A VARIABLE METRIC TECHNIQUE FOR THE SOLUTION OF AFFINELY PARAMETRIZED NONDIFFERENTIABLE OPTIMAL DESIGN PROBLEMS

by

E. Polak and E. J. Wiest

Memorandum No. UCB/ERL M88/42

10 June 1988
(Revised May 22, 1989)

A VARIABLE METRIC TECHNIQUE FOR THE SOLUTION OF AFFINELY PARAMETRIZED NONDIFFERENTIABLE OPTIMAL DESIGN PROBLEMS

by

E. Polak and E. J. Wiest

Memorandum No. UCB/ERL M88/42

10 June 1988
(Revised May 22, 1989)

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

TITLE PAGE

A VARIABLE METRIC TECHNIQUE FOR THE SOLUTION OF AFFINELY PARAMETRIZED NONDIFFERENTIABLE OPTIMAL DESIGN PROBLEMS

by

E. Polak and E. J. Wiest

Memorandum No. UCB/ERL M88/42

10 June 1988
(Revised May 22, 1989)

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

A Variable Metric Technique for the Solution of Affinely Parametrized Nondifferentiable Optimal Design Problems[1,2]

E. Polak[3] and E. J. Wiest[4]

Abstract. The composite functions which appear in various optimal feedback system design problems, as well as in open loop optimal control problems, can lead to severely ill-conditioned minimax problems. This ill-conditioning can cause first-order minimax algorithms to converge very slowly. We propose a variable metric technique which substantially mitigates this ill-conditioning. The technique does not require the evaluation of second derivatives and can be used to speed the convergence of any first-order minimax algorithm which produces estimates of the optimal multipliers. Numerical experiments are presented which show that the variable metric technique increases the speed of two algorithms.

Key Words. Minimax, nonsmooth optimization, rate of convergence, variable metric, scaling, conditioning.

---

[1]The research reported herein was sponsored in part by the National Science Foundation grant ECS-8713334, the Air Force Office of Scientific Research contract AFOSR-86-0116 and the State of California MICRO Program grant 532410-19900.

[2]The authors would like to thank an anonymous referee for his careful scrutiny of the manuscript and of the numerical experiments.

[3]Professor, Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, CA 94720, U.S.A.

[4]Graduate student, Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, CA 94720, U.S.A.

1. Introduction

The term *variable metric method* is commonly used to describe a number of algorithms, such as those discussed in Refs. 1 and 2, which emulate the behavior of the Newton method. The term can be applied to any optimization algorithm which uses a sequence of linear transformations of the variables to convert the original optimization problem into a sequence of equivalent problems, to each of which it applies one iteration of a "standard" method and uses a transformed result as a starting point for the iteration on the next problem. Variable metric methods are effective when there is a linear transformation which transforms an optimization problem into a better conditioned form. Since the desired transformation is not known a priori, an approximating sequence of transformations is constructed as the computation progresses.

In the past, variable metric techniques were used as a means of improving the conditioning of an optimization problem with respect with respect to a particular algorithm. For example, the Armijo-Newton method (Ref. 3) can be viewed as a combination of a sequence of transformations with the Armijo gradient method. Consider the problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{1a}$$

where $f: \mathbb{R}^n \to \mathbb{R}$ is strictly convex and twice Lipschitz continuously differentiable. Given an estimate x_i of the solution $\hat{x}$, at iteration i, the Armijo-Newton method uses the transformation $x = F(x_i)^{-1/2} y$ to construct the equivalent problem

$$\min_{y \in \mathbb{R}^n} f(F(x_i)^{-1/2} y), \tag{1b}$$

to which it applies one iteration of the Armijo gradient method. Thus, (i) it computes the search direction[1] $d_i = -\nabla_y f(F(x_i)^{-1/2} y_i)$, in the new coordinates, (ii) then it transforms this search direction back to the original space by the formula $h_i = F(x_i)^{-1/2} d_i = -F(x_i)^{-1} \nabla f(x_i)$, and (iii) computes the step size, which is unity near the solution, using a suitably transformed Armijo step size rule (see Ref. 3) completing the construction of x_{i+1}. As is well known, the result is a quadratically converging algorithm. Similarly, it

[1] It is interesting to observe that the Newton search direction h_i is the solution of the problem

$$\min_{h \in \mathbb{R}^n} \langle \nabla f(x_i), h \rangle + \tfrac{1}{2} \| h \|^2_{F(x_i)},$$

Theorem 2.1: (Refs. 13 and 9) *If* $\hat{x}$ is a solution to (3), then there exists a $\hat{\mu} \in \Sigma_p$ such that

$$\nabla_x l(\hat{x},\hat{\mu}) = \sum_{j \in \underline{p}} \hat{\mu}^j \nabla f^j(\hat{x}) = 0, \tag{5a}$$

$$\sum_{j \in \underline{p}} \hat{\mu}^j [f^j(\hat{x}) - \max_{j \in \underline{p}} f^j(\hat{x})] = 0. \tag{5b}$$

■

Now suppose that the functions $f^j(\cdot)$ are strictly convex. Then it follows from (5a) and the convexity of $l(\cdot,\hat{\mu})$ that, if $\hat{x}$ is a solution to (3), then it must also be a minimizer of the function $\phi_{\hat{\mu}}(\cdot) \triangleq l(\cdot,\hat{\mu})$. Now, as we have seen in the Section 1, the conditioning of the problem $\min_{x \in \mathbb{R}^n} \phi_{\hat{\mu}}(x)$ can be improved by a linear domain transformation based on the Hessian of $\phi(\cdot)$. Our method originated in the conjecture that this transformation would also improve the conditioning of problem (3). Han used this matrix as the basis for a variable metric method for problem (3) in Ref. 10.

We now return to problem (2). The Lagrangian for problem (2) is given by $l(x,\mu) = \sum_{j \in \underline{p}} \mu^j g^j(A_j x)$. Hence its Hessian with respect to x is given by

$$L(x,\mu) = \sum_{j \in \underline{p}} \mu^j A_j^T G^j(A_j x) A_j\,, \tag{6}$$

where $G^j(\cdot)$ denotes the second derivative matrix of $g^j(\cdot)$. In many engineering optimization problems, such as those mentioned in the introduction, the functions $g^j(\cdot)$ do not contribute to the ill-conditioning of the matrix $L(\hat{x},\hat{\mu})$, at a solution. Furthermore, their Hessians may be very difficult to compute. Hence we propose to replace the Hessian matrices $G^j(A_j x_i)$ in (6) by $l_j \times l_j$ identity matrices. Thus, for any $\mu \in \Sigma_p$, let

$$R(\mu) \triangleq \sum_{j \in \underline{p}} \mu^j A_j^T A_j\,. \tag{7}$$

We will show that a sequential transformation method based on the matrix $R(\mu)$ can compensate for the ill-conditioning introduced by the matrices A_j.

To ensure that a sequential domain transformation method does not destroy the convergence properties of the algorithm which it uses, there must be both an upper bound and a strictly positive lower bound on the eigenvalues of the domain transformation matrices. However, the minimum positive

In this paper, we present a sequential linear transformation technique[2], which is intended to mitigate the ill-conditioning caused by the matrices A_j. Our technique was inspired by the observation that when all the functions $g^j(\cdot)$ in (2) are convex, any solution $\hat{x}$ to (2) is an unconstrained minimizer of the corresponding Lagrangian, $l(x,\hat{\mu}) \triangleq \sum_{j \in \underline{p}} \hat{\mu}^j g^j(A_j x)$, where the $\hat{\mu}^j$ are optimal multipliers. Although the Hessian of this Lagrangian is usually singular, a restriction to a suitable subspace can be used to recondition the problem $\min_{x \in \mathbb{R}^n} l(x,\hat{\mu})$. Since, in many engineering applications, only the matrices A_j cause ill-conditioning and since second order derivatives of the $g^j(\cdot)$ can be very costly to compute, we replace the Hessians of the $g^j(\cdot)$ by identity matrices and use linear transformations to improve the conditioning of approximations to the matrix $\sum_{j \in \underline{p}} \hat{\mu}_j A_j^T A_j$. The resulting sequential linear transformation technique can be used in conjunction with any first-order minimax algorithm which produces estimates of the optimal multipliers. Our variable metric technique is developed in Section 2. In Sections 3 and 4, we present theoretical results which show that our variable metric technique can improve the speed of convergence of the Pshenichnyi method (Ref. 11). In Section 5, we present numerical experiments which show that our variable metric technique reconditions problems with respect to both the Pshenichnyi method and a new interior point method (Ref. 12).

2. Development of the Variable Metric Technique

We begin by providing a heuristic rationale for our method. Consider the *general* minimax problem,

$$\min_{x \in \mathbb{R}^n} \max_{j \in \underline{p}} f^j(x), \tag{3}$$

where the functions $f^j(\cdot)$ are twice continuously differentiable. We will denote the standard unit simplex in $\mathbb{R}^n$ by $\Sigma_p \triangleq \{ \mu \in \mathbb{R}^p \mid \sum_{j \in \underline{p}} \mu^j = 1 , \mu \geq 0 \}$, and the second derivative matrix of $f^j(\cdot)$ by $F^j(\cdot)$. We can associate with problem (3) the Lagrangian $l{:}\mathbb{R}^n \times \Sigma_p \to \mathbb{R}$, defined by

$$l(x,\mu) = \sum_{j \in \underline{p}} \mu^j f^j(x) . \tag{4}$$

We recall the following result.

