MAXIMIZING A SECOND-DEGREE POLYNOMIAL ON THE UNIT SPHERE

BY

GEORGE E. FORSYTHE and GENE H. GOLUB

TECHNICAL REPORT CS16

FEBRUARY 5, 1965

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

MAXIMIZING A SECOND-DEGREE POLYNOMIAL ON THE UNIT SPHERE */

by

George E. Forsythe

and

Gene H. Golub

ABSTRACT-

Let A be a hermitian matrix of order n, and b a known vector in C^n. The problem is to determine which vectors make $\Phi(x) = (x-b)^H A(x-b)$ a maximum or minimum on the unit sphere $U = \{x : {}^H xx = 1\}$. The problem is reduced to the determination of a finite point set, the spectrum of (A,b). The theory reduces to the usual theory of hermitian forms when b = 0.

 This report was supported in part by Office of Naval Research Contract Nonr-225(37)(NR-044-211) at Stanford University.

1. The problem.

Let A be a hermitian square matrix of complex elements and order n. Let b be a known n-vector of complex numbers. For each complex n-vector x, the nonhomogeneous quadratic expression

$$\Phi(x) = (x-b)^H A(x-b) \tag{1.1}$$

(H denotes complex conjugate transpose) is a real number. The problem, suggested to us by C. R. Rao of the Indian Statistical Institute, Calcutta, is to maximize (or minimize) $\Phi(x)$ for complex x on the unit sphere $S = \{x: x^H x = 1\}$. Since Φ is a continuous function on the compact set S, such maxima and minima always exist.

In summary, our problem is:

$$\text{maximize or minimize } \Phi(x) \text{ subject to } x^H x = 1 . \tag{1.2}$$

The purpose of this note is to reduce the problem (1.2) to the determination of a certain finite real point set which we shall call the spectrum of the system (A,b) (defined at end of Sec. 1), and show that a unique number λ in the spectrum determines the one or more $x = x^\lambda$ which maximize $\Phi(x)$ for given b. Theorem (4.1) is the main result. The development is an extension to general b of the familiar theory for the homogeneous case when $b = \theta$, the zero vector. No consideration to a practical computer algorithm is given here.

In Sec. 7 we show that determining the least-squares solution of an arbitrary system of linear equations $Cy = f$, subject to the quadratic

constraint $y^H y = 1$, is a special case of problem (1.2).

Phillips (9.2) and Twomey (9.3) begin the actual numerical solution of certain integral equations by approximating them with algebraic problems very closely related to the minimum problem (1.2).

Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be the (necessarily real) eigenvalues of A, and let $\{u_1,\ldots,u_n\}$ be a corresponding real orthonormal set of eigenvectors, with $Au_i = \lambda_i u_i$ $(i=1,\ldots,n)$.

Let a given b be written

$$b = \sum_{i=1}^{n} b_i u_i \quad . \tag{1.3}$$

(1.4) Theorem. If x is any vector in S for which $\Phi(x)$ is stationary with respect to S, then there exists a real number $\lambda = X(x)$ such that

$$A(x-b) = Xx \quad , \tag{1.5}$$

$$x^H x = 1 \quad . \tag{1.6}$$

Conversely, if any real λ and vector x satisfy (1.5, 1.6), then x renders $\Phi(x)$ stationary.

Proof. Let xo be a point of S. Now, as shown in lemma (8.7), $\Phi(x)$ is stationary at x_o with respect to x in S, if and only if there exists a real Lagrange multiplier λ such that $\psi(x) = (x-b)^H A(x-b) - \lambda\, x^H x$ is stationary at x_o with respect to all neighboring complex vectors x. Since

$$0 = \tfrac{1}{2} \operatorname{grad} \psi(x_o) = A(x_o-b) - \lambda x_0 \quad ,$$

the theorem is proved.

To see what conditions are satisfied by the λ of theorem (1.4), we note that the system (1.5,1.6) is equivalent to the system

$$(A-\lambda I)x = Ab \ , \tag{1.7}$$

$$x^H x = 1 \ . \tag{1.8}$$

Let

$$x = \sum_{i=1}^{n} x_i u_i \ .$$

Then (1.7) is equivalent to

$$\sum_{i=1}^{n} (\lambda_i - \lambda) x_i u_i = \sum_{i=1}^{n} \lambda_i b_i u_i \ . \tag{1.9}$$

Definition. By the spectrum of the pair (A,b) we mean the set of all real λ for which there exists an x such that (1.7) and (1.8) are satisfied.

Given any λ, x satisfying (1.7) and (1.8), we shall say that x belongs to λ, and frequently write x in the form x^{λ}.

Note that the spectrum of (A,θ) is the ordinary spectrum $\{\lambda_i\}$ of A.

2. Special case: no $\lambda_i b_i = 0$.

Assume for the present section that $\lambda_i b_i \neq 0$ (all i). This implies that all $\lambda_i \neq 0$, i.e., that A is nonsingular. If λ is in the spectrum of (A,b), (1.9) implies that $\lambda \neq \lambda_i$ for all i, and also that

(2.1) $$X_i = \frac{\lambda_i b_i}{\lambda_i - \lambda} \qquad (i=1,\ldots,n) \ .$$

Then the requirement that

(2.2) $$x^H x = \sum_{i=1}^{n} |x_i|^2 = 1$$

is equivalent to the condition

(2.3) $$g(\lambda) = \sum_{i=1}^{n} \frac{\lambda_i^2 \, |b_i|^2}{|\lambda_i - \lambda|^2} = 1.$$

Although all λ corresponding to stationary values of $\Phi(x)$ are known by theorem (1.4) to be real, it is useful to define $g(\lambda)$ by (2.3) for all complex λ not in $\{\lambda_i\}$.

Let G be the set of complex numbers λ such that $g(X) = 1$. For small enough $\exists\, |b_i|$, G is the union of n simple closed curves in the complex plane, the k-th of which surrounds λ_k. As the $|b_i|$ grow, adjacent curves first coalesce in double points, and then merge into single curves. For very large values of all $|b_i|$, G is one simple closed curve including all $\{\lambda_i\}$ in its interior. The family of sets G resembles the family of lemniscates $\prod_{i=1}^{n} |\lambda - \lambda_i| = \text{const.}$

Note, moreover, that $g(X) > 1$ for λ inside any component curve G_j of G, while $g(X) < 1$ in the exterior of all components G_j of G.

