



INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Analysis of a Quadratic Programming Decomposition Algorithm

W. W. Hager — G. Bencteux — E. Cancès — C. Le Bris

N° 6288 — version 2

version initiale September 2007 — version révisée June 2009

Thème NUM

R
apport
de recherche



Analysis of a Quadratic Programming Decomposition Algorithm *

W. W. Hager[†], G. Bencteux[‡], E. Cancès[§], C. Le Bris[¶]

Thème NUM — Systèmes numériques
Projets MicMac

Rapport de recherche n° 6288 — version 2 — version initiale September 2007 — version révisée June 2009 — 28 pages

Abstract: We analyze a decomposition algorithm for minimizing a quadratic objective function, separable in \mathbf{x}_1 and \mathbf{x}_2 , subject to the constraint that \mathbf{x}_1 and \mathbf{x}_2 are orthogonal vectors on the unit sphere. Our algorithm consists of a local step where we minimize the objective function in either variable separately, while enforcing the constraints, followed by a global step where we minimize over a subspace generated by solutions to the local subproblems. We establish a local convergence result when the global minimizers nondegenerate. Our analysis employs necessary and sufficient conditions and continuity properties for a global optimum of a quadratic objective function subject to a sphere constraint and a linear constraint. The analysis is connected with a new domain decomposition algorithm for electronic structure calculations.

Key-words: quadratic programming, orthogonality constraints, domain decomposition method, electronic structure calculations

* August 29, 2007. This material is based upon work supported by the National Science Foundation under Grants 0619080 and 0620286.

[†] hager@math.ufl.edu, <http://www.math.ufl.edu/~hager>, PO Box 118105, Department of Mathematics, University of Florida, Gainesville, FL 32611-8105. Phone (352) 392-0281. Fax (352) 392-8357.

[‡] EDF R&D, 1 Avenue du Général de Gaulle, 92141 Clamart Cedex, France and INRIA, MICMAC team project, BP 105, 78153 Rocquencourt, France.

[§] CERMICS, École Nationale des Ponts et Chaussées, 6 & 8, Avenue Blaise Pascal, Cité Descartes, 77455 Marne-La-Vallée Cedex 2, France and INRIA, MICMAC team project, BP 105, 78153 Rocquencourt, France.

[¶] CERMICS, École Nationale des Ponts et Chaussées, 6 & 8, Avenue Blaise Pascal, Cité Descartes, 77455 Marne-La-Vallée Cedex 2, France and INRIA, MICMAC team project, BP 105, 78153 Rocquencourt, France.

Analyse d'un algorithme de décomposition pour un problème de programmation quadratique

Résumé : On présente l'analyse numérique d'un algorithme de décomposition pour la minimisation d'une fonction coût quadratique, séparable en \mathbf{x}_1 et \mathbf{x}_2 , sous la contrainte que \mathbf{x}_1 et \mathbf{x}_2 sont orthogonaux sur la sphère unité. Notre algorithme consiste en une étape locale où la fonction coût est minimisée séparément en chacune de ses deux variables, en respectant les contraintes. Cette première étape est suivie d'une étape globale où on minimise la fonction coût sur un sous-espace généré par les solutions de l'étape locale. Un théorème de convergence locale est établi quand les minimiseurs globaux ne sont pas dégénérés. Notre analyse utilise les conditions nécessaires et suffisantes et les propriétés de continuité du minimum global d'une fonction coût quadratique minimisée sous une double contrainte sphérique et linéaire. Cette analyse est reliée à un nouvel algorithme de décomposition de domaine pour les calculs de structure électronique.

Mots-clés : programmation quadratique, contraintes d'orthogonalité, méthode de décomposition de domaine, calculs de structure électronique

1 Introduction

In [2] and [3] we develop a multilevel domain decomposition algorithm for electronic structure calculations which has been extremely effective in computing electronic structure for large, linear polymer chains. Both the computational cost and memory requirement scale linearly with the number of atoms. Although this algorithm has been very effective in practice, a theory establishing convergence has not yet been developed. The algorithm in [2, 3] was motivated by a related decomposition algorithm for a quadratic programming problem with an orthogonality constraint. In this paper, we develop a convergence theory for the decomposition algorithm.

Let \mathbf{H}_1 and \mathbf{H}_2 be symmetric n by n matrices. We consider the following quadratic optimization problem:

$$\begin{aligned} \min F(\mathbf{x}_1, \mathbf{x}_2) &:= \mathbf{x}_1^\top \mathbf{H}_1 \mathbf{x}_1 + \mathbf{x}_2^\top \mathbf{H}_2 \mathbf{x}_2 \\ \text{subject to } \mathbf{x}_1^\top \mathbf{x}_1 &= 1 = \mathbf{x}_2^\top \mathbf{x}_2, \quad \mathbf{x}_1^\top \mathbf{x}_2 = 0. \end{aligned} \quad (1)$$

In other words, find orthogonal unit vectors \mathbf{x}_1 and \mathbf{x}_2 which minimize the separable quadratic objective function. Our algorithm for (1) consists of a “local step” where we minimize F over each variable separately, while enforcing the constraints, followed by a “global step” where we optimize over a subspace generated by the iterates of the local step. There are two modes of the local step, a “forward” and a “reverse” mode. In consecutive iterations, we employ the forward mode followed by the reverse mode. If $\mathbf{x}_k = (\mathbf{x}_{k1}, \mathbf{x}_{k2})$ is the iterate at step k , then the forward and the reverse modes of the local step are the following:

$$\begin{aligned} \text{forward} \quad & \begin{cases} \mathbf{y}_{k1} \in \arg \min \{F(\mathbf{z}, \mathbf{x}_{k2}) : \|\mathbf{z}\| = 1, \mathbf{z}^\top \mathbf{x}_{k2} = 0\}, \\ \mathbf{y}_{k2} \in \arg \min \{F(\mathbf{y}_{k1}, \mathbf{z}) : \|\mathbf{z}\| = 1, \mathbf{z}^\top \mathbf{y}_{k1} = 0\}, \end{cases} \\ \text{reverse} \quad & \begin{cases} \mathbf{y}_{k2} \in \arg \min \{F(\mathbf{x}_{k1}, \mathbf{z}) : \|\mathbf{z}\| = 1, \mathbf{z}^\top \mathbf{x}_{k1} = 0\}, \\ \mathbf{y}_{k1} \in \arg \min \{F(\mathbf{z}, \mathbf{y}_{k2}) : \|\mathbf{z}\| = 1, \mathbf{z}^\top \mathbf{y}_{k2} = 0\}. \end{cases} \end{aligned} \quad (2)$$

Here and throughout the paper, $\|\cdot\|$ denotes the Euclidean norm.

The problem which must be solved in electronic structure calculations is more general than (1) and the multilevel algorithm developed in [2, 3] is more complex than (2). For example, in electronic structure calculations, \mathbf{H}_1 and \mathbf{H}_2 could be of different dimensions and the orthogonality condition $\mathbf{x}_1^\top \mathbf{x}_2 = 0$ in (1) would be replaced by the more general condition $\mathbf{x}_1^\top \mathbf{P} \mathbf{x}_2 = 0$ where \mathbf{P} is rectangular. Nonetheless, the algorithm studied in this paper was the basis for the more general algorithm developed in [2, 3], and our analysis is an initial step towards justifying and understanding the convergence properties of the more general algorithm.

One can think of either the forward or reverse modes as a block Gauss-Seidel iteration [5, p. 323]. In the forward mode, we first hold the second block of variables \mathbf{x}_{k2} fixed and we optimize over the first block of variables to obtain \mathbf{y}_{k1} ; in the second step, we hold the first block of variables fixed at \mathbf{y}_{k1} and we optimize over the second block to obtain \mathbf{y}_{k2} .

In general, the local steps converge to a limit which may not be a stationary point of (1). To achieve convergence to a stationary point for (1), each local step, either forward or reverse, is followed by a “global step” where we minimize F over the subspace spanned by the following 4 vectors, while enforcing the constraints of (1):

$$\begin{bmatrix} \mathbf{y}_{k1} \\ \mathbf{0} \end{bmatrix}, \quad \begin{bmatrix} \mathbf{y}_{k2} \\ \mathbf{0} \end{bmatrix}, \quad \begin{bmatrix} \mathbf{0} \\ \mathbf{y}_{k1} \end{bmatrix}, \quad \begin{bmatrix} \mathbf{0} \\ \mathbf{y}_{k2} \end{bmatrix} \quad (3)$$

After imposing the two normalization condition $\mathbf{x}_1^\top \mathbf{x}_1 = 1$ and $\mathbf{x}_2^\top \mathbf{x}_2 = 1$ and the orthogonality condition $\mathbf{x}_1^\top \mathbf{x}_2 = 0$ on the subspace, we are left with a one dimensional curve of feasible points in the subspace spanned by the 4 vectors (3). This curve can be expressed in the following way:

$$\mathbf{z}_k(s) = \frac{1}{\sqrt{1+s^2}}(\mathbf{y}_k + s\mathbf{d}), \quad (4)$$

where

$$\mathbf{d} = \pm \begin{bmatrix} \mathbf{y}_{k2} \\ -\mathbf{y}_{k1} \end{bmatrix} \quad \text{and} \quad \mathbf{y}_k = \begin{bmatrix} \mathbf{y}_{k1} \\ \mathbf{y}_{k2} \end{bmatrix}.$$

The vector $\mathbf{z}_k(s)$ lies in the space spanned by the 4 vectors in (3) for each choice of s ; the orthogonality condition holds since

$$(1+s^2)\mathbf{z}_{k1}^\top \mathbf{z}_{k2} = (\mathbf{y}_{k1} \pm s\mathbf{y}_{k2})^\top (\mathbf{y}_{k2} \mp s\mathbf{y}_{k1})^\top = \pm s \mp s = 0;$$

and the factor $1/\sqrt{1+s^2}$ ensures that the two components of \mathbf{z}_k are unit vectors.

Let $F_k(s) = F(\mathbf{z}_k(s))$ be the objective function evaluated along the search direction. For convenience, the sign in the definition of \mathbf{d} in the global step is chosen so that $F'_k(0) \leq 0$. At iteration k in the global step, we set

$$\mathbf{x}_{k+1} = \mathbf{z}(s_k), \quad (5)$$

where s_k is the stepsize.

The motivation for optimizing over the subspace spanned by the 4 vectors (3) is the following: First, the subspace should include the original vectors $(\mathbf{y}_{k1}, \mathbf{0})$ and $(\mathbf{0}, \mathbf{y}_{k2})$ to ensure that the objective function value decreases. In order to further broaden the search space, we should consider vectors orthogonal to the original vectors. Since the vectors $(\mathbf{y}_{k2}, \mathbf{0})$ and $(\mathbf{0}, \mathbf{y}_{k1})$ are orthogonal to the original vectors, they are suitable for inclusion in the subspace. Finally, as we will see in Section 4, the optimality condition associated with the global step and with these 4 vectors provides a link between the subproblems which is exploited to obtain convergence.

