frac{1}{2}}AM^{-\frac{1}{2}}(M^{\frac{1}{2}}x) = \lambda(M^{\frac{1}{2}}x)$$

and the generalized Rayleigh Quotient is invariant with respect to this transformation it suffices to show that the statement is true for the case when the matrix M is identity. Besides, the lemma shows that without loss of generality we can assume that the columns of X are orthonormal. We can append X by $n-m$ columns such that the resulting matrix $\tilde{X}$ is unitary.

The Rayleigh Quotient $\rho_m(X)$ is equal to the trace of the $m \times m$ block of $\tilde{A} = \tilde{X}^T A \tilde{X}$. Since the eigenvalues of this block when in increasing order are not smaller than the corresponding eigenvalues of $\tilde{A}$ (see for example section 10.1 of Parlett's book [26]) and since the spectrum of $\tilde{A}$ matches that of A we have

$$\rho_{(m)}(X) \geq \sum_{l=1}^{m} \lambda_l.$$

If X^* is formed by the eigenvectors corresponding to the lowest m eigenvalues the above inequality becomes equality. □

3.2.2 Subspace Correction for Simultaneous Computation of Several Eigenvalues

In this section we adjust the subspace correction algorithms presented is section 3.1 to the case of finding several lowest eigenmodes at once. Let $X^{k+\frac{i-1}{J}}$ be the current approximation to the solution of (3.39), then the next iterate is of the form

$$X^{k+\frac{i}{J}} = X^{k+\frac{i-1}{J}} + P_i D_i^k,$$

where D_i^k is the $n_i \times m$ matrix of full column rank that solves the subspace problem

$$\rho_{(m)}(X^{k+\frac{i}{J}}) = \min_{D_i \in R^{n_i \times m}} \rho_{(m)}(X^{k+\frac{i-1}{J}} + P_i D_i). \tag{3.40}$$

As we did in Chapter 3 we introduce auxiliary matrices

$$\tilde{D}_i^k = \begin{pmatrix} D_i^k \\ \\ I_m \end{pmatrix} \qquad P_i^k = \begin{pmatrix} P_i & X^{k+\frac{i-1}{J}} \end{pmatrix} \tag{3.41}$$

and since we can write the new iterate as

$$X^{k+\frac{i}{J}} = P_i^k \tilde{D}_i^k$$

the minimization problem (3.40) becomes

$$\begin{aligned} \rho_{(m)}(X^{k+\frac{i}{J}}) &= \min_{\tilde{D}_i} tr \left(\left((P_i^k \tilde{D}_i^k)^T M P_i^k \tilde{D}_i^k \right)^{-1} \left((P_i^k \tilde{D}_i^k)^T A P_i^k \tilde{D}_i^k \right) \right) \\ &= \min_{\tilde{D}_i} tr \left((\tilde{D}_i^{k^T} M_i^k \tilde{D}_i^k)^{-1} (\tilde{D}_i^{k^T} A_i^k \tilde{D}_i^k) \right) \end{aligned} \tag{3.42}$$

with

$$A_i^k = P_i^{k^T} A P_i^k \text{ and } M_i^k = P_i^{k^T} M P_i^k.$$

The final expression in (3.42) is just the generalized Rayleigh Quotient for the pair (A_i^k, M_i^k) and therefore the columns of $\tilde{D}_i^k$ that minimize (3.42) are eigenvectors corresponding to m lowest eigenvectors of the corresponding restricted problem.

Remark 3.4 *Since $\rho_{(m)}(X)$ is invariant with respect to linear transformations of X the identity block I_m in the definition of $\tilde{D}_i^k$ (3.41) is chosen just for the sake of simplicity and can be replaced by any nonsingular $m \times m$ matrix. Therefore the minimization (3.42) can be carried over all $n \times (n_i + m)$ matrices $\tilde{D}_i$ of full column rank.*

We conclude this section pointing out that the alternative minimization formulation for the eigenvalue problem (3.21) also has a natural generalization for finding several lowest eigenmodes.

$$\begin{aligned} \sigma(X) &= \min_{X \in R^{n\times m}} tr\,(X^T A X + \mu(I_m - X^T M X)^2) \\ &= \min_{X \in R^{n\times m}} \left(tr(X^T A X) + \mu \|I_m - X^T M X\|_F^2\right), \end{aligned} \tag{3.43}$$

where $\|\cdot\|_F$ is the Frobenius matrix norm.

3.2.3 Numerical Examples on Unstructured Grids

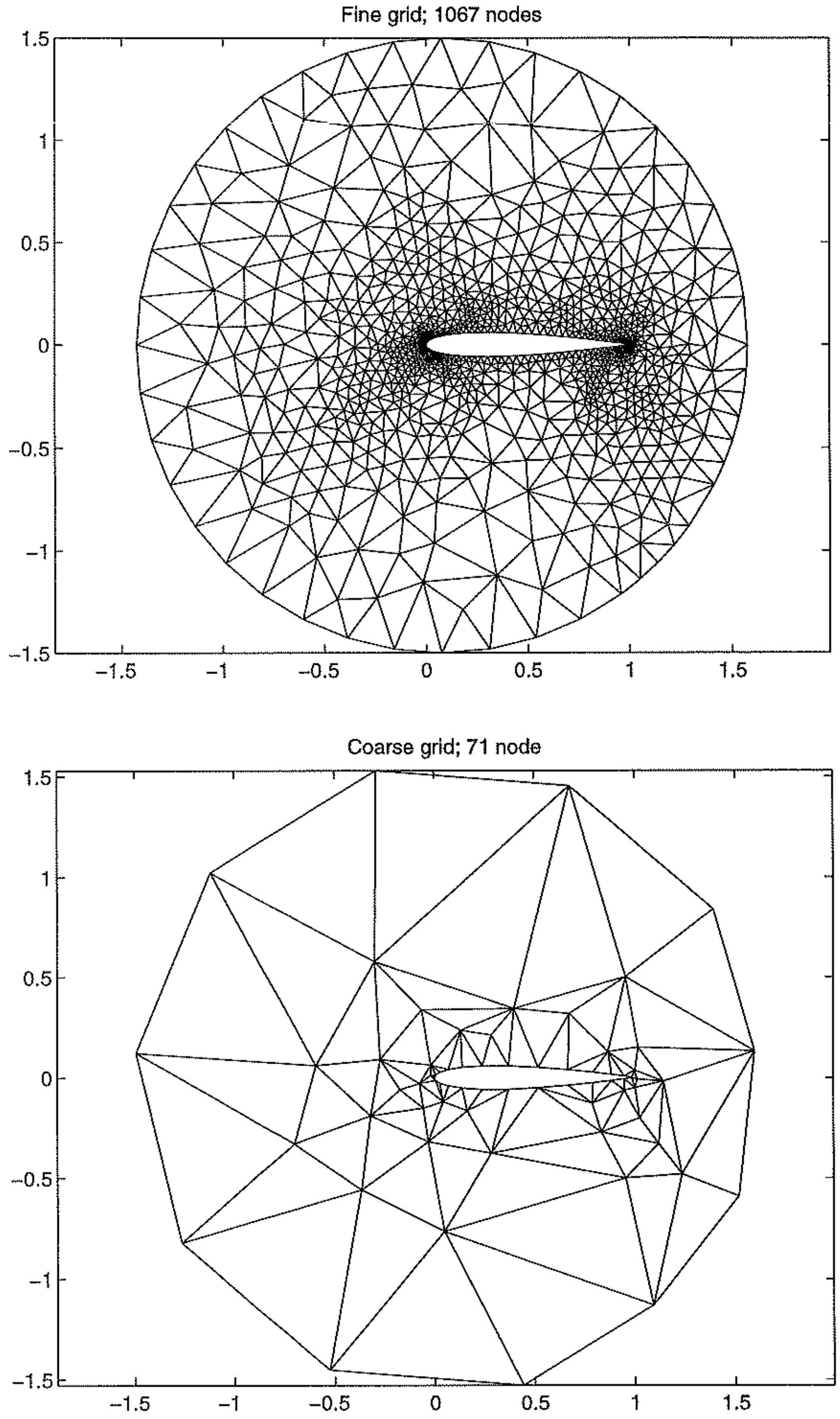
As an example of applications discussed in the previous section we apply the Multiplicative Schwarz method for simultaneous computation of four lowest eigenmodes of 2-D Laplacian over a region whose fine and coarse triangulations are shown in Figure 3.3. These triangulations are produced by Susie Go[1] using SIMPLEX2D [4] package.

[1] sgo@math.ucla.edu

We perform iterations using the multiplicative Schwarz algorithm with subspace correction discussed in the previous section. We used a greedy algorithm described in [7] to generate overlapping subdomains that generate subspaces V_i. The error reduction for all four eigenmodes for the cases of 35 and 78 overlapping subdomains (maximum number of nodes per subdomain is 31 and 21 respectively) is shown in Figures 3.4 and 3.5. We can see that the effect of coarse grid correction is the same as it is in the experiments of Section 3.1, it eliminates the dependence of the convergence rate on the number of subdomains. To demonstrate independence of the meshsize we perform the same computation using the next level of coarsening. The grid containing 264 nodes is partitioned into 14 overlapping subdomains with the largest subdomain containing 24 nodes. The error reduction per iteration is shown in Figure 3.6. Comparing this example with the previous ones we observe that the convergence rate of the algorithm is independent of the meshsize when the coarse grid correction is added. Finally, the computed solution is shown in Figure 3.7.

3.3 Graph Partitioning Using Spectral Bisection

As an application of the eigenvalue problem we consider the spectral bisection graph partitioning method. This heuristic approach is designed to approximate the solution of the problem of bisecting the connected graph into two parts with



Fine grid; 1067 nodes

Coarse grid; 71 node

Figure 3.3: Airfoil grids for simultaneous computation of several eigenmodes.

