Optimization of a Constrained Quadratic Function

Charles Hamaker

Department of Mathematics and Computer Science

PO Box 3517 Saint Mary's College of California

Moraga, CA,94575

chamaker@stmarys-ca.edu

Dedicated on the occasion of his 60th birthday to Neculai Andrei in appreciation of a lifetime of contributions to mathematics, as a productive researcher, an author of valuable texts and software, and a leader in the mathematical community.

Abstract: For A a positive definite, symmetric n x n matrix and b a real n-vector, the objective function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^t A\mathbf{x} + \mathbf{b}^t\mathbf{x}$ is optimized over the unit sphere. The proposed iterative methods, based on the gradient of f, converge in general for maximization and for large |b| for minimization with the principal computational cost being one or two matrix-vector multiplications per iteration. The rate of convergence improves as |b| increases, becoming computationally competitive in that case with algorithms developed for the more general problem wherein A may be indefinite.

Keywords: constrained optimization, quadratic functions, iterated gradients, acceleration of convergence.

Charles Hamaker works in applicable analysis, particularly the application of Radon transforms to mathematical tomography and differential equations. He is also the director of his college's core undergraduate program in reading classical texts of western culture.

1. Introduction and Preliminaries

Let A be a real-valued, positive definite, symmetric n x n matrix, and let **b** be a real-valued n-vector. The problem under consideration here is:

Optimize $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^t A\mathbf{x} + \mathbf{b}^t\mathbf{x}$ over $S^{n-1} = \{\mathbf{x} \in R^n : |\mathbf{x}| = 1\}$.

For small dimensions, the well-established approach involving diagonalization of A, outlined in [Golub and Van Loan 1996] and sketched in the remark preceding Lemma 1, suffices for both maximization and minimization. For large dimensions, the computational cost of diagonalization is excessive.

The maximization problem was encountered for n=4 in computations for lattice gauge simulations [Montero, 1999]. With A allowed to be indefinite, the minimization problem is the trust region subproblem which must be solved in each iteration of a trust region method [Conn, Gould, Toint, 2000] and [Hager, 2001]. The subproblem has been the subject of considerable attention, resulting in sophisticated iterative algorithms (see [Hager, 2001], [Sorenson, 1997], and [Rendl and Wolkowicz, 1997]). For instance, the SSM algorithm described in [Hager, 2001] converges quadratically, using the Lanczos process for startup and with each iterate being the solution of the problem on a low-dimensional subspace. That subspace is determined in part by applying Newton's method at the previous iterate to the associated Lagrange problem. As is typical of these algorithms, the tradeoff for the small number of iterations required is that initialization and each iteration require a significant number of matrix-vector multiplications. See [Hager, 2001] for a comparison of the computational effectiveness of some of these algorithms. These minimization algorithms are adaptable to the maximization problem presumably with similar computational cost.

Denoting the eigenvalues of A by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$, direct the corresponding orthonormal basis of eigenvectors $\{\mathbf{e^1}, \cdots, \mathbf{e^n}\}$ so that

$b_j = \mathbf{b}^t\mathbf{e}^j \geq 0, j = 1, \cdots, n,$

and in the case that $\lambda_i = \lambda_{i+1} = \ldots = \lambda_j$, so that

$$b_i \ge 0 \ \ and \ \ b_{i+1} = \ldots = b_j = 0$$

and let E denote the orthogonal matrix having the e^j as columns. Then, since the change of coordinates $\mathbf{y} = E^t\mathbf{x}$ diagonalizes A and preserves the constraint set S^{n-1}, it can be observed that

$$f(\mathbf{x}) = \frac{1}{2}(E\mathbf{y})^t A(E\mathbf{y}) + \mathbf{b}^t(E\mathbf{y})$$

$$= \sum_{j=1}^{n} \left(\frac{1}{2}\lambda_j y_j^2 + b_j y_j\right).$$

Thus, with respect to the coordinates $x_j = \mathbf{x}^t e_j$, the problem takes the diagonalized form:

Optimize $f(\mathbf{x}) = \sum_{j=1}^{n} (\frac{1}{2}\lambda_j x_j^2 + b_j x_j)$ over S^{n-1}

where λ_j and b_j are as above.

It is convenient to adopt the following notations. For $\mathbf{x} \in R^n$, representation in coordinates will be with respect to the orthonormal eigenbasis $\{\mathbf{e}^1, \cdots, \mathbf{e}^n\}$, and will be denoted by $\mathbf{x} = [x_i]_i$. The outer subscript will be omitted when no ambiguity results, e.g. $A\mathbf{x} = [\lambda_i x_i]$ while $[x_i y_j]_j = x_i[y_j]$. The euclidean norm of a vector **x** will be denoted by |**x**|. The subset of S^{n-1} consisting of all vectors with non-negative (non-positive) coordinates will be denoted by $S_+^{n-1}(S_-^{n-1})$.

Geometric insight into the problem follows from completing the square in the diagonalized form of $f(x)$ to obtain

$$f(\mathbf{x}) = \sum_{j=1}^{n} \frac{1}{2}\lambda_j (x_j + \frac{b_j}{\lambda_j})^2 - k \qquad \text{where}$$

$$k = \sum_{j=1}^{n} \frac{b_j^2}{2\lambda_j} \ge 0.$$

Thus, for *c*>-*k*, the level hypersurface *f*(**x**) = *c* is an ellipsoid with center at $[-\frac{b_j}{\lambda_j}]$.