Notice that when \mathbf{H}_1 and \mathbf{H}_2 are 2 by 2 matrices, the 4 vectors in (3) span \mathbb{R}^4 . Hence, in the 2 by 2 case, the global step yields a global optimum for (1). More generally, we find that the local steps steer the iterates into a subspace associated with the eigenvectors of \mathbf{H}_1

for $k = 0, 1, 2, \dots$

If k is even, perform a forward step:

$$\begin{aligned} \mathbf{y}_{k1} &\in \arg \min \{F(\mathbf{z}, \mathbf{x}_{k2}) : \|\mathbf{z}\| = 1, \mathbf{z}^\top \mathbf{x}_{k2} = 0\}, \\ \mathbf{y}_{k2} &\in \arg \min \{F(\mathbf{y}_{k1}, \mathbf{z}) : \|\mathbf{z}\| = 1, \mathbf{z}^\top \mathbf{y}_{k1} = 0\}, \end{aligned}$$

Else perform a reverse step:

$$\begin{aligned} \mathbf{y}_{k2} &\in \arg \min \{F(\mathbf{x}_{k1}, \mathbf{z}) : \|\mathbf{z}\| = 1, \mathbf{z}^\top \mathbf{x}_{k1} = 0\}, \\ \mathbf{y}_{k1} &\in \arg \min \{F(\mathbf{z}, \mathbf{y}_{k2}) : \|\mathbf{z}\| = 1, \mathbf{z}^\top \mathbf{y}_{k2} = 0\}. \end{aligned}$$

Global step: Set $\mathbf{x}_{k+1} = \mathbf{z}(s_k)$ where

$$\mathbf{z}_k(s) = \frac{1}{\sqrt{1+s^2}}(\mathbf{y}_k + s\mathbf{d}), \quad \mathbf{d} = \pm \begin{bmatrix} \mathbf{y}_{k2} \\ -\mathbf{y}_{k1} \end{bmatrix},$$

and

$$s_k \in \arg \min \{F_k(s) : s \in [0, -\rho F'_k(0)]\}, \quad F_k(s) = F(\mathbf{z}_k(s)).$$

The sign of \mathbf{d} is chosen so that $F'_k(0) \leq 0$.

end

Figure 1: The decomposition algorithm.

and \mathbf{H}_2 corresponding to the smallest eigenvalues, while the global step finds the best point within this low dimensional subspace.

Ideally, the stepsize s_k is the global minimum of $F_k(s)$ over all s . However, the convergence analysis for this ‘‘optimal step’’ is not easy since $\|\mathbf{y}_k - \mathbf{x}_{k+1}\|$ could be on the order of 1 for all k . For example, if $\mathbf{H}_1 = \mathbf{H}_2$, then $F_k(s)$ is constant, independent of s ; consequently, any choice of s is optimal, and there is no control over the iteration change. To ensure global convergence of the algorithm, we restrict the stepsize to an interval $[0, -\rho F'_k(0)]$, where ρ is a fixed positive scalar. In other words, we take

$$s_k \in \arg \min \{F_k(s) : s \in [0, -\rho F'_k(0)]\}. \quad (6)$$

Notice that $s_k = 0$ when $F'_k(0) = 0$ and the global step is skipped. In practice, we observe convergence when s_k is a global minimizer of F_k . The constraint on the stepsize is needed to rigorously prove convergence of the iteration. For reference, the complete algorithm is recapped in Figure 1.

As we show later in (41), $F'_k(0)$ tends to zero. Hence, the constraint $s \in [0, -\rho F'_k(0)]$ on the stepsize in the line search (6) implies that the iteration difference $\mathbf{x}_{k+1} - \mathbf{y}_k$ tends to zero. Another approach for controlling the stepsize is to employ a trust region scheme [4, p. 129] where we minimize F in the subspace (3) and a ball of radius ρ_k centered at \mathbf{y}_k . If ρ_k tends to zero, then the change $\mathbf{x}_{k+1} - \mathbf{y}_k$ again tends to zero. The update (6) amounts to a trust region step with a special choice for the trust region radius.

Since F is a pure quadratic, the objective function satisfies

$$F(\mathbf{x}_1, \mathbf{x}_2) = F(-\mathbf{x}_1, \mathbf{x}_2) = F(\mathbf{x}_1, -\mathbf{x}_2) = F(-\mathbf{x}_1, -\mathbf{x}_2).$$

Hence, if \mathbf{y}_{kj} is a minimum in a subproblem at iteration k , then so is $-\mathbf{y}_{kj}$. In order to carry out the analysis, it is convenient to choose the signs so that following inequalities hold:

$$\begin{cases} \mathbf{x}_{k2}^\top \mathbf{H}_1 \mathbf{y}_{k1} \geq 0 & \text{and} & \mathbf{y}_{k1}^\top \mathbf{H}_2 \mathbf{y}_{k2} \geq 0 & \text{(forward mode)} \\ \mathbf{x}_{k1}^\top \mathbf{H}_2 \mathbf{y}_{k2} \geq 0 & \text{and} & \mathbf{y}_{k2}^\top \mathbf{H}_1 \mathbf{y}_{k1} \geq 0 & \text{(reverse mode)} \end{cases} \quad (7)$$

With this sign convention, the multipliers associated with the orthogonality constraints in the local step are always nonnegative as shown in Section 4.

Our analysis establishes local, and in some cases global, convergence of the decomposition algorithm of Figure 1 to a stationary point. In Corollary 1 we show that if $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$ is a local minimizer for (1), then there exist scalars λ_1 , λ_2 , and μ satisfying the first-order condition

$$\begin{bmatrix} \mathbf{H}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_2 \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 \mathbf{I} & \mu \mathbf{I} \\ \mu \mathbf{I} & \lambda_2 \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}, \quad (8)$$

where λ_1 and λ_2 lie between the smallest and second smallest eigenvalues of \mathbf{H}_1 and \mathbf{H}_2 respectively. The condition (8) together with the requirement that \mathbf{y}_1 and \mathbf{y}_2 are feasible in (1) form the KKT (Karush-Kuhn-Tucker) conditions.

Solutions of the subproblems (2) satisfy the following conditions: There exist scalars λ_{k1} , μ_{k1} , λ_{k2} , and μ_{k2} such that

$$\begin{array}{l} \text{forward} \\ \text{reverse} \end{array} \begin{bmatrix} \mathbf{H}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_2 \end{bmatrix} \begin{bmatrix} \mathbf{y}_{k1} \\ \mathbf{y}_{k2} \end{bmatrix} = \begin{bmatrix} \lambda_{k1} \mathbf{y}_{k1} + \mu_{k1} \mathbf{x}_{k2} \\ \mu_{k2} \mathbf{y}_{k1} + \lambda_{k2} \mathbf{y}_{k2} \end{bmatrix}, \quad (9)$$

A fundamental difference between the first-order optimality conditions for the original optimization problem (1) and the subproblems (2) is that μ_{k1} may not equal μ_{k2} in the subproblems. Hence, a key objective in the analysis is to show that the multipliers in the subproblems approach a common limit. As will be seen in the analysis that follows, we are able to bound the difference $\mu_{k1} - \mu_{k2}$ in terms of $F'_k(0)$, which tends to zero. Hence, as $F'_k(0)$ tends to zero, both the iteration change $\mathbf{x}_{k+1} - \mathbf{y}_k$ tends to zero, according to (6), and the multiplier difference $\mu_{k1} - \mu_{k2}$ tends to zero.

The local step in the decomposition algorithm requires the solution of a quadratic program of the following form:

$$\min \mathbf{x}^\top \mathbf{H} \mathbf{x} \quad \text{subject to} \quad \|\mathbf{x}\| = 1, \quad \mathbf{a}^\top \mathbf{x} = 0, \quad (10)$$

where $\mathbf{a} \in \mathbb{R}^n$ with $\|\mathbf{a}\| = 1$ and \mathbf{H} is symmetric. In the decomposition algorithm, \mathbf{H} is \mathbf{H}_i and \mathbf{a} is either \mathbf{x}_{ki} or \mathbf{y}_{ki} , $i = 1$ or 2 . Since \mathbf{H} is symmetric, we can perform an orthogonal change of variables to diagonalize \mathbf{H} . Hence, *without loss of generality*, we can assume that \mathbf{H} is diagonal with the ordered eigenvalues

$$\epsilon_1 \leq \epsilon_2 \leq \dots \leq \epsilon_n \quad (11)$$

The analysis of the decomposition algorithm is based on an analysis of how the multiplier for the constraint $\mathbf{a}^\top \mathbf{x} = 0$ in (10) depends on \mathbf{a} . If \mathbf{H} is a multiple of the identity, then this multiplier vanishes, and the dependence of the multiplier on \mathbf{a} is trivial. Except in this special case, the dependence of the multiplier on \mathbf{a} is nontrivial. For almost every choice of \mathbf{a} , the multiplier is unique and depends continuously on \mathbf{a} . Suppose that \mathbf{H} is not a multiple of the identity and let ϵ_s denote the smallest eigenvalue of \mathbf{H} which is strictly larger than ϵ_1 . The degenerate choices of \mathbf{a} , where uniqueness and continuity are lost correspond to those $\mathbf{a} \neq \mathbf{0}$ which satisfy the equations

$$\sum_{\epsilon_i = \epsilon_1} \frac{a_i^2}{\epsilon_s - \epsilon_1} = \sum_{\epsilon_i > \epsilon_s} \frac{a_i^2}{\epsilon_i - \epsilon_s}, \quad a_i = 0 \text{ when } \epsilon_i = \epsilon_s. \quad (12)$$

We say that \mathbf{a} is degenerate for \mathbf{H} if (12) holds, and conversely, \mathbf{a} is nondegenerate if (12) is violated or \mathbf{H} is a multiple of the identity. The degenerate choices of \mathbf{a} compose a set of measure 0. We say that $(\mathbf{y}_1, \mathbf{y}_2)$ is nondegenerate for (1) if \mathbf{y}_1 is nondegenerate for \mathbf{H}_2 and \mathbf{y}_2 is nondegenerate for \mathbf{H}_1 . If \mathbf{H}_1 and \mathbf{H}_2 commute, then the solution to (1), given in Section 3, is nondegenerate.

Our main result is the following:

Theorem 1. *If the global minimizers of (1) are all nondegenerate, then for any starting guess sufficiently close to the solution set, there exists a subsequence of the iterates of the decomposition algorithm of Figure 1 that approaches a stationary point for (1).*

The proof of Theorem 1 will be given in Section 4.

Remark 1. In the special case where $\mathbf{H}_1 = \mathbf{H}_2$ and $\epsilon_3 - \epsilon_2 \geq \epsilon_2 - \epsilon_1$, it is shown in [1] that the decomposition algorithm is globally convergent for any starting point. On the other hand, we observe in Section 5 that when $\epsilon_3 - \epsilon_2 < \epsilon_2 - \epsilon_1$, then for specially chosen starting points, the algorithm could converge to a stationary point which is not a global minimum.

In our local convergence result Theorem 1, the requirement for the starting point ensures that the iterates avoid degenerate points for either \mathbf{H}_1 or \mathbf{H}_2 . Let C_d denote the minimum value for the objective function of (1) subject to the additional constraint that either \mathbf{x}_1 is degenerate for \mathbf{H}_2 or \mathbf{x}_2 is degenerate for \mathbf{H}_1 . Since the global minimizers of (1) are nondegenerate, C_d is strictly larger than the minimum value for the objective function. If the objective function at the starting point is strictly less than C_d , then the iterates are bounded away from degenerate points for either \mathbf{H}_1 or \mathbf{H}_2 .

The paper is organized as follows. In Section 2 we develop necessary and sufficient optimality conditions for a quadratic optimization problem with both a sphere and an affine constraint, and we develop necessary optimality conditions for (1). In Section 3 we apply the optimality theory to obtain an optimal solution for the local subproblem, and we show that the multipliers in the subproblems possess a continuity property. The optimality theory also yields the solution to the original problem (1) when \mathbf{H}_1 and \mathbf{H}_2 commute. In Section 4 we prove our local convergence result Theorem 1. In Section 5 we investigate the global convergence of the decomposition algorithm using a series of numerical examples.