[2] Our technique is related to one used implicitly by Han in Ref. 10.

Theorem 2.1: (Refs. 13 and 9) *If* $\hat{x}$ is a solution to (3), then there exists a $\hat{\mu} \in \Sigma_p$ such that

$$\nabla_x l(\hat{x},\hat{\mu}) = \sum_{j \in \underline{p}} \hat{\mu}^j \nabla f^j(\hat{x}) = 0, \tag{5a}$$

$$\sum_{j \in \underline{p}} \hat{\mu}^j [f^j(\hat{x}) - \max_{j \in \underline{p}} f^j(\hat{x})] = 0 . \tag{5b}$$

■

Now suppose that the functions $f^j(\cdot)$ are strictly convex. Then it follows from (5a) and the convexity of $l(\cdot,\hat{\mu})$ that, if $\hat{x}$ is a solution to (3), then it must also be a minimizer of the function $\phi_{\hat{\mu}}(\cdot) \triangleq l(\cdot,\hat{\mu})$. Now, as we have seen in the Section 1, the conditioning of the problem $\min_{x \in \mathbb{R}^n} \phi_{\hat{\mu}}(x)$ can be improved by a linear domain transformation based on the Hessian of $\phi(\cdot)$. Our method originated in the conjecture that this transformation would also improve the conditioning of problem (3). Han used this matrix as the basis for a variable metric method for problem (3) in Ref. 10.

We now return to problem (2). The Lagrangian for problem (2) is given by $l(x,\mu) = \sum_{j \in \underline{p}} \mu^j g^j(A_j x)$. Hence its Hessian with respect to x is given by

$$L(x,\mu) = \sum_{j \in \underline{p}} \mu^j A_j^T G^j(A_j x) A_j , \tag{6}$$

where $G^j(\cdot)$ denotes the second derivative matrix of $g^j(\cdot)$. In many engineering optimization problems, such as those mentioned in the introduction, the functions $g^j(\cdot)$ do not contribute to the ill-conditioning of the matrix $L(\hat{x},\hat{\mu})$, at a solution. Furthermore, their Hessians may be very difficult to compute. Hence we propose to replace the Hessian matrices $G^j(A_j x_i)$ in (6) by $l_j \times l_j$ identity matrices. Thus, for any $\mu \in \Sigma_p$, let

$$R(\mu) \triangleq \sum_{j \in \underline{p}} \mu^j A_j^T A_j . \tag{7}$$

We will show that a sequential transformation method based on the matrix $R(\mu)$ can compensate for the ill-conditioning introduced by the matrices A_j.

To ensure that a sequential domain transformation method does not destroy the convergence properties of the algorithm which it uses, there must be both an upper bound and a strictly positive lower bound on the eigenvalues of the domain transformation matrices. However, the minimum positive eigenvalue of $R(\mu_i)$ may decrease to zero as $\mu_i \to \hat{\mu}$. Hence, we propose to modify $R(\mu_i)$ by augmenting the small eigenvalues in its spectral decomposition, as follows. For any $\mu \in \Sigma_p$, let $\lambda_1(\mu) \geq \lambda_2(\mu) \geq \ldots \geq \lambda_n(\mu)$ be the eigenvalues of $R(\mu)$. Let U_μ be any real unitary matrix such that $R(\mu) = U_\mu \operatorname{diag}(\lambda_1(\mu), \ldots, \lambda_n(\mu)) U_\mu^T$, and let $\tilde{\lambda}_j(\mu) \triangleq \max\{\lambda_j(\mu), \varepsilon\}$, where $\varepsilon > 0$ is a small fixed number. Then we define

$$Q(\mu) \triangleq U_\mu \operatorname{diag}(\tilde{\lambda}_1(\mu), \ldots, \tilde{\lambda}_n(\mu)) U_\mu^T. \tag{8}$$

Proposition 2.1: *The matrix-valued function $Q(\cdot)$ is well defined and continuous in μ.*

Proof: We begin by showing that $Q(\mu)$ is well defined even though the selection of the eigenvector matrix U_μ is not unique when $R(\mu)$ has multiple eigenvalues. Letting $\lambda_0 \triangleq \infty$, we define

$$D(\mu) \triangleq \{ j \in \underline{n} \mid \lambda_{j-1}(\mu) > \lambda_j(\mu) = \lambda_{j+1}(\mu) \cdots = \lambda_{j+m_j(\mu)-1}(\mu) > \lambda_{j+m_j(\mu)}(\mu) \}, \tag{9}$$

so that $\{ \lambda_j(\mu) \}_{j \in D(\mu)}$ is the set of distinct eigenvalues of $R(\mu)$, with multiplicities $m_j(\mu)$. Next, let u_j denote the j–*th* column of U_μ, $j \in \underline{n}$. Then,

$$R(\mu) = \sum_{j \in D(\mu)} \lambda_j(\mu) \left[\sum_{k \in \underline{m_j(\mu)}} u_{j+k-1} u_{j+k-1}^T \right] \tag{10}$$

is a spectral decomposition of $R(\mu)$. The matrix $\sum_{k \in \underline{m_j(\mu)}} u_{j+k-1} u_{j+k-1}^T$ represents a projection operator which projects onto the eigenspace corresponding to $\lambda_j(\mu)$, and hence it is independent of the selection of U_μ. Since

$$Q(\mu) \triangleq U_\mu \operatorname{diag}(\tilde{\lambda}_1(\mu), \ldots, \tilde{\lambda}_n(\mu)) U_\mu^T = \sum_{j \in D(\mu)} \tilde{\lambda}_j(\mu) \left[\sum_{k \in \underline{m_j(\mu)}} u_{j+k-1} u_{j+k-1}^T \right], \tag{11}$$

we see that it, too, is independent of the selection of U_μ.

Now, suppose that the sequence $\{\mu_i\}_{i=0}^{\infty} \subset \Sigma_p$ converges to some $\hat{\mu} \in \Sigma_p$ as $i \to \infty$. For each μ_i, let $U_i = [\, u_{1,i}, \ldots, u_{n,i} \,]$ be a unitary matrix of eigenvectors of $R(\mu_i)$, so that

$$Q(\mu_i) = \sum_{j \in D(\mu_i)} \tilde{\lambda}_j(\mu_i) \left[\sum_{k \in \underline{m_j(\mu_i)}} u_{j+k-1,i}\, u_{j+k-1,i}^T \right]. \tag{12}$$

The sequences $\{ Q(\mu_i) \}_{i=0}^{\infty}$ and $\{ U_i \}_{i=0}^{\infty}$ are bounded, since the eigenvalues are continuous and the matrices U_i are unitary. Therefore, there exists an infinite subset $K \subset \mathbb{N}$, and matrices $\overline{Q}$ and $\overline{U} = [\overline{u}_1, \ldots, \overline{u}_n]$, such that $Q(\mu_i) \xrightarrow{K} \overline{Q}$ and $U_i \to \overline{U}$ as $i \to \infty$. Because $U_i U_i^T = I$, for all $i \in \mathbb{N}$, $\overline{U}\,\overline{U}^T = I$. Since $[R(\mu_i) - \lambda_j(\mu_i)I]u_{j,i} = 0$, for $j \in \underline{n}$, and since the eigenvalues, $\lambda_j(\cdot)$, are continuous, $[R(\overline{\mu}) - \lambda_j(\overline{\mu})I]\overline{u}_j = 0$ for $j \in \underline{n}$. Thus, $\overline{U}$ is a unitary matrix of eigenvectors for $R(\overline{\mu})$. The matrix $Q(\mu_i)$ can also be written in the form

$$Q(\mu_i) = \sum_{j \in D(\overline{\mu})} \left[\sum_{k \in \underline{m_j(\overline{\mu})}} \tilde{\lambda}_{j+k-1}(\mu_i) u_{j+k-1,i}\, u_{j+k-1,i}^T \right] . \tag{13}$$

Taking limits in (13) as $i \to \infty$, $i \in K$, yields

$$\begin{aligned} \overline{Q} &= \sum_{j \in D(\overline{\mu})} \left[\sum_{k \in \underline{m_j(\overline{\mu})}} \tilde{\lambda}_{j+k-1}(\overline{\mu})\, \overline{u}_{j+k-1} \overline{u}_{j+k-1}^T \right] \\ &= \sum_{j \in D(\overline{\mu})} \tilde{\lambda}_j(\overline{\mu}) \left[\sum_{k \in \underline{m_j(\overline{\mu})}} \overline{u}_{j+k-1} \overline{u}_{j+k-1}^T \right] = Q(\overline{\mu}) . \end{aligned} \tag{14}$$

Since the sequence $\{ Q(\mu_i) \}_{i=0}^{\infty}$ is bounded and any accumulation point of this sequence equals $Q(\overline{\mu})$, it follows that $\lim_{i \to \infty} Q(\mu_i) = Q(\overline{\mu})$, and hence $Q(\cdot)$ is continuous. ■

We now provide an algorithm model which shows how to combine our variable metric technique with any one-step, first-order minimax algorithm which produces multiplier estimates. To simplify notation, we rewrite problem (2) as

$$\min_{x \in \mathbb{R}^n} \psi(x), \tag{15}$$

where

$$\psi(x) \triangleq \max_{j \in \underline{p}} g^j(A_j x) . \tag{16}$$

Now consider any first-order minimax algorithm for solving (15) which generates estimates of the optimal multipliers at each iteration. We can associate with the algorithm a ψ-dependent, set-valued iteration map $M_\psi : \mathbb{R}^n \to \mathbb{R}^n \times 2^{\Sigma_p}$ such that, if $\{ (x_i, \mu_i) \}_{i=1}^{\infty}$ is a sequence generated by the algorithm on the problem $\min_{x \in \mathbb{R}^n} \psi(x)$, then

$$(x_{i+1}, \mu_{i+1}) \in M_\psi(x_i) , \tag{17}$$

for all $i \in \mathbb{N}$.