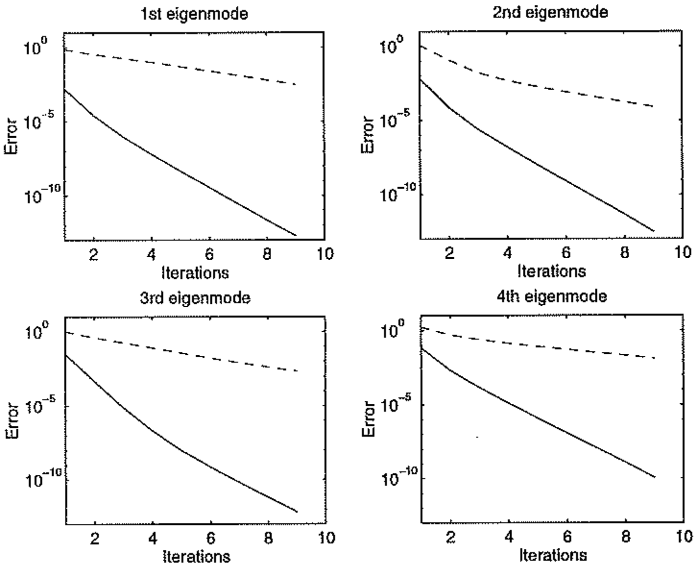


Figure 3.4: Simultaneous computation of four eigenmodes. 1067 nodes 35 subdomains.

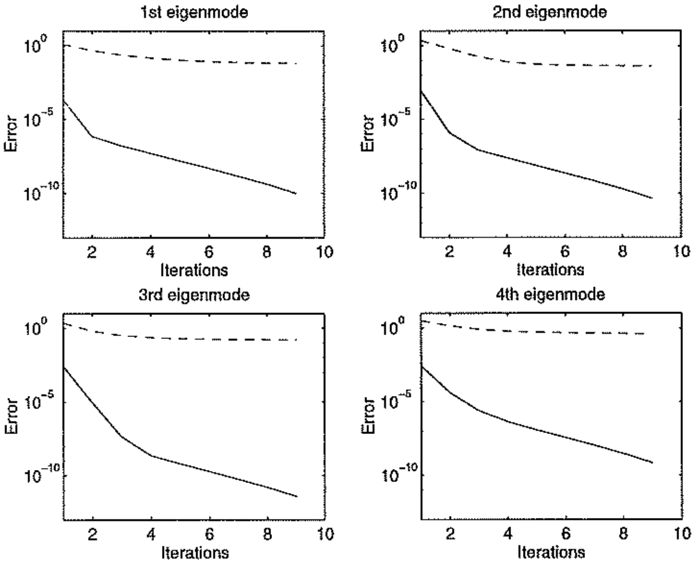


Figure 3.5: Simultaneous computation of four eigenmodes. 1067 nodes, 78 subdomains.

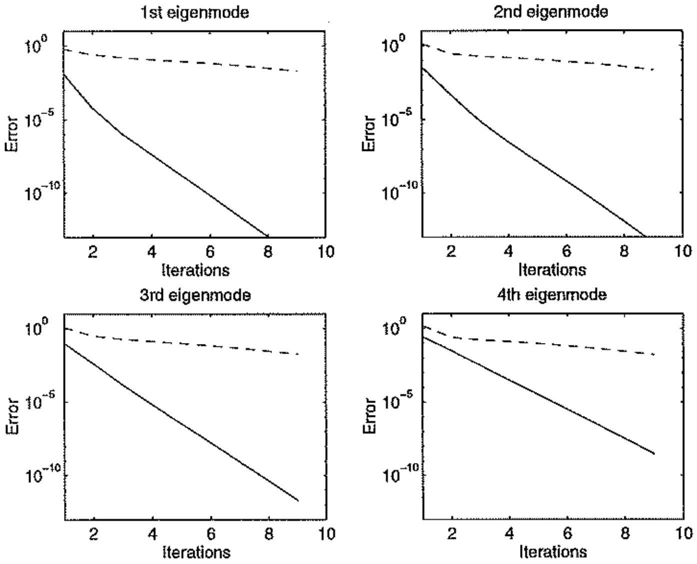


Figure 3.6: Coarser discretization of the problem. 264 nodes, 14 subdomains.

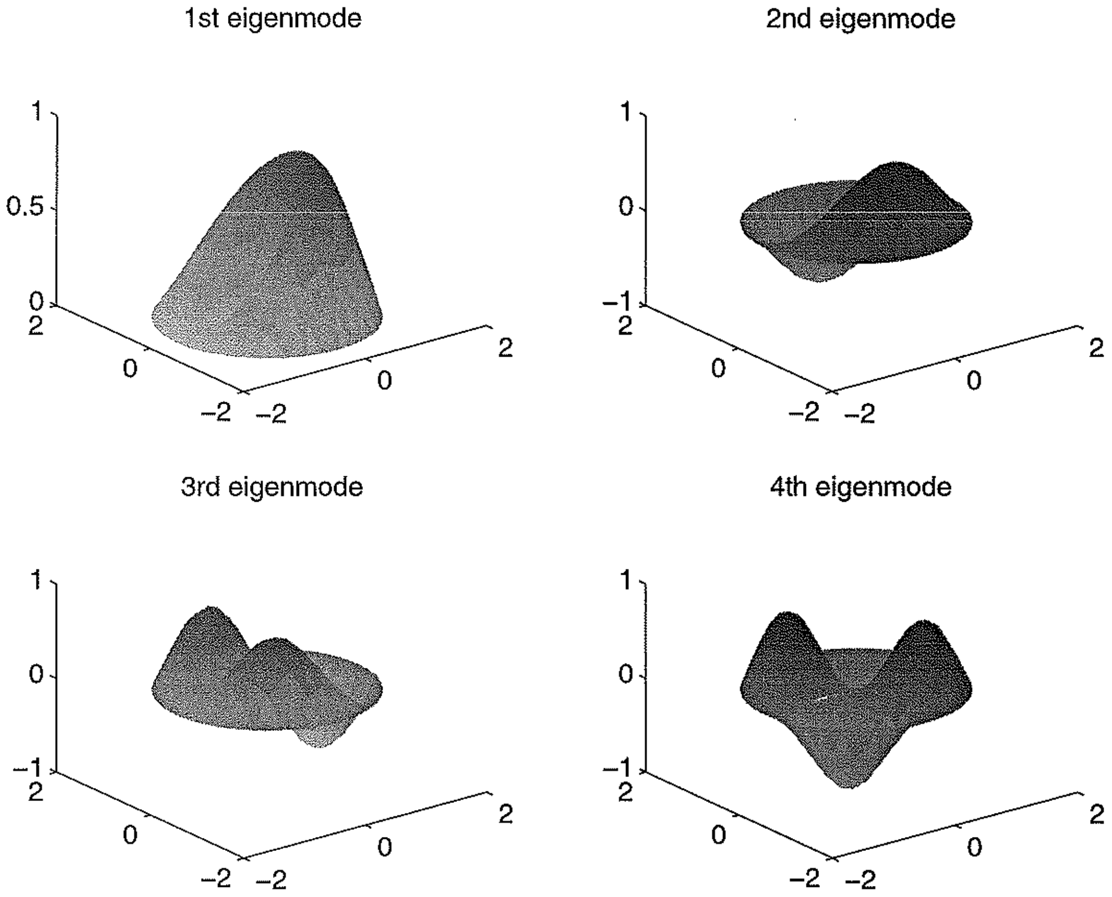


Figure 3.7: Four eigenmodes of the Laplacian on the airfoil grid.

the same number of vertices minimizing the number of cut edges. The spectral bisection method relates this problem to the second eigenvector x_2 of the Laplacian of the graph defined as $Q = (q_{ij})$ with

$$q_{ij} = \begin{cases} -1, \text{ if vertices } i \text{ and } j \text{ are connected} \\ \text{degre of vertex } i, \text{ if } i = j \\ 0 \text{ otherwise} \end{cases} .$$

After the eigenvector x_2 is computed its entries are sorted by magnitude and the median induces the partition. Futher details and motivation behind the spectral bisection graph partitioning method can be found in [9].

To illustrate the method we apply it to the airfoil grids used in the previous section. Since the first eigenvector of the graph Laplacian Q is a constant vector we can find x_2 as the lowest eigenvector of Q restricted to the space orthogonal to constants using Algorithm 3.1. As in the previous section the subdomains are chosen using a greedy algorithm described in [7].

The resulting bisection and the reduction in the number of cut edges are shown in Figures 3.8 and 3.9.