2 Optimality Conditions

Each step of the domain decomposition algorithm requires the solution of a sphere constrained, quadratic programming problem with a linear constraint. This leads us to consider a problem with the structure

$$\min f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top \mathbf{H}\mathbf{x} - \mathbf{h}^\top \mathbf{x} \quad \text{subject to } \mathbf{x}^\top \mathbf{x} = 1, \quad \mathbf{A}\mathbf{x} = \mathbf{b}, \quad (13)$$

where \mathbf{A} is m by n , $\mathbf{h} \in \mathbb{R}^n$, and $\mathbf{b} \in \mathbb{R}^m$. The local steps of our decomposition algorithm correspond to the case $\mathbf{h} = \mathbf{0}$, $m = 1$, and $b = 0$. Our analysis in this section, however, applies to the more general quadratic cost function and linear constraints appearing in (13).

The following result gives necessary and sufficient conditions for a point to be a global minimum. Without the linear constraint, this result is known (see [7]). We give a slightly different analysis which also takes into account linear constraints. Recall that at a local minimizer where a constraint qualification holds, the Hessian of the Lagrangian is typically positive semidefinite over the tangent space associated with all the constraints, both the linear constraint $\mathbf{A}\mathbf{x} = \mathbf{b}$ and the sphere constraint $\mathbf{x}^\top \mathbf{x} = 1$. If \mathbf{y} is a global minimizer for (13) and λ is the Lagrange multiplier for the sphere constraint, then the second-order necessary optimality condition is that the first-order condition (8) holds and

$$\mathbf{d}^\top (\mathbf{H} - \lambda \mathbf{I}) \mathbf{d} \geq 0 \quad \text{whenever } \mathbf{A}\mathbf{d} = \mathbf{0} \quad \text{and} \quad \mathbf{y}^\top \mathbf{d} = 0.$$

In (15), we claim that the condition $\mathbf{y}^\top \mathbf{d} = 0$ can be dropped and the Hessian of the Lagrangian is positive semidefinite over a larger space, the null space of \mathbf{A} .

Proposition 1. *Suppose that \mathbf{y} is feasible in (13). A necessary and sufficient condition for \mathbf{y} to be a global minimizer is that there exist $\lambda \in \mathbb{R}$ and $\boldsymbol{\mu} \in \mathbb{R}^m$ such that*

$$\mathbf{H}\mathbf{y} = \mathbf{h} + \mathbf{y}\lambda + \mathbf{A}^\top \boldsymbol{\mu} \quad (14)$$

and

$$\mathbf{d}^\top (\mathbf{H} - \lambda \mathbf{I}) \mathbf{d} \geq 0 \quad \text{whenever } \mathbf{A}\mathbf{d} = \mathbf{0}. \quad (15)$$

Proof. Let $\mathcal{L} : \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}$ be the Lagrangian defined by

$$\mathcal{L}(\lambda, \boldsymbol{\mu}, \mathbf{x}) = f(\mathbf{x}) + \frac{\lambda}{2}(1 - \mathbf{x}^\top \mathbf{x}) + \boldsymbol{\mu}^\top (\mathbf{b} - \mathbf{A}\mathbf{x}).$$

First, suppose that there exist $\lambda \in \mathbb{R}$ and $\boldsymbol{\mu} \in \mathbb{R}^m$ such that (14) and (15) hold. For any feasible \mathbf{x} for (13), a Taylor expansion of \mathcal{L} around \mathbf{y} yields

$$\begin{aligned} f(\mathbf{x}) &= \mathcal{L}(\lambda, \boldsymbol{\mu}, \mathbf{x}) \\ &= \mathcal{L}(\lambda, \boldsymbol{\mu}, \mathbf{y}) + \nabla_x \mathcal{L}(\lambda, \boldsymbol{\mu}, \mathbf{y})(\mathbf{x} - \mathbf{y}) + \frac{1}{2}(\mathbf{x} - \mathbf{y})^\top (\mathbf{H} - \lambda \mathbf{I})(\mathbf{x} - \mathbf{y}) \\ &= f(\mathbf{y}) + \frac{1}{2}(\mathbf{x} - \mathbf{y})^\top (\mathbf{H} - \lambda \mathbf{I})(\mathbf{x} - \mathbf{y}). \end{aligned} \quad (16)$$

The first-derivative term in (16) vanishes due to (14). Since \mathbf{x} is feasible, $\mathbf{A}(\mathbf{x} - \mathbf{y}) = \mathbf{0}$. Hence, (15) and (16) imply that $f(\mathbf{x}) \geq f(\mathbf{y})$, which shows that \mathbf{y} is a global minimizer for (13).

Conversely, suppose that \mathbf{y} is a global minimizer for (13). Condition (14) is the usual first-order optimality condition at \mathbf{y} . This condition holds if the following ‘‘constraint qualification’’ is satisfied (e. g. see [6]): For each vector \mathbf{d} in the tangent space \mathcal{T} at \mathbf{y} , there exists a feasible curve approaching \mathbf{y} along the direction \mathbf{d} , where

$$\mathcal{T} = \{\mathbf{d} \in \mathbb{R}^n : \mathbf{y}^\top \mathbf{d} = 0, \mathbf{A}\mathbf{d} = \mathbf{0}\}.$$

Given $\mathbf{d} \in \mathcal{T}$, such a feasible curve is given by the formula

$$\mathbf{x}(t) = \mathbf{x}_0 + \left(\frac{\mathbf{y} + t\mathbf{d} - \mathbf{x}_0}{\|\mathbf{y} + t\mathbf{d} - \mathbf{x}_0\|} \right) \|\mathbf{x}_0 - \mathbf{y}\|, \quad (17)$$

where t is a scalar and \mathbf{x}_0 is the point satisfying the linear equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ which is closest to the origin. Since the expression in parentheses in (17) lies in the null space of \mathbf{A} and since $\mathbf{A}\mathbf{x}_0 = \mathbf{b}$, it follows that $\mathbf{A}\mathbf{x}(t) = \mathbf{b}$ for each choice of t . Since \mathbf{x}_0 is orthogonal to the null space of \mathbf{A} , it follows from the Pythagorean theorem that $\mathbf{x}(t)$ is a unit vector for each choice of t . Differentiating $\mathbf{x}(t)$, we obtain

$$\mathbf{x}'(0) = \mathbf{d} - (\mathbf{y} - \mathbf{x}_0) \left(\frac{(\mathbf{y} - \mathbf{x}_0)^\top \mathbf{d}}{\|\mathbf{y} - \mathbf{x}_0\|^2} \right).$$

Since $\mathbf{d} \in \mathcal{T}$, $\mathbf{y}^\top \mathbf{d} = 0$. Since \mathbf{x}_0 is orthogonal to the null space of \mathbf{A} and $\mathbf{A}\mathbf{d} = \mathbf{0}$, we have $\mathbf{x}_0^\top \mathbf{d} = \mathbf{0}$. Hence, $\mathbf{x}'(0) = \mathbf{d}$ and there exists a feasible curve approaching \mathbf{y} in the direction \mathbf{d} . This verifies the constraint qualification for (13); consequently, the first-order condition (14) is satisfied for some $\lambda \in \mathbb{R}$ and $\boldsymbol{\mu} \in \mathbb{R}^m$.

By (16), the first-order optimality condition (14), and the global optimality of \mathbf{y} , we have

$$(\mathbf{x} - \mathbf{y})^\top (\mathbf{H} - \lambda \mathbf{I})(\mathbf{x} - \mathbf{y}) = 2(f(\mathbf{x}) - f(\mathbf{y})) \geq 0 \quad (18)$$

whenever \mathbf{x} is feasible in (13). Suppose that $\mathbf{A}\mathbf{d} = \mathbf{0}$. If in addition, $\mathbf{d}^\top \mathbf{y} = 0$, then $\mathbf{d} \in \mathcal{T}$. Earlier we observed that when $\mathbf{d} \in \mathcal{T}$, $\mathbf{x}(t)$ is feasible in (13) for all choices of t . Since $\mathbf{x}(t)$ is feasible, we can substitute $\mathbf{x} = \mathbf{x}(t)$ in (18). Since $\mathbf{x}(t) - \mathbf{y} = t\mathbf{d} + O(t^2)$, it follows from (18), after dividing by t^2 and letting t approach 0^+ , that (15) holds. If $\mathbf{d}^\top \mathbf{y} \neq 0$, then $\mathbf{d} \notin \mathcal{T}$, and $\|\mathbf{y} + t\mathbf{d}\| < 1$ for a suitable choice of t near 0. Increase the magnitude of t until $\|\mathbf{y} + t\mathbf{d}\| = 1$. Substituting $\mathbf{x} = \mathbf{y} + t\mathbf{d}$ in (18) gives (15). \square

We now obtain bounds on the location of the multiplier λ associated with (13).

Proposition 2. *If the eigenvalues of \mathbf{H} are arranged in increasing order as in (11) and if \mathbf{A} has rank $k \geq 1$, then $\lambda \leq \epsilon_{k+1}$ when (15) holds. Moreover, if $\mathbf{h} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$ and \mathbf{y} is a global minimizer in (13), then $\lambda \geq \epsilon_1$.*

Proof. If \mathcal{W} is the $k + 1$ dimensional space spanned by the eigenvectors associated with the $k + 1$ smallest eigenvalues of \mathbf{H} , then we have

$$\epsilon_{k+1} = \max\{\mathbf{v}^\top \mathbf{H} \mathbf{v} : \mathbf{v} \in \mathcal{W}, \|\mathbf{v}\| = 1\}. \quad (19)$$

Since \mathbf{A} has rank $k \geq 1$, the dimension of the null space of \mathbf{A} is $n - k$, and there exists a unit vector \mathbf{v} which lies both in the null space of \mathbf{A} and in \mathcal{W} . Since $\mathbf{A} \mathbf{v} = \mathbf{0}$, the second-order condition (15) and (19) yield

$$\epsilon_{k+1} \geq \mathbf{v}^\top \mathbf{H} \mathbf{v} \geq \lambda.$$

If $\mathbf{h} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$, then the first-order condition (14) implies that

$$\lambda = \mathbf{y}^\top \mathbf{H} \mathbf{y} \geq \min\{\mathbf{v}^\top \mathbf{H} \mathbf{v} : \|\mathbf{v}\| = 1\} = \epsilon_1.$$

□

Next, we focus on the original 2-variable problem (1).