Now we shall show for the special case of Sec. 2 that each λ in G determines a unique x^λ which satisfies (1.7, 1.8). For that x^λ

(2.4) $$\Phi(x^{\lambda}) = f(h) \ ,$$

where we define f by

(2.5) $$f(\lambda) = |\lambda|^2 \sum_{i=1}^{n} \frac{\lambda_i |b_i|^2}{|\lambda_i - \lambda|^2} \ .$$

Fix λ, and drop the superscript λ on x. To prove (2.4), note that (1.7) says $(\lambda_i - \lambda)x_i = \lambda_i b_i$. Thus

$$(\lambda_i - \lambda)(x_i - b_i) = \lambda_i b_i - b_i(\lambda_i - \lambda)$$

$$= \lambda b_i \ .$$

Hence

$$x_i - b_i = \frac{\lambda b_i}{\lambda_i - \lambda} \ ,$$

and

$$\Phi(x) = \sum_{i=1}^{n} \lambda_i |x_i - b_i|^2$$

$$= |\lambda|^2 \sum_{i=1}^{n} \frac{\lambda_i |b_i|^2}{|\lambda_i - \lambda|^2}$$

$$= f(\lambda) \ ,$$

proving (2.4).

Since the Lagrange multipliers λ must be real, the spectrum of (A,b) is the intersection of G with the real axis. This consists of from

2 to 2n distinct real numbers. How many numbers are actually in the spectrum depends on b; this will be discussed in Sec. 5 for n = 2.

We wish to determine which λ in the spectrum corresponds to the maximum [minimum] value of $f(\lambda)$. Let G_j be any component curve of the set G.

(2.6) Theorem. The maximum and minimum real parts of λ, for λ in any one G_j, both occur for λ on the real axis.

Proof. Let $\lambda = \sigma + i\tau$, with σ,τ real. Then

$$g(\lambda) = g_1(\sigma,\tau) = \sum_{i=1}^{n} \frac{\lambda_i^2 |b_i|^2}{(\sigma-\lambda_i)^2 + \tau^2} .$$

Hence, for $\tau > 0$ and fixed σ, $g(\lambda)$ strictly decreases as τ increases. Then in the upper half plane $\tau > 0$, any line σ = constant intersects G_j in exactly one point. The theorem follows from this.

Definition. Let $\Lambda_R[\Lambda_L]$ denote the unique real value of λ of maximum [minimum] real part in the set G.

(2.7) Theorem, Under the assumptions that A is regular (i.e., $\lambda_i \neq 0$ for all i) and $b_i \neq 0$ (i=1,2,...,n), for all λ in G such that $\lambda \neq \Lambda_R$, $\lambda \neq \Lambda_L$ we have

$$f(\Lambda_L) < f(\lambda) < f(\Lambda_R) .$$

Proof. Let $a_i = \lambda_i^2 |b_i|^2$ (i=1,...,n). Introduce two independent complex variables λ,μ, where μ will later be set equal to $\bar{\lambda}$. In order to study the gradients of the functions g, f, and h (defined below) for

complex λ, we shall use the tools of Sec. 8. This requires extending these functions into the space of λ and μ.

Let $\lambda = \sigma + i\tau$ (σ,τ real). For all complex $\lambda \neq \lambda_i$, define the functions g_1 and g_2 by

$$g(\lambda) = g_1(\sigma,\tau) = g_2(\lambda,\bar{\lambda}) \quad ,$$

where

$$g_2(\lambda,\mu) = \sum_{i=1}^{n} \frac{a_1}{(\lambda_i-\lambda)(\lambda_i-\mu)} \quad .$$

(This definition is consistent with (2.3).) Then

$$\frac{1}{2}\left(\frac{\partial g_1}{\partial \sigma} + i\,\frac{\partial g_1}{\partial \tau}\right) = \left[\frac{\partial g_2}{\partial \mu}\right]_{\mu=\bar{\lambda}} \quad , \text{ by lemma (8.1)}$$

$$= \sum_{i=1}^{n} \frac{a_1}{(\lambda_i-\lambda)(\lambda_i-\bar{\lambda})^2} \tag{2.8}$$

$$= \sum_{i=1}^{n} \frac{a_1}{|\lambda_i-\lambda|^2\cdot(\lambda_i-\bar{\lambda})} \quad .$$

For all complex $\lambda \neq \lambda_i$, define $f(X)$ by (2.5). We then define the functions f_1 and f_2 by

$$f(\lambda) = f_1(\sigma,\tau) = f_2(\lambda,\bar{\lambda}) \quad ,$$

where

$$f_2(\lambda,\mu) = \lambda\mu \sum_{i=1}^{n} \frac{a_i}{\lambda_i(\lambda_i-\lambda)(\lambda_i-\mu)}$$

Then

$$\tfrac{1}{2}\left(\frac{\partial f_1}{\partial \sigma} + i\,\frac{\partial f_1}{\partial \tau}\right) = \left[\frac{\partial f_2}{\partial \mu}\right]_{\mu=\bar{\lambda}}, \quad \text{by lemma (8.1)}$$

$$= \lambda \sum_{i=1}^{n} \frac{a_i}{\lambda_i|\lambda_i-\lambda|^2} + \lambda\bar{\lambda} \sum_{i=1}^{n} \frac{a_i}{\lambda_i(\lambda_i-\lambda)(\lambda_i-\bar{\lambda})^2}$$

$$= \lambda \sum_{i=1}^{n} \frac{a_i}{\lambda_i|\lambda_i-\lambda|^2}\left[1 + \frac{\bar{\lambda}}{\lambda_i-\bar{\lambda}}\right]$$

$$= \lambda \sum_{i=1}^{n} \frac{a_i}{|\lambda_i-\lambda|^2(\lambda_i-\bar{\lambda})}$$

$$= \lambda\left[\frac{\partial g_2}{\partial \mu}\right]_{\mu=\bar{\lambda}}, \qquad \text{by (2.8)} \ .$$