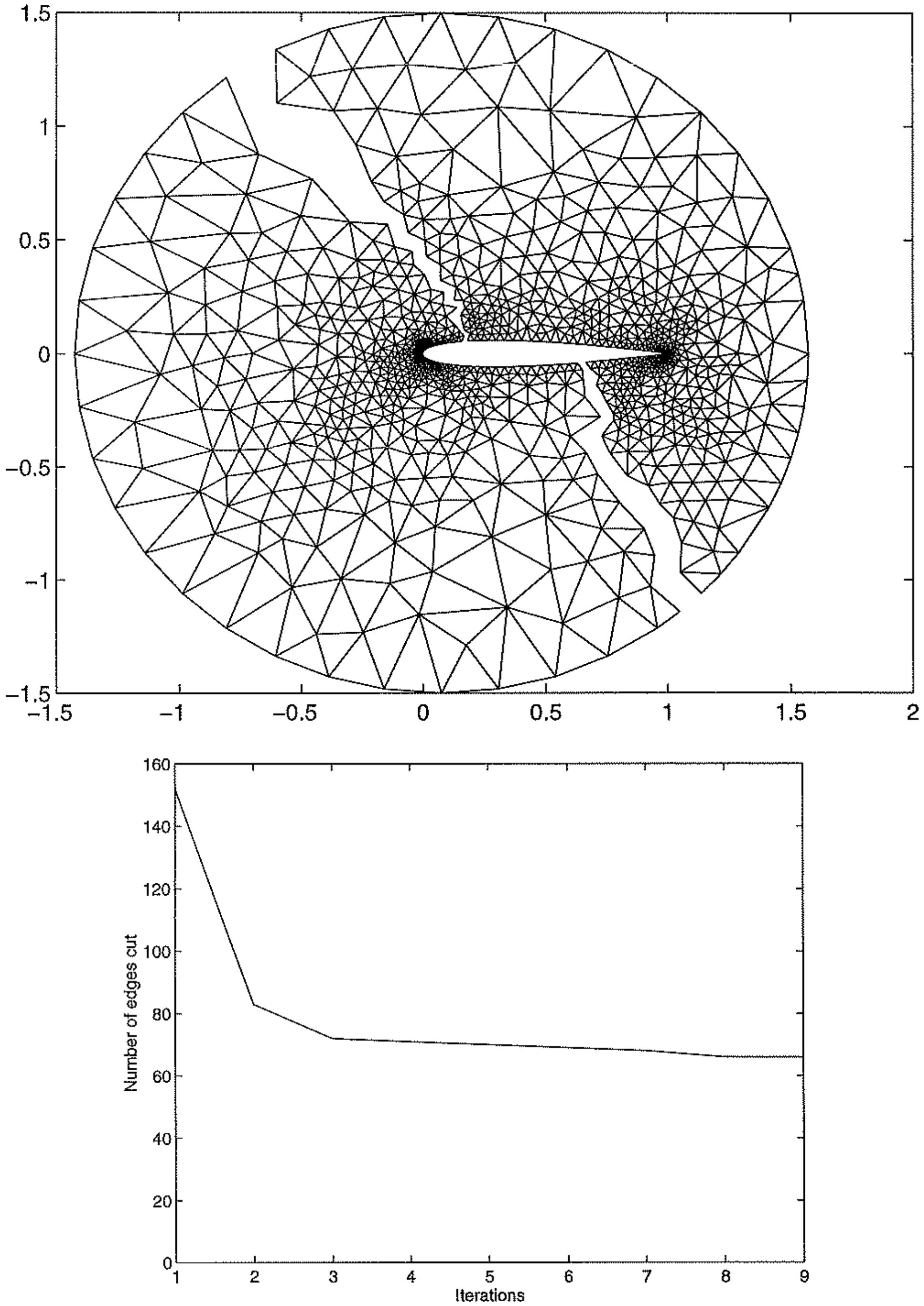


Figure 3.8: Bisection of the airfoil grid with 1067 nodes.

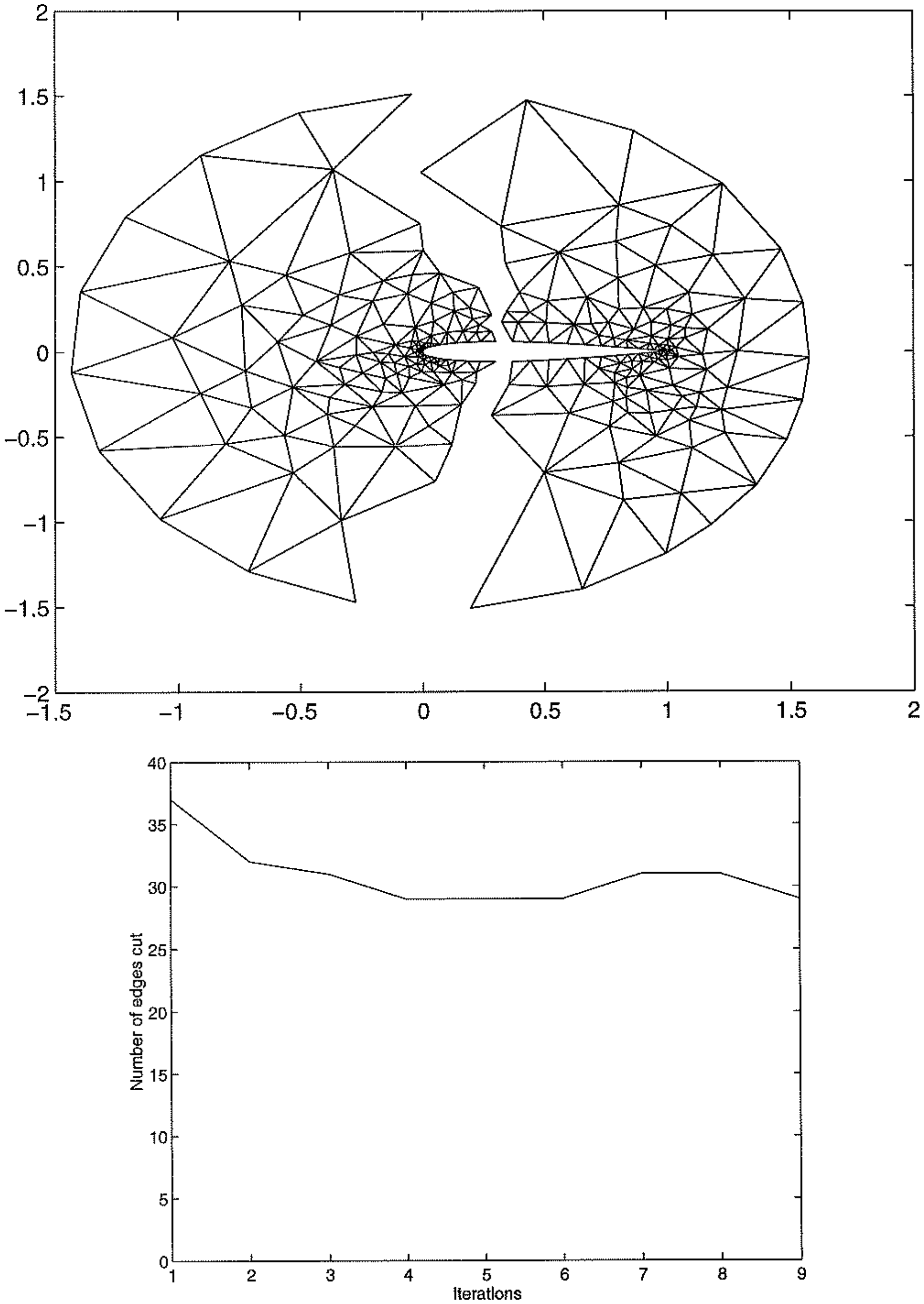


Figure 3.9: Bisection of the airfoil grid with 264 nodes.

CHAPTER 4

Applications to Optimization Problems Arising in Mathematical Finance

In this chapter we consider applications of the subspace correction methods to constrained optimization problems. As we did in the previous chapters, we expect the resulting subspace problems to be computationally manageable even if the original problem is big. Besides, as we will see in the first two examples, the subspace problem can be conceptually simpler than the reduced version of the original problem, which provides another advantage of using these methods.

Unlike the examples of the PDE-based optimization we considered, the problems addressed in this chapter are purely algebraic. Therefore, to handle this problems using the subspace correction framework we, instead of partitioning a continuous physical domain, group together the unknowns and perform the optimization steps over these groups keeping the remaining unknowns fixed. In this case Algorithm 2.1 becomes analogous to the block Gauss-Seidel method.

In this chapter we numerically analyze the application of this algorithm to the following problems:

- **Minimization of the Frobenius norm.**

Given a symmetric indefinite matrix A with positive diagonal entries, find the symmetric positive semi-definite M that coincides with A on the diagonal and minimizes the Frobenius distance from A:

$$\min_{\substack{M=M^T\geq 0\\ diag(M)=diag(A)}} \|A-M\|_F.$$

- **Factor analysis problem.**

  Given a symmetric positive definite matrix A, find its best Frobenius approximation by a sum of a symmetric low rank and a diagonal positive semi-definite matrices:

$$\min_{\substack{rank(M)\leq m\\ M=M^T\geq 0\\ D=diag(d)\geq 0}} \|A-M-D\|_F.$$

- **Constrained l_1-norm minimization.**

  Given $n \times m$ matrix B and an n-vector f, find x that solves

$$\min_{\substack{x\in R^m\\ \Sigma(Bx)_i=\mu\\ \Sigma(x)_j=1}} \|Bx-f\|_1.$$

4.1 Frobenius Norm Minimization for Multivariate GARCH Estimation

4.1.1 Problem Formulation and Numerical Solution

In this section we consider a problem that is related to the estimation of the covariance matrix for the stock data using the multivariate version of the popular GARCH[1] model. If the estimation of different entries is based on different observations then the resulting approximation to the covariance matrix can be indefinite with positive diagonal entries that correspond to the observed individual variances [17]. We will be looking for the positive semi-definite matrix that is the closest to the observed one and preserves its diagonal.

Given an indefinite symmetric $n \times n$ matrix A with the property $diag(A) > 0$ we are looking for M that solves

$$\min_{\substack{M=M^T \geq 0 \\ diag(M)=diag(A)}} \|A - M\|_F. \tag{4.1}$$

We will attempt to solve this problem iteratively. At each step we will be altering one column and the corresponding row of a feasible approximation M to the solution of (4.1). Let the block representation of A and the current feasible approximation to the solution of the problem above be

$$A = \begin{pmatrix} \alpha_{11} & a_{21}^T \\ a_{21} & A_{22} \end{pmatrix} \qquad M = \begin{pmatrix} \alpha_{11} & m_{21}^T \\ m_{21} & M_{22} \end{pmatrix},$$

where a_{21} and m_{21} are $(n-1)$-vectors.