Corollary 1. *If $(\mathbf{y}_1, \mathbf{y}_2)$ is a local minimizer for (1), then there exist λ_1, λ_2 , and μ such that (8) holds. If \mathbf{y} is a global minimizer for (1), then for $i = 1, 2$, we have $\lambda_i \in [\epsilon_{i1}, \epsilon_{i2}]$, where ϵ_{ij} is the j -smallest eigenvalue of \mathbf{H}_i ,*

$$\epsilon_{i1} \leq \epsilon_{i2} \leq \dots \leq \epsilon_{in}. \quad (20)$$

Proof. The gradients of the constraints for (1) at \mathbf{y} are multiples of the 3 vectors

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{0} \end{bmatrix}, \quad \begin{bmatrix} \mathbf{0} \\ \mathbf{y}_2 \end{bmatrix}, \quad \begin{bmatrix} \mathbf{y}_2 \\ \mathbf{y}_1 \end{bmatrix}.$$

Since these vectors are orthogonal, they are linearly independent. Since the “linear independence constraint qualification” is satisfied, the first-order optimality condition (8) holds for suitable choices of λ_1, λ_2 , and μ . If \mathbf{y} is a global minimizer of (1), then \mathbf{y}_i is a global minimizer of the problem

$$\min \mathbf{x}^\top \mathbf{H}_i \mathbf{x} \quad \text{subject to } \mathbf{x}^\top \mathbf{x} = 1, \quad \mathbf{z}^\top \mathbf{x} = 0,$$

where $\mathbf{z} = \mathbf{y}_2$ when $i = 1$ and $\mathbf{z} = \mathbf{y}_1$ when $i = 2$. We apply Proposition 2 with $k = 1$ to obtain $\lambda_i \in [\epsilon_{i1}, \epsilon_{i2}]$, $i = 1, 2$. □

3 The local step and continuity

In each step of the domain decomposition algorithm, we must solve a quadratic programming problem of the form (10). After an orthogonal change of variables, we can assume, without loss of generality, that \mathbf{H} is diagonal with the ordered eigenvalues (11) on the diagonal and $\|\mathbf{a}\| = 1$. Using Propositions 1 and 2, we now determine the optimal solutions to (10). In the special case $\mathbf{h} = \mathbf{0}$ and $\mathbf{A} = \mathbf{a}^\top$, the first-order optimality conditions (14) reduce to

$$\mathbf{H}\mathbf{y} = \mathbf{y}\lambda + \mathbf{a}\mu \quad (21)$$

Case 1: $\epsilon_1 = \epsilon_2$. By Proposition 2, the multiplier λ of Proposition 1 is $\lambda = \epsilon_1 = \epsilon_2$. Define the set

$$\mathcal{E}_i = \{j : \epsilon_j = \epsilon_i\}.$$

If $\mu \neq 0$, then by (21) we must have $a_i = 0$ for all $i \in \mathcal{E}_1$. If $i \notin \mathcal{E}_1$, then

$$y_i = \frac{\mu a_i}{\epsilon_i - \epsilon_1} \quad \text{and} \quad \mathbf{a}^\top \mathbf{y} = \mu \sum_{i \notin \mathcal{E}_1} \frac{a_i^2}{\epsilon_i - \epsilon_1} \neq 0, \quad (22)$$

which violates the orthogonality condition $\mathbf{a}^\top \mathbf{y} = 0$. Hence, $\mu = 0$ and all \mathbf{y} satisfying the following conditions are solutions to (10):

$$y_i = 0 \text{ if } i \notin \mathcal{E}_1, \quad \mathbf{a}^\top \mathbf{y} = 0, \quad \|\mathbf{y}\| = 1. \quad (23)$$

Observe that there is an infinite set of solutions \mathbf{y} while the multipliers λ and μ are unique.

Case 2: $\epsilon_1 < \epsilon_2$ and $\mathbf{a}_1 = \mathbf{0}$. If $\lambda > \epsilon_1$, then the second-order condition (15) is violated by the vector $d_1 = 1$ and $d_i = 0$ for $i > 1$. Hence, $\lambda = \epsilon_1$. As in Case 1, the orthogonality condition $\mathbf{a}^\top \mathbf{y} = 0$ is violated unless $\mu = 0$. The solution is again given by (23) and the multipliers are $\lambda = \lambda_1$ and $\mu = 0$.

Case 3: $\epsilon_1 < \epsilon_2$ and $\mathbf{a}_1 \neq \mathbf{0}$. We first show that $\lambda > \epsilon_1$. Suppose, to the contrary, that $\lambda = \epsilon_1$. The first component of (21) implies that $\mu = 0$. Hence, (21) reduces to $\mathbf{H}\mathbf{y} = \epsilon_1 \mathbf{y}$. Since \mathbf{H} is diagonal and $\epsilon_i > \epsilon_1$ for $i > 1$, we conclude that $y_i = 0$ for $i > 1$. Hence, $y_1 = \pm 1$ since \mathbf{y} is a unit vector. However, a vector of this form violates the orthogonality condition $\mathbf{a}^\top \mathbf{y} = 0$ when $a_1 \neq 0$. This gives a contradiction, so we have $\lambda > \epsilon_1$.

(a) **$\mathbf{a}_i \neq \mathbf{0}$ for some $i \in \mathcal{E}_2$.** We show that $\lambda < \epsilon_2$. Suppose, to the contrary, that $\lambda = \epsilon_2$. Since $a_1 \neq 0$ and $a_i \neq 0$ for some $i \in \mathcal{E}_2$, the second-order condition (15) is violated by taking \mathbf{d} to be completely zero except for components 1 and i . Since $\epsilon_1 < \lambda < \epsilon_2$, (21) can be solved for \mathbf{y} :

$$\mathbf{y} = \mu(\mathbf{H} - \lambda \mathbf{I})^{-1} \mathbf{a}. \quad (24)$$

If $\mu = 0$, then $\mathbf{y} = \mathbf{0}$, which violates the constraint $\mathbf{y}^\top \mathbf{y} = 1$. We combine the expression (24), with the orthogonality condition $\mathbf{a}^\top \mathbf{y} = 0$, and the fact that $\mu \neq 0$ to obtain the equation

$$g(\lambda) := \sum_{i=1}^n \frac{a_i^2}{\epsilon_i - \lambda} = 0. \quad (25)$$

Observe that g is strictly monotone increasing on the interval (ϵ_1, ϵ_2) and $g(\epsilon_1^+) = -\infty$ since $a_1 \neq 0$ while $g(\epsilon_2^-) = +\infty$ since $a_2 \neq 0$. There exists a unique zero λ of g in (ϵ_1, ϵ_2) . The solution to (10) is

$$y_i = \frac{\mu a_i}{\epsilon_i - \lambda} \quad \text{where} \quad \mu^2 = \left(\sum_{i=1}^n \frac{a_i^2}{(\epsilon_i - \lambda)^2} \right)^{-1} = g'(\lambda)^{-1}. \quad (26)$$

The equation for μ^2 is obtained from the requirement that $\mathbf{y}^\top \mathbf{y} = 1$. Notice that both the solution \mathbf{y} and the multiplier μ are unique to within sign.

- (b) $\mathbf{a}_i = \mathbf{0}$ for all $i \in \mathcal{E}_2$ and $g(\epsilon_2) < 0$. We show that $\lambda = \epsilon_2$ and $\mu = 0$. By (21), we have

$$y_i = \frac{\mu a_i}{\epsilon_i - \lambda} \quad \text{when } i \notin \mathcal{E}_2.$$

If $\mu \neq 0$, then the orthogonality condition $\mathbf{a}^\top \mathbf{y} = 0$ reduces to (25), which has no solution on (ϵ_1, ϵ_2) since g is monotone on this interval, $g(\epsilon_1^+) = -\infty$, and $g(\epsilon_2) < 0$. Hence, $\mu = 0$. If $\lambda < \epsilon_2$, then (24) implies that $\mathbf{y} = \mathbf{0}$, which violates the constraint $\mathbf{y}^\top \mathbf{y} = 1$. Hence, $\lambda = \epsilon_2$ and $\mu = 0$. The solution consists of all vectors \mathbf{y} satisfying

$$y_i = 0 \text{ if } i \notin \mathcal{E}_2, \quad \|\mathbf{y}\| = 1. \quad (27)$$

Notice that λ and μ are again unique.

- (c) $\mathbf{a}_i = \mathbf{0}$ for all $i \in \mathcal{E}_2$ and $g(\epsilon_2) > 0$. First, suppose that $\mu \neq 0$. Since $g(\epsilon_1^+) = -\infty$ while $g(\epsilon_2) > 0$, g in (25) has a unique zero on (ϵ_1, ϵ_2) . Hence, one solution to (21) is given by (26). We now consider the possibility that $\mu = 0$ at a global minimum. We will show that this leads to a contradiction. Consequently, there is a unique (to within sign) global minimizer for (10) given by (26). If $\mu = 0$, then by (21), $(\epsilon_i - \lambda)y_i = 0$ for all i , which implies that $y_i = 0$ for $i \notin \mathcal{E}_2$ since $\epsilon_1 < \lambda \leq \epsilon_2$. Since $\|\mathbf{y}\| = 1$, it follows that $\lambda = \epsilon_2$ (or else $\mathbf{y} = \mathbf{0}$, violating the condition $\|\mathbf{y}\| = 1$). We now show that the second-order condition (15) is violated for the choice

$$d_i = \frac{a_i}{\epsilon_i - \gamma},$$

where γ is the unique zero of g on the interval (ϵ_1, ϵ_2) . This choice for \mathbf{d} satisfies the condition $\mathbf{a}^\top \mathbf{d} = 0$ since $g(\gamma) = 0$. Since $\gamma < \epsilon_2$, we have

$$\begin{aligned} \mathbf{d}^\top (\mathbf{H} - \lambda \mathbf{I}) \mathbf{d} &= \mathbf{d}^\top (\mathbf{H} - \epsilon_2 \mathbf{I}) \mathbf{d} = \sum_{i=1}^n d_i^2 (\epsilon_i - \epsilon_2) \\ &= \sum_{i \notin \mathcal{E}_2} \frac{a_i^2 (\epsilon_i - \epsilon_2)}{(\epsilon_i - \gamma)^2} < \sum_{i \notin \mathcal{E}_2} \frac{a_i^2 (\epsilon_i - \gamma)}{(\epsilon_i - \gamma)^2} = g(\gamma) = 0. \end{aligned}$$

This violates the second-order condition (15).

- (d) $\mathbf{a}_i = \mathbf{0}$ for all $i \in \mathcal{E}_2$ and $g(\epsilon_2) = 0$. This is the degenerate case introduced in Section 1. We first observe that $\lambda = \epsilon_2$. Suppose, to the contrary, that $\lambda < \epsilon_2$. By (21), \mathbf{y} is given by (24). If $\mu \neq 0$, then the orthogonality condition gives (25), which has no solution on (ϵ_1, ϵ_2) since g is monotone and $g(\epsilon_2) = 0$. Consequently, $\mu = 0$ and (24) implies that $\mathbf{y} = \mathbf{0}$, violating the constraint $\mathbf{y}^\top \mathbf{y} = 1$. Thus $\lambda = \epsilon_2$. By (21),

$$y_i = \frac{\mu a_i}{\epsilon_i - \epsilon_2} \quad \text{for } i \notin \mathcal{E}_2. \quad (28)$$

For $i \in \mathcal{E}_2$, the first-order condition (14) provides no information concerning y_i since both sides of the equation vanish identically:

$$(\epsilon_i - \epsilon_2) y_i = \mu a_i = 0.$$

The general solution is the following. First, choose any value for y_i , $i \in \mathcal{E}_2$, such that

$$\sum_{i \in \mathcal{E}_2} y_i^2 \leq 1.$$

Then choose μ in (28) such that $\|\mathbf{y}\| = 1$. In other words, we choose μ so that

$$\mu^2 = \frac{1 - \sum_{i \in \mathcal{E}_2} y_i^2}{g'(\epsilon_2)}.$$

Notice that λ is unique in the degenerate case, while both μ and \mathbf{y} are not unique.