I.e.,

$$\left[\frac{\partial f_2}{\partial \mu}\right]_{\mu=\bar{\lambda}} = \lambda \left[\frac{\partial g_2}{\partial \mu}\right]_{\mu=\bar{\lambda}} \tag{2.9}$$

While it is possible to use (2.9) to study the behavior of $f(\lambda)$ on the set G where $g(\lambda) - 1 = 0$, it is more convenient here and in Sec. 3 to introduce a new function $h(\lambda)$, which agrees with $f(\lambda)$ on G. For all complex $\lambda \neq \lambda_i$, define

(2.10) $$h(\lambda) = f(X) + \frac{\lambda+\lambda}{2}[1-g(\lambda)],$$

and note that

(2.11) $$h(\lambda) = f(\lambda), \quad \text{for } \lambda \in G .$$

As with f and g, we introduce functions h_1 and h_2 so that

$$h(\lambda) = h_1(\sigma,\tau) = h_2(\lambda,\mu) ,$$

where

$$h_2(\lambda,\mu) = f_2(\lambda,\mu) + \frac{\lambda+\mu}{2}[1-g_2(\lambda,\mu)] .$$

-Then

(2.12) $$\frac{\partial h_2}{\partial \mu} = \frac{\partial f_2}{\partial \mu} + \tfrac{1}{2}[1-g_2(\lambda,\mu)] - \frac{\lambda+\mu}{2}\frac{\partial g_2}{\partial \mu}$$

$$= \frac{\lambda-\mu}{2}\frac{\partial g_2}{\partial \mu} + \tfrac{1}{2}[1-g_2(\lambda,\mu)] , \text{ by (2.9)} .$$

Hence

$$\tfrac{1}{2}\left(\frac{\partial h_1}{\partial \sigma} + i\frac{\partial h_1}{\partial \tau}\right) = \left[\frac{\partial h_2}{\partial \mu}\right]_{\mu=\bar{\lambda}} , \text{ by (8.1)}$$

(2.13)

$$= \frac{\lambda-\bar{\lambda}}{2}\left[\frac{\partial g_2}{\partial \mu}\right]_{\mu=\bar{\lambda}} + \tfrac{1}{2}[1-g_2(\lambda,\bar{\lambda})]$$

$$= \frac{\tau i}{2}\left(\frac{\partial g_1}{\partial \sigma} + i\frac{\partial g_1}{\partial \tau}\right) + \tfrac{1}{2}[1-g(\lambda)] .$$

Now any component G_j of the set G where $g(h) = 1$ encloses a region where $g(x) > 1$. On G the gradient vector of g,

$$\frac{\partial g_1}{\partial \sigma} + i \frac{\partial g_1}{\partial \tau} \quad ,$$

is non-zero, is normal to G_j, and points to the interior of G_j. Then, by (2.12), the gradient vector of h on G_j, namely

$$\frac{\partial h_1}{\partial \sigma} + i \frac{\partial h_1}{\partial \tau} = i\tau \left(\frac{\partial g_1}{\partial \sigma} + i \frac{\partial g_1}{\partial \tau} \right) \quad ,$$

is non-zero for $\tau \neq 0$ and points along the tangent to G_j in the direction of increasing σ. Hence

(2.14) $h(X)$ is strictly increasing, as λ traces G_j in the direction of increasing σ.

From (2.14) it follows that $h(h)$ assumes its maximum value, for each separate component curve G_j of G, at the point β_j on G_j of maximum real part. By theorem (2.6), β_j is on the axis of real λ.

Note that setting $\mu = \bar{\lambda} = \lambda$ in (2.12) yields the result that

(2.15) $$h'(X) = 1 - g(X), \quad \text{for real } \lambda \ .$$

To complete the proof of the present theorem, we much show that $f(X)$ is larger at the point α_j of least real part on the component G_j of G

than it is at the right-most point β_{j-1} of the component G_{j-1} of G immediately to the left of G_j.

Note that g is continuous for $\lambda \in [\beta_{j-1},\alpha_j]$, and that $g(\beta_{j-1}) = g(\alpha_j) = 1$ but $g(h) < 1$ for $\beta_{j-1} < \lambda < \alpha_j$. Then

$$h(\alpha_j) = h(\beta_{j-1}) + \int_{\beta_{j-1}}^{\alpha_j} h'(\lambda)d\lambda$$

$$= h(\beta_{j-1}) + \int_{\beta_{j-1}}^{\alpha_j} [1-g(\lambda)]d\lambda \quad , \quad \text{by (2.15)}$$

$$> h(\beta_{j-1}), \qquad \text{since} \quad g(\lambda) < 1 \ .$$

Thus

$$h(\beta_{j-1}) < h(\alpha_j) \quad , \tag{2.16}$$

as was to be proved.

We conclude that $h(X)$ increases, as λ increases along the real axis between adjacent components of G. Since $h(h) = f(X)$ on G, we see from (2.14) and (2.16) that

$$\max_{\lambda \in G} f(\lambda) = f(\Lambda_R) \quad ,$$

$$\min_{\lambda \in G} f(h) = f(\Lambda_L) \ .$$

It follows trivially from theorem (2.7) that the maximum and minimum values of $f(\lambda)$ over the real numbers in G (i.e., over the spectrum of (A,b) in the present case) are also $f(\Lambda_R)$ and $f(\Lambda_L)$, respectively.

By (1.9), our condition that no $\lambda_i b_i = 0$ implies that $\lambda \neq \lambda_i$ for all i and for all λ in G. Hence Λ_L and Λ_R are not eigenvalues of A, and so neither $A - \Lambda_L I$ nor $A - \Lambda_R I$ is a singular matrix. Therefore we can solve equation (1.7) uniquely for x_{max} and x_{min}:

$$x_{max} = x^{\Lambda_R} = (A-\Lambda_R I)^{-1}Ab \ ,$$

$$x_{min} = x^{\Lambda_L}, = (A-\Lambda_L I)^{-1}Ab \ .$$

These equations give unique solutions to the problem of minimizing and maximizing $\Phi(x) = (x-b)^H A(x-b)$, for nonsingular A and b such that no $b_i = 0$.