For any $\nu \in \Sigma_p$, let $S(\nu) \triangleq Q(\nu)^{-\frac{1}{2}}$. Then the function $\psi(S(\nu)y)$ can be written in the alternative form $(\psi \circ S)(y)$, which leads to the notation $M_{\psi \circ S(\nu)}$ for the iteration map defined for the problem transformed by $S(\nu)$. Hence a variable metric algorithm for solving problem (2), based on the the iteration map M_ψ and the transformation matrix $S(\nu)$ has the form

Variable Metric Algorithm Model 2.1:

Data: $x_0 \in \mathbb{R}^n$, $\mu_{-1} \in \Sigma_p$, $i = 0$.

Step 1: Set $y_i = S(\mu_i)^{-1} x_i$,

Step 2: Compute $(y_{i+1}, \mu_{i+1}) \in M_{\psi \circ S(\mu_i)}(y_i)$,

Step 3: Set $x_{i+1} = S(\mu_i) y_{i+1}$.

Step 4: Replace i by $i+1$ and go to Step 1. ■

Note that the multiplier vectors do not require transformation because, for any invertible matrix S, $(\hat{x}, \hat{\mu}) \in \mathbb{R}^n \times \Sigma_p$ satisfies equations (5a, 5b) with respect to problem (2) if and only if $(S^{-1}\hat{x}, \hat{\mu})$ satisfies these equations with respect to problem $\min_{y \in \mathbb{R}^n} \psi(Sy)$.

3. Rate of Convergence of the Pshenichnyi Algorithm

We will now summarize a number of results, established in Refs. 9 and 14, for a version of the Pshenichnyi minimax algorithm (Ref. 11) which uses an exact minimizing line search. When applied to problem (2), with $\psi(\cdot)$ defined in (16), this algorithm has the following form:

Algorithm 3.1 : (see Algorithm 5.2 and Corollary 5.1 in Ref. 9)

Data: $x_0;\ \gamma > 0$.

Step 0: Set $i = 0$.

Step 1: Compute the search direction

$$h_i = \arg \min_{h \in \mathbb{R}^n} \max_{j \in \underline{p}} g^j(A_j x_i) + \langle A_j^T \nabla g^j(A_j x_i), h \rangle + \tfrac{1}{2}\gamma \|h\|^2 . \tag{18}$$

Step 2: Compute a minimizing step size, $\lambda_i \in \arg\min_{\lambda \in \mathbb{R}} \psi(x_i + \lambda h_i)$.

Step 3: Set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i + 1$ and go to Step 1. ■

Theorem 3.1: (Ref. 9) *If the functions $g^j(\cdot)$ in problem (2) are continuously differentiable, then any accumulation point $\bar{x}$ of a sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 3.1 satisfies the first-order optimality conditions (5a, 5b).* ■

To show that the algorithm converges linearly, we need to introduce more restrictive assumptions. Let the set of minimizers for problem (2) be denoted by $\hat{G} \triangleq \arg\min_{x \in \mathbb{R}^n} \psi(x)$. By analogy with nonlinear programming convention, we say that *strict complementary slackness* holds at $\hat{x} \in \hat{G}$ if, for every $\hat{\mu} \in \Sigma_p$ such that the pair $(\hat{x}, \hat{\mu})$ satisfies (5a, 5b), we have $\hat{\mu}^j > 0$ if and only if $g^j(A_j\hat{x}) = \psi(\hat{x})$.

Hypothesis 3.1: *Suppose that*

(i) the functions $g^j(\cdot)$ are twice continuously differentiable,

(ii) there exist $0 < l \leq L < \infty$ such that, for all $j \in \underline{p}$,

$$l\, \|h\|^2 \leq \langle h, G^j(z)\, h \rangle \leq L\, \|h\|^2 , \quad \forall\ z, h \in \mathbb{R}^{l_j} , \tag{19}$$

(iii) strict complementary slackness holds for all $\hat{x} \in \hat{G}$. ■

It follows from Hypothesis 3.1 that (i) for any $\hat{x} \in \hat{G}$ there is a unique optimal multiplier $\hat{\mu} \in \Sigma_p$ satisfying equations (5a, 5b), (ii) the set of optimal multipliers, associated with the set of optimal solutions, $\hat{G}$, is a singleton, $\{\hat{\mu}\}$, and (iii) the set of indices of functions maximal at $\hat{x}$, $\hat{J} \triangleq \{ j \in \underline{p} \mid g^j(A_j\hat{x}) = \psi(\hat{x}) \}$, is independent of $\hat{x} \in \hat{G}$ (see Ref. 13).

Let $j_1 < \ldots < j_b$ be the indices comprising $\hat{J}$, then we define the matrix $\hat{A}^T \triangleq [A_{j_1}^T, \ldots, A_{j_b}^T]$. Let $a \triangleq Rank(\hat{A}^T)$ and let Z be an $n \times a$ matrix, the columns of which form an orthonormal basis for $Range(\hat{A}^T)$. Then we have the following result.

Theorem 3.2: (Ref. 14) *Suppose that Hypothesis 3.1 holds with respect to problem (2) and, in addition, (iv) l and L are chosen so that the scaling parameter, γ, in Algorithm 3.1 satisfies*

$$l\sigma^+[\sum_{j\in p} \hat{\mu}^j A_j^T A_j] \; < \; \gamma \; < \; L\max_{j\in p} \|Z^T A_j^T A_j Z\| \; , \tag{20}$$

where $\sigma^+[X]$ denotes the minimum positive eigenvalue of the symmetric matrix X. If Algorithm 3.1 constructs an infinite sequence $\{x_i\}_{i=0}^{\infty}$, then, (a) $x_i \to \hat{x}$ as $i \to \infty$ with $\hat{x} \in \hat{G}$, and (b) either there exists an $i_0 \in \mathbf{N}$ such that $x_i = \hat{x}$ for all $i \geq i_0$ or

$$\lim_{i\to\infty} \sup \frac{\psi(x_{i+1}) - \hat{\psi}}{\psi(x_i) - \hat{\psi}} \leq \rho \; , \tag{21}$$

where

$$\rho \; \triangleq \; 1 - \frac{l}{L} \, \frac{\sigma^+[\sum_{j\in p} \hat{\mu}^j A_j^T A_j]}{\max_{j\in p} \|Z^T A_j^T A_j Z\|} \; . \tag{22}$$

■

Following Ref. 6, we refer to the quantity $\lim_{i\to\infty} \sup\, (\psi(x_{i+1}) - \hat{\psi})/(\psi(x_i) - \hat{\psi})$ as the *convergence ratio* of the sequence $\{\, \psi(x_i) \,\}_{i=0}^{\infty}$. The quantity ρ in (21) bounds the convergence ratio of any sequence constructed by Algorithm 3.1 in solving any problem in the class defined by (2) and the assumptions stated.

4. Rate of Convergence of Variable-Metric-Pshenichnyi Algorithm

We will refer to the algorithm obtained by combining our sequential transformation method with the iteration map of Algorithm 3.1, in the manner stated in Algorithm Model 2.1, as the Variable-Metric-Pshenichnyi Algorithm. We will now show that the Variable-Metric-Pshenichnyi Algorithm converges faster on problems of the form (2) than Algorithm 3.1.

For the transformed problem

$$\min_{x\in \mathbb{R}^n} \psi(S(v)y) \; , \tag{23}$$

given a point $y = S(v)^{-1}x$ and a $v \in \Sigma^p$, the search direction computation (18) has the form

$$d(y,v) \; \triangleq \; \arg \min_{d\in \mathbb{R}^n} \max_{j\in p} g^j(A_jS(v)y) + \langle (A_jS(v))^T \nabla g^j(A_jS(v)y), d\rangle + \tfrac{1}{2}\gamma\|d\|^2 \; . \tag{24}$$

The result can be transformed back to the original space using the formula

$$h(x,\nu) \triangleq S(\nu)d(y,\nu)\,. \tag{25}$$

Equivalently, $h(x,\mu)$ can be computed directly using the variable metric defined by $S(\nu)$, as follows:

$$h(x,\nu) = \arg \min_{h \in R^n} \max_{j \in \underline{p}} g^j(A_jx) + \langle A_j^T \nabla g^j(A_jx),\, h\rangle - \psi(x) \; + \tfrac{1}{2}\gamma \langle h, Q(\nu)h\rangle\,. \tag{26}$$

Since the max function in (26) is strictly convex in h, $h(x,\nu)$ is unique, and hence it also follows that $h(\cdot,\cdot)$ is continuous.