Lemma 1. *For the optimization problem (10) and a global minimizer \mathbf{y} , the multiplier λ associated with the constraint $\mathbf{y}^\top \mathbf{y} = 1$ is a Lipschitz continuous function of \mathbf{a} on the unit sphere. With appropriate sign, the corresponding multiplier μ associated with the orthogonality constraint $\mathbf{a}^\top \mathbf{y} = 0$ is continuous at any nondegenerate \mathbf{a} .*

Proof. In Case 1, Lipschitz continuity is trivially satisfied, so we focus on the situation where $\epsilon_1 < \epsilon_2$. Since the intersection of the hyperplanes $a_1 = 0$ or $a_2 = 0$ with the unit sphere are sets of measure zero on the surface of the sphere, Lipschitz continuity over the complement implies Lipschitz continuity over the entire sphere (by continuity). Hence, we restrict our attention to $\epsilon_1 < \epsilon_2$, $a_1 \neq 0$, and $a_2 \neq 0$. In this case, λ is the unique solution to (25) on the interval (ϵ_1, ϵ_2) . Differentiating (25) gives

$$\frac{\partial \lambda}{\partial a_i} = \frac{2a_i}{(\lambda - \epsilon_i)g'(\lambda)}. \quad (29)$$

If for some i , we have

$$\sum_{j \in \mathcal{E}_i} a_j^2 \geq 1/2, \quad (30)$$

then

$$\begin{aligned} |\epsilon_i - \lambda|g'(\lambda) &= \left(\frac{1}{|\epsilon_i - \lambda|} \right) \left(\sum_{j \in \mathcal{E}_i} a_j^2 \right) + |\epsilon_i - \lambda| \sum_{j \notin \mathcal{E}_i} \frac{a_j^2}{(\epsilon_j - \lambda)^2} \\ &\geq \left(\frac{1}{|\epsilon_i - \lambda|} \right) \sum_{j \in \mathcal{E}_i} a_j^2 \geq \frac{1}{2|\epsilon_i - \lambda|} \geq \frac{1}{2(\epsilon_n - \epsilon_1)}. \end{aligned} \quad (31)$$

It follows from (29) that when (30) holds,

$$\left| \frac{\partial \lambda}{\partial a_i} \right| \leq 4(\epsilon_n - \epsilon_1).$$

If $a_i = 0$, then $\partial \lambda / \partial a_i = 0$ by (29). Now, suppose that $a_i \neq 0$ and (30) is violated. By (29) and (31), we have

$$\begin{aligned} \left| \frac{\partial \lambda}{\partial a_i} \right| &= \frac{2|a_i|}{\frac{1}{|\epsilon_i - \lambda|} \left(\sum_{j \in \mathcal{E}_i} a_j^2 \right) + |\epsilon_i - \lambda| \left(\sum_{j \notin \mathcal{E}_i} \frac{a_j^2}{(\epsilon_j - \lambda)^2} \right)} \\ &\leq \frac{2}{\left(\frac{|a_i|}{|\epsilon_i - \lambda|} \right) + \frac{|\epsilon_i - \lambda|}{|a_i|} \left(\sum_{j \notin \mathcal{E}_i} \frac{a_j^2}{(\epsilon_n - \epsilon_1)^2} \right)} \\ &\leq \frac{2}{\left(\frac{|a_i|}{|\epsilon_i - \lambda|} \right) + \left(\frac{|\epsilon_i - \lambda|}{|a_i|} \right) \left(\frac{1}{2(\epsilon_n - \epsilon_1)^2} \right)}. \end{aligned} \quad (32)$$

The last inequality is due to the assumption that (30) is violated, which implies that

$$\sum_{j \notin \mathcal{E}_i} a_j^2 \geq 1/2.$$

The denominator contains both $|a_i|/|\epsilon_i - \lambda|$ and its reciprocal $|\epsilon_i - \lambda|/|a_i|$. Suppose that

$$\frac{|a_i|}{|\epsilon_i - \lambda|} \geq 1. \quad (33)$$

By (32), the partial derivative $\partial\lambda/\partial a_i$ is bounded by 2 in magnitude since both terms in the denominator of (32) are positive and one of the terms is greater than or equal to 1. Conversely, if (33) is violated, then we drop the first term in the denominator of (32) to obtain

$$\left| \frac{\partial\lambda}{\partial a_i} \right| \leq 4(\epsilon_n - \epsilon_1)^2.$$

At this point, we have shown that there exists a constant β with the property that if \mathbf{a} lies on the unit sphere with $a_1 \neq 0$ and $a_2 \neq 0$, then

$$\left| \frac{\partial\lambda}{\partial a_i} \right| \leq \beta$$

for each i . Observe that the solution λ to (25) does not change if \mathbf{a} is multiplied by a nonzero scalar. Hence, for any nonzero \mathbf{a} (not necessarily on the unit sphere) with $a_1 \neq 0$ and $a_2 \neq 0$, we have

$$\left| \frac{\partial\lambda}{\partial a_i} \right| \leq \beta/\|\mathbf{a}\|.$$

Consequently, there exists constants $r_1 < 1 < r_2$ with the property that

$$\left| \frac{\partial\lambda}{\partial a_i} \right| \leq 2\beta$$

whenever $r_1 \leq \|\mathbf{a}\| \leq r_2$, $a_1 \neq 0$, and $a_2 \neq 0$. Given two arbitrary points on the unit sphere, we can construct a piecewise linear path between them with the property that the line segments all lie within the shell formed by the spheres of radius r_1 and r_2 . Due to the bound on the partial derivatives of λ , the change in λ across each line segment is bounded by 2β times the length of the line segment. Since the number of line segments is bounded, independent of the location of the points, we deduce that λ is a Lipschitz continuous function of \mathbf{a} on the unit sphere.

Now consider the multiplier μ . If $\epsilon_1 = \epsilon_2$, then $\mu = 0$ (Case 1) and there is nothing to prove. Next, we focus on Case 3a where μ is given by (26). For $\lambda \in (\epsilon_1, \epsilon_2)$, g' is bounded away from zero. For example,

$$g'(\lambda) = \sum_{i=1}^n \frac{a_i^2}{(\epsilon_i - \lambda)^2} \geq \frac{1}{(\epsilon_n - \epsilon_1)^2}.$$

Let $\lambda(\mathbf{a})$ denote the unique multiplier associated with any given \mathbf{a} . Since λ is a Lipschitz continuous function of \mathbf{a} , it follows that μ is continuous at any point \mathbf{a} where $\lambda(\mathbf{a}) \neq \epsilon_i$ for all i . Since $\lambda \in [\epsilon_1, \epsilon_2]$, the only potential points of discontinuity are those points \mathbf{b} for which $\lambda(\mathbf{b}) = \epsilon_i$, $i = 1$ or $i \in \mathcal{E}_2$. If $b_1 \neq 0$ or $b_i \neq 0$ for any $i \in \mathcal{E}_2$, then $\mu(\mathbf{a})$ approaches 0 as \mathbf{a} approaches \mathbf{b} due to the pole in the denominator of g' . Hence, μ is continuous at \mathbf{b} .

If $b_1 = 0$ and $\lambda(\mathbf{b}) = \epsilon_1$, then we use (25) to solve for $a_1^2/(\epsilon_1 - \lambda)$:

$$\frac{a_1^2}{\lambda - \epsilon_1} = \sum_{i>1} \frac{a_i^2}{\epsilon_i - \lambda} \quad (34)$$

By assumption, $\lambda(\mathbf{a})$ approaches ϵ_1 as \mathbf{a} approaches \mathbf{b} . Since the right side of (34) is continuous when λ is near ϵ_1 , the limit of the left side as \mathbf{a} approaches \mathbf{b} is

$$\sum_{i>1} \frac{b_i^2}{\epsilon_i - \epsilon_1} > 0$$

since $b_1 = 0$ and $\|\mathbf{b}\| = 1$. Consequently, by (26), μ tends to 0 as \mathbf{a} approaches \mathbf{b} , the same limit given in Case 2.

Finally, suppose that $b_i = 0$ for all $i \in \mathcal{E}_2$ and $\lambda(\mathbf{b}) = \epsilon_2$. Again, by (25), we have

$$\left(\frac{1}{\lambda - \epsilon_2}\right) \sum_{i \in \mathcal{E}_2} a_i^2 = \sum_{i \notin \mathcal{E}_2} \frac{a_i^2}{\epsilon_i - \lambda} \quad (35)$$

According to the statement of the lemma, we only need to prove continuity at nondegenerate \mathbf{b} , in which case the right side does not vanish at $\lambda = \epsilon_2$. Hence, as \mathbf{a} approaches \mathbf{b} , the right side approaches the limit

$$\sum_{i \notin \mathcal{E}_2} \frac{b_i^2}{\epsilon_i - \epsilon_2} \neq 0.$$

Consequently, by (26), μ tends to 0 as \mathbf{a} approaches \mathbf{b} , the same limit given in Case 3b. Note that Case 3c is not a point of discontinuity of μ since $\lambda < \epsilon_2$. This completes the proof. \square

Now let us consider the original problem (1). If \mathbf{H}_1 and \mathbf{H}_2 commute, then they are simultaneously diagonalizable by the same eigenvector matrix [8, p. 249]. In this case, we can perform an orthogonal change of variables to reduce \mathbf{H}_1 and \mathbf{H}_2 to diagonal matrices. The solution is as follows:

Corollary 2. *Suppose \mathbf{H}_i , $i = 1$ and 2 , are diagonal with diagonal element ϵ_{ij} , $1 \leq j \leq n$, arranged in increasing order. The minimum cost in (1) is*

$$\epsilon_{11} + \epsilon_{22} \quad \text{or} \quad \epsilon_{12} + \epsilon_{21},$$

whichever is smaller. In the first case, an associated solution to (1) is

$$y_{11} = 1, \quad y_{22} = 1, \quad \text{and} \quad y_{ij} = 0 \text{ otherwise.}$$

In the latter case, an associated solution to (1) is

$$y_{12} = 1, \quad y_{21} = 1, \quad \text{and} \quad y_{ij} = 0 \text{ otherwise.}$$

Proof. First, let us consider the case where the diagonal elements of \mathbf{H}_i are strictly separated:

$$\epsilon_{i1} < \epsilon_{i2} < \dots < \epsilon_{in}$$

for $i = 1, 2$. By the first-order optimality conditions (8) and by the diagonal structure of the \mathbf{H}_i , we have

$$(\epsilon_{1j} - \lambda_1)y_{1j} = \mu y_{2j} \quad \text{and} \quad (\epsilon_{2j} - \lambda_2)y_{2j} = \mu y_{1j}.$$

We combine these equations to obtain

$$\left[(\epsilon_{1j} - \lambda_1)(\epsilon_{2j} - \lambda_2) - \mu^2 \right] y_{1j} = 0 = \left[(\epsilon_{1j} - \lambda_1)(\epsilon_{2j} - \lambda_2) - \mu^2 \right] y_{2j}. \quad (36)$$

By Corollary 1, the multipliers λ_1 and λ_2 satisfy $\lambda_i \in [\epsilon_{i1}, \epsilon_{i2}]$. Hence, the coefficients of y_{1j} and y_{2j} in (36) are strictly increasing functions of $j \in [2, n]$. It follows that these coefficients can vanish for at most one $j \in [2, n]$ and possibly for $j = 1$. When the coefficients of y_{1j} and y_{2j} do not vanish in (36), we must have $y_{1j} = y_{2j} = 0$. In summary, at the global optimum, all the components of y_{ij} vanish except possibly y_{11} , y_{21} , y_{1j} , and y_{2j} for some $j \in [2, n]$. We focus on the case $j = 2$ since $j > 2$ leads to a larger cost.