It is the simplicity of this geometric characterization that suggests the iterative algorithm for the maximization problem: $\mathbf{x}^{i+1} = \mathbf{T}(\mathbf{x}^i)$ where $\mathbf{T}(\mathbf{x})$ is the normalization of the gradient $\nabla f(x)$. Note that the solution is a fixed point for **T**. While convergence in general will be established, it is reasonable to expect that the rate will improve as |**b**| increases, moving the common center of the ellipsoids further from the origin and thus resulting in less variation of $\nabla f(x)$ over the constraint set S^{n-1}. The algorithm for the minimization problem (See section 4) is based on similar heuristic considerations.

Characterization of the maximizing vector $\mathbf{w} \in S^{n-1}$ can begin by observing in the diagonalized form of the problem that it must have non-negative coordinates because the b_j / λ_j are positive, and furthermore must satisfy the Lagrange multiplier condition:

$$\nabla f(x) = \mu w \quad \text{for some} \quad \mu \ne 0 \qquad (1)$$

Since $\nabla f(\mathbf{x}) = A\mathbf{x} + \mathbf{b} = [\lambda_j x_j + b_j]$, this implies

$$w_j = \frac{b_j}{\mu - \lambda_j} \quad unless \quad \mu = \lambda_j. \qquad (2)$$

This leads to defining

$$g(t) = \sum_{j=1}^{n} \frac{b_j^2}{(t - \lambda_j)^2}$$

and observing that the requirement that $\mathbf{w} \in S^{n-1}$ means that μ must satisfy the secular equation g(t)=1.

If b_1>0, then the requirement that $w_1 \ge 0$ implies that $\mu > \lambda_1$, and hence that all the w_i are defined by (2). Because g(t) is unbounded and decreasing on $[\lambda_1, \infty]$ and approaches 0 as $t \to \infty$, there exists a unique μ_0 satisfying $g(\mu_0) = 1, \mu_0 > \lambda_1$. Thus there is a unique maximizing point w.

In the case that $b_1 = \ldots = b_{k-1} = 0$ and $b_k > 0$, the requirement that $w_k \ge 0$ implies $\mu > \lambda_k$, and there is a unique μ_0 satisfying

$g(\mu_0) = 1, \mu_0 > \lambda_k$. Then (2) can be satisfied for $i < k$ in one of two ways. First, if $\mu \neq \lambda_i$, then $w_i = 0$. Second, if $\mu = \lambda_i$ and $g(\lambda_i) < 1$ (in which case $\mu > \mu_0$) the unit vectors $\mathbf{w}^i \in S_+^{n-1}$ determined by the Lagrange condition have coordinates w_j^i determined by (2) for $\lambda_j \neq \lambda_i$, while $\sum_{\lambda_j = \lambda_i} (w_j^i)^2 = 1 - g(\lambda_i)$. Thus $f(\mathbf{w}^i) = h(\lambda_i)$, where

$$h(\lambda) = \frac{1}{2}(\lambda + \sum_{j=k}^{n} \frac{b_j^2}{\lambda - \lambda_j}) \ for\ \lambda \geq \mu_0.$$

For $\mathbf{w}^0$ the unit vector determined by μ_0 in (2), $f(\mathbf{w}^0) = h(\mu_0)$, and $h'(\lambda) = \frac{1}{2}(1 - g(\lambda))$ is positive for $\lambda > \mu_0$. Thus the desired maximum corresponds to the largest value of μ satisfying (1), namely the larger of μ_0 and λ_1.

Remark: Assuming that the problem has been diagonalized, i.e. the eigenvectors $\mathbf{e}^i$ have been found, and that $b_1 > 0$, μ in equation (1) can be found by solving the secular equation g(t)=1 and then used in equation (2) to find w. This approach is outlined in [Golub and Van Loan, 1996] for the minimization problem and used in [Montero, 1999] to solve the maximization problem for $n = 4$.

The resulting characterization of the maximizing vector is summarized as follows.

Lemma 1

1. If $b_1 > 0$, then the maximizing vector w satisfies $w_1 > 0$ and the corresponding Lagrange multiplier satisfies $\mu > \lambda_1$.

2. If $b_1 = 0$ and $\mu_0 \geq \lambda_1$, then $w_1 = 0$ and $\mu = \mu_0$.

3. If $b_1 = 0$ and $\mu_0 < \lambda_1$, then $\mu = \lambda_1$ with the set S of maximizing vectors consisting of all w with w_j determined by (2) for $\lambda_j < \lambda_1$ and satisfying $\sum_{\lambda_j = \lambda_1} (w_j)^2 = 1 - g(\lambda_1)$.

Characterization of the minimizing vector v as corresponding via (2) to the smallest value ν which satisfies (1) is completely similar, and is given in Lemma 2.2 of [Hager 2001]. In that paper, the counterparts to cases 1, 2, and 3. are referred to as *non-degenerate, non-degenerate degenerate*, and *degenerate*, and the literature on the trust region subproblem refers to cases 2 and 3 collectively as the *hard case*.

It is useful to characterize the optimizing vectors and their multipliers for |**b**| large.

Lemma 2 *For $\mathbf{u} \in S_+^{n-1}$, let b= τ u, τ > 0. Then, for fixed A,*

1. $\lim_{\tau\to\infty} \frac{\mu}{\tau} = 1$ *and* $\lim_{\tau\to\infty} w = u$.