[1] Generalized Autoregressive Conditional Heteroskedasticity

Using a matrix of the form

$$P = \begin{pmatrix} \rho & x^T \\ 0 & I_{n-1} \end{pmatrix},$$

where I_{n-1} is the identity block and x is an $(n-1)$-vector, we can introduce the next iterate by

$$\tilde{M} = PMP^T = \begin{pmatrix} \rho^2\alpha_{11} + 2\rho x^T m_{21} + x^T M_{22} x & \rho m_{21}^T + x^T M_{22} \\ \rho m_{21} + M_{22} x & M_{22} \end{pmatrix}. \tag{4.2}$$

If we enforce the condition

$$\rho^2\alpha_{11} + 2\rho x^T m_{21} + x^T M_{22} x = \alpha_{11} \tag{4.3}$$

then the new approximation $\tilde{M}$ satisfies the constraints of the problem: $\tilde{M} = \tilde{M}^T \geq 0$ and $diag(\tilde{M}) = diag(A)$ and we have

$$\|A - \tilde{M}\|_F^2 - \|A - M\|_F^2 = 2\|a_{21} - (\rho m_{21} + M_{22}x)\|_2^2 - 2\|a_{21} - m_{21}\|_2^2.$$

Therefore, choosing x and ρ that minimize $\|a - (\rho m_{21} + M_{22}x)\|_2^2$, we get the optimal $\tilde{M}$ of the form (4.2). It minimizes the objective function of (4.1) over the matrices obtained form the previous approximation M by changing its first row and column (4.2) and satisfies the constraints of the problem. This procedure can be naturally generalized to changing the i-th column and row of the current iterate.

Remark 4.1 *The convexity of the problem implies that the solution matrix M is on the boundary of the feasible region, i.e. is singular. Since $det(\tilde{M}) = \rho^2 det(M)$*

we can make the iterates stay within the interior of the feasible region by initializing the process with a nonsingular matrix and choosing ρ to be bounded away from zero. Later on we treat ρ as a chosen constant between zero and one, in this case the iterates approach a singular matrix not faster than exponentially.

One step of the iterative procedure becomes

$$\min_x \|a_{21} - (\rho m_{21} + M_{22}x)\|_2^2$$

subject to (4.3), or introducing

$$b = a_{21} - \rho m_{21}$$

we can formulate the subspace problem

$$\min_x \|M_{22}x - b\|_2^2 \tag{4.4}$$

still subject to (4.3).

The Lagrangian of this problem is

$$L(x, \lambda) = \|M_{22}x - b\|_2^2 + \lambda(\rho^2\alpha_{11} + 2\rho x^T m_{21} + x^T M x - \alpha_{11})$$

and the optimality conditions are

$$F(x) = \rho^2\alpha_{11} + 2\rho x^T m_{21} + x^T M x - \alpha_{11} = 0 \tag{4.5}$$

and

$$\nabla_x L(x, \lambda) = \bar{0},$$

which can be written as

$$M^2x - Mb + \lambda\rho m_{21} + \lambda Mx = 0. \tag{4.6}$$

For any λ (4.6) can be solved for x

$$x(\lambda) = (M^2 + \lambda M)^{-1}(Mb - \lambda\rho m_{21}) \tag{4.7}$$

and

$$F(\lambda) \equiv F(x(\lambda)) = 0$$

can be solved using the Newton's method

$$\lambda \Leftarrow \lambda - \frac{F(\lambda)}{F_\lambda(\lambda)}. \tag{4.8}$$

To obtain the analytic expression for $F_\lambda(\lambda)$ we can differentiate (4.5)

$$F_\lambda(\lambda) = \nabla_x F(x) \cdot x_\lambda = 2(\rho m_{21} + Mx)^T x_\lambda \tag{4.9}$$

and differentiate (4.6) in λ

$$M^2x_\lambda + \rho m_{21} + Mx + xMx_\lambda = 0,$$

to get

$$x_\lambda = -(M^2 + \lambda M)^{-1}(\rho m_{21} + Mx),$$

which we substitute in (4.9)

$$F_\lambda(\lambda) = -2(\rho m_{21} + Mx)^T(M^2 + \lambda M)^{-1}(\rho m_{21} + Mx). \tag{4.10}$$

We can summarize the solution of the subproblem:

Alg. 4.1 - ***Newton's Method for solving (4.4), (4.3)***

```
    Initialize $\lambda$ (say $\lambda = 0$)
        Compute $x$ by (4.7)
        Compute $F(\lambda)$ and $F_\lambda(\lambda)$ using (4.5) and (4.10)
        Update $\lambda$ using Newton's step (4.8)
    Repeat the Newton's procedure
end
```

Remark 4.2 *The expressions (4.7) and (4.10) involve the inverse of $M^2 + \lambda M$ which is singular if M is. Restricting ρ to be a nonzero constant and choosing the initial guess to be positive definite we prevent M from being singular. A natural choice for the initial guess might be the diagonal of A.*

Now we can formulate the entire procedure.

Alg. 4.2 - ***Block Gauss-Seidel method. Frobenius minimization***

```
    Choose $\rho \in (0,1)$ and initialize $M^0 = diag(A)$
    for $k = 0$ until convergence
      for $i = 1 : n$
          Use Newton's Method to find $x$ that solves subproblem
          (4.4),(4.3) for the i-th row/column of $M^{k+(i-1)/n}$.
          Update the iterate:
              $M^{k+i/n} = \begin{pmatrix} \rho & x^T \\ 0 & I_{n-1} \end{pmatrix} M^{k+(i-1)/n} \begin{pmatrix} \rho & 0 \\ x & I_{n-1} \end{pmatrix}$
      end
    end
end
```

4.1.2 Numerical Examples

We illustrate the above algorithm with 20 × 20 and 50 × 50 examples. In both cases we initialize M with the diagonal of A and continue the procedure until the convergence is suspected. In Figure 4.1 we see that the convergence is slower for higher values of ρ. It happens because the solution M is singular and large values of ρ limit the rate at which the iterates approach a singular matrix. If ρ is cosen to be small the convergence is faster but this choise can be disadvantageous because, as the iterates become ill-conditioned, the numerical errors that arise in the solution of equations (4.7) and (4.10) can become significant.

4.2 Factor Analysis Problem

In this section we consider a problem of approximating a symmetric positive definite matrix by a sum of a low-rank and a diagonal positive semi-definite matrices. Usually this problem is applied to the covariance matrix of a system of random variables, for example stock returns. The low-rank part can be represented as the product of factors that capture the common sources of variation, while the diagonal corrections correspond to the residuals and represent the individual sources of variation.

Given a symmetric positive definite $n \times n$ matrix A we are looking for the solution of

$$\min_{M,D} \|A - M - D\|_F, \tag{4.11}$$

where M is symmetric positive semi-definite matrix of rank at most $m < n$ and D is non-negative diagonal.

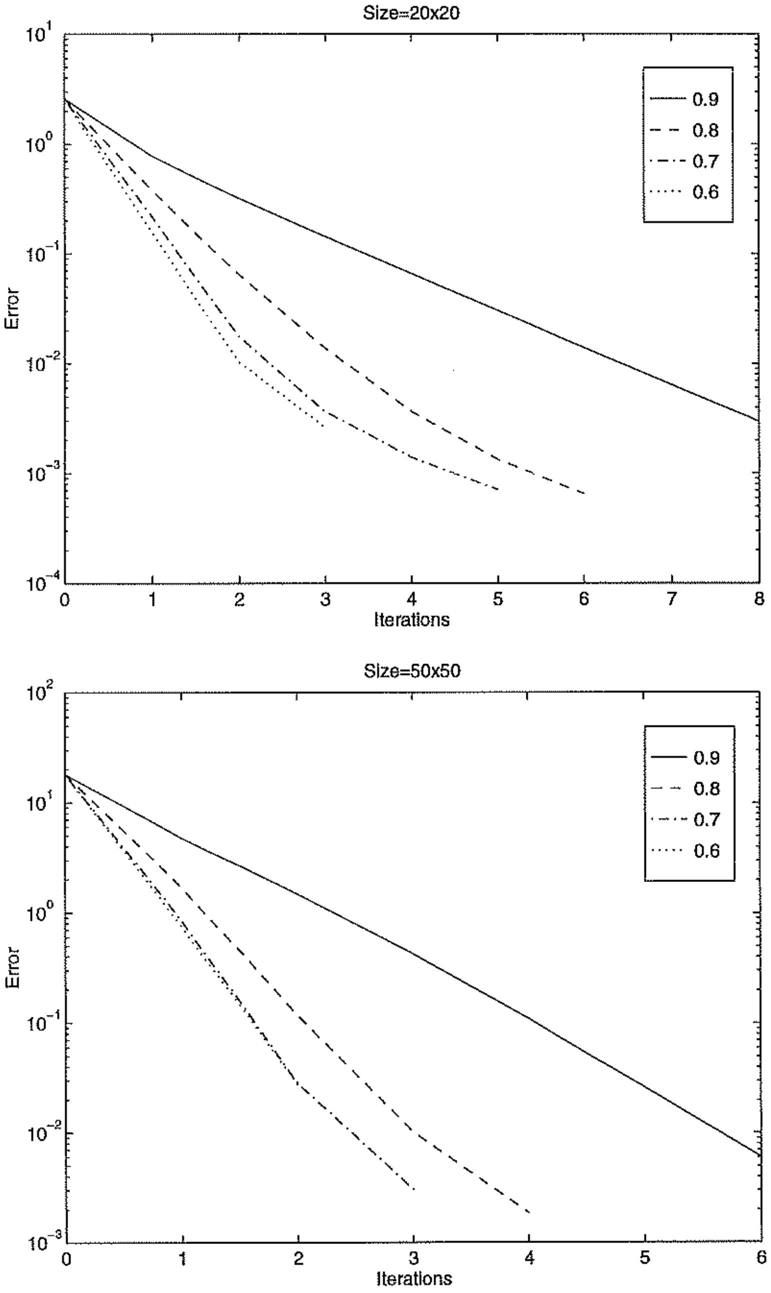


Figure 4.1: Reduction in Frobenius distance for $n = 20$ and $n = 50$ with different values of ρ.