Define $x_{1j}^2 = v_j$ and $x_{2j}^2 = w_j$ for $j = 1, 2$. The optimization problem (1) with \mathbf{H}_i diagonal and $x_{ij} = 0$ for $j > 2$ reduces to

$$\begin{aligned} \min \quad & v_1 \epsilon_{11} + v_2 \epsilon_{12} + w_1 \epsilon_{21} + w_2 \epsilon_{22} \\ \text{subject to} \quad & v_1 + v_2 = 1 = w_1 + w_2, \\ & v_1 w_1 = v_2 w_2, \quad v_1, v_2, w_1, w_2 \geq 0. \end{aligned}$$

The equation $v_1 w_1 = v_2 w_2$ is the orthogonality condition $x_{11} x_{21} = -x_{12} x_{22}$ squared. We substitute $v_1 = 1 - v_2$ and $w_1 = 1 - w_2$ to reduce the optimization problem to

$$\begin{aligned} \min \quad & \epsilon_{11} + \epsilon_{21} + v_2(\epsilon_{12} - \epsilon_{11}) + w_2(\epsilon_{22} - \epsilon_{21}) \\ \text{subject to} \quad & v_2 + w_2 = 1, \quad v_2 \geq 0, \quad w_2 \geq 0. \end{aligned} \quad (37)$$

We substitute $v_2 = 1 - w_2$ in the objective function to further reduce the optimization problem to

$$\begin{aligned} \min \quad & \epsilon_{21} + \epsilon_{12} + w_2(\epsilon_{11} + \epsilon_{22} - \epsilon_{21} - \epsilon_{12}) \\ \text{subject to} \quad & 0 \leq w_2 \leq 1. \end{aligned}$$

Since the cost function is linear in w_2 , the minimum is achieved at either $w_2 = 0$ ($w_1 = 1, v_1 = 0, v_2 = 1$) with objective function value $\epsilon_{21} + \epsilon_{12}$ or $w_2 = 1$ ($w_1 = 0, v_1 = 1, v_2 = 0$) with objective function value $\epsilon_{11} + \epsilon_{22}$.

When the diagonal elements are not strictly separated, the solution given in the statement of the corollary remains valid. This can be proved as follows: First, perturb the diagonal elements to make them strictly separated. By the previous analysis, we know that the solution given in the statement of the corollary is valid. Next, let the perturbation tend to zero. The limit of these perturbed solutions is a solution of the original unperturbed problem. \square

Remark 2. Assuming the eigenvalues are all distinct, then the degenerate choices for \mathbf{a} in (10) correspond to those vectors \mathbf{a} for which

$$\frac{a_1^2}{\epsilon_2 - \epsilon_1} = \sum_{i=3}^n \frac{a_i^2}{\epsilon_i - \epsilon_2}, \quad a_2 = 0, \quad \sum_{i \neq 2} a_i^2 = 1. \quad (38)$$

The solution to (1) given by Corollary 2 has the property that the nonzeros lie in the first two components of the vectors, while a degenerate \mathbf{a} must have nonzero in components greater than or equal to 3. In fact, it follows from (38) that

$$\sum_{i=3}^n a_i^2 \geq \frac{\epsilon_3 - \epsilon_2}{\epsilon_3 - \epsilon_1}.$$

Remark 3. Let us consider the special case $\mathbf{H}_1 = \mathbf{H}_2 = \mathbf{H}$. Since \mathbf{H}_1 and \mathbf{H}_2 commute, we can apply Corollary 2. Let ϵ_j denote the j -th smallest eigenvalue of \mathbf{H} . As shown in (37), the optimization problem (1) reduces to

$$\begin{aligned} \min \quad & 2\epsilon_1 + (v_2 + w_2)(\epsilon_2 - \epsilon_1) \\ \text{subject to} \quad & v_2 + w_2 = 1, \quad v_2 \geq 0, \quad w_2 \geq 0. \end{aligned}$$

Since $v_2 + w_2 = 1$, the objective function value is $\epsilon_1 + \epsilon_2$, independent of the choice of v_2 and w_2 satisfying the constraints. Hence, when $\mathbf{H}_1 = \mathbf{H}_2$ there are an infinite number of solutions to (1). \mathbf{y}_1 is any unit vector in the span of the eigenvectors associated with ϵ_1 and ϵ_2 , and \mathbf{y}_2 is any orthogonal unit vector in the same eigenspace. Note that if \mathbf{H} is 2 by 2, then all feasible points are optimal and $F(\mathbf{x}_1, \mathbf{x}_2)$ is the trace of \mathbf{H} whenever \mathbf{x}_1 and \mathbf{x}_2 are feasible in (1).

4 Convergence of the decomposition algorithm

The proof of Theorem 1 is organized into four steps. In Step 1, we analyze the global step and show that the iteration difference $\|\mathbf{x}_{k+1} - \mathbf{y}_k\|$ tends to 0. In Step 2, we show that the multipliers μ_{kj} in the local step are almost monotone decreasing since the violation in

monotonicity decays to zero as the iteration number k tends to infinity. Step 3 and 4 focus on the limit of the multipliers μ_{kj} as k tends to infinity. Step 3 considers the limit 0, while Step 4 considers a positive limit.

Step 1. Analysis of the global step

Suppose that iteration k corresponds to the forward mode. Let \mathbf{y}_k denote the result of the local step, and let \mathbf{x}_{k+1} be the result of the global step based on the starting point \mathbf{y}_k . Since \mathbf{x}_{k1} is feasible in the first subproblem of the forward mode (2), we have

$$F(\mathbf{y}_{k1}, \mathbf{x}_{k2}) \leq F(\mathbf{x}_{k1}, \mathbf{x}_{k2}).$$

Since \mathbf{x}_{k2} is feasible in the second subproblem, we have

$$F(\mathbf{y}_{k1}, \mathbf{y}_{k2}) \leq F(\mathbf{y}_{k1}, \mathbf{x}_{k2}).$$

Combining these relations gives

$$F(\mathbf{y}_k) \leq F(\mathbf{x}_k). \quad (39)$$

A similar analysis for the reverse mode also gives $F(\mathbf{y}_k) \leq F(\mathbf{x}_k)$.

The components of $\mathbf{z}_k(s)$ lie on the unit sphere for all choices of s . Consequently, $F_k''(s)$ is bounded by a finite constant M , uniformly in k and s . Define the constants

$$\delta = \min\{\rho, 1/M\} \quad \text{and} \quad \bar{s}_k = -\delta F_k'(0).$$

Since \bar{s}_k lies on the interval $[0, -\rho F_k'(0)]$ appearing in (6), we have

$$F(\mathbf{x}_{k+1}) = F_k(s_k) \leq F_k(\bar{s}_k). \quad (40)$$

Expanding in a Taylor series around $s = 0$, there exists $\xi_k \in [0, \bar{s}_k]$ such that

$$\begin{aligned} F_k(\bar{s}_k) &= F_k(0) + F_k'(0)\bar{s}_k + \frac{1}{2}\bar{s}_k^2 F_k''(\xi_k) \\ &\leq F_k(0) + F_k'(0)\bar{s}_k + \frac{1}{2}\bar{s}_k^2 M \\ &= F_k(0) + \delta F_k'(0)^2 (\frac{1}{2}\delta M - 1) \leq F_k(0) - \frac{\delta}{2} F_k'(0)^2 \\ &= F(\mathbf{y}_k) - \frac{\delta}{2} F_k'(0)^2 \leq F(\mathbf{x}_k) - \frac{\delta}{2} F_k'(0)^2, \end{aligned}$$

where the last inequality is (39). Combining this with (40) gives

$$F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) - \left(\frac{\delta}{2}\right) F_k'(0)^2.$$

Summing this inequality over k yields

$$F(\mathbf{x}_k) \leq F(\mathbf{x}_0) - \left(\frac{\delta}{2}\right) \sum_{i=0}^{k-1} F_i'(0)^2.$$

Since the feasible points for (1) lie on the unit sphere, the objective function value is bounded from below. Hence, we have

$$\lim_{k \rightarrow \infty} F'_k(0) = 0. \quad (41)$$

By the definition of \mathbf{z}_k and the fact that $s_k \in [0, -\rho F'_k(0)]$ where $F'_k(0)$ approaches 0, we also conclude that

$$\|\mathbf{x}_{k+1} - \mathbf{y}_k\| \leq c|F'_k(0)| \quad (42)$$

where c is a constant which is independent of k .

Step 2. The change in the multiplier μ .

By the orthogonality between \mathbf{y}_{k1} and \mathbf{x}_{k2} (forward), between \mathbf{y}_{k1} and \mathbf{y}_{k2} (forward and reverse), and between \mathbf{x}_{k1} and \mathbf{y}_{k2} (reverse), the first-order optimality conditions (9) for the subproblems (2) yield:

$$\text{forward} \begin{cases} \lambda_{k1} = \mathbf{y}_{k1}^\top \mathbf{H}_1 \mathbf{y}_{k1} \\ \mu_{k1} = \mathbf{x}_{k2}^\top \mathbf{H}_1 \mathbf{y}_{k1} \\ \lambda_{k2} = \mathbf{y}_{k2}^\top \mathbf{H}_2 \mathbf{y}_{k2} \\ \mu_{k2} = \mathbf{y}_{k1}^\top \mathbf{H}_2 \mathbf{y}_{k2} \end{cases} \quad \text{reverse} \begin{cases} \lambda_{k1} = \mathbf{y}_{k1}^\top \mathbf{H}_1 \mathbf{y}_{k1} \\ \mu_{k1} = \mathbf{y}_{k2}^\top \mathbf{H}_1 \mathbf{y}_{k1} \\ \lambda_{k2} = \mathbf{y}_{k2}^\top \mathbf{H}_2 \mathbf{y}_{k2} \\ \mu_{k2} = \mathbf{x}_{k1}^\top \mathbf{H}_2 \mathbf{y}_{k2} \end{cases} \quad (43)$$

By our sign convention (7), the multipliers μ_{kj} are nonnegative.

By the definition of $F_k(s)$, we have

$$F'_k(0) = \pm 2(\mathbf{y}_{k1}^\top \mathbf{H}_1 \mathbf{y}_{k2} - \mathbf{y}_{k2}^\top \mathbf{H}_2 \mathbf{y}_{k1}). \quad (44)$$

We multiply the first equation in (9) by \mathbf{y}_{k2}^\top to obtain $\mathbf{y}_{k1}^\top \mathbf{H}_1 \mathbf{y}_{k2} = \mu_{k1} \mathbf{y}_{k2}^\top \mathbf{x}_{k2}$. We multiply the second equation by \mathbf{y}_{k1}^\top to obtain $\mathbf{y}_{k1}^\top \mathbf{H}_2 \mathbf{y}_{k2} = \mu_{k2}$. Hence, in the forward mode, it follows from (44) that

$$\begin{aligned} \mu_{k2} &= \mathbf{y}_{k1}^\top \mathbf{H}_2 \mathbf{y}_{k2} \\ &= \mathbf{y}_{k1}^\top \mathbf{H}_1 \mathbf{y}_{k2} \mp F'_k(0)/2 \\ &= \mu_{k1} \mathbf{y}_{k2}^\top \mathbf{x}_{k2} \mp F'_k(0)/2, \end{aligned} \quad (45)$$

which implies that

$$\mu_{k2} \leq |\mu_{k1} \mathbf{y}_{k2}^\top \mathbf{x}_{k2}| + |F'_k(0)|/2 \leq \mu_{k1} + |F'_k(0)|/2 \quad (46)$$

since \mathbf{y}_{k2} and \mathbf{x}_{k2} are unit vectors. In a similar fashion, for the reverse mode at iteration $k+1$, we have

$$\mu_{k+1,1} \leq \mu_{k+1,2} + |F'_{k+1}(0)|/2. \quad (47)$$

If iteration k corresponds to the forward mode, then the multiplier μ_{k1} corresponds to $\mathbf{a} = \mathbf{x}_{k2}$ and $\mathbf{H} = \mathbf{H}_1$ in (10). The multiplier $\mu_{k-1,1}$ corresponds to $\mathbf{a} = \mathbf{y}_{k-1,2}$ and $\mathbf{H} = \mathbf{H}_1$ in (10). By (42) $\|\mathbf{x}_{k2} - \mathbf{y}_{k-1,2}\| \leq c|F'_{k-1}(0)|$. We apply Lemma 1 and (41). For any (small) $\eta > 0$, we have

$$|\mu_{k1} - \mu_{k-1,1}| \leq \eta \quad (48)$$

when k is sufficiently large, which implies that

$$\mu_{k1} \leq \mu_{k-1,1} + \eta. \quad (49)$$

The analogous result for the reverse mode is

$$\mu_{k+1,2} \leq \mu_{k2} + \eta \quad (50)$$

for k sufficiently large.