2. $\lim_{\tau\to\infty} \frac{\nu}{\tau} = -1$ *and* $\lim_{\tau\to\infty} \nu = -u.$

Proof: The secular equation *g(t)* = 1 becomes

$$\sum_{j=1}^{n} \frac{u_j^2}{(t - \lambda_j)^2} = \tau^{-2}.$$

To establish case 1, note first that μ_0, the largest solution of the secular equation, satisfies $\lim_{\tau\to\infty} \mu_0 = \infty$. Thus for τ large enough, $\mu = \mu_0$, and so $\lim_{\tau\to\infty}(u_j/(1 - \frac{\lambda_j}{\mu})) = u_j$. Setting $t = \mu$ in the secular equation, multiplying by μ^2, and taking limits gives

$$\lim_{\tau\to\infty}\left(\frac{\mu}{\tau}\right)^2 = \lim_{\tau\to\infty}\sum_{j=1}^{n}\left(\frac{u_j}{1 - \frac{\lambda_j}{\mu}}\right)^2 = 1.$$

Then

$$\lim_{\tau\to\infty} w_j = \lim_{\tau\to\infty} \frac{\tau u_j}{\mu - \lambda_j} = \lim_{\tau\to\infty} \frac{\tau}{\mu} \frac{u_j}{1 - \frac{\lambda_j}{\mu}} = u_j.$$

Case 2 is established similarly. □

2. Maximizing

The main results pertain to the convergence to a maximizing vector of the iterates of **T**, where

$$\mathbf{T}(\mathbf{x}) = \frac{\nabla f(\mathbf{x})}{|\nabla f(\mathbf{x})|} \, for \ \mathbf{x} \in S^{n-1}.$$

See figure 1. Note that the fixed points of **T** are those on S^{n-1} which satisfy the Lagrange multiplier condition and also that computation of **T** requires only multiplication by A and simple vector operations. In general, the initial value is $\mathbf{x}^0 = \mathbf{b}/|\mathbf{b}|$.

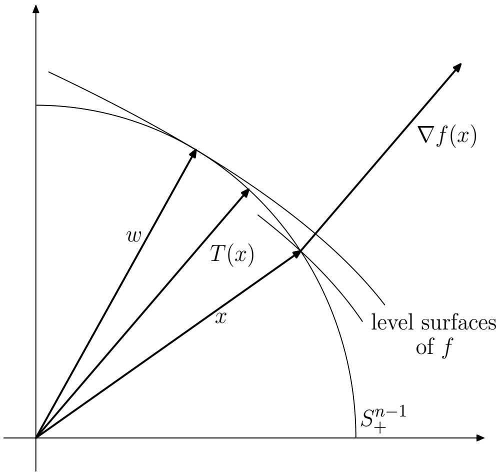

Figure 1. An iteration of T

Lemma 3 *For* $\mathbf{x} \in S^{n-1}$ *not a fixed point of* **T**, $f(T(x)) > f(x)$

Proof: Let $\mathbf{x} \in S^{n-1}$ and **y**=**T**(**x**). Then by Taylor's Theorem,

$$f(y) = f(\mathbf{x}) + \nabla f(\mathbf{x})^t(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^t A(\mathbf{y}-\mathbf{x})$$
$$= f(\mathbf{x}) + (|\nabla f(\mathbf{x})| - \nabla f(\mathbf{x})^t \mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^t A(\mathbf{y}-\mathbf{x}).$$

The second and third terms in the sum above are non-negative, the former by the Cauchy-Schwarz inequality, equaling 0 only if **y**=**x**, i.e. if **x** is a fixed point of **T**. □

Proposition 1 *Let* $\mathbf{w} \in S^{n-1}$ *satisfy equation (1) with multiplier* μ, *and let P be the orthogonal projection onto* $\mathbf{w}^{\perp}$, *the orthogonal complement of* **w**. *Then for* $\mathbf{x} \in S^{n-1}$ *near* **w**,

$$\mathbf{T}(\mathbf{x}) - \mathbf{w} = \frac{1}{\mu}P(A(\mathbf{x}-\mathbf{w})) + \mathbf{O}(|\,\mathbf{x}-\mathbf{w}\,|^2). \quad (3)$$

Proof: It straightforward to check that

$$\frac{\partial \mathbf{T}_i}{\partial x_j}(\mathbf{x}) = \frac{\lambda_i}{|\nabla f(\mathbf{x})|}\delta_{ij} -$$
$$\frac{(\lambda_i x_i + b_i)}{|\nabla f(\mathbf{x})|^3}\lambda_j(\lambda_j x_j + b_j)$$

where δ_{ij} is the Kronecker delta. Then, since $|\nabla f(\mathbf{w})|=\mu$ and $\lambda_i w_i + b_i = \mu w_i$,

$$\frac{\partial T_i}{\partial x_j}(w) = \frac{1}{\mu}\left(\lambda_i \delta_{ij} - \lambda_j w_i w_j\right).$$

Applying Taylor's Theorem to $\mathbf{T}_i$ at **w** and noting that $\mathbf{T}_i(\mathbf{w})=w_i$ yields

$$T_i(x) - w_i = \frac{1}{\mu}(\lambda_i(x_i - w_i) - w^t A(x-w)w_i) + O|x-w|^2 \quad (4)$$

Therefore,

$$T(x) - w = \frac{1}{\mu}\{A(x-w) - (A(x-w)^t w)w\}$$
$$+ O(|x-w|^2) = \frac{1}{\mu}P(A(x-w)) + O(|x-w|^2)$$

□

The following theorem characterizes the convergence of the iterates of **T** in each of three cases. In case 1, which includes the non-degenerate case and the non-degenerate degenerate case of [Hager 2001], the rate of convergence is linear.

Theorem 1 *Let* $\mathbf{x}^0 \in S^{n-1}_{+}$ *with* $x_1^0 > 0$, *and let* $\mathbf{x}^i = \mathbf{T}^i(\mathbf{x}^0)$.