Problem (26) can be solved by converting it to dual form as follows. Let $\theta(x,\nu)$ denote the minimum value in (26). Then for any $x \in \mathbb{R}^n$ and $\nu \in \Sigma_p$, the search direction problem can be written in the following equivalent forms:

$$\theta(x,\nu) \triangleq \min_{h \in R^n} \max_{j \in \underline{p}} g^j(A_jx) + \langle A_j^T \nabla g^j(A_jx),\, h\rangle - \psi(x) \; + \tfrac{1}{2}\gamma \langle h, Q(\nu)h\rangle$$

$$= \min_{h \in R^n} \max_{\mu \in \Sigma_p} \sum_{j \in \underline{p}} \mu^j \left[g^j(A_jx) + \langle A_j^T \nabla g^j(A_jx),\, h\rangle - \psi(x) \right] + \tfrac{1}{2}\gamma \langle h, Q(\nu)h\rangle\,, \tag{27}$$

By an extension of von Neumann's Minimax Theorem (Ref. 9), the max and min in (27) can be interchanged. Hence,

$$\theta(x,\nu) = \max_{\mu \in \Sigma_p} \min_{h \in \mathbb{R}^n} \sum_{j \in \underline{p}} \mu^j \left[g^j(A_jx) + \langle A_j^T \nabla g^j(A_jx),\, h\rangle - \psi(x) \right] + \tfrac{1}{2}\gamma \langle h, Q(\nu)h\rangle$$

$$= \max_{\mu \in \Sigma_p} \sum_{j \in \underline{p}} \mu^j \left[g^j(A_jx) - \psi(x) \right] - \frac{1}{2\gamma} \Big\| \sum_{j \in \underline{p}} \mu^j A_j^T \nabla g^j(A_jx) \Big\|^2_{Q(\nu)^{-1}}\,, \tag{28}$$

where the last expression is obtained by solving the inner minimization problem[3]. Since the solution to (28) is usually not unique, we denote the solution set by

$$U(x,\nu) \triangleq \arg \max_{\mu \in \Sigma_p} \sum_{j \in \underline{p}} \mu^j \left[g^j(A_jx) - \psi(x) \right] - \frac{1}{2\gamma} \Big\| \sum_{j \in \underline{p}} \mu^j A_j^T \nabla g^j(A_jx) \Big\|^2_{Q(\nu)^{-1}}\,. \tag{29}$$

The set-valued function $U(\cdot,\cdot)$ has the following properties: (i) it is upper semi-continuous in the sense of Berge (Ref. 21); (ii) for any minimizer $\hat{x}$ of (1) and any $\nu \in \Sigma_p$, $U(\hat{x},\nu)$ is the set of multiplier vectors which, together with $\hat{x}$, satisfy equations (5a) and (5b), (iii) under Hypothesis 3.1, $U(\hat{G},\Sigma_p) = \{\, \hat{\mu} \,\}$, a singleton, (iv) any multiplier vector $\mu \in U(x,\nu)$ yields the *unique* solution to the

[3] Several methods exist (see, for example, Refs. 15-20) for solving the positive semi-definite quadratic program (28).

primal problem (26), according to the formula

$$h(x,\nu) = -\frac{1}{\gamma}Q(\nu)^{-1}\sum_{j\in p}\mu^j A_j^T \nabla g^j(A_j x)\,. \tag{30}$$

Steps 2 and 3 of the Variable Metric Algorithm 2.1, using the iteration map of Algorithm 3.1, can also be performed in the original space without affecting the sequence of iterates produced. We therefore present it in this form to simplify proofs.

Algorithm 4.1:

Data: $x_0;\ \gamma > 0\ ,\ \mu_{-1} \in \Sigma_p\ ,\ \varepsilon > 0,\ i = 0.$

Step 1: Compute the multiplier vector, $\mu_i \in U(x_i,\mu_{i-1})$.

Step 2: Compute $h_i = h(x_i,\mu_{i-1})$ using (30).

Step 3: Compute the minimizing step size, $\lambda_i = \arg\min_{\lambda\in\mathbb{R}} \psi(x_i + \lambda h_i)$.

Step 4: Set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i+1$ and go to Step 1. ■

We will now establish several properties of Algorithm 4.1.

Theorem 4.1: *If the functions $g^j(\cdot)$ in problem (2) are continuously differentiable, then any accumulation point $\hat{x}$ of a sequence $\{x_i\}_{i=0}^{\infty}$ generated by Algorithm 4.1 satisfies the necessary conditions (5a, 5b).*

Proof: This follows from the proof of convergence for Algorithm 3.1 in Ref. 9 and the fact that the scaling matrices $S(\mu_{i-1})$ are uniformly bounded, *i.e.* - that for all $\nu \in \Sigma_p$,

$$(\max_{j\in p} \|A_j^T A_j\|)^{-1/2}\ \|h\|^2 \le \|S(\nu)h\|^2 \le \varepsilon^{-1/2}\ \|h\|^2\,. \tag{31}$$

■

Next we will show, under assumptions of convexity and complementary slackness, that the sequence of iterates, $\{x_i\}_{i=0}^{\infty}$, constructed by Algorithm 4.1 converges to the solution set $\hat{G}$ and that the corresponding sequence of multiplier vectors, $\{\mu_i\}_{i=-1}^{\infty}$ converges to $\hat{\mu}$, the unique optimal multiplier associated with the solution set $\hat{G}$. We will use the notation $z_i \to Z$ to represent the convergence of a sequence $\{z_i\}_{i=0}^{\infty} \subset \mathbb{R}^n$ to a set $Z \subset \mathbb{R}^n$, *i.e.* - $\lim_{i\to\infty} \min_{y\in Z} \|z_i - y\| = 0$.

Theorem 4.2: *Suppose that Hypothesis 3.1 holds and that Algorithm 4.1 generates sequences of iterates* $\{ x_i \}_{i=0}^{\infty}$ *and of multiplier vectors* $\{ \mu_i \}_{i=0}^{\infty}$. *Then,*

(a) there exists an open set $W \supset \hat{G}$ *such that* $\mu^j = 0$ *for every* $j \notin \hat{J}$ *and all* $\mu \in U(W, \Sigma_p)$,

(b) there exists $\hat{x} \in \hat{G}$ *such that* $x_i \to \hat{x}$ *as* $i \to \infty$,

(c) there exists $i_0 \in \mathbb{N}$ *such that* $x_i \in \hat{x} + Range(\hat{A}^T)$ *for all* $i \geq i_0$,

(d) $\mu_i \to \hat{\mu}$ *as* $i \to \infty$.

Proof: (a) Since $h(\cdot,\cdot)$ defined in (25) and $\theta(\cdot,\cdot)$ defined in (27) are uniformly continuous in (x,ν) on compact sets in $\mathbb{R}^n \times \Sigma_p$, and since both functions are zero on the set $\hat{G} \times \Sigma_p$, (a) follows from the same argument as Proposition 5.2 in Ref. 14.

(b) Let $A^T \triangleq [A_1^T, \ldots, A_p^T]$. Equation (30) and the fact that $Range(A^T)$ is invariant under $S(\nu)$ for all $\nu \in \Sigma_p$ imply that the sequence of search directions $\{ h_i \}_{i=0}^{\infty}$ is contained in the range of A^T. Therefore, the sequence of iterates $\{ x_i \}_{i=0}^{\infty}$ is contained in the set

$$V \triangleq (x_0 + Range(A^T)) \cap \{ x \in \mathbb{R}^n \mid \psi(x) \leq \psi(x_0) \} . \tag{32}$$

The set V is compact by the same argument as in the proof of Theorem 5.1 in Ref. 14, and therefore the set $\{ x_i \}_{i=0}^{\infty}$ converges to the set of its accumulation points. By Theorem 4.1, these must satisfy the optimality condition (5a, 5b). Since $\psi(\cdot)$ is convex, these necessary conditions are sufficient for optimality, implying that $x_i \to \hat{G}$.

From part (a) of this theorem, $h(x,\nu) \in Range(\hat{A}^T)$ for all $x \in W$ and all $\nu \in \Sigma_p$. Because $x_i \to \hat{G}$, there exists $i_0 \in \mathbb{N}$ such that $x_i \in W$ for all $i > i_0$. Hence, $\{ x_i \}_{i=i_0}^{\infty} \subset x_{i_0} + Range(\hat{A}^T)$, and $x_i \to (x_{i_0} + Range(\hat{A}^T)) \cap \hat{G}$.