4.2.1 Numerical Solution

We can represent M and D as

$$M = BB^T \tag{4.12}$$

and

$$D = diag(d)^2, \tag{4.13}$$

where B is an $n \times m$ matrix and d is an n-vector. Representation (4.12) is unique up to similarity transformations $B \leftarrow BU$ and (4.13) is unique up to the signs of the components of d.

Similarly to the approach we chose in the previous section we will be looking for M and D iteratively by altering the i-th row and column of M and the i-th component of D during each step. In terms of B and d it amounts to changing the i-th row of B and i-th component of d. For $i = 1$ let

$$B = \begin{pmatrix} b^T \\ B_2 \end{pmatrix} \qquad \text{and} \qquad d = \begin{pmatrix} \delta \\ d_2 \end{pmatrix}$$

and let the corresponding representation of A be

$$A = \begin{pmatrix} \alpha_{11} & a_{21}^T \\ a_{21} & A_{22} \end{pmatrix}.$$

With B_2 and d_2 being held fixed the original minimization problem (4.11) results in

$$\min_{b,\delta} \left((\alpha_{11} - b^T b - \delta^2)^2 + 2\|a_{21} - B_2 b\|_2^2 \right). \tag{4.14}$$

We can solve this problem by first solving the least squares problem

$$\min_b \|a_{21} - B_2 b\|_2^2 \tag{4.15}$$

to find

$$b = (B_2^T B_2)^{-1} a_{21} \tag{4.16}$$

and if $b^T b \leq \alpha_{11}$ then this b and

$$\delta^2 = \alpha_{11} - b^T b \tag{4.17}$$

solve (4.14). Alternatively, if b defined by (4.16) satisfies

$$b^T b > \alpha_{11} \tag{4.18}$$

then $\delta = 0$ and b that solves (4.14) also solves

$$\min_b \left((\alpha_{11} - b^T b)^2 + 2\|a_{21} - B_2 b\|_2^2 \right). \tag{4.19}$$

Since the objective function in (4.19) is smooth, its gradient should be equal to zero at the solution

$$4(b^T b - \alpha_{11})b + 4(B_2^T B_2 b - B_2^T a_{21}) = 0$$

or

$$((b^T b - \alpha_{11})I + B_2^T B_2)b = B_2^T a_{21}. \tag{4.20}$$

This problem can be solved using the Newton's method. For $\sigma \geq 0$ we consider

$$b_\sigma = (\sigma I + B_2^T B_2)^{-1} B_2^T a_{21} \tag{4.21}$$

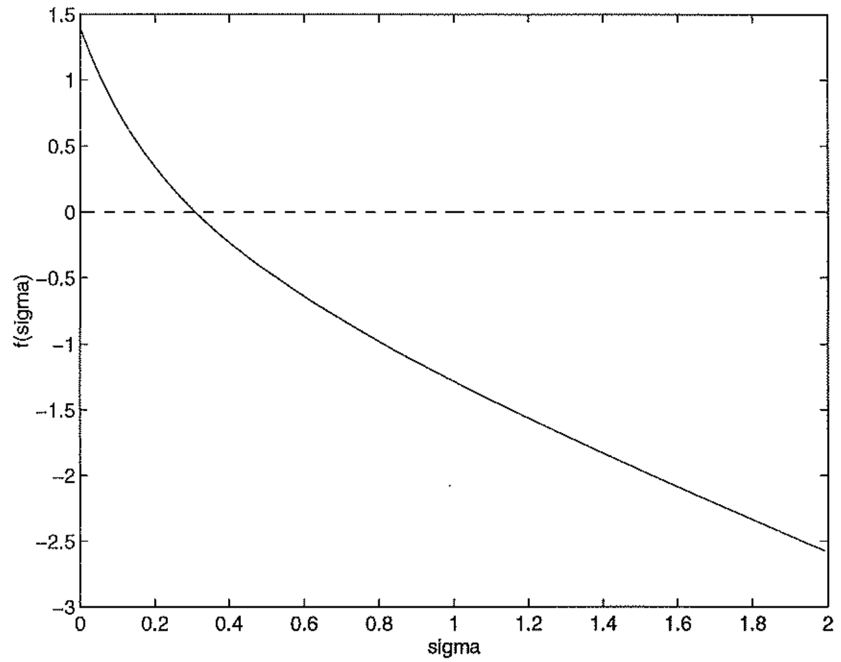


Figure 4.2: Monotonicity of $f(\sigma)$ for the factor analysis problem.

and

$$f(\sigma) = b_\sigma^T b_\sigma - \alpha_{11} - \sigma. \tag{4.22}$$

It is easy to see that if $f(\sigma) = 0$ then the corresponding b_σ solves (4.20). This solution exists because (4.18) implies $f(0) > 0$ and $\lim_{\sigma \to +\infty} = -\infty$. Uniqueness follows from the monotonicity of $f(\sigma)$ that is shown below. From (4.21) we obtain

$$\frac{\partial b_\sigma}{\partial \sigma} = -(\sigma I + B_2^T B_2)^{-1} b_\sigma$$

and therefore

$$\begin{aligned}\frac{\partial f}{\partial \sigma} &= 2b_\sigma^T \frac{\partial b_\sigma}{\partial \sigma} - 1 \\ &= -2b_\sigma^T(\sigma I + B_2^T B_2)^{-1} b_\sigma - 1 \qquad (4.23) \\ &< 0.\end{aligned}$$

A typical example of $f(\sigma)$ is shown on Figure 4.2.

We can apply the Newton's method for solving $f(b(\sigma)) = 0$:

```
Alg. 4.3 - Newton's Method for solving (4.21), (4.22)
    Initialize $\sigma$ (say $\sigma = 0$)
        Compute $b_\sigma$ by (4.21)
        Compute $f(\sigma)$ and $\frac{\partial f(\sigma)}{\partial \sigma}$ from (4.22) and (4.23).
        Update $\sigma$ using Newton's step $\sigma \leftarrow \sigma - f(\sigma)(\frac{\partial f(\sigma)}{\partial \sigma})^{-1}$
    Repeat the Newton's procedure
end
```

The entire procedure for solving the problem (4.11) becomes:

```
Alg. 4.4 - Block Gauss-Seidel method. Factor Analysis Problem
    Initialize $M^0 = B^0 {B^0}^T$ and $D^0 = diag(d^0)^2$
    for $k = 0$ until convergence
        for $i = 1 : n$
            Update i-th row $\beta$ of $B^{k+(i-1)/n}$ and $\delta = d_i^{k+(i-1)/n}$ solving
            (4.14) either by (4.16)-(4.17) or using Newton's Method
            to get $B^{k+i/n}$ and $d^{k+i/n}$. Update the iterates:

                $M^{k+i/n} = B^{k+i/n} {B^{k+i/n}}^T \qquad D = diag(d^{k+i/n})^2$

        end
    end
end
```

The reduction in the objective function for problems of sizes 20×20 and 40×40 are shown in Figure 4.3.

4.3 Gain-Loss Optimization

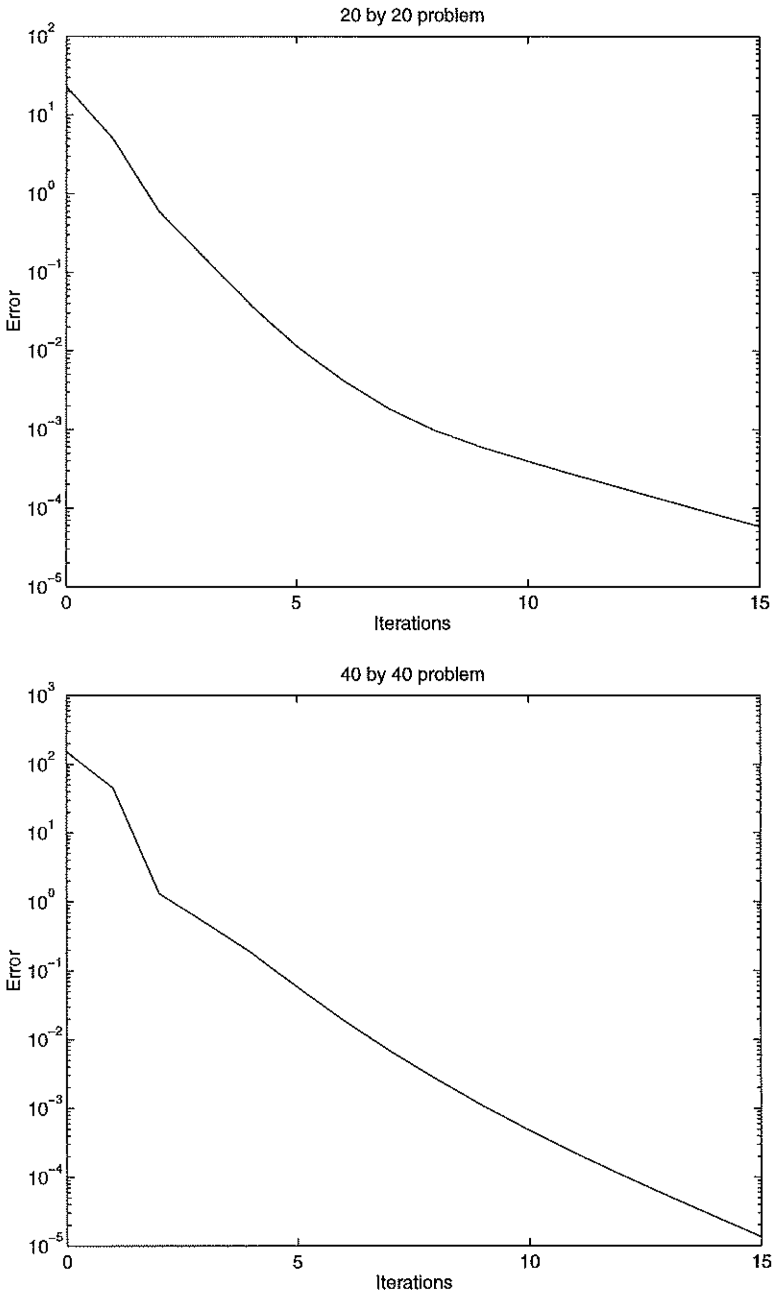


Figure 4.3: Error reduction for the factor analysis problem with $n = 20$ and $n = 40$.