1. If $b_1 > 0$ *or* $b_1 = 0$ *and* $\mu_0 > \lambda_1$, *then the iterates* $\mathbf{x}^i$ *converge to the unique maximizing* $\mathbf{w} \in S^{n-1}_{+}$. *Furthermore,*

$$\mathbf{x}^{i+1} - \mathbf{w} = |\,\mathbf{x}^i - \mathbf{w}\,|\,B(\mathbf{u}) + \mathbf{O}(|\,\mathbf{x}^i - \mathbf{w}\,|^2)$$

where **u** *is the normalization of* $P(\mathbf{x}^i - \mathbf{w})$ *and* $B(\mathbf{x}) = \frac{1}{\mu}P(A\mathbf{x})$ *is symmetric and positive definite on* $w^{\perp}$ *with operator norm less than* λ_1/μ.

2. If $b_1 = 0$, λ_1 *has multiplicity 1, and* $\mu_0 \le \lambda_1$, *then the iterates* $\mathbf{x}^i$ *converge to the unique maximizing* $\mathbf{w} \in S^{n-1}_{+}$.

3. If b_1*=0,* λ_1 *has multiplicity m>1, and* $\mu_0 \le \lambda_1$*, then the iterates* $\mathbf{x}^i$ *converge to a maximizing* $\mathbf{w} \in S_+^{n-1}$.

Proof: Each of the cases will be considered separately. Let μ denote the multiplier corresponding to a maximizing vector.

As a preliminary observation, note that Lemma 3 implies that any convergent subsequence of $\{\mathbf{x}^i\}$ must converge to a fixed point of **T**. Note also that, because

$$T(x)_1 = \frac{\lambda_1 x_1 + b_1}{|\nabla f(x)|} ,$$

$x_1^0 > 0$ implies that $x_1^i > 0$.

In case 1, $\mu > \lambda_1$, and so the maximizing vector **w** is the unique vector in S_+^{n-1} which satisfies the Lagrange condition (1). Since each $\mathbf{x}^i \in S_+^{n-1}$ and Lemma 3 shows that the sequence $f(\mathbf{x}^i)$ is increasing, the limit of that sequence is $f(\mathbf{w})$. Therefore, $\lim \mathbf{x}^i = \mathbf{w}$. That the rate of convergence is linear follows from Proposition 1, whose main result can be rewritten

$$T(x) - w = |x - w| \left(\frac{1}{\mu} P(A) \frac{x - w}{|x - w|} \right) \qquad (5)$$
$$+ O\left(|x - w|^2\right).$$

Let $u^1 = P(\frac{x - w}{|x - w|})$.

Then, since $(x - w)^t w = \frac{-1}{2}|x - w|^2$ for any unit vectors **x** and **w**, it follows that $x - w = |x - w| u^1 + O(|x - w|^2)$. In fact, since $|\mathbf{u}^1| = 1 + O(|\mathbf{x}-\mathbf{w}|^2)$, $\mathbf{u}^1$ can be replaced by its normalization **u** in the previous statement. Therefore

$$\mathbf{T}(\mathbf{x}) - \mathbf{w} = |\mathbf{x}-\mathbf{w}| B(\mathbf{u}) + \mathbf{O}(|\mathbf{x}-\mathbf{w}|^2) \quad . \qquad (6)$$

It is straightforward to verify that the restriction of B to $\mathbf{w}^\perp$ is symmetric and positive definite and that the norm of B is less than or equal to $\frac{\lambda_1}{\mu}$, which in this case is less than 1.

In case 2, $\mu = \lambda_1$, and the maximizing vector **w** is unique.

If $\mu_0 = \lambda_1$, then the maximizing vector **w** is the unique element of S_+^{n-1} which satisfies the Lagrange condition. Therefore, the iterates converge to **w** as in the previous case.

If $\mu_0 < \lambda_1$, then there are other $z \in S_+^{n-1}$ which satisfy the Lagrange condition of (1). For such a vector, Lemma 1 shows that $z_1 = 0$ and that its multiplier ρ in (1) is less than λ_1. However, no subsequence of the iterates x^i can converge to such **z**. This follows from the equation (4) with *i*=1 and **w**=**z**, which implies

$$\mathbf{T}_1(\mathbf{x}) - z_1 = \frac{\lambda_1}{\rho} x_1 + O(|\,\mathbf{x} - \mathbf{z}\,|^2).$$

Therefore, since $x_1^i > 0$, no subsequence of x_1^i can converge to 0. By the preliminary observation, $\lim_{i\to\infty} \mathbf{x}^i = \mathbf{w}$.

In case 3, recall from Lemma 1 that the $m-1$ dimensional sphere S of maximizing vectors consists of all **w** satisfying $w_1^2 + \ldots + w_m^2 = 1 - g(\lambda_1)$, having multiplier $\mu = \lambda_1$, and having $w_j, j > m$, uniquely determined by equation (2). By the argument given in the previous case, $\lim_{j\to\infty} f(\mathbf{x}^i) = f(\mathbf{w}^0)$ for some $w^0 \in S$. Observe that, for $j \le m$, $\mathbf{T}_j(\mathbf{x}) = s x_j$ where $s = \lambda_1 / |\nabla f(x)|$. Let **w** now denote the vector determined by equation (2) for $j > m$, and by $w_j = l x_j^0$ where

$$l = \left((1 - g(\lambda_1)) / \sum_{j=k}^{m} (x_k^0)^2 \right)^{\frac{1}{2}}$$

for $j \le m$. Then any convergent subsequence of the $\mathbf{x}^i$ must converge to some $\mathbf{w} \in S$. By straightforward compactness considerations, $\lim_{i\to\infty} \mathbf{x}^i = \mathbf{w}$. □

Note for A fixed that as a consequence of the Lemma 2, part 1, the larger $|\mathbf{b}|$ is, the smaller the bound λ_1 / μ in case 1 of Theorem 1. Thus the rate of convergence for the iterates of **T** should improve as $|\mathbf{b}| \to \infty$. This observation is consonant with the recognition that ∇f is nearly constant on S^{n-1} for sufficiently large **b**, i.e that $\nabla f(\mathbf{x}) \approx \nabla f(\mathbf{w})$ and so $\mathbf{T}(\mathbf{x}) \approx \mathbf{w}$.