We show that this limit set is a singleton. Suppose $x_1, x_2 \in (x_{i_0} + Range(\hat{A}^T)) \cap \hat{G}$. Then, since $\psi(\cdot)$ is convex, the entire line segment between x_1 and x_2, $[x_1, x_2]$, is contained in this set. Now, $U([x_1,x_2], \Sigma_p) = \{ \hat{\mu} \}$ and $\hat{\mu}^j > 0$ for all $j \in \hat{J}$. Hence, $g^j(A_j x) = \psi(x) = \hat{\psi}$ for all $x \in [x_1, x_2]$ and all $j \in \hat{J}$, by equation (5b). Since the functions $g^j(\cdot)$ are strictly convex, this implies that $A_j(x_1 - x_2) = 0$

for all $j \in \hat{J}$. Since $x_1 - x_2 \in Range(\hat{A}^T)$, this implies that $x_1 - x_2 \in Range(\hat{A}^T) \cap Null(\hat{A}) = \{ 0 \}$, *i.e.* - that $x_1 = x_2$. Thus, $\hat{G} \cap (x_{i_0} + Range(\hat{A}^T)) = \{ \hat{x} \}$ for some $\hat{x}$.

(c) From the proof of (b), $x_i \in x_{i_0} + Range(\hat{A}^T) = \hat{x} + Range(\hat{A}^T)$, for all $i \geq i_0$.

(d) The set-valued map $U(\cdot,\cdot)$ defined in (29) is upper semicontinuous in the sense of Berge (Ref. 21), uniformly on compact sets in $\mathbb{R}^n \times \Sigma_p$. Since $x_i \to \hat{x} \in \hat{G}$ by (b) and $U(\hat{G},\Sigma_p) = \{ \hat{\mu} \}$, this implies that $\mu_i \to \hat{\mu}$. ∎

We define the function $\rho : \mathbb{R}^{n \times n} \to \mathbb{R}$ by

$$\rho(S) \triangleq 1 - \frac{l}{L} \frac{\sigma^+[\sum_{j \in \underline{p}} \hat{\mu}^j S^T A_j^T A_j S]}{\max\limits_{j \in \underline{p}} \| Z^T S^T A_j^T A_j S Z \|} . \tag{33}$$

Note that ρ in equation (22) equals $\rho(I)$, where I is the $n \times n$ identity matrix.

Theorem 4.3: *Suppose that Hypothesis 3.1 holds and, in addition, (iv) l and L are chosen so that the scaling parameter, γ, satisfies*

$$l \sigma^+[\sum_{j \in \underline{p}} \hat{\mu}^j S(\hat{\mu}) A_j^T A_j S(\hat{\mu})] < \gamma < L \max_{j \in \underline{p}} \| Z^T S(\hat{\mu})^T A_j^T A_j S(\hat{\mu}) Z \| . \tag{34a}$$

If $\{ x_i \}_{i=0}^{\infty}$ is an infinite sequence generated by Algorithm 4.1, then, either there exists an $i_0 \in \mathbb{N}$ and $\hat{x} \in \hat{G}$ such that $x_i = \hat{x}$ for all $i \geq i_0$, or

$$\limsup_{i \to \infty} \frac{\psi(x_{i+1}) - \hat{\psi}}{\psi(x_i) - \hat{\psi}} \leq \rho(S(\hat{\mu})) . \tag{34b}$$

Proof: By Theorem 4.1(b), the sequence of iterates has a limit point $\hat{x} \in \hat{G}$. Assume that $x_i \neq \hat{x}$ for all $i \in \mathbb{N}$. Hypothesis 3.1 and assumption (iv) of this theorem ensure that the assumptions of Theorem 3.2 are met for the transformed problem,

$$\min_{y \in \mathbb{R}^n} \psi(S(\hat{\mu})y) . \tag{35}$$

Since the range of $\hat{A}^T$ is invariant under $S(\hat{\mu})$, the columns of Z form a basis for the range of $S(\hat{\mu})\hat{A}^T$. This fact and assumption (iv) of this theorem imply that assumption (iv) of Theorem 3.2 holds with respect to problem (35). This and Hypothesis 3.1 ensure that the assumptions of Theorem 3.2 are satisfied with respect to problem (35). The following result, slightly stronger than Theorem 3.2, but valid under the same assumptions, is stated in Theorem 5.1 of Ref. 14:

$$\limsup_{\substack{y \to \hat{y} \\ y \neq \hat{y} \\ y \in Range(Z)}} \min_{\lambda \in \mathbb{R}} \frac{\psi(\, S(\hat{\mu})(y + \lambda d(y,\hat{\mu}))\,) - \hat{\psi}}{\psi(\, S(\hat{\mu})y\,) - \hat{\psi}} \leq \rho(S(\hat{\mu}))\,, \tag{36}$$

where $\hat{y} \triangleq S(\hat{\mu})^{-1}\hat{x}$ for an arbitrary $\hat{x} \in \hat{G}$. Using the substitution $y = S(\hat{\mu})^{-1}x$ and the fact that $h(S(\hat{\mu})y,\hat{\mu}) = S(\hat{\mu})d(y,\hat{\mu})$, (36) can be rephrased as follows. For any $\delta > 0$, there exists a set $V' \subset \hat{x} + Range(Z)$, which is open in the affine space $\hat{x} + Range(Z)$, such that

$$\min_{\lambda \in \mathbb{R}} \frac{\psi(\, x + \lambda h(x,\hat{\mu})\,) - \hat{\psi}}{\psi(x) - \hat{\psi}} \leq (1 + \delta)\rho(S(\hat{\mu}))\,, \tag{37}$$

for all $x \in V'$, $x \neq \hat{x}$. Since $\psi(\cdot)$ is strongly convex, the min over $\mathbb{R}$ in (37) and in Step 3 of Algorithm 4.1 can be replaced by a min over a closed interval C. With this modification, the left hand side of (37) is continuous in (x,μ), since $h(\cdot,\cdot)$ is continuous. This implies that there exists a neighborhood of $\hat{\mu}$, $D \subset \Sigma_p$, such that

$$\min_{\lambda \in C} \frac{\psi(x + \lambda h(x,\mu)\,) - \hat{\psi}}{\psi(x) - \hat{\psi}} \leq (1 + 2\delta)\rho(S(\hat{\mu}))\,, \tag{38}$$

for all $x \in V'$ and $\mu \in D$. Of course, since δ was arbitrary,

$$\limsup_{\substack{x \to \hat{x} \\ x \neq \hat{x} \\ x \in \hat{x} + Range(Z)}} \min_{\lambda \in C} \frac{\psi(x + \lambda h(x,\mu)\,) - \hat{\psi}}{\psi(x) - \hat{\psi}} \leq \rho(S(\hat{\mu}))\,. \tag{39}$$

By Theorem 4.2(c), $x_i \in \hat{x} + Range(\hat{A}^T) = \hat{x} + Range(Z)$ for large i. Then, since $x_i \to \hat{x}$, $\mu_i \to \hat{\mu}$ and $\psi(x_i + \lambda_i h_i) = \min_{\lambda \in C} \psi(x_i + \lambda h_i)$, (34b) holds. ■

The following comparison of the two convergence ratio bounds, ρ given by (22) and $\rho(S(\hat{\mu}))$ given by (33), suggests that our variable metric technique results in a faster algorithm than the original

Pshenichnyi Algorithm 3.1.

Proposition 4.1: *If* $\sigma^+[R(\hat{\mu})] > \varepsilon$, *then* $\rho(I) \geq \rho(S(\hat{\mu}))$.

Proof: Consider a spectral decomposition $R(\hat{\mu}) = U \Lambda U^T$, where U is unitary, $\Lambda \triangleq diag(\ \lambda_1(\hat{\mu}),\dots,\lambda_n(\hat{\mu})\ )$ and $\tilde{\Lambda} \triangleq diag(\ \tilde{\lambda}_1(\hat{\mu}),\dots,\tilde{\lambda}_n(\hat{\mu})\ )$. We have that

$$\sum_{j \in \underline{p}} \hat{\mu}^j\, S(\hat{\mu})^T A_j^T A_j\, S(\hat{\mu}) = S(\hat{\mu})^T R(\hat{\mu})\, S(\hat{\mu}) \tag{40}$$

$$= U \tilde{\Lambda}^{-1/2} U^T (\ U \Lambda U^T\ )\ U \tilde{\Lambda}^{-1/2} U^T$$

$$= U \tilde{\Lambda}^{-1} \Lambda U^T .$$

Since $\sigma^+[R(\hat{\mu})] > \varepsilon$, we have that, for each $j \in \underline{n}$, either $\tilde{\lambda}_j(\hat{\mu}) = \lambda_j(\hat{\mu})$ or $\lambda_j(\hat{\mu}) = 0$. Hence, $\sum_{j \in \underline{p}} \hat{\mu}^j\, S(\hat{\mu})^T A_j^T A_j S(\hat{\mu}) = U\ diag(1,\dots,1,0,\dots,0)\ U$, which implies that $\sigma^+[\sum_{j \in \underline{p}} \hat{\mu}^j\, S(\hat{\mu})^T A_j^T A_j S(\hat{\mu})] = 1$, and

$$\rho(S(\hat{\mu})) \triangleq 1 - \frac{l}{L} \frac{1}{\max_{j \in \underline{p}} \| Z^T S(\hat{\mu})^T A_j^T A_j S(\hat{\mu}) Z \|} . \tag{41}$$