4.3.1 Problem Formulation

The gain-loss optimization constitutes an alternative to the mean-variance optimization for portfolio selection [22]. Let B be an $n \times m$ matrix B each of whose columns contain returns for a particular stock and whose rows correspond to different times when this returns are computed. For any m-vector x whose components represent the weights of the stocks in a portfolio the components of the product $r_i = (Bx)_i$ give the returns for the portfolio at different times. The positive part r_i^+ and the negative part r_i^- of the return components correspond to the gain and the loss respectively.

We consider the gain-loss optimization problem of finding a portfolio x that minimizes the loss $\sum_{i=1}^m r_i^-$ while the gain $\sum_{i=1}^m r_i^+$ is kept fixed.

Because of the relations

$$\begin{aligned} r_i^+ - r_i^- &= (Bx)_i \\ r_i^+ + r_i^- &= |(Bx)_i| \end{aligned} \tag{4.24}$$

and consequently

$$\begin{aligned} r_i^+ &= \frac{|(Bx)_i| + (Bx)_i}{2} \\ r_i^- &= \frac{|(Bx)_i| - (Bx)_i}{2} \end{aligned}$$

the gain-loss optimization problem can be put in an equivalent form of minimizing the sum of the gain and the loss

$$\min_x f(x) = \min_x \|Bx\|_1 \tag{4.25}$$

subject to the overall return $\sum_{i=1}^{m} r_i$ being held fixed

$$\sum_{i=1}^{n}(Bx)_i = \mu. \tag{4.26}$$

We assume that the portfolio x is normalized, i.e. the weights (which can be negative) add up to one

$$\sum_{j=1}^{m}(x)_j = 1. \tag{4.27}$$

Remark 4.3 *To avoid multiple solutions of the problem (4.25)-(4.27) we assume that the dimensions of B satisfy $m < n+2$ and that the columns of B are linearly independent. For many practical problems we have $m \ll n$.*

In order to allow the future derivation of the block Gauss-Seidel method for this problem we generalize the objective function (4.25)

$$\min_x F(x) = \min_x \|Bx - f\|_1, \tag{4.28}$$

where f is a given n-vector. We will also normalize the condition (4.26) replacing B by μB:

$$\sum_{i=1}^{n}(Bx)_i = 1. \tag{4.29}$$

Introducing an m-vector e of ones and an m-vector b_c of column-sums of B we can represent the constrains (4.27) and (4.29) as

$$e^T x = 1 \qquad \text{and} \qquad b_c^T x = 1. \tag{4.30}$$

The problem (4.28), (4.30) can be transformed to a linear programming formulation (cf. [10]). Representing the residual vector as a difference of two vectors with nonnegative components $(Bx-f)_i = r_i^+ - r_i^-$ we can formulate the equivalent problem as

$$\min_{x,r^\pm}\left(\sum_{i=1}^{n} r_i^+ + \sum_{i=1}^{n} r_i^-\right)$$

Subject to

$$\begin{aligned} Bx \quad +r^+ - r^- &= f \\ e^T x \qquad &= 1 \\ b_c^T x \qquad &= 1 \\ r^\pm \qquad &\geq 0 \end{aligned}$$

Several algorithms based on the simplex method were proposed for this problem ([2], [3]), but in this work we will apply an interior point method analogous to the one discussed in [31].

We can represent x as $x_j = x_j^+ - x_j^-$ and put the problem in canonical form

$$\min_{x^\pm,r^\pm}\left(\sum_{i=1}^{n} r_i^+ + \sum_{i=1}^{n} r_i^-\right), \tag{4.31}$$

subject to

$$\begin{aligned} Bx^+ - Bx^- \quad +r^+ - r^- &= f \\ e^T x^+ - e^T x^- \qquad &= 1 \\ b_c^T x^+ - b_c^T x^- \qquad &= 1 \\ x_j^\pm, r_i^\pm \qquad &\geq 0. \end{aligned} \tag{4.32}$$

The matrix of the constraints for this problem is

$$A_p = \begin{pmatrix} B & -B & I & -I \\ e^T & -e^T & 0 & 0 \\ b_c^T & -b_c^T & 0 & 0 \end{pmatrix}. \tag{4.33}$$

4.3.2 Dual Problem and Interior Point Algorithm

The dual for the problem (4.31), (4.32) is

$$\max_{y \in R^n, \alpha, \beta} f^T y + \alpha + \beta$$

subject to

$$A_p^T \begin{pmatrix} y \\ \alpha \\ \beta \end{pmatrix} \le \begin{pmatrix} 0_m \\ 0_m \\ e_n \\ e_n \end{pmatrix},$$

which can be written as

$$\min_{y \in R^n, \alpha, \beta} -f^T y - \alpha - \beta$$

with

$$B^T y + e^T \alpha + b_c^T \beta = 0_n$$

$$-1 \le y_j \le 1 \qquad j = 1 \ldots n.$$

Changing

$$y_j + 1 \to y_j$$

and adding slack variables z_j , $j = 1 \ldots n$ we get the problem in canonical form.

$$\min_{y,z,\alpha^{\pm},\beta^{\pm}} -f^T y - \alpha^+ - \beta^+ + \alpha^- + \beta^-$$

subject to

$$B^T y + e^T \alpha^+ + b_c^T \beta^+ - e^T \alpha^- - b_c^T \beta^- = b_c$$

$$y_j + z_j = 2 \qquad j = 1 \ldots n$$

$$y_j, z_j, \alpha^{\pm}, \beta^{\pm} \geq 0$$

The constraint matrix of this problem is

$$A_d = \begin{pmatrix} B^T & C & 0 \\ I & 0 & I \end{pmatrix}, \tag{4.34}$$

where the block C is given by

$$C = \begin{pmatrix} e & b_c & -e & -b_c \end{pmatrix}. \tag{4.35}$$

We can rewrite this problem in a more compact form

$$\min_{\tilde{y}} \tilde{c}^T \tilde{y} \tag{4.36}$$

subject to

$$A_d \tilde{y} = \tilde{b}, \qquad \tilde{y} \geq 0, \tag{4.37}$$

where

$$\tilde{y} = \begin{pmatrix} y \\ a \\ z \end{pmatrix}, \quad a = \begin{pmatrix} \alpha^+ \\ \beta^+ \\ \alpha^- \\ \beta^- \end{pmatrix}, \quad \tilde{b} = \begin{pmatrix} b_c \\ 2e_n \end{pmatrix}, \quad \tilde{c} = \begin{pmatrix} -f \\ c \\ 0_n \end{pmatrix}, \quad c = \begin{pmatrix} -1 \\ 1 \\ -1 \\ 1 \end{pmatrix}.$$

Since this problem is equivalent to the dual of of (4.25)-(4.26), its dual is quivalent to the original problem and it is easy to verify that the solution to the dual of (4.36)-(4.37) is

$$\tilde{x} = -\begin{pmatrix} x \\ \\ r^+ \end{pmatrix}.$$

To recover the solution x of the original problem (4.28), (4.30), we can apply the iterior point algorithm (see for example [1] p.84) to (4.25)-(4.26):

```
Alg. 4.5 - Interior Point Algorithm for solving LP-problem (4.36)-(4.37)

  1. Set $k = 0$, start with a feasible interior $\tilde{y}^0$ (i.e. $A_d\tilde{y}^0 = \tilde{b},\ \tilde{y}^0 > 0$)
  2. Define
        $D = D^k = diag(y_1^k, y_2^k, \dots y_n^k)$
  3. Find the primal estimate $\tilde{x}^k = -\begin{pmatrix} x^k \\ r^{+k} \end{pmatrix}$ by solving
        $(A_d D^2 A_d^T)\tilde{x}^k = A_d D^2 \tilde{c}$        (4.38)
  4. Find $z^k = \tilde{c} - A_d^T \tilde{x}^k$ and the step direction
        $d\tilde{y}^k = -D^2 z^k$
  5. Update the dual estimate by
        $\tilde{y}^{k+1} = \tilde{y}^k + \rho\alpha d\tilde{y}^k$,   where   $0 < \rho < 1$   and
        $\alpha = \min_i \left[ -\frac{\tilde{y}_i^k}{d\tilde{y}_i^k} : \quad for \quad d\tilde{y}_i^k > 0 \right]$
  6. Terminate if stopping criteria are satisfied, otherwise go to 2.
end
```

Remark 4.4 *The structure of (4.36)-(4.37) allows to initialize the method with a $(2n+4)$-vector of ones*

$$\tilde{y}^0 = e_{2n+4}$$

which is feasible for this problem.