3. Acceleration

In the case of linear convergence, i.e. $\mu > \lambda_1$, this section introduces an acceleration of the iterates which is an extrapolation based on heuristic considerations, and offers a partial analysis of its characteristics.

From Theorem 1, it can be verified that, for **x** near **w**, **u** the normalization of $P(\mathbf{x}-\mathbf{w})$, and $k \in Z^+$,

$$\mathbf{T}^k(\mathbf{x})-\mathbf{w}=|\mathbf{x}-\mathbf{w}|\,B^k(\mathbf{u})+\mathbf{O}(|\mathbf{x}-\mathbf{w}|^2).$$

Thus, the convergence of the iterates of **T** to **w** is equivalent to that of the iterates of B to **0**. Since the restriction of B to $w^{\perp}$ is linear, symmetric, and positive definite and has operator norm less than 1, it has a complete set $1 > \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{n-1} > 0$ of eigenvalues with corresponding orthonormal eigenvectors $\{v^1, \ldots, v^{n-1}\}$ which are a basis for $w^{\perp}$. Then, for $u = \sum_{i=1}^{n-1} c_i v^i$, it follows that

$$B^k(u) = \sum_{i=1}^{n-1} \sigma_i^k c_i v^i, \text{ where } \sum_{i=1}^{n-1} c_i^2 = 1$$

Let $\mathbf{u}^k = B^k(\mathbf{u})$ and $\mathbf{d}^k = \mathbf{u}^{k+1} - \mathbf{u}^k$. Accepting the approximation $u^{k+1} \approx ru^k$ for some constant $0<r<1$, it follows that $d^k \approx (r-1)u^k$ and that $d^{k+1} \approx rd^k$. Then $r \approx |d^1|/|d^0|$. From $u^k = \frac{1}{r-1} d^k$, it follows that

$$u^2 \approx \frac{1}{r-1} d^2 = \frac{1}{r-1}(rd^1 + (d^2 - rd^1))$$

$$\approx \frac{r}{r-1}(d^1 + (d^1 - rd^0))$$

The term $\mathbf{d}^1 - r\mathbf{d}^0$ can be viewed as a correction to the implicit, but false assumption of the collinearity of the $\mathbf{u}^k$. The approximation above suggests that

$$z = u^2 + \frac{r}{1-r}(2d^1 - rd^0) \approx 0$$

be used as an acceleration of $\{u^0, u^1, u^2\}$. Expressing this approximation in terms of iterates of **T** yields an acceleration for their convergence to **w**:

$$(\mathbf{T}^2(\mathbf{x}) - \mathbf{w}) + \frac{r}{1-r}\left(2(\mathbf{T}^2(\mathbf{x}) - \mathbf{T}(\mathbf{x})) - r(\mathbf{T}(\mathbf{x}) - \mathbf{x})\right) \approx 0$$

or

$$\mathbf{w} \approx \frac{1}{1-r}\left((1+r)\mathbf{T}^2(\mathbf{x}) - r(r+2)\mathbf{T}(\mathbf{x}) + r^2\mathbf{x}\right) \tag{7}$$

where

$$r = \frac{|T^2(x) - T(x)|}{|T(x) - x|} .$$

The approximation to **w** given by the right hand side of (7) should be normalized before continuing to the next iteration.

The comparison of **z** to $\mathbf{u}^2$ begins by noting that

$$z = \frac{1}{1-r}\sum_{i=1}^{n-1}\left((1+r)\sigma_i^2 - (2r+r^2)\sigma_i + r^2\right)c_i \mathbf{v}^i$$

$$= \sum_{i=1}^{n-1} h(\sigma_i)\sigma_i^2 c_i v^i$$

where

$$h(\sigma) = \frac{1}{1-r}\left((1+r) - (2r+r^2)\sigma_i^{-1} + r^2\sigma^{-2}\right)$$

Then $h\left(\frac{r}{2}\right) = 1$, $h(r) = h\left(\frac{r}{r+1}\right) = 0$ and the minimum of $h(\sigma)$ occurs at $\sigma = \frac{2r}{r+2}$ and equals $\frac{-r^2}{4(1-r)}$. Now several useful observations can be made.

1. Since $r^2 = \frac{\sum_{i=1}^{n-1} \sigma_i^2 (1-\sigma_i)^2 c_i^2}{\sum_{i=1}^{n-1} (1-\sigma_i)^2 c_i^2}$,

it follows that $0 < r < \sigma_1 \leq \frac{\lambda_1}{\mu} < 1$.

2. For $\sigma_i > r/2$, the i^{th} component of **z** is smaller than that of $\mathbf{u}^2$, dramatically so for σ_i near $r/(r+1)$ or r. For $\sigma_i < r/2$, the i^{th} component of **z** is larger than that of $\mathbf{u}^2$, but

smaller than that of $\frac{r^2}{1-r}u^0$. Since r can be expected to be closest to the σ_i for which the c_i are largest, |**z**| can be expected to be smaller than |**u**^{2}|.