Now,

$$\| Z^T S(\hat{\mu})^T A_j^T A_j S(\hat{\mu}) Z \| = \max_{y \in \mathbb{R}^a} \frac{\langle y, Z^T S(\hat{\mu})^T A_j^T A_j S(\hat{\mu}) Z y \rangle}{\langle y, y \rangle}$$

$$= \max_{y \in \mathbb{R}^a} \frac{\| A_j S(\hat{\mu}) Z y \|^2}{\| y \|^2}$$

$$= \max_{y \in \mathbb{R}^a} \frac{\| A_j S(\hat{\mu}) Z y \|^2}{\| Z y \|^2} , \tag{42}$$

since the orthonormality of the columns of Z implies that $\|Zy\| = \|y\|$ for the Euclidean norm. Making the substitution $z = S(\hat{\mu}) Z y$ yields

$$\| Z^T S(\hat{\mu})^T A_j^T A_j S(\hat{\mu}) Z \| = \max_{z \in Range(Z)} \frac{\| A_j z \|^2}{\| S(\hat{\mu})^{-1} z \|^2}$$

$$= \max_{z \in Range(Z)} \frac{\| A_j z \|^2}{\| Q(\hat{\mu})^{1/2} z \|^2}$$

$$= \max_{z \in Range(Z)} \frac{|A_j z|^2}{\langle z, Q(\hat{\mu})z \rangle} . \tag{43}$$

Substituting (43) into (41) yields

$$\rho(S(\hat{\mu})) \triangleq 1 - \frac{l}{L} \min \left\{ \frac{\langle z, Q(\hat{\mu})z \rangle}{|A_j z|^2} \,\middle|\, j \in \underline{p} \,,\, z \in Range(Z) \right\} . \tag{44}$$

By inspection, $\rho(S(\hat{\mu}))$ is never greater than

$$\rho(I) = 1 - \frac{l}{L} \min \left\{ \frac{\langle z, Q(\hat{\mu})z \rangle}{|A_j y|^2} \,\middle|\, j \in \underline{p} \,,\, y,z \in Range(Z) \,,\, |y| = |z| = 1 \right\} . \tag{45}$$

■

The difference between $\rho(I)$ and $\rho(S(\hat{\mu}))$ can be quite significant, as the following example shows.

Example 3.1: Suppose that a minimax problem involves two scaling matrices,

$$\begin{bmatrix} 1 & 0 \\ 0 & 10^{-2} \end{bmatrix} , \quad \begin{bmatrix} 10^{-2} & 0 \\ 0 & 10 \end{bmatrix} , \tag{46}$$

and that $\mu^1 = \mu^2 = \frac{1}{2}$ and $l = L = 1$. The rate constant for the unscaled Algorithm 3.1 is $\rho(I) = 0.995$, whereas, under rescaling, it is $\rho(S(\hat{\mu})) = 0.5$. This suggests that $\lceil \log 10^{-1} / \log \rho(I) \rceil = 460$ iterations of the Pshenichnyi Algorithm 3.1 would be required to achieve a ten-fold reduction in $\psi(x) - \hat{\psi}$ near the solution, while only $\lceil \log 10^{-1} / \log \rho(S(\hat{\mu})) \rceil = 4$ iterations of the Variable-Metric Pshenichnyi Algorithm 4.1 would be required. ■

5. Results of Numerical Experiments

Since rate of convergence results are indicative only of the terminal behavior of an algorithm, we performed a number of numerical experiments to evaluate the overall behavior of the variable metric technique. We compare the performance of the Pshenichnyi Algorithm 3.1 with that of the Variable-Metric-Pshenichnyi Algorithm 4.1 and with Han's method (Ref. 10), which uses full second order information. In addition, we compared the performance of the *barrier function minimax algorithm* in Ref. 12 with a corresponding variable-metric-barrier-function method which we constructed in accordance with the Variable Metric Algorithm Model 2.1. The barrier function method is based on the penalty function

$$\sum_{j \in \underline{p}} \frac{1}{\alpha - g^{j}(A_j x)} \ , \tag{47}$$

which is differentiable at all x for which $\psi(x) < \alpha$. An iteration of the barrier method involves an indefinite number of inner cycles, each of which requires the evaluation of all functions and first order derivatives. The rate of convergence of this algorithm has not been established and hence we can only evaluate the effect of our sequential transformation technique on it through numerical experiments. The five algorithms were applied to the two problems below. An Armijo-like step size rule (Ref. 9),

$$\lambda_i = \max\{\ \beta^k \tilde{\lambda}_i \mid \psi(x_i + \tilde{\lambda}_i \beta^k h_i) - \psi(x_i) \le \alpha \tilde{\lambda}_i \beta^k \theta(x_i, \mu_i)\ \}\ , \tag{48}$$

with $\alpha, \beta \in (0,1)$, was substituted for the exact minimizing line search in Algorithms 3.1 and 4.1, since problem (48) can be solved in a finite number of steps. Quadratic interpolation was used to determine a trial step size $\tilde{\lambda}_i$. In all of the experiments, the algorithm parameters were set to $\alpha = 0.7$, $\beta = 0.9$, $\gamma = 1.0$, and $\varepsilon = 10^{-10}$ (in the definition of the matrices $Q(\mu)$). Since in engineering applications, gradients and Hessians are frequently computed using finite differences, the evaluation counts in Tables 1 and 2 are tabulated as though the gradients and Hessians of the functions $g^{j}(\cdot)$ were evaluated by finite differences. The evaluation of a single function $g^{j}(z)$ incurs one function evaluation, and the gradient $\nabla g^{j}(z)$ incurs an additional l_j evaluations. Thus, the total number of function evaluations required to obtain the information to compute a search direction for the Pshenichnyi and Variable-Metric-Pshenichnyi Algorithms is $\sum_{j \in \underline{p}} (\ l_j + 1)$. The evaluation of Hessians for use by the Han algorithm incurs an additional $\frac{1}{2}(l_j^2 + 1)$ evaluations per function $g^{j}(\cdot)$.

Problem 5.1: Consider the simple problem $\min_{x \in \mathbb{R}^4} \max\{\ g^1(A_1 x)\ ,\ g^2(A_2 x)\ \}$, where

$$g^1(y) = y_1^2 + y_2^2 + (y_3 - 1)^2 - 1\ , \tag{49a}$$

$$g^2(y) = y_1^2 + y_2^2 + (y_3 + 1)^2 - 1\ , \tag{49b}$$

and the matrices A_1, A_2 are given by

$$A_1 = \begin{bmatrix} 10 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 10^{-1} & 0 \end{bmatrix} \quad A_2 = \begin{bmatrix} 10^2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{49c}$$

An initial point of $(10^{-3},0,10,0)$ was used. The minimum value of 0 is achieved on the subspace spanned by the vector $(0,0,0,1)$. Table 1 shows the work required for the five algorithms to achieve two given levels of accuracy in the value of $\hat{\psi}$. The units of work listed are number of iterations, the number of function evaluations and the CPU time. Figure 1 plots the function values $\{ \psi(x_i) \}$ versus the number of function evaluations for the Pshenichnyi and Variable-Metric-Pshenichnyi Algorithms. ■

Problem 5.2: (Ref. 22) Consider the problem of designing a controller for the feedback system in Figure 2 with plant,

$$P(s) = \frac{1}{(s+2)^2(s+3)} \begin{bmatrix} s^2+8s+10 & 3s^2+7s+4 \\ 2s+2 & 3s^2+9s+8 \end{bmatrix}. \tag{50}$$

Since the plant is stable, we can parametrize the controller by $C(x) = (I - R(x,s)P(s))^{-1}R(x,s)$ where $R(x,s)$ is a 2×2 matrix of rational polynomials in the complex variable s, which are bounded and analytic for $Re(s) \leq 0$. We chose to shape frequency domain tracking error by solving the problem

$$\min_x \tfrac{1}{2} \max_{\omega \in \Omega} \| H_{e_1 u}(j\omega, R(x,j\omega)) \|_F^2 , \tag{51}$$

where Ω consists of six frequency points, equally spaced on a logarithmic scale, $\{ 0.010, 0.029, 0.080, 0.240, 0.693, 2.0 \}$, and $\|\cdot\|_F$ denotes the Frobenius norm. For this system, $H_{e_1 u}(x,s) = I - P(s)R(x,s)$. We used the following first order expansion of $R(x,s)$,

$$R(x,s) = \begin{bmatrix} x1 & x2 \\ x3 & x4 \end{bmatrix} \frac{1}{(s+10)} + \begin{bmatrix} x5 & x6 \\ x7 & x8 \end{bmatrix}. \tag{52}$$

The initial point $x_0 = (0,0,0,0,1,0,0,1)$ was chosen, and our computations converged to the minimum value of 0.0255085 at the point

$$\hat{x} = \Big(-80.308718709, -4.4337113582, 84.132574000, -31.534025985,$$

$$9.2348949849, -0.0051528236, -8.9338039187, 4.8550280952 \Big)$$

The work required for the algorithms to achieve two given levels of accuracy is recorded in Table 2.