It would be desirable to be able to prove that $h(\alpha_j) < h(\beta_j)$, in the notation of theorem (2.7), without analyzing h(X) and g(h) for complex values of λ.

3. General case: Some $\lambda_i b_i = 0$.

We now study the general case where one or more $\lambda_i b_i = 0$. To be explicit, let $\mathcal{C} = \{a: \lambda_\alpha b_\alpha = 0\}$, a set of integers. We wish to examine the spectrum of (A,b).

Define $\mathcal{L}$ as the set $\{\lambda_\alpha: \alpha \in \mathcal{C}\}$.

First, given one $\alpha \in \mathcal{C}$, if $\lambda \neq \lambda_\alpha$ and λ is in the spectrum of (A,b) with corresponding vector x^λ, then (1.9) shows that $x_\alpha^\lambda = 0$. Thus, if λ is in the spectrum but not in the set $\mathcal{L}$, then $x_\alpha^\lambda = 0$ for all $\alpha \in \mathcal{C}$ and, just as in the derivation of (2.3), λ will satisfy the equation

$$\sum_{i=1}^{n} \frac{\lambda_i^2 |b_i|^2}{|\lambda_i - \lambda|^2} = 1 \quad . \tag{3.1}$$

Conversely, any real solution λ of (3.1) -which is not in the set $\mathcal{L}$ will be in the spectrum of (A,b), and its corresponding vector x^λ will have $x_\alpha^\lambda = 0$ for all $\alpha \in \mathcal{C}$. If we interpret O/O in (3.1) as 0, then it is possible that some eigenvalue λ_k in $\mathcal{L}$ will also satisfy (3.1). If so, we will show that this λ_k is also in the spectrum of (A,b). However, the spectrum may also contain eigenvalues λ_k in $\mathcal{L}$ which do not satisfy (3.1), as we shall now show. No eigenvalue λ_k not in $\mathcal{L}$ can be in the spectrum, because $\lambda = \lambda_k$ would make the left side of (3.1) equal to ∞.

Fix attention on one λ_k for $k \in \mathcal{C}$. We wish to examine the possibility that this λ_k is in the spectrum of (A,b). Let m be the multiplicity of λ_k as an eigenvalue of A. Let $\mathcal{D}_k = \{i: \lambda_i = \lambda_k\}$ so that card $(\mathcal{D}_k) = m$. If λ_k is in the spectrum, then (1.9) shows that $\lambda_\alpha b_\alpha = 0$ for all α in $\mathcal{D}_k$. Moreover, if λ_k is in the spectrum, then the corresponding vector $x = x^{\lambda_k}$ has the properties

$$x_i = \frac{\lambda_i b_i}{\lambda_i - \lambda_k} \qquad (i \notin \mathcal{D}_k) \ ,$$

and, by (2.1), also

$$(3.2) \qquad \sum_{i \notin \mathcal{N}_k} |x_i|^2 = \sum_{i \notin \mathcal{N}_k} \frac{\lambda_i^2 |b_i|^2}{|\lambda_i - \lambda_k|^2} < 1 \ .$$

Conversely, if (3.2) holds then we can always define x_i for all $i \in \mathcal{N}_k$ in such a way that

$$(3.3) \qquad \sum_{i \in \mathcal{N}_k} |x_i|^2 = 1 - \sum_{i \notin \mathcal{N}_k} \frac{\lambda_i |b_i|^2}{|\lambda_i - \lambda_k|^2} \ .$$

Hence, by (3.2) and (3.3), equation (1.8) holds and, since (1.7) is satisfied, λ_k is in the spectrum of (A,b).

If equality holds in (3.2) then x_i must be 0 for all $i \in \mathcal{N}_k$; i.e., λ_k satisfies (3.1), and x^{λ_k} is unique. But if inequality $<$ holds in (3.2), then there is an $(m-1)$-dimensional sphere $\mathcal{V}$ of values of $\{x_i\}$, for $i \in \mathcal{N}_k$, which satisfy (3.3). For, if a point $(x_{i_1},\ldots,x_{i_m})$ is in $\mathcal{V}$, then so are all points of form

$$(x_{i_1} e^{i\theta_1},\ldots,x_{i_m} e^{i\theta_m}) \quad \text{(all } \theta_i \text{ real)} \ ,$$

since $\sum_{i \in \mathcal{N}_k} |x_i|^2$ is constant for all of these. In this case uniqueness of x^{λ_k} is lost. The sphere is analogous to (in fact is a generalization of) the sphere of unit eigenvectors of a hermitian matrix A belonging to an eigenvalue of multiplicity m.

Note that an inequality $<$ in (3.2) states that λ_k is in the exterior of the graph

$$G = \left\{ \lambda : \sum_{i \notin \mathcal{O}_k} \frac{\lambda_i^2 |b_i|^2}{|\lambda_i - \lambda|^2} = 1 \right\} ,$$

i.e., λ_k can be joined to ∞ by an arc not cutting G. Thus, in brief, the spectrum of (A,b) consists of the union of all real numbers in the set

$$G = \left\{ \lambda : \sum_{i=1}^{n} \frac{\lambda_i^2 |b_i|^2}{|\lambda_i - \lambda|^2} = 1 \right\} \tag{3.4}$$

where we interpret $0/0$ as 0, with those numbers λ_k which are exterior to the graph G. (If G is the null set, then $b = \theta$ and the spectrum of (A,θ) consists of all eigenvalues λ_k.)

We must now examine $\Phi(x^\lambda)$ for λ in the spectrum of (A,b). The study of $\Phi(x^\lambda)$, for real $\lambda \in G$ in (3.4) is the same as in Sec. 2, and yields the same results (2.4) and (2.5): First, for $\lambda \in G$, $\Phi(x^\lambda) = f(X) = |\lambda|^2 \sum_{i=1}^{n} \frac{\lambda_i |b_i|^2}{|\lambda_i - \lambda|^2}$, where $0/0 = 0$. Second, let μ_R, μ_L be the right-most [resp. left-most] points of G. Then $f(\mu_R)$ maximizes [resp. $f(\mu_L)$ minimizes] $f(X)$ for $\lambda \in G$. It remains to consider $\Phi(x^{\lambda_k})$, for eigenvalues λ_k outside G.