The only computationally involved stage of Algorithm 4.5 is step 3 which requires forming the matrix $A_d D^2 A_d^T$ and solving a linear system (4.38) with it. The advantage of the dual formulation described in this section is that for the matrix A_d given by (4.34) we have

$$A_d D^2 A_d^T = \begin{pmatrix} B^T D_y^2 B + C D_a^2 C^T & D_y^2 B^T \\ B D_y^2 & D_y^2 + D_z^2 \end{pmatrix},$$

where D_y, D_z are the blocks of D corresponding to y, z variables and D_a to α, β respectively. Since this matrix has a big diagonal block $D_y^2 + D_z^2$ we can use its Schur complement

$$S = B(D_y^2 - D_y^2(D_y^2 + D_z^2)^{-1} D_y^2)B^T + C D_a^2 C^T \tag{4.39}$$

to compute the action of its inverse.

The matrix (4.39) is easily computable and the cost of computation for a dense matrix B is $O(nm^2)$. Direct computation of the inverse of S which is of the size $m \times m$ requires $O(m^3) < O(nm^2)$ operations. Therefore, we can conclude that the cost one iteration of the algorithm is $O(mn^2)$. This makes it more advantageous to use the dual formulation of the linear programming problem with the matrix of constraints A_d (4.34) rather than the primal one with the matrix A_p (4.33) whose structure doesn't allow to use fewer than $O(n^3)$ operations per iteration when applying the interior point iteration.

4.3.3 Subspace Correction and Gauss-Seidel Method

As we did in previous sections we can apply the subspace correction approach to the problem (4.28), (4.30). Let x be the current feasible approximation to the solution and P be an $m \times n_i$ prolongation matrix ($n_i < m$). The best correction $x \leftarrow x + Pd$, where d is a n_i-vector satisfies

$$\min_{d \in R^k} \|BPd + Bx - f\|_1 \tag{4.40}$$

subject to

$$\sum_{i=1}^{m} (BPd)_i = 1 - \sum_{i=1}^{m} (Bx)_i = 0 \tag{4.41}$$

and

$$\sum_{j=1}^{n} (d)_j = \sum_{j=1}^{n} (Pd)_j = 1 - \sum_{j=1}^{n} (x)_j = 0. \tag{4.42}$$

The problem (4.40)-(4.42) is of the same type as (4.28), (4.30) and therefore can be solved by the same interior point procedure with the following adjustments:

$$
\begin{aligned}
x &\rightarrow d \\
B &\rightarrow BP \qquad (4.43) \\
C &\rightarrow CP \qquad (4.44) \\
f &\rightarrow f - Bx \\
c &\rightarrow 0_4.
\end{aligned}
$$

We can see that because of (4.43) and (4.44) the Schur complement (4.39) for the subspace problem is related to the original one as

$$S \rightarrow PSP^T.$$

Therefore, performing one iteration of the interior point algorithm applied to a subspace problem corresponds to performing one step of the Gauss-Seidel method for the linear problem (4.38) of Algorithm 4.5 applied directly to (4.36)-(4.37).

As we pointed out in Section 2.3, Algorithm 2.4 can converge to a boundary point of the feasible set which is not the solution of the problem but further improvement over the subspaces V_i is not possible.

This situation is illustrated in Figure 4.4. We consider two examples with matrices B of sizes 128×32 and 1000×10 that contain actual stock data. In both cases we represent the unknown variables as three overlapping groups and apply Algorithm 2.4 to the corresponding problems (4.28), (4.30) using one iteration of Algorithm 4.5 to approximate the solution of the subspace problems. From the plots we can see that if we choose the value of the parameter ρ in step 5 of Algorithm 4.5 large the iterates converge to a point that does not optimize the objective function. To explain this we should point out that the parameter $\rho < 1$ controls the rate with which the iterates approach the boundary of the feasible region. Small values of ρ postpone hitting the boundary at the point where further improvement in the objective along the chosen subspaces is not possible.

Nevertheless, the subspace correction method that produces the iterates that converge to the true solution can still be applied to (4.28), (4.30). Rather than using Algorithm 4.5 as a method for solving the subspace problems of Algorithm 2.4 we can use it directly to solve (4.28), (4.30) but apply the subspace correction method (in this case Gauss-Seidel) to its crucial step (4.38).

We conclude this chapter illustrating the solution of (4.25)-(4.27) for different values of μ and generating the efficient frontier shown in Figure 4.5. The point on the frontier that maximizes the ratio

$$\frac{\sum_{i=1}^{n}(Bx)_i}{\|Bx\|_1}$$

can be found by solving (4.25) using just one normalization constraint (4.27). Finally, the efficient frontier in terms of gain-loss variables $\sum_{i=1}^{m} r_i^+$ and $\sum_{i=1}^{m} r_i^-$ can be found using the transformation (4.24) and is shown in Figure 4.6.

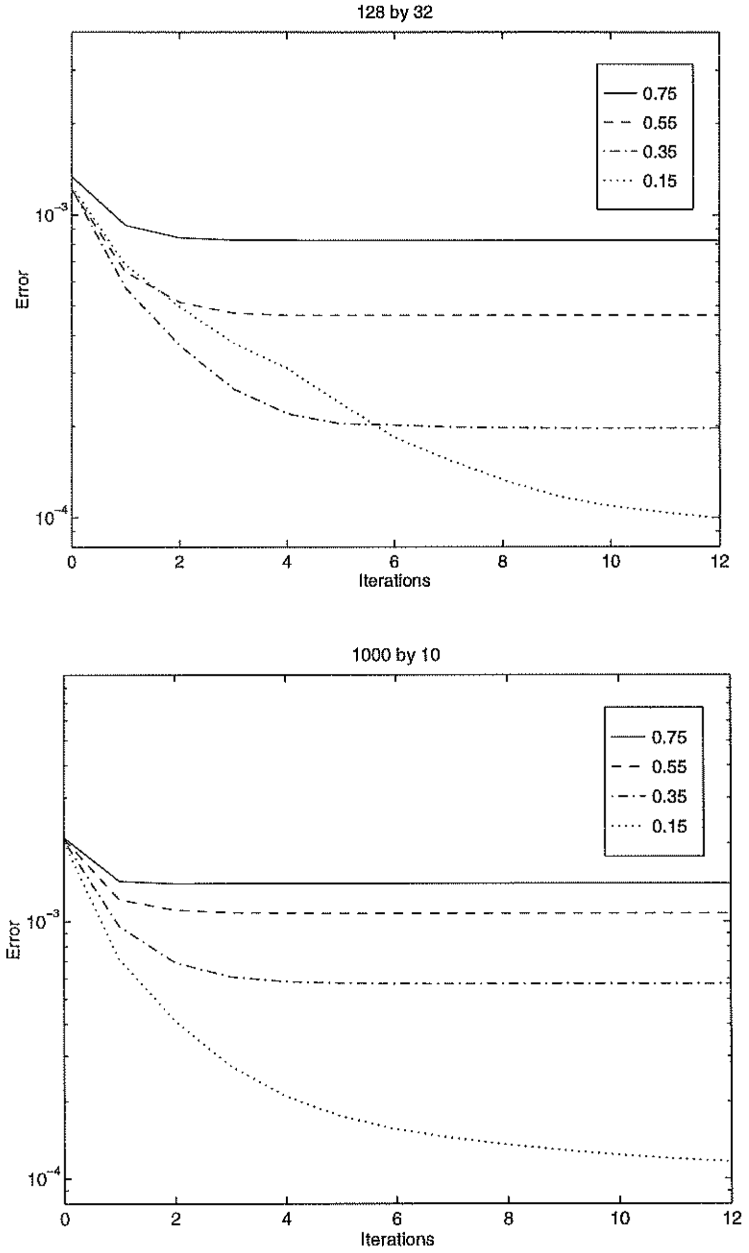


Figure 4.4: Error reduction for the l_1-norm minimization problem with 3 subdomains and different values of ρ.