Combining (46)–(50), it follows that when k is large enough that $|F'_j(0)| \leq \eta$ for all $j \geq k$, we have

$$\mu_{k-1,1} \succeq \mu_{k1} \succeq \mu_{k2} \succeq \mu_{k+1,2} \succeq \mu_{k+1,1} \quad (51)$$

where the notation $\mu_{k1} \succeq \mu_{k2}$ means that $\mu_{k2} \leq \mu_{k1} + \eta$. Hence, in each iteration, the μ multiplier either decreases or makes an increase which is bounded by η . By (43), the multipliers are bounded by the largest absolute eigenvalues of \mathbf{H}_1 and \mathbf{H}_2 .

Step 3. The case $\liminf \mu_{k1} = 0$.

When $\liminf \mu_{k1} = 0$, there exists a subsequence of the iterates with the property that μ_{k1} tends to 0. By (51) and the fact that η can be taken arbitrarily small, we conclude that the corresponding subsequence of the multipliers μ_{k2} also approaches 0. Since \mathbf{y}_k lies in a compact set, we can extract subsequences converging to a limit that we denote by \mathbf{y} . By (43), the corresponding subsequence of multipliers λ_{k1} and λ_{k2} also approach limits denoted λ_1 and λ_2 . By (9), we have

$$\begin{bmatrix} \mathbf{H}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_2 \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 \mathbf{y}_1 \\ \lambda_2 \mathbf{y}_2 \end{bmatrix}.$$

Since \mathbf{y}_1 and \mathbf{y}_2 are orthogonal unit vectors, we conclude that \mathbf{y} is a stationary point for (1) corresponding to the multiplier $\mu = 0$.

Step 4. The case $\mu = \liminf \mu_{k1} > 0$.

We extract a subsequence, denoted ν_{j1} , of the multiplier sequence μ_{k1} which converges to μ :

$$\lim_{j \rightarrow \infty} \nu_{j1} = \mu.$$

Given any $\eta > 0$, choose K large enough that (51) holds for all $k \geq K$. Also, choose K larger, if necessary, so that

$$|\mu - \nu_{j1}| \leq \eta \text{ for all } j \geq K. \quad (52)$$

Hence, for any $j \geq K$, we have

$$\mu - \eta \leq \nu_{j1} \leq \mu + \eta. \quad (53)$$

By (51), (53), and the fact that the ν_{j1} form a subsequence of the μ_{k1} , it follows that

$$\nu_{j2} \leq \nu_{j1} + \eta \leq \mu + 2\eta.$$

Let k denote an index in the original sequence with the property that $\mu_{k2} = \nu_{j2}$. By (51), we have

$$\nu_{j2} = \mu_{k2} \geq \mu_{k+1,2} - \eta \geq \mu_{k+1,1} - 2\eta.$$

By (48) and (52), it follows that

$$\mu_{k+1,1} \geq \mu_{k1} - \eta = \nu_{j1} - \eta \geq \mu - 2\eta.$$

Combining these inequalities gives

$$\mu - 4\eta \leq \nu_{j2} \leq \mu + 2\eta \quad \text{and} \quad \mu - \eta \leq \nu_{j1} \leq \mu + \eta.$$

Since η is arbitrary, it follows that ν_{j1} and ν_{j2} approach the same limit μ .

Again, by extracting subsequences, there exist limits \mathbf{y}_1 , \mathbf{y}_2 , λ_1 , and λ_2 such that

$$\begin{bmatrix} \mathbf{H}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_2 \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 \mathbf{y}_1 + \mu \mathbf{x}_2 \\ \mu \mathbf{y}_1 + \lambda_2 \mathbf{y}_2 \end{bmatrix}. \quad (54)$$

By (45), we have

$$\mu = \mu \mathbf{y}_2^\top \mathbf{x}_2,$$

where $\mu > 0$. Since \mathbf{y}_2 and \mathbf{x}_2 are unit vectors, we deduce that $\mathbf{x}_2 = \mathbf{y}_2$. When \mathbf{x}_2 is replaced by \mathbf{y}_2 in (54), we see that \mathbf{y} is a stationary point.

Remark 4. In the special case $\mathbf{H}_1 = \mathbf{H}_2 = \mathbf{H}$, both the analysis and the algorithm simplify. As noted earlier, $F'_k(0) = 0$ in this case so the global step is skipped. Moreover, the monotonicity property (51) holds without the reverse iteration. Hence, the decomposition algorithm can simply employ forward steps, for which the associated first-order optimality condition is

$$\begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{k1} \\ \mathbf{y}_{k2} \end{bmatrix} = \begin{bmatrix} \lambda_{k1} \mathbf{y}_{k1} + \mu_{k1} \mathbf{y}_{k-1,2} \\ \mu_{k2} \mathbf{y}_{k1} + \lambda_{k2} \mathbf{y}_{k2} \end{bmatrix}. \quad (55)$$

Multiplying the first equation by \mathbf{y}_{k2}^\top gives

$$\begin{aligned} \mu_{k1} \mathbf{y}_{k2}^\top \mathbf{y}_{k-1,2} &= \mathbf{y}_{k2}^\top \mathbf{H} \mathbf{y}_{k1} \\ &= \mathbf{y}_{k1}^\top (\mu_{k2} \mathbf{y}_{k1} + \lambda_{k2} \mathbf{y}_{k2}) = \mu_{k2}. \end{aligned}$$

Since \mathbf{y}_{k2} and $\mathbf{y}_{k-1,2}$ are unit vectors, this shows that $\mu_{k2} \leq \mu_{k1}$. Multiplying the first equation in (55) by $\mathbf{y}_{k-1,2}$ gives

$$\begin{aligned}\mu_{k1} &= \mathbf{y}_{k-1,2}^\top \mathbf{H} \mathbf{y}_{k1} \\ &= \mathbf{y}_{k1}^\top (\mu_{k-1,2} \mathbf{y}_{k-1,1} + \lambda_{k-1,2} \mathbf{y}_{k-1,2}) \\ &= \mu_{k-1,2} \mathbf{y}_{k1}^\top \mathbf{y}_{k-1,1}.\end{aligned}$$

Since \mathbf{y}_{k1} and $\mathbf{y}_{k-1,1}$ are unit vectors, we deduce that $\mu_{k1} \leq \mu_{k-1,2}$. Hence, (51) holds with \succeq replaced by \geq .

5 Numerical experiments

A series of numerical experiments were performed to investigate the convergence rate of the decomposition algorithm and to explore the connections between the theoretical analysis and the practical convergence. The experiments we describe were performed using Scilab (www.scilab.org). The solution of each local step (10) was obtained by computing an eigenvector associated with the smallest non-zero eigenvalue of the matrix $\mathbf{P}\mathbf{H}\mathbf{P}$, where $\mathbf{P} = \mathbf{I} - \mathbf{a}\mathbf{a}^\top$ is the projection into the subspace perpendicular to \mathbf{a} . The global step was implemented using the Scilab routine “optim” with default parameter values.

Recall that in our theoretical analysis, the stepsize was restricted to an interval $[0, -\rho F'_k(0)]$, for some fixed $\rho > 0$, to ensure that the iterates approach each other in the limit. However, in all our numerical experiments, we found that there was no need to restrict the stepsize to obtain convergence. Hence, it appears that the restriction on the stepsize in (6) is an artifact of the analysis presented in this paper.

In (51) we show that the multipliers associated with the local steps in the decomposition algorithm almost decay monotonically. In Remark 4, we show that the decay is monotone when $\mathbf{H}_1 = \mathbf{H}_2$. Numerically, we found that when $\mathbf{H}_1 \neq \mathbf{H}_2$, the convergence of the multipliers may not be monotone. An illustration is given in Figure 2 where we randomly generate two 100 by 100 diagonal matrices \mathbf{H}_1 and \mathbf{H}_2 with entries between -1 and $+1$, and we plot the multipliers as a function of the iteration number. Due to the initial growth in the multipliers during the first 300 iterations, the convergence is not monotone and the inequalities \succeq in (51) can not be replaced by \geq in general.

In our experiments, the convergence speed when $\mathbf{H}_1 = \mathbf{H}_2$ was closely related to the distribution of the smallest 3 eigenvalues of the matrix, independent of the matrix dimension. To illustrate the typical convergence, we consider the 3 by 3 diagonal matrix

$$\mathbf{H}_1 = \mathbf{H}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 + \alpha \end{bmatrix}, \quad (56)$$

where $\alpha > 0$ is a parameter which we vary to explore the convergence. The starting guess is

$$\mathbf{x}_{02}^\top = \frac{1}{\sqrt{3}} [1 \ 1 \ 1]^\top. \quad (57)$$

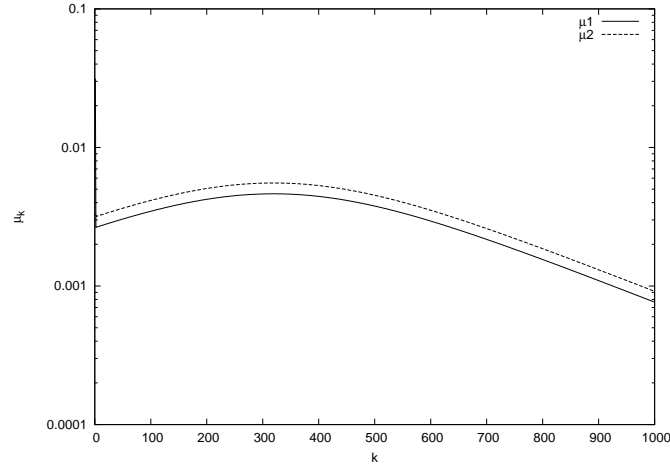


Figure 2: Convergence of the multipliers, random 100 by 100 diagonal matrices \mathbf{H}_1 and \mathbf{H}_2 , $\rho = 1$.

Since $\mathbf{H}_1 = \mathbf{H}_2$, the reverse step can be skipped, and decomposition algorithm operates in “forward mode” without a global step. The components of the iterates are always nonzero, and the iterates given by (26) can be expressed in the form

$$x_{k1i} = \frac{\mu_{k1} x_{k-1,2i}}{\epsilon_i - \lambda_{k1}}, \quad x_{k2i} = \frac{\mu_{k2} x_{k,1i}}{\epsilon_i - \lambda_{k2}}. \quad (58)$$

The iterates converge to a pair $(\mathbf{x}_1, \mathbf{x}_2)$ of the form

$$\mathbf{x}_1^\top = [a \ b \ 0]^\top \quad \text{and} \quad \mathbf{x}_2^\top = [-b \ a \ 0]^\top$$

with $a^2 + b^2 = 1$, which is a valid solution to (1) according to Corollary 2 and Remark 3.