(3.5) Theorem. For any λ in the spectrum of (A,b) we have

$$\Phi(x^{\lambda}) = h(X) = f(X) + \lambda[1-g(\lambda)], \tag{3.6}$$

where $f(X)$ is given by (2.5), with $0/0$ intepreted as 0.

Proof. Take any λ in the spectrum of (A,b).

If $\lambda \neq \lambda_k$ $(k=1,\ldots,n)$, then $\lambda \in G$, and everything proceeds as in the proof of (2.4), showing that $\Phi(x^{\lambda}) = f(X)$. Since $g(X) = 1$, we have proved (3.6) when $\lambda \neq \lambda_k$.

If $\lambda = \lambda_k$, an eigenvalue of A, let x_i denote the i-th coordinate of any x^{λ_k} which satisfies (1.5) and (1.6) (and hence (3.2) and (3.3)). Since λ_k is in the spectrum of (A,b), we have $\lambda_i b_i = 0$ for all $i \in \mathcal{O}_k$, where $\mathcal{O}_k$ is defined above after (3.1), and as $\lambda_i|x_i-b_i|^2 = \lambda_i|x_i|^2 = \lambda_k|x_i|^2$, for all $i \in \mathcal{O}_k$. Then, by (3.3),

$$\begin{aligned} \sum_{i \in \mathcal{O}_k} \lambda_i|x_i-b_i|^2 &= \lambda_k \left[1 - \sum_{i \notin \mathcal{O}_k} \frac{\lambda_i|b_i|^2}{(\lambda_i-\lambda_k)^2}\right] \\ &= \lambda_k \left[1 - \sum_{i=1}^{n} \frac{\lambda_i|b_i|^2}{(\lambda_i-\lambda_k)^2}\right], \quad \text{where } o/o = 0 . \end{aligned} \tag{3.7}$$

Moreover, like (2.4) we can prove

$$\sum_{i \notin \mathcal{O}_k} \lambda_i |x_i - b_i|^2 = \lambda_k^2 \sum_{i \notin \mathcal{O}_k} \frac{\lambda_i |b_i|^2}{(\lambda_i - \lambda_k)^2}$$

(3.8)

$$= \lambda_k^2 \sum_{i=1}^{n} \frac{\lambda_i |b_i|^2}{(\lambda_i - \lambda_k)^2} \quad , \quad \text{where} \quad 0/0 = 0 \ .$$

Adding (3.7) to (3.8), we get

$$\Phi(x^\lambda) = \sum_{i=1}^{n} \lambda_i |x_i - b_i|^2$$

(3.9)

$$= \lambda_k^2 \sum_{i=1}^{n} \frac{\lambda_i |b_i|^2}{(\lambda_i - \lambda_k)^2} + \lambda_k \left[1 - \sum_{i=1}^{n} \frac{\lambda_i |b_i|^2}{(\lambda_i - \lambda_k)^2} \right]$$

$$= f(\lambda_k) + \lambda_k [1 - g(\lambda_k)]$$

$$= h(\lambda_k) \ .$$

This proves (3.6) when $\lambda = \lambda_k$.

It is property (3.6) of h which motivated our use of h in Sec. 2.

Note. It is easily shown from (3.6) or (3.9) that, for all λ in the spectrum of (A,b),

(3.10) $$h(h) = \lambda + \lambda \sum_{i=1}^{n} \frac{\lambda_i |b_i|^2}{\lambda - \lambda_i}, \quad \text{where } 0/0 = 0 \ .$$

If λ is in the spectrum of (A,b), but is not an eigenvalue of A, we can derive (3.10) as follows. Let x belong to λ. Then

$$\Phi(x) = (x-b)^H A(x-b)$$

$$= (x-b)^H \lambda x \ , \qquad \text{by (1.5)}$$

$$= \lambda\, x^H x - \lambda\, b^H x$$

$$= \lambda - \lambda\, b^H x \ , \qquad \text{by (1.6)}$$

$$= \lambda - \lambda\, b^H (A-\lambda I)^{-1} A b \ , \qquad \text{by (1.7)}$$

$$\lambda - \lambda \sum_{i=1}^{n} \frac{\lambda_i |b_i|^2}{\lambda_i - \lambda} \ .$$

We shall not make use of (3.10) here.

We now use formula (3.6) to extend the domain of h to all real λ where $g(X) < 00$, i.e., to all λ except where, for some i, $\lambda = \lambda_i$ and $\lambda_i b_i \neq 0$.

As stated before (3.5), we know that the largest value of $\Phi(x^{\lambda}) = h(X)$ for λ in G occurs at the right-most point μ_R of G. It remains to see whether $h(\lambda_k)$ may be still larger for any λ_k in the spectrum of (A,b), if $\mu_R < \lambda_k$.

The answer is furnished by formula (2.15), which is valid for the general case of Sec. 3 with the understanding that O/O = 0. Thus h is increasing on all segments of the real axis between or exterior to components of the curve G. It follows that h(h) takes its maximum at the rightmost point Λ_R of the spectrum of (A,b) and its minimum value at the left-most point Λ_L of the spectrum of (A,b), whether or not these are eigenvaluee of A.

From the considerations following (3.3), we see that the maximizing x is unique if $\Lambda_R \in G$. If, however, Λ_R is not in G and is an eigenvalue of A of multiplicity m, then the maximizing x include all points of an (m-1)-sphere of nonzero radius, whose center is not at θ when $b \neq \theta$.

The above result about Λ_R and Λ_L for the case where some $\lambda_i b_i = 0$ can be obtained by continuity from the case where no $\lambda_i b_i = 0$. It is not clear that we could use continuity to deduce the nature of the maximizing and minimizing vectors, for multiple roots

4. The main result.

In Secs. 2 and 3 we have proved our result:

(4.1) Theorem. Given A, hermitian with eigenvalues $\{\lambda_i\}$, and b, arbitrary, define $\{b_i\}$ as in (1.3). Then the spectrum of (A,b) consists of all real λ such that

$$g(\lambda) = \sum_{i=1}^{n} \frac{\lambda_i^2 |b_i|^2}{|\lambda_i - \lambda|^2} = 1 \quad (o/o = 0;\ 1/0 = \infty),$$

together with each eigenvalue λ_k of A for which $g(\lambda_k) < 1$.