3. If the slower convergence for components of **u** corresponding to small σ_i is problematical, several iterations of B (or equivalently **T**) would yield rapid convergence in those components. Applying acceleration then would improve convergence for components corresponding to large σ_i.

Another approach, suggested by the SSM algorithm of [Hager 2001] to producing an improved approximation to the maximizing vector **w** from the triple $\{\mathbf{x}, \mathbf{T}(\mathbf{x}), \mathbf{T}^2(\mathbf{x})\}$ would be to find the maximizing vector **z** for the restriction of *f* to the subspace *W* spanned by the triple. Specifically, if $\{v^1, v^2, v^3\}$ is an orthonormal basis for *W* and *V*=[$\mathbf{v}^1\mathbf{v}^2\mathbf{v}^3$] is the matrix with those vectors as columns, one would

maximize $g(y) = \frac{1}{2}y^t By + c^t y$ over S^2,

where $B = V^t AV, \mathbf{c} = V^t\mathbf{b}$ and $\mathbf{y} \in R^3$. This three dimensional problem can be then solved by complete diagonalization, yielding a better approximation **z**=*V***y** than that of the method previously discussed in this section.

This approach has two drawbacks. First, the triple $\{\mathbf{x}, \mathbf{T}(\mathbf{x}), \mathbf{T}^2(\mathbf{x})\}$, being members of a convergent sequence, will be nearly collinear. Thus calculating the orthonormal basis for *W* will be numerically unstable. Second, the computation of B will require one matrix-vector multiplication in addition to the two needed to compute the original triple, thus raising the computational cost. The numerical tests in Section 5 indicate that improvement in the rate of convergence from subspace optimization, if any, is outweighed by the additional computational cost.

4. Minimization

To minimize $f(\mathbf{x}), \mathbf{x} \in S^{n-1}$, a similar iterative algorithm is proposed, one that is effective mainly for large **b**, one that also requires matrix-vector multiplications and simple vector operations. An iterate *R*(**x**) for S_-^{n-1} is defined as follows. In general, the initial value is x^0 = -b/│b│.

1. Let *t*(**x**) be the value that minimizes *f*(**x**+*t*∇*f*(**x**)). It is straightforward to show that

$$t(x) = \frac{-\nabla f(x)^t \nabla f(x)}{\nabla f(x)^t A \nabla f(x)}.$$

2. Fix *r*>0, and let **V**(**x**)=**x**+*r t*(**x**)∇*f*(**x**). Note that $\mathbf{V}(\mathbf{x}) \in S_-^{n-1}$.

3. Set $R(x) = \frac{V(x)}{|V(x)|}$ (See figure 2.)

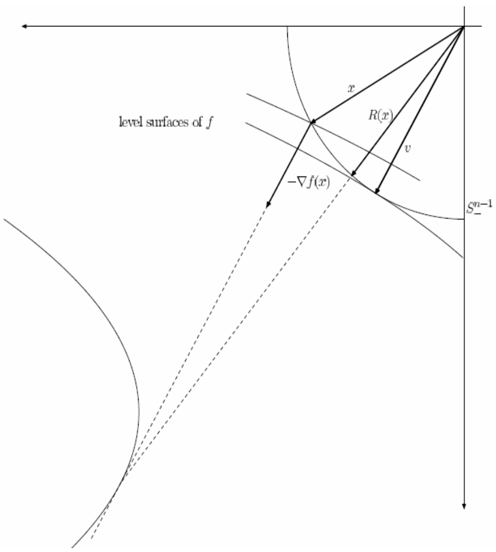

Figure 2. An iteration of R

Iterates of −**T** were not used due to ineffectiveness when |**b**| is not large. First if the common center of the elliptical level surfaces of *f* is inside S^{n-1}, then $-T \in S_+^{n-1}$, while the minimizer $v \in S_-^{n-1}$. More importantly, since the level surfaces of *f* are concave with respect to the origin near S_-^{n-1}, their curvature for intermediate values of |**b**| will cause excessive variation of −**T**, whereas iterates of **R** will vary less. It may be that the respective iterates behave similarly when |**b**| is large.

The discussion below of the convergence of iterates of **R** will consider the case where ν<λ$_n$, which assures that the minimizing vector **v** is the only fixed point of **R** in S_-^{n-1}.

As was done for **T**, the convergence characteristics of **R** will be examined via Taylor's formula applied at **v**, which is a fixed point of **R**. The result is:

Proposition 2 *Let* $\mathbf{v} \in S_{-}^{n-1}$ *satisfy the Lagrange condition (1) with multiplier* ν, *and let P be the orthogonal projection onto* $\mathbf{v}^{\perp}$. *Then with* $\mathbf{x} \in S_{-}^{n-1}$ *near* **v**, k=$\mathbf{v}^t A\mathbf{v}$, *and r>0 chosen such that* $l = 1 - \frac{r}{k}\nu$ *is positive,*

$$R(x) - v = \frac{1}{l}P\left((I - \frac{r}{k}A)(x - v) \right) + O\left(|x - v|^2\right)$$

Proof: Observe that

$$\frac{\partial \mathbf{R}}{\partial x_i} = |\mathbf{V}|^{-3}\left(-(\mathbf{V}^t \frac{\partial \mathbf{V}}{\partial x_i})\mathbf{V} + |\mathbf{V}|^2 \frac{\partial \mathbf{V}}{\partial x_i} \right)$$

and

$$\frac{\partial \mathbf{V}}{\partial x_i} = \mathbf{e}_i + rt\frac{\partial}{\partial x_i}(\nabla f) + r\frac{\partial t}{\partial x_i}\nabla f.$$