The values of $\psi(\cdot)$ are plotted versus the number of function evaluations for the Pshenichnyi and Variable-Metric-Pshenichnyi Algorithms in Figure 3. ■

Theorems 3.1 and 4.2 apply under the same assumptions to versions of Algorithms 3.1 and 4.1 employing an Armijo-like step size rule, except that the convergence ratio bounds are given by

$$\rho \triangleq 1 - \alpha\beta\frac{l}{L}\,\frac{\sigma^+[\sum_{j \in \underline{p}} \hat{\mu}^j A_j^T A_j]}{\max\limits_{j \in \underline{p}} \| Z^T A_j^T A_j Z \|}\,, \tag{53}$$

and

$$\rho(S(\hat{\mu})) \triangleq 1 - \alpha\beta\frac{l}{L}\,\frac{\sigma^+[\sum_{j \in \underline{p}} \hat{\mu}^j S(\hat{\mu})^T A_j^T A_j S(\hat{\mu})]}{\max\limits_{j \in \underline{p}} \| Z^T S(\hat{\mu})^T A_j^T A_j S(\hat{\mu}) Z \|}\,. \tag{54}$$

Table 3 presents the convergence ratios of the *sequences* constructed by the algorithms under comparison on Problems 5.1 and 5.2, as well as the convergence ratio bounds derived above. Table 3 shows that the variable metric technique reduces both quantities. The reduction in computational effort corresponding to the decrease in the convergence ratios of the observed sequences is evident from Tables 1 and 2. The reduction in effort entailed by even the slight reductions in the convergence ratio bounds is also large. To show this, we have included in Table 3 the number of iterations which the algorithm convergence ratio bounds suggest would be required to reduce $\psi(x) - \hat{\psi}$ by a factor of 10 near a solution, *i.e.* - $\lceil \log 0.1 / \log \rho \rceil$.

If a variable metric $Q_H(\mu)$ is based on $R_H(\mu) \triangleq \sum_{j \in \underline{p}} \mu^j A_j^T G^j(A_j x) A_j$, rather than $R(\mu)$ as in Section 2, and if $\sigma^+[R_H(\hat{\mu})] > \varepsilon$, the search direction of Algorithm 4.1 coincides with that of Han's algorithm near $\hat{G}$. A result similar to Theorem 4.3 holds for this algorithm with

$$\rho_H = 1 - \min\left\{ \frac{\langle z, Q_H(\hat{\mu}) z\rangle}{\langle z, A_j^T G^j(A_j\hat{x}) A_j z\rangle} \;\middle|\; j \in \underline{p}\;,\; z \in Range(Z) \right\}. \tag{55}$$

In general, $\rho_H > 0$, suggesting that only linear convergence is achieved despite the use of second order information. This is born out by the strictly positive convergence ratios observed for versions of the Han algorithm using an exact minimizing line search. While a sequence $\{x_i\}$ constructed by the Han algorithm with a fixed step size of 1 converges superlinearly to a minimizer, some iterations may produce an *increase* in $\psi(\cdot)$. It is likely that a *descent* algorithm based on Han's search direction will not be superlinearly convergent without the use of devices analogous to the feasibility enhancing corrections of some algorithms for nonlinear programming (see, for example, Refs. 23 and 24).

6. Conclusion

We have introduced a variable metric technique which substantially mitigates the ill-conditioning produced in the composite minimax problem by the A_j matrices. The technique does not require the evaluation of second derivatives and can be used as described in Algorithm Model 2.1 to speed the convergence of any first-order minimax algorithm which produces estimates of the optimal multipliers. We have analyzed the effect of the technique on the rate of convergence of the Pshenichnyi minimax algorithm. An upper bound on the convergence ratio was obtained for the variable metric version of the algorithm which can be considerably smaller than for the unscaled version. Numerical experiments verify the improvement suggested by the decrease in the convergence ratio bounds. The variable metric technique yielded a dramatic acceleration in convergence. The experiments also confirmed that the technique can speed convergence of another minimax algorithm. The variable metric technique can be applied without modification to a version of the Pshenchinyi algorithm for solving minimax problems involving semi-infinite composite max functions (Ref. 9) of the form $\max_{y \in Y} \phi(x,y)$, where $Y \subset \mathbb{R}^r$ is a compact, but infinite set. The convergence rate analysis extends to this case as well.

7. References

1. DENNIS, J. E. and MORE, J. J., *Quasi-Newton Methods: Motivation and Theory*, SIAM Review, Vol. 19, pp. 46-89, 1977.
2. BYRD, R. H., NOCEDAL, J. and YUAN, Y.-X., *Global Convergence of a Class of Quasi-Newton Methods on Convex Problems*, SIAM Journal of Numerical Analysis, Vol. 24, No. 5, pp. 1171-1190, 1987.
3. GOLDSTEIN, A. A., *Constructive Real Analysis*, Harper and Row, New York, N. Y., 1967.
4. DIXON, L. C. W., *Variable Metric Algorithms: Necessary and Sufficient Conditions for Identical Behavior on Nonquadratic Functions*, Journal of Optimization Theory and Applications, Vol. 10, pp. 34-40, 1972.
5. NEMIROVSKY, A. S. and YUDIN, D. B., *Problem Complexity and Method Efficiency in Optimization*, John Wiley, New York, N. Y., 1979.
6. LUENBERGER, D. G., *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, Massachusetts, 1973.

7. POLAK, E. and SALCUDEAN, S. E., *On The Design of Linear Multivariable Feedback Systems via Constrained Nondifferentiable Optimization in H^∞ Spaces*, IEEE Transactions on Automatic Control, Vol. 34, No.3, pp 268-276, 1989.

8. BAKER, T. E., *Algorithms for Optimal Control of Systems Described by Partial and Ordinary Differential Equations*, Ph.D. Dissertation, Department of Electrical Engineering and Computer Science, University of California, Berkeley, 1988.

9. POLAK, E., *On the Mathematical Foundations of Nondifferentiable Optimization in Engineering Design*, SIAM Review, Vol. 29, No.1, pp. 21-91, March 1987.

10. HAN, S. P., *Variable Metric Methods for Minimizing a Class of Nondifferentiable Functions*, Mathematical Programming, Vol. 20, pp. 1-13, 1981.

11. PSHENICHNYI, B. N. and DANILIN, YU. M., *Numerical Methods in Extremal Problems*, Nauka, Moscow, 1975.

12. POLAK, E., HIGGINS, J. E. and MAYNE, D. Q., *A Barrier Function Method for Minimax Problems*, University of California, Berkeley, Electronics Research Laboratory Memo No. UCB/ERL M88/64, 20 October 1988.

13. CLARKE, F. H., *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, New York, 1983.

14. WIEST, E. J. and POLAK, E., *On the Linear Convergence of the Pshenichnyi Method of Linearizations*, University of California, Berkeley, Electronics Research Laboratory, Memo No. UCB/ERL M89/55, May 1989.

15. GILL, P. E., HAMMARLING, J., MURRAY, W., SAUNDERS, M. A. and WRIGHT, M. H., *User's Guide for LSSOL (Version 1.0): A Fortran Package for Constrained Linear Least-squares and Convex Quadratic Programming*, Technical report SOL 86-1, Department of Operations Research, Stanford University, Stanford, January 1986.

16. VON HOHENBALKEN, B., *Simplical Decomposition in Nonlinear Programming Algorithms*, Mathematical Programming, Vol. 13, pp. 49-68, 1977

17. KIWIEL, K. C., *A Method for Solving Certain Quadratic Programming Problems Arising in Nonsmooth Optimization*, IMA Journal of Numerical Analysis, Vol. 6, pp. 137-152, 1986.

18. KIWIEL, K. C., *A Dual Method for Solving Certain Positive Semi-Definite Quadratic Programming Problems*, SIAM Journal of Scientific and Statistical Computing, Vol. 10, 1989.

19. RUSCZYNSKI, A., *A Regularized Decomposition Method for Minimizing a Sum of Polyhedral Functions*, Mathematical Programming, Vol. 35, pp. 46-61, 1986.

20. HIGGINS, J. E. and POLAK, E., *Minimizing Pseudo-Convex Functions on Convex Compact Sets*, University of California, Berkeley, Electronics Research Laboratory Memo No. UCB/ERL M88/22, March 1988. To appear in this Journal.

21. BERGE, C., *Topological Spaces*, Macmillan, New York, New York, 1963.

22 HOANG, P. D., *Computational Results on the Optimal Design of Feedback Compensators*, Masters Thesis, Department of Electrical Engineering and Computer Science, University of California, Berkeley, 1987.

23. MAYNE, D. Q. and POLAK, E., *A Superlinearly Convergent Algorithm for Constrained Optimization Problems*, Mathematical Programming Study 16, pp. 45-61, 1982.

24. PANIER, E. R., TITS, A. L. and HERSKOVITS, J. N., *A QP-Free, Globally Convergent, Locally Superlinearly Convergent Algorithm for Inequality Constrained Optimization*, SIAM J. Control and Opt., Vol. 26, No. 4, pp. 788-811, 1988.

Table 1: Numerical Results for Problem 5.1.

Algorithm	$\psi_i \leq \hat{\psi} + 10^{-2}$			$\psi_i \leq \hat{\psi} + 10^{-4}$		
	Iterations	Function evaluations	Time (sec.)	Iterations	Function evaluations	Time (sec.)
Pshenichnyi	291	5,246	11.6	397	7,154	15.9
VM-Pshenichnyi	4	80	0.3	6	116	0.4
Han	3	98	0.2	5	152	0.3
Barrier	45	6,806	9.4	48	16,640	22.1
VM-Barrier	40	2,276	4.8	43	2,812	5.7

Table 2: Numerical results for Problem 5.2.