For each λ in the spectrum with $g(X) = 1$, a unique x^λ is found by solving (1.7, 1.8). For each λ in the spectrum with $g(h) < 1$, there exists an (m-1)-sphere of x^λ satisfying (1.7, 1.8), where $m = \text{card}\ (X_j : \lambda_j = \lambda_k)$.

Each x^{λ} so found renders $Q(x)$ stationary on S. Let $\Lambda_R = \max\{\lambda : \lambda \in \text{spectrum}\}$; let $\Lambda_L = \min\{\lambda : \lambda \in \text{spectrum}\}$. Let $x_{min} = \text{any } x^{\Lambda_L}$; let $x_{max} = \text{any } x^{\Lambda_R}$. Then $\Phi(x_{min})$ minimizes $\Phi(x)$ on S, and $\Phi(x_{max})$ maximizes $\Phi(x)$ on S.

5. The number of points in the spectrum.

As we noted in Sec. 2, if A is of order n, then the spectrum of (A,b) contains anywhere from 2 to 2n real numbers. When does it have the full amount 2n? If any $\lambda_i b_i = 0$, then the discussion of Sec. 3 showed that the spectrum necessarily has fewer than 2n points. So we are limited to the case where all $\lambda_i b_i \neq 0$. But then, as shown in Sec. 2, we know that the spectrum is the intersection of the graph of

$$\mu = \sum_{i=1}^{n} \frac{\lambda_i^2 |b_i|^2}{|\lambda_i - \lambda|^2} \tag{5.1}$$

for real λ with the line $\mu = 1$.

The graph of (5.1) for real λ consists of n + 1 branches between the n vertical asymptotes $\lambda = \lambda_i$ (i=1,...,n). Since $\mu > 0$ for all λ, and $\mu \to 0$ as $\lambda \to \infty$ and $\lambda \to -\infty$, the right-most and left-most branches necessarily cut $\mu = 1$. The spectrum has the full number 2n of points if and only if each of the n - 2 interior branches of the curve reaches its minimum with $\mu < 1$. For general n a condition for this is probably too complicated to derive. For n = 2, however, we can answer the question, as follows:

(5.2) Theorem.

(5.3) Let $n = 2$, and assume A is in diagonal form with $\lambda_1 < \lambda_2$. If the spectrum of (A,b) consists of 4 distinct numbers, then

$$0 < |b_1\lambda_1| \quad \text{and} \quad 0 < |b_2\lambda_2| , \tag{5.4}$$

and also

$$|b_1\lambda_1|^{\frac{2}{3}} + |b_2\lambda_2|^{\frac{2}{3}} < (\lambda_2-\lambda_1)^{\frac{2}{3}} . \tag{5.5}$$

(5.6) Conversely, if (5.4) and (5.5) hold, then the spectrum of (A,b) consists of 4 distinct numbers.

Proof of (5.3). Let $a_i = |\lambda_i b_i|^2$ $(i=1,2)$. If either a_1 or a_2 were zero, then the development in Sec. 3 shows that the spectrum would consists of at most 3 points. Hence $a_1 > 0$ and $a_2 > 0$; i.e., (5.4) holds. Let $M = (a_2/a_1)^{1/3}$. Now the development in Sec. 2 shows that the spectrum of (A,b) consists precisely of the real roots λ of the equation

$$g(\lambda) = \frac{a_1}{(\lambda-\lambda_1)^2} + \frac{"2}{(\lambda-\lambda_2)^2} = 1 . \tag{5.7}$$

Since (5.7) has 4 real roots, we know that two roots must lie in the internal (λ_1,λ_2). Now let μ be the unique real root of

$$g'(\lambda) = \frac{-2a_1}{(\lambda-\lambda_1)^3} - \frac{2a_2}{(\lambda-\lambda_2)^3} = 0 .$$

Then, because there are two roots of (5.7) in (λ_1,λ_2),

$$g(\mu) < 1 \quad . \tag{5.8}$$

We now show that (5.8) implies (5.4).

Solving $g'(\mu) = 0$ shows that

$$\frac{\lambda_2-\mu}{\mu-\lambda_1} = M \quad ,$$

whence

$$\mu - \lambda_1 = \frac{1}{1+M}(\lambda_2-\lambda_1) \quad ,$$

$$\lambda_2 - \mu = \frac{M}{1+M}(\lambda_2-\lambda_1) \quad .$$

Hence

$$g(\mu) = \frac{a_1(1+M)^2}{(\lambda_2-\lambda_1)^2} + \frac{a_2(1+M)^2}{M^2(\lambda_2-\lambda_1)^2}$$

$$= \frac{(1+M)^2}{(\lambda_2-\lambda_1)^2}\left[a_1 + \frac{a_2}{M^2}\right.$$

$$= \frac{(1+M)^2\ a_1^{\frac{2}{3}}}{(\lambda_2-\lambda_1)^2}\left[a_1^{\frac{1}{3}} + \right.$$

$$= \frac{\left(a_1^{\frac{1}{3}} + a_2^{\frac{1}{3}}\right)^2}{(\lambda_2-\lambda_1)^2}\left(a_1^{\frac{1}{3}} + a_2^{\frac{1}{3}}\right)$$

$$= \frac{\left(a_1^{\frac{1}{3}} + a_2^{\frac{1}{3}}\right)^3}{(\lambda_2-\lambda_1)^2} \quad .$$

Thus $g(\mu) < 1$ implies

$$a_1^{\frac{1}{3}} + a_2^{\frac{1}{3}} < (\lambda_2 - \lambda_1)^{\frac{2}{3}} , \tag{5.9}$$

which implies (5.5). Thus (5.3) is proved.

Proof of (5.6). We have $a_1 > 0$, $a_2 > 0$, and (5.9). The above steps are reversable, and so $g(\mu) < 1$, whence there are 4 real roots of $g(\mu) = 1$.

Thus theorem (5.2) is completely proved.