Evaluating at **x**=**v** and noting that ∇*f*(**v**)=ν**v** yield $t = \frac{-1}{k}$ and

$$\frac{\partial V}{\partial x_i} = (1 - r\frac{\lambda_i}{k})e^i + r\frac{\partial t}{\partial x_i}\nu \ \mathbf{v}.$$

Since **V**=**v**+*rt*(ν**v**)=*l***v** and $\mathbf{V} \in S_{-}^{n-1}$, it follows that *l* > 0. Then

$$\frac{\partial \mathbf{R}}{\partial x_i} = \frac{1}{l}\left(-(\mathbf{v}^t \frac{\partial \mathbf{V}}{\partial x_i})\mathbf{v} + \frac{\partial \mathbf{V}}{\partial x_i} \right) =$$

$$= \frac{1}{l}(1 - r\frac{\lambda_i}{k})(\mathbf{e}^i - v_i \mathbf{v}).$$

Thus,

$$\frac{\partial R_j}{\partial x_i} = \frac{1}{l}(1 - r\frac{\lambda_i}{k})(\delta_{ij} - v_i v_j), \ \text{ and so}$$

$$R(x) - v = \left[\sum_{i=1}^{n} \left(\frac{\partial R_j}{\partial x_i}(x_i - v_i) \right) \right]_j + O\left(|x - v|^2\right)$$

$$= \frac{1}{l}\left((I - \frac{r}{k}A)(x - v) - (v^t (I - \frac{r}{k}A)(x - v)) \right)$$

$$+ O(|x - v|^2)$$

$$= \frac{1}{l}P\left((I - \frac{r}{k}A)(x - v) \right) + O(|x - v|^2) \ .□$$

In practice, the iterates of **R**(**x**) converge, slowly for small |**b**| and rapidly for large |**b**|. The latter is confirmed as follows.

Lemma 4 *For* $\mathbf{u} \in S_{+}^{n-1}$, *let* $\mathbf{b} = \tau\,\mathbf{u}$, $\tau > 0$. *Then* $\lim_{\tau\to\infty} l = \infty$.

Proof:

$$\lim_{\tau\to\infty} \frac{l}{\tau} = \lim_{\tau\to\infty}\left(\frac{1}{\tau} - r\frac{\nu}{k\tau} \right) =$$

$$\lim_{\tau\to\infty} \frac{r}{v^t A v} = \frac{r}{u^t A u},$$

with the latter two equalities following from part 2 of Lemma 2. □

It is then clear from Proposition 2 that the convergence of the iterates of *R* is linear with rate decreasing to 0 as |**b**| increases.

5. Numerical Results

Preliminary tests confirmed that convergence properties of both algorithms were independent of choice of coordinates, i.e. of whether or not the matrix A was diagonalized. Thus, the algorithms were tested using diagonalizing coordinates. The spectrum of *A* was generated randomly within the interval (0,2], with the exception that λ_1=2.02 and the coordinates of **b** were generated randomly with |**b**| then being set to a desired value. The termination criterion was $|\mathbf{x}^{i+1} - \mathbf{x}^i| < \varepsilon$ where $\varepsilon = 10^{-10}$ for **T** and $\varepsilon = 10^{-9}$ for **R**. (The less stringent tolerance for **R** was an expedient due to unstable convergence when |**b**| was small.) Since the problems were diagonalized, the optimizing vector **w** (or **v**) could be computed by solving the secular equation *g*(*t*)=1 for μ (or ν) and using equation (2). Except when |**b**|≤0.1, the terminal iterates differed in norm from those corresponding solutions by no more than 10^{-8}.

Results for iterates of T:

Iteration started with x^0 = b/│b│, two iterations of **T** were computed initially, and thereafter two successive iterates were used in equation (7) to compute an accelerated iterate. Note that the major computational cost of a simple iteration is one matrix-vector multiplication, while an extrapolated iterate requires two.

n=1000, number of trials = 100

case	|**b**|	0.3	0.5	1	3	5	10	20
general	extrapolation MV	151.9	68.7	32.1	16.0	12.0	10.0	8.0
"	subspace MV	395.7	144.7	53.4	20.4	17.0	11.0	11.0
hard	extrapolation MV	158.4	73.1	34.5	17.3	14.0	10.0	8.0
"	subspace MV	503.6	167.1	59.4	23.0	17.0	14.0	11.0

Table 1. Average matrix-vector products (MV) per trial for convergence of T iterates.

1. The simple iterates of **T** and the accelerated iterates monotonically converged coordinatewise to the limit, except possibly in the case where b_1 was very small. The use of accelerated iteration by extrapolation decreased the number of matrix-vector multiplications, compared to simple iteration, with the difference being less as |**b**| increased.

2. The dimension n had only a small effect on the rate of convergence. For example, over 10 problems with |**b**|=1 the average number of accelerated iterates was 31.6 for n=10, 40.2 for n=50, and 44.2 for n=100. This suggests that convergence with respect to the uniform norm, i.e. coordinatewise, was essentially independent of n.

3. The dominant factor affecting the rate of convergence is |**b**|. See Table 1, which was generated by solving 100 problems with n=1000. The number of matrix-vector multiplications was dramatically low for |**b**| large and prohibitively high for |**b**| small.

4. As Table 1 indicates computational cost of solving the three dimensional problem on the subspace determined by a pair of iterates (as describe at the end of Section 3) is worse than that of accelerating the pair.

5. A series of "hard" problems was generated by setting b_1=0, b_2 to be the largest coordinate of $\mathbf{b}, \lambda_1 = 2.02$, $\lambda_2 = 2.018$, and $n = 1000$ The results, given in Table 1, indicate that the computational cost is only slightly higher, again being comparatively worst when |**b**| is small.