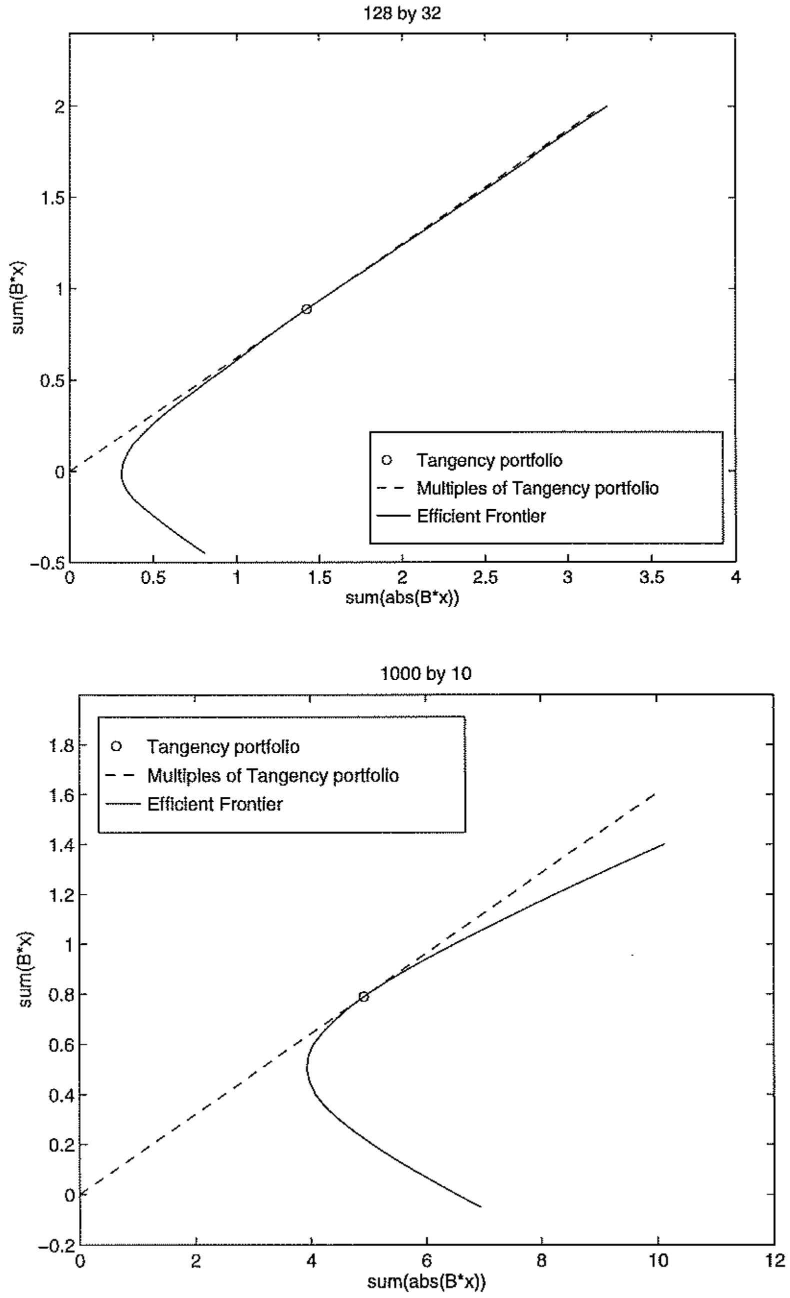


Figure 4.5: Efficient frontier for the minimization of the l_1-norm.

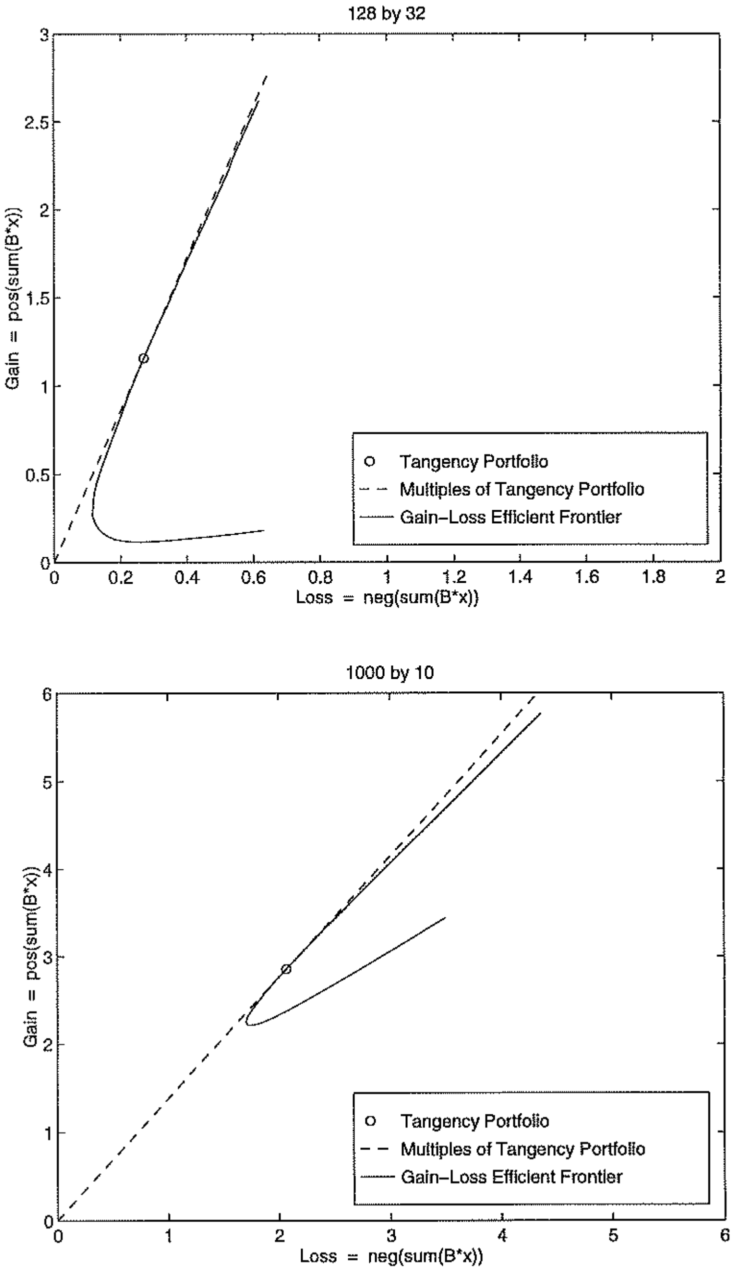


Figure 4.6: Gain-loss efficient frontier.

Bibliography

[1] A. Arbel. *Exploring Interior-Point Linear Programming.* MIT Press, 1993.

[2] R.D. Armstrong and J.P. Godfrey. Two linear programming algorithms for the discrete l_1 norm problem. *Mathematics of Computation*, 33:289–300, 1979.

[3] I. Barrodale and F.D.K. Roberts. An improved algorithm for discrete l_1 linear approximation. *SIAM J. Numer. Anal.*, 10:839–848, 1973.

[4] T.J. Barth. SIMPLEX2D user guide. Available at http://oldwww.nas.nasa.gov/~barth/c++/guide.ps.Z, 1995.

[5] F. Bourquin and F. Hennezel. Application of domain decomposition techniques to modal synthesis for eigenvalue problems. In *Fifth International Symposium on Domain Decomposition Methods for Partial Differential Equations (Norfolk, VA, 1991)*, pages 214–223, Philadelphia, 1992. SIAM.

[6] J.H. Bramble, J.E. Pasciak, J. Wang, and J. Xu. Convergence estimates for product iterative methods with applications to domain decomposition. *Math. Comp.*, 57(195):1–21, 1991.

[7] T.F. Chan, S. Go, and L. Zikatanov. Lecture notes on multilevel methods for elliptic problems on unstructured grids. Technical Report (CAM) 97-11, Department of Mathematics, University of California, Los Angeles, CA 90095-1555, 1997.

[8] T.F. Chan and T.P. Mathew. Domain decomposition algorithms. In *Acta Numerica*, pages 61–143. Cambridge Univ. Press, Cambridge, 1994.

[9] T.F. Chan and W.K. Szeto. On the optimality of the median cut spectral bisection graph partitioning method. *SIAM J. Sci. Comput.*, 18(3):943–948, 1997.

[10] A. Charnes, W.W. Cooper, and R. Ferguson. Optimal estimation of executive compensation by linear programming. *Management Science*, 2:138–151, 1955.

[11] D.K. Fadeev and V.N. Fadeeva. *Computational Methods of Linear Algebra.* W.H. Freeman and Company, San Francisco, 1963.

[12] C. Farhat and M. Géradin. On a component mode synthesis method and its application to incompatible substructures. *Computers and Structures*, 51:459–473, 1994.

[13] E. Gelman and J. Mandel. On multilevel iterative methods for optimization problems. *Math. Progr.*, 48:1–17, 1990.

[14] D. Gilbarg and N.S. Trudinger. *Elliptic Partial Differential Equations of Second Order.* Springer-Verlag, 1983.

[15] W. Hackbusch. *Multigrid Methods.* Springer-Verlag, 1984.

[16] M.S. Kaschiev. An iterative method for minimization of the Rayleigh-Ritz functional. In *Computational processes and systems, No. 6 (Russian)*, pages 160–170. Nauka, Moscow, 1988.

[17] O. Ledoit and P. Santa-Clara. Estimating large conditional covariance matrices with an application to risk management. Technical report, The Anderson School at UCLA, 1997.

[18] P.L. Lions. On the schwarz alternating method. i. In *First International Symposium on Domain Decomposition Methods for Partial Differential Equations*, pages 1–42, Philadelphia, 1988. SIAM.

[19] S.H. Lui. Domain decomposition for eigenvalue problems (preprint). Hong Kong Univ. of Science and Tech., 1995.

[20] S.H. Lui. On two Schwarz alternating methods for the symmetric eigenvalue problem (preprint). Hong Kong Univ. of Science and Tech., 1996.

[21] S.Yu. Maliassov. On the analog of Schwarz method for spectral problems. In *Numerical methods and mathematical modeling (Russian)*, pages 70–79. Otdel Vychisl. Mat., Moscow, 1992.

[22] H.M. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

[23] G. Mathew and V.U. Reddy. Development and analysis of a neural network approach to Pisarenko's harmonic retrieval method. *IEEE Trans. Sig. Proc.*, 42(3):663 – 673, 1994.

[24] S.F. McCormick. *Multilevel Projection Methods for Partial Differential Equations.* SIAM, Philadelphia, 1992.

[25] J.M. Ortega. *Matrix Theory.* Plenum Press, New York, 1987.

[26] B.N. Parlett. *The Symmetric Eigenvalue Problem.* Prentice-Hall, Inc., Englewood Cliffs, NJ, 1980.

[27] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:449–458, 1992.

[28] A. Stathopoulos, Y. Saad, and C.F. Fischer. A Schur complement method for eigenvalue problems. In *Proceedings of the Seventh Copper Mountain Conference on Multigrid Methods*, April 2-7 1995.

[29] X.-C. Tai and M. Espedal. Rate of convergence of some space decomposition methods for linear and nonlinear problems. *SIAM J. Numer. Anal. (to appear)*, 1997.

[30] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Rev.*, 34:581–613, 1992.

[31] Y. Zhang. Primal-dual interior point approach for computing l_1-solutions and l_∞-solutions of overdetermined linear systems. *J. of Optim. Theor. and Appl.*, 77(2):323–341, 1993.