Condition (5.4) says that neither λ_1 nor λ_2 is 0, and that the point $b = (b_1, b_2)$ does not lie on an axis of the (x_1, x_2)-plane. Condition (5.5) requires that (b_1, b_2) be inside a curve Γ which depends only on the ratio λ_2/λ_1. If $\lambda_2/\lambda_1 = 2$, for example, the curve Γ is $|b_1|^{2/3} + |2b_2|^{2/3} = 1$.

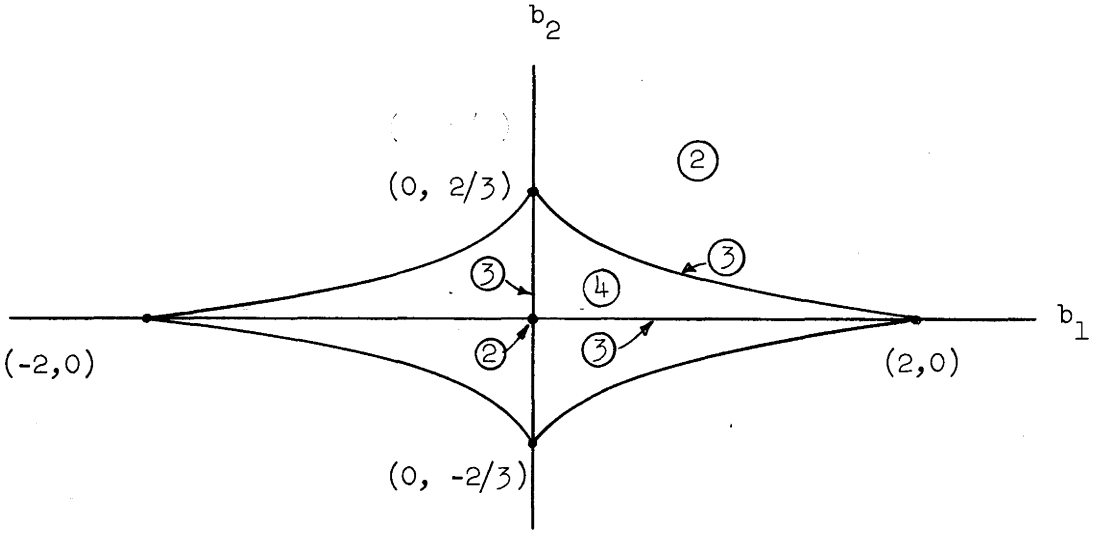

Figure 1

In Figure 1 the number of points in the spectrum of (A,b) is indicated for different b in the first quadrant by integers in circles.

If $|\lambda_2/\lambda_1| > 2$, the curve Γ includes values $|b_1| > 1$. But $|\lambda_2/\lambda_1| > 1$ implies that, on Γ, $|b_2| < 1$.

6. Geometrical interpretation.

Thesurfaces $\Phi(x) = k$ are similar conic surfaces with center b in the euclidean n-space $\mathcal{E}_n$ of vectors x. The maximum problem (1.2) is to find the conic surface with maximum k which touches the constraint surface S, the unit sphere in $\mathcal{E}_n$. The rotation of A to diagonal form is a rotation of $\mathcal{E}_n$ (leaving S "invariant, of course) which causes principal axes of the conic surfaces to coincide with the axes of $\mathcal{E}_n$.

The vector Ax - b is half the gradient of $\Phi(x)$, and x is the radius vector. Condition (1.5) merely states that at a point where $\Phi(x)$ is stationary, for x on S, the surface $\Phi(x) = k$ is tangent to S.

Fix x at a solution of (1.5), and let t be real. If the constant λ of (1.5) is positive, the value of $\Phi(tx)$ increases as t increases from 1; if λ is negative, $\Phi(tx)$ decreases as t increases from 1.

The main result of Secs. 2 and 3 is that the maximum problem of Sec. 1 is solved for the largest value: of λ satisfying (1.5), for x on S. The authors see no obvious geometrical reason why this should be so.

If all $b_i\lambda_i \neq 0$, then Sec. 2 shows that any vector $x = x^\lambda$ which makes $\Phi(x)$ stationary on S is uniquely determined by λ.

Figure 2 shows, for $n = 2$ and $0 < \lambda_1 < \lambda_2$, a case where there are

4 distinct points of tangency of an ellipse with the unit circle. All ellipses have center b and common value of $\lambda_2/\lambda_1 > 2$. Since

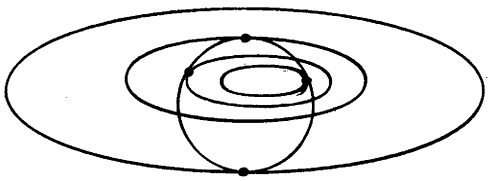

Figure 2

$\lambda_2/\lambda_1 > 2$, it was shown in Sec. 5 that 4 distinct tangencies were possible for certain b outside S.

Whenever some $b_k = 0$, then, provided that (3.2) holds with the inequality sign <, we get more than one x belonging to a given λ. That is illustrated in Fig. 3, where n = 2 and k = 1. What is not obvious to the authors is a geometrical reason why necessarily $\lambda = \lambda_k$ in this case.

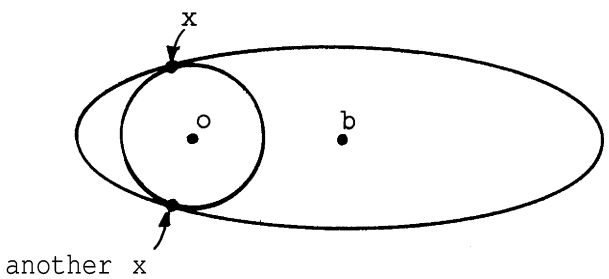

Figure 3

7. A constrained least squares problem.

Let C be an m X n matrix (m 3 n) and f an m-vector, both over the complex field. We wish to study the set of complex n-vectors y of euclidean length $\|y\| = (y^H y)^{1/2} = 1$ such that

$$\|Cy-f\|^2 = (Cy-f)^H(Cy-f) = \min. \tag{7.1}$$

The constraint is

$$\|y\|^2 = y^H y = 1 \ . \tag{7.2}$$

Because euclidean length is invariant under unitary transformations, it is useful to rotate coordinates in both the space of y and the space of f. To do this, let r = rank(C), and write