Algorithm	$\psi_i \le \hat{\psi} + 10^{-2}$			$\psi_i \le \hat{\psi} + 10^{-4}$		
	Iterations	Function evaluations	Time (sec.)	Iterations	Function evaluations	Time (sec.)
Pshenichnyi	11628	976806	2317.9	11976	100603	2391.9
VM-Pshenichnyi	4	390	2.0	6	558	2.8
Han	4	1350	2.2	6	1902	3.1
Barrier	15	2,314,548	2,788.0	21	10,904,772	13,008.5
VM-Barrier	4	1422	7.1	11	4962	13.2

Table 3: Convergence ratios and bounds for Problems 5.1 and 5.2.

Algorithm	Problem 5.1			Problem 5.2		
	Convergence ratio	Convergence ratio bound	Iterations	Convergence ratio	Convergence ratio bound	Iterations
Pshenichnyi	.83	.999979	109,646	.9969	.999994	383,763
VM-Pshenichnyi	.67	.697697	7	.0805	.937000	36
Han	.0840	.697697	7	.0805	.937000	36

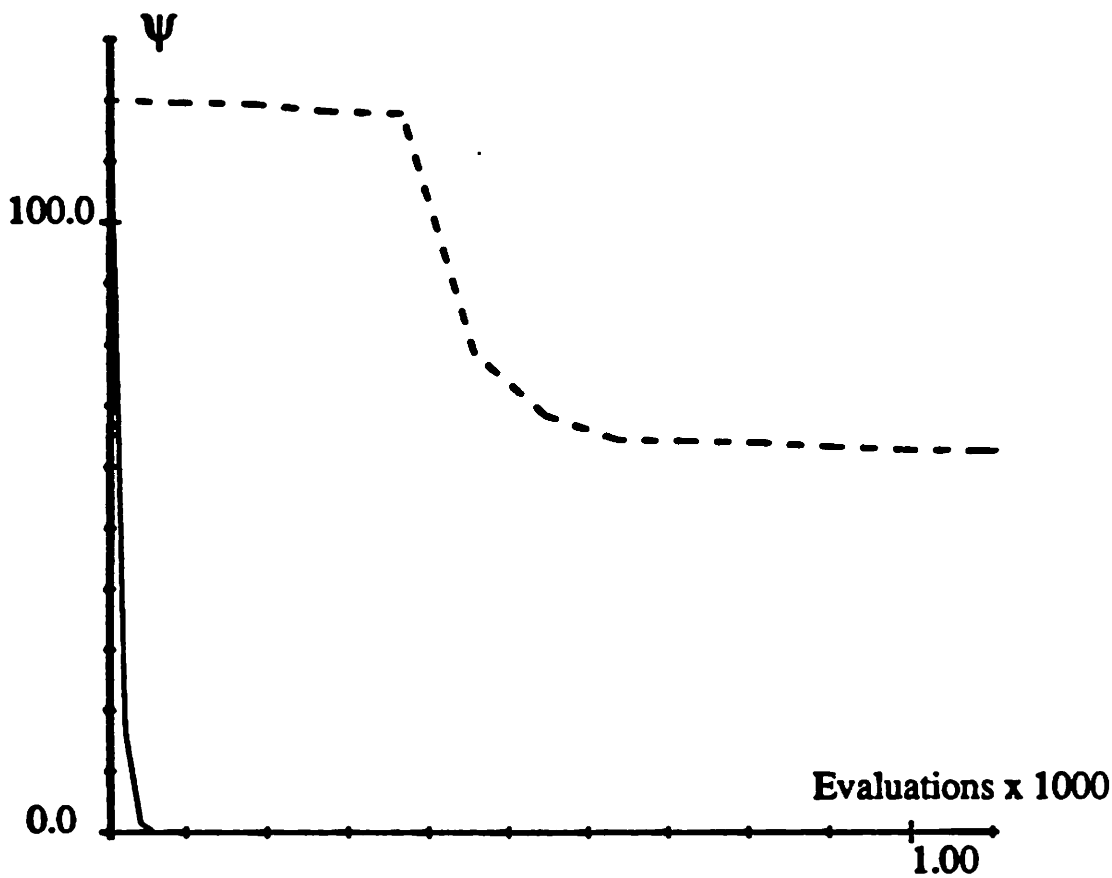

Figure 1: **Peformance of the Pshenichnyi Algorithm (dashed) and the Variable-Metric-Pshenichnyi Algorithm (solid) on Problem 5.1.**

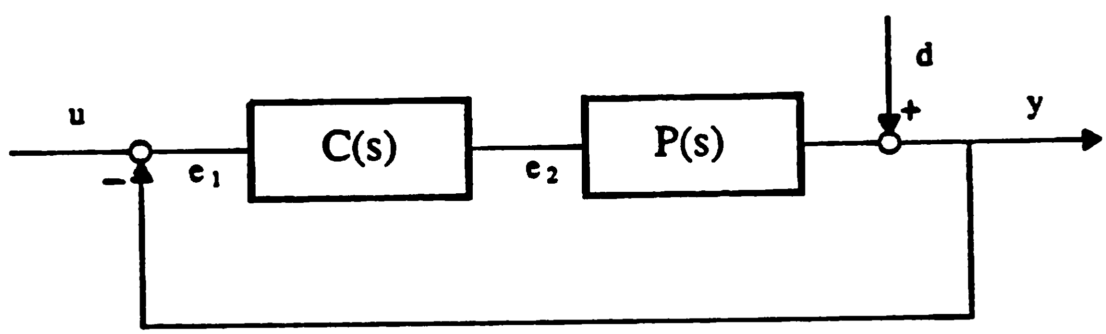

Figure 2: A feedback system.

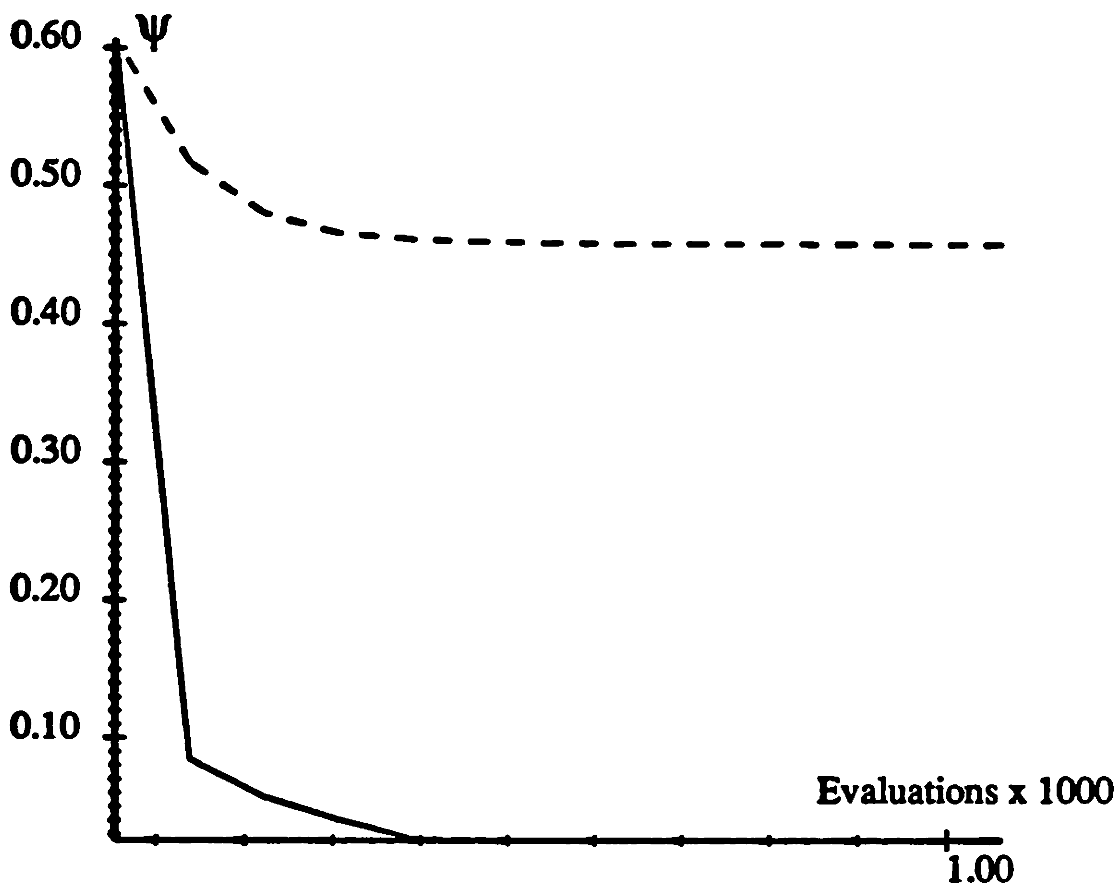

Figure 3: Peformance of the Pshenichnyi Algorithm (dashed) and the Variable-Metric-Pshenichnyi Algorithm (solid) on Problem 5.2.