Results for iterates of R:

Iteration started with $\mathbf{x}^0$=−**b**/|**b**|, two iterations of **R** were computed initially, and thereafter two successive iterates were used in equation (7) to compute an accelerated iterate. Note that the major computational cost of a simple iteration is two matrix-vector multiplications, while an extrapolated iterate requires four.

1. As for the iterates of **T**, |**b**| was the dominant factor for the rate of convergence of the iterates of **R**, with the rate increasing with |**b**|. See Table 2.

2. The rate of convergence was affected by the choice of the mollifier r, but did not vary greatly as r varied. In Table 2, the mollifier that resulted in the minimum of the average number of matrix-vector multiplications per trial is given, but the average varied only slightly over a significant range for r.

3. Subspace optimization, as an alternative to extrapolation, did not perform well, presumably due to numerical instability. Note however that for |**b**|=0.3 convergence was successful using this alternative, though it failed using extrapolation.

4. As might be expected from Proposition 2, for small |**b**|, the diagonalized coordinates of the simple iterates alternated above and below the limit. In those circumstances, the extrapolated iterates either converged slowly or not at all, while simple iteration converged very slowly. For intermediate

values of |**b**|, decreasing the mollifier *r* eliminated this alternation.

5. A series of hard problems was generated by setting b_n=0, λ_n=0.02, and randomly generating $\lambda_i \in (0.02,2.02]$, $1 \le i < n$.. No appreciable degradation of performance, compared to the general case, occurred.

n=1000, number of trials = 100

case	|**b**|	0.3	0.5	1	3	5	10	20
general	extrapolation MV	*	265.6	70.2	28.04	24.0	20.0	16.0
	mollifier r	*	1.0	1.5	1.0	1.0	1.0	1.0
"	subspace MV	477.7	206.2	82.3	34.0	29.0	24.0	29.4
	mollifier r	0.8	0.8	1.0	1.0	1.0	0.2	1.2
hard	extrapolation MV	*	178.7	60.0	28.0	24.0	20.0	16.0
	mollifier r	*	0.8	0.5	1.0	1.0	1.0	1.0
"	subspace MV	450.7	202.4	82.1	34.0	29.0	24.0	19.0
	mollifier r	0.8	0.8	1.0	0.5	0.6	0.2	0.2

Table 2. Average matrix-vector products (MV) per trial for convergence of R iterates. (*: Convergence was not successful.)

6. Conclusion

For the positive-definite quadratic problem that has been treated, the iterative algorithms, based on **T** for maximization and on **R** for minimization, converge reliably to a solution except when |**b**| is small relative to the matrix norm of *A*. The linear rate of convergence, even when enhanced by extrapolation, results in more iterations than the superlinear or quadratic algorithms developed for the trust region subproblem (i.e minimization with *A* indefinite), but the computational cost per iteration, one or two matrix-vector multiplications, is significantly lower than the more substantial cost of initialization and iteration of those algorithms. When |**b**| is large relative to the matrix norm of *A*, the proposed algorithms for the problem treated here perform well, when compared to those algorithms.

Application of iterates of **R**, −**T**, or a similar step employing ∇f to the trust region subproblem may have more promise than is initially evident, particularly when |**b**| is large. When the spectrum of *A* has a number of values near 0, the common center of the level elliptical hyperboloids of *f* will be far from the constraint set S^{n-1}. As noted in the remark ending section 2, the small variation of ∇f over S^{n-1} suggests effective convergence. While no analysis has been done, simple numerical tests suggests that convergence occurs. In that case, an effective acceleration algorithm could be developed. It may also be that such a method could be used in conjunction with existing algorithms. For instance the SSM algorithm of [Hager, 2001] converges quadratically, but has an expensive initialization before the iteration phase. If that initialization could be replaced by a few gradient-based iterations, considerable computational savings would result.

7. Acknowledgements

Preliminary results for the maximization algorithm were presented at the Santa Clara University Mathematics Colloquium. I wish to thank Roy Wensley for introducing me to the maximization problem, Aaron Melman for helpful comments, and Paul Dupuis for directing me to the literature on the trust region subproblem. The referee made a number of useful suggestions, about both exposition and substance, for which I am appreciative. The initial work on this project was supported by a sabbatical grant from Saint Mary's College of California.

REFERENCES

1. CONN, A., R., GOULD, N.I.M., TOINT, P.T., [Conn, Gould, Toint, 2000] **Trust Region Methods**, MPS-SIAM Series on Optimization, Philadelphia, 2000.

2. GOLUB, G. and VAN LOAN, C., [Golub and Van Loan, 1996] **Matrix Computations,** 3rd edition, Johns Hopkins University Press, 1996.

3. HAGER, W., [Hager, 2001] **Minimizing a quadratic over a sphere**, SIAM J. Optim., V. 12, nr. 1 (2001), pp. 188 – 208.

4. MONTERO, A., [Montero, 1999] **Study of SU(3) vortex-like configurations with a new maximal center gauge fixing method**, Phys. Letters B467, 1999, pp. 106-111 (also FTUAM-99-16).

5. RENDL, F. and WOLKOWICZ, H., [Rendl and Wolkowicz, 1997] **A semidefinite framework for trust region subproblems with applications to large scale minimization**, Math. Programming, V. 77, nr. 2 series B (1997), pp. 273-299.

6. D. SORENSON, [Sorenson, 1997], **Minimization of a large-scale quadratic function subject to a spherical constraint,** SIAM J. Optim., V. 7, nr. 1 (1997), pp. 141-161.