Since the third component of the solution vanishes, we will study how quickly the third component approaches 0. By (58), the ratio between the second and third components can be expressed as

$$\frac{x_{k13}}{x_{k12}} = \frac{x_{k-1,23}}{x_{k-1,22}} \left(\frac{\epsilon_2 - \lambda_{1k}}{\epsilon_3 - \lambda_{1k}} \right) = \frac{x_{k-1,13}}{x_{k-1,12}} \left(\frac{\epsilon_2 - \lambda_{2,k-1}}{\epsilon_3 - \lambda_{2,k-1}} \right) \left(\frac{\epsilon_2 - \lambda_{1k}}{\epsilon_3 - \lambda_{1k}} \right).$$

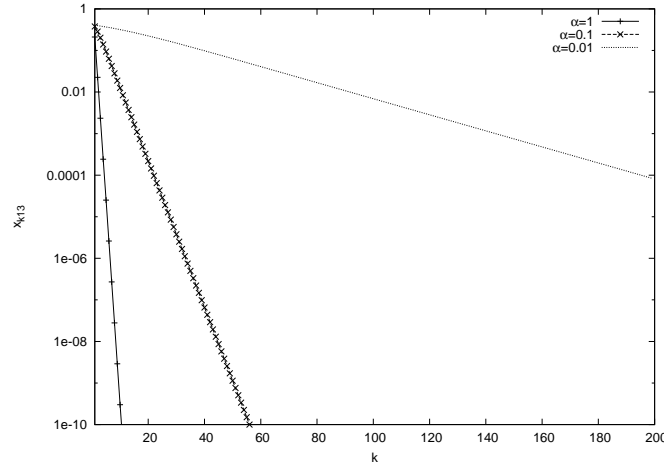


Figure 3: Convergence of x_{k13} for matrices (56) and starting point (57).

Since $\epsilon_2 > \epsilon_3$ for the matrix (56), the rational function $(\epsilon_2 - \lambda)/(\epsilon_3 - \lambda)$, $\lambda \in [1, 2]$, attains its maximum value $1/(1 + \alpha)$ at $\lambda = 1$. Hence, by induction, we have

$$\frac{x_{k13}}{x_{k12}} \leq \left(\frac{1}{1 + \alpha} \right)^{2k}. \quad (59)$$

Thus as α approaches 0, the bound on the rate at which the third component approaches 0, relative to the second component, grows, and the convergence could be much slower. And as α becomes large, the bound decreases and the convergence rate increases. In summary, the convergence speed seems to depend on the ratio of the gap between ϵ_2 and ϵ_3 relative to the gap between ϵ_1 and ϵ_2 . As the ratio approaches 0, the convergence could be slower, as seen in (59).

In Figure 3 we show the convergence of \mathbf{x}_{k13} as a function of the iteration number k for various choices of α . Notice that as α approaches 0, the optimization problem (1) becomes more poorly conditioned since the points

$$\mathbf{x}_1^\top = [1 \ 0 \ 0] \quad \text{and} \quad \mathbf{x}_2^\top = [0 \ 0 \ 1]$$

are feasible with objective function value $3 + \alpha \approx 3$, when $\alpha \approx 0$. Thus there are feasible points which are separated from the optimal solution, but with nearly the same cost as the optimal solution. In the extreme case where $\alpha = 0$, the algorithm finds the global minimum

of the objective function at the first iteration. The two vectors \mathbf{x}_1 and \mathbf{x}_2 have their three components different from zero.

We now give an example where the decomposition algorithm does not converge to the global minimum when the starting guess is sufficiently poor. In Remark 1, we point out that the decomposition algorithm is convergent for any starting guess when $\mathbf{H}_1 = \mathbf{H}_2$ and $\epsilon_3 - \epsilon_2 \geq \epsilon_2 - \epsilon_1$. Suppose that $\epsilon_3 - \epsilon_2 < \epsilon_2 - \epsilon_1$, and the starting guess is

$$\mathbf{x}_{02}^\top = \frac{1}{\sqrt{2}}[1 \ 0 \ 1 \ 0 \ 0 \ \dots]^\top.$$

According to case 3c of Section 3,

$$\mathbf{x}_{11}^\top = \frac{1}{\sqrt{2}}[-1 \ 0 \ 1 \ 0 \ 0 \ \dots]^\top \quad \text{and} \quad \mathbf{x}_{12} = \mathbf{x}_{02}.$$

Hence, the algorithm converges after one iteration to the stationary point

$$\mathbf{x}_1^\top = \frac{1}{\sqrt{2}}[-1 \ 0 \ 1 \ 0 \ 0 \ \dots]^\top \quad \text{and} \quad \mathbf{x}_2 = \frac{1}{\sqrt{2}}[1 \ 0 \ 1 \ 0 \ 0 \ \dots]^\top.$$

This starting guess, however, is exceptional. If the second component of \mathbf{x}_{02} is changed to any nonzero value α , then the iterates quickly converge to the global minimum. For example, if $\alpha = 10^{-14}$ and

$$\mathbf{H}_1 = \mathbf{H}_2 = \text{diag}[-0.9, -0.5, -0.4, -0.3, -0.2, -0.1, 0.0, +0.1, +0.2, +0.3],$$

then the error is reduced to 10^{-10} within 44 iterations.

For the case $\mathbf{H}_1 \neq \mathbf{H}_2$, the decomposition algorithm may converge to a stationary point which is not the global minimum when the starting guess is degenerate and the degenerate iterates are chosen in a very special way. As an example, suppose that \mathbf{H}_1 and \mathbf{H}_2 are diagonal matrices, with diagonal elements arranged in increasing order, and that all the components of the starting point \mathbf{x}_{02} are nonzero except for component 2 which is 0. The nonzero components of \mathbf{x}_{02} are chosen to make it degenerate for \mathbf{H}_1 . The initial iterate \mathbf{y}_{11} is described by case 3d of Section 3. There are an infinite number of solutions to the local subproblem. We choose the solution for which the second component is zero (the remaining components are nonzero). Take ϵ_{11} close enough to ϵ_{12} to ensure that y_{111} is near 1 in magnitude. In this case, $g(\epsilon_{22}) < 0$ in the second local step. By case 3b of Section 3, all the components of the iterate \mathbf{y}_{12} are zero except for the second component which is 1. Since $F_1'(0) = 0$, the global step has no effect; we have $\mathbf{x}_{21} = \mathbf{y}_{11}$ and $\mathbf{x}_{22} = \mathbf{y}_{12}$. Thereafter, \mathbf{x}_{k1} is the first column of the identity and \mathbf{x}_{k2} is the second column of the identity. If \mathbf{H}_1 and \mathbf{H}_2 are chosen so that

$$\epsilon_{11} + \epsilon_{22} > \epsilon_{21} + \epsilon_{12},$$

then the iteration has reached a stationary point which is not the global minimum. In contrast, with any perturbation in the second component of \mathbf{x}_{02} , we obtain convergence to the global minimum.

If the eigenvectors corresponding to the smallest eigenvalues of \mathbf{H}_1 and \mathbf{H}_2 are orthogonal, then these orthogonal eigenvectors are the solution of (1). By randomly choosing the remaining orthogonal eigenvectors of \mathbf{H}_1 and \mathbf{H}_2 , we obtain noncommuting matrices for which the solution of (1) is known. As a specific example, we took $\mathbf{H}_i = \mathbf{Q}_i \mathbf{D}_i \mathbf{Q}_i^\top$ where \mathbf{D}_i is a diagonal matrix with diagonal elements chosen randomly on $[-1, 1]$, and \mathbf{Q}_i , $i = 1$ or 2 , is an orthogonal matrix of the form

$$\mathbf{Q}_1 = \begin{bmatrix} \mathbf{e}_1 & \mathbf{U}_1 \end{bmatrix} \quad \text{and} \quad \mathbf{Q}_2 = \begin{bmatrix} \mathbf{e}_2 & \mathbf{U}_2 \end{bmatrix}.$$

Here \mathbf{e}_i denotes the i -th column of the identity matrix, and \mathbf{U}_i is an n by $n - 1$ matrix with randomly chosen entries such that \mathbf{Q}_i is orthogonal, $i = 1$ or 2 . For all starting points, we observed convergence to the global minimum. Convergence to local, non global, minima has also been observed in the case where the matrices \mathbf{H}_1 and \mathbf{H}_2 do not commute, but have the same eigenvector associated with the smallest eigenvalue.

6 Conclusions

A decomposition algorithm is developed for a quadratic programming problem with sphere and orthogonality constraints. The algorithm consists of local steps, both forward and reverse, and a global step where we minimize over a subspace. Without the global step, any limit $(\mathbf{y}_1, \mathbf{y}_2)$ of the local step satisfies

$$\begin{bmatrix} \mathbf{H}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_2 \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 \mathbf{I} & \mu_1 \mathbf{I} \\ \mu_2 \mathbf{I} & \lambda_2 \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}.$$

This differs from the first-order optimality conditions (8) associated with the original optimization problem (1) because μ_1 may not equal μ_2 . If the local step is followed by the global step, then according to the analysis of Section 4, $F'_k(0)$ tends to zero (see (41)), which implies that (see (44)) $\mathbf{y}_1^\top \mathbf{H}_1 \mathbf{y}_2 = \mathbf{y}_2^\top \mathbf{H}_2 \mathbf{y}_1$. Since $\mu_1 = \mathbf{y}_1^\top \mathbf{H}_1 \mathbf{y}_2$ and $\mu_2 = \mathbf{y}_2^\top \mathbf{H}_2 \mathbf{y}_1$, the global step ensures that $\mu_1 = \mu_2$. Consequently, the first-order optimality condition for (1) is satisfied.

The complexity of the analysis is connected with the proof of convergence. To show that the iterate \mathbf{x}_{k2} converges to the same limit as \mathbf{y}_{k2} , we studied the properties of multipliers for the subproblem (10). We showed that the multiplier λ for the sphere constraint lies between ϵ_1 and ϵ_2 , the two smallest eigenvalues of \mathbf{H} ; moreover, λ depends Lipschitz continuously on \mathbf{a} . In contrast, the multiplier μ associated with the orthogonality constraint is continuous at *nondegenerate* choices for \mathbf{a} .

A less technical explanation for the performance of the decomposition algorithm is that the local steps steer the iterates into a low dimensional subspace associated with the eigenvectors of the smallest eigenvalues of \mathbf{H}_1 or \mathbf{H}_2 , while the global step finds the best point in this subspace.

Acknowledgments

Constructive comments by the reviewers are gratefully acknowledged. Their comments significantly improved the paper.

References

- [1] M. BARRAULT, *Développement de méthodes rapides pour le calcul de structures électroniques*, PhD thesis, Ecole Nationale des Ponts et Chaussées, 2005.
- [2] M. BARRAULT, E. CANCES, W. W. HAGER, AND C. LE BRIS, *Multilevel domain decomposition for electronic structure calculations*, J. Comput. Phys, 222 (2007), pp. 86–109.
- [3] G. BENCTEUX, M. BARRAULT, E. CANCES, C. LE BRIS, AND W. W. HAGER, *Domain decomposition and electronic structure computations: a promising approach*, in Numerical analysis and scientific computing for PDEs and their challenging applications, R. Glowinski and P. Neittaanmäki, eds, Springer, 2008, pp. 147–164.
- [4] J. E. DENNIS AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
- [5] W. W. HAGER, *Applied Numerical Linear Algebra*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
- [6] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, Springer, New York, 1999.
- [7] D. C. SORENSEN, *Newton's method with a model trust region modification*, SIAM J. Numer. Anal., 19 (1982), pp. 409–426.
- [8] G. STRANG, *Linear Algebra and Its Applications*, Thomson, Belmont, CA, 4th ed., 2006.



Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)

Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
<http://www.inria.fr>
ISSN 0249-6399