$$C = U^H D V \ , \tag{7.3}$$

where U, V are unitary, and where the only non-zero elements of D are the first r elements of the leading diagonal, which we may arrange so that

$$d_1 \geq d_2 \geq \bullet \ \_*e>dr>0 \ .$$

Now let Vy = x and Uf = g. Then

$$Cy - f = U^H DVy - U^H Uf$$

$$= U^H(Dx-g) \quad .$$

Hence

$$\|Cy-f\|^2 = \|Dx-g\|^2$$

(7.4)

$$= \sum_{i=1}^{r} |d_i x_i - g_i|^2 + \sum_{i=r+1}^{m} |g_i|^2 \quad .$$

The problem (7.1,7.2) is to minimize

$$\sum_{i=1}^{r} I\, d_i x_i - g_i |^2 = \sum_{i=1}^{r} d_i^2 \, |x_i - g_i/d_i|^2$$

-subject to the constraint

(7.5)

$$\sum_{i=1}^{n} |X_i|^2 = 1 \quad .$$

Now let $\lambda_i = \begin{cases} 0 & (i = 1,2,\ldots,n-r) \\ d^2_{n+1-i} & (i = n-r+1,\ldots,n) \end{cases}$,

and let $b_i = \begin{cases} 0 & (i = 1,2,\ldots,n-r) \\ \dfrac{g_{n+1-i}}{d_{n+1-i}} & (i = n-r+1,\ldots,n) \end{cases}$.

We then have changed our problem to one of minimizing

(7.6) $$\sum_{i=1}^{n} \lambda_i |x_i - b_i|^2 \quad ,$$

subject to the constraint (7.5), where

(7.7) $$\lambda_1 \leq \cdots \leq \lambda_{n-r} < \lambda_{n\,r+1} \leq \cdots \leq \lambda_n \quad .$$

This is precisely the minimum problem (1.2) of Sec. 1. The special role of the $n - r$ zero eigenvalues of $C^H C$ becomes evident.

Thus the general problem of the least-squares solution of $Cy = f$ with constraint (7.2) is a special case of our minimum problem (1.2).

8. Lemmas from complex function theory.

In this final section we state and prove three lemmas relating partial derivatives of certain regular analytic functions of several complex variables to gradients of real-valued functions of vector variables. This technique is common in the study of second-order partial differential equations; for example, see (9.1). We include the material mainly to keep our treatment self-contained, and partly to call explicit attention to the fact that the Lagrange multiplier λ must be real even though complex variables are used.

(8.1) Lemma. Let $\Phi(\lambda,\mu)$ be a regular analytic function of two complex variables λ, μ such that, for all real x, y,

(8.2) $$F(x,y) = \Phi(x+iy,\ x-iy)$$

is real-valued. Then

$$\frac{\partial F}{\partial x} + i\,\frac{\partial F}{\partial y} = 2\,\frac{\partial \Phi}{\partial \mu}\Big|_{\substack{\lambda=x+iy \\ \mu=x-iy}}$$

Proof. Differentiate (8.2):

(8.3) $$\frac{\partial F}{\partial x} = \frac{\partial \Phi}{\partial \lambda}\cdot 1 + \frac{\partial \Phi}{\partial \mu}\cdot 1 \quad ,$$

(8.4) $$\frac{\partial F}{\partial y} = \frac{\partial \Phi}{\partial \lambda}\cdot i - \frac{\partial \Phi}{\partial \mu}\cdot i \quad .$$

Add (8.3) to (8.4) x i:

$$\frac{\partial F}{\partial x} + i\,\frac{\partial F}{\partial y} = 2\,\frac{\partial \Phi}{\partial \mu} \quad .$$

(8.5) **Lemma.** Let F and G be real-valued differentiable functions of real variables $x_1,y_1,\ldots,x_n,y_n$. For abbreviation, let $z_k = x_k + iy_k$, and let $z = (z_1,\ldots,z_n)$. Then, for F(z) to be stationary at z = a with respect to all neighboring z such that G(z) = G(a), it is necessary and sufficient that there exist a real Lagrange constant λ such that

(8.6) $$\frac{\partial F}{\partial x_k} + i\,\frac{\partial F}{\partial y_k} - \lambda\left(\frac{\partial G}{\partial x_k} + i\,\frac{\partial G}{\partial y_k}\right) = 0$$

for z = a and k = 1,...,n.

Proof. Condition (8.6) is nothing but the usual condition that the real gradient vector

$$\left(\frac{\partial F}{\partial x_1} , \frac{\partial F}{\partial y_1} , \frac{\partial F}{\partial k_2} , \frac{\partial F}{\partial y_2} , \cdots , \frac{\partial F}{\partial x_n} , \frac{\partial F}{\partial y_n} \right)$$

be parallel to the vector

$$\left(\frac{\partial G}{\partial x_1} , \frac{\partial G}{\partial y_1} , \frac{\partial G}{\partial x_2} , \frac{\partial G}{\partial y_2} , \cdots , \frac{\partial G}{\partial x_n} , \frac{\partial G}{\partial y_n} \right) .$$

The use of the complex variables z_k is unessential.

Given any vector $z = (z_1,\ldots,z_n)$, we let $\bar{z}$ denote the vector of complex conjugates $(\bar{z}_1,\ldots,\bar{z}_n)$.

(8.7) Lemma. Let $\Phi(z,w)$ and $\psi(z,w)$ be regular analytic functions of the two complex vector variables $z = (z_1,\ldots,z_n)$ and $w = (w_1,\ldots,w_n)$ with the property that $\Phi(z,\bar{z})$ and $\psi(z,\bar{z})$ are real. Then $\Phi(z,\bar{z})$ is stationary at $z = a$ with respect to all z such that $\psi(z,\bar{z}) = \psi(a,\bar{a})$, if and only if there exists a real Lagrange constant λ such that

$$\frac{\partial \Phi}{\partial w_k} - \lambda \frac{\partial \psi}{\partial w_k} = 0$$

for $z = a$ and $w = \bar{a}$ and $k = 1,2,\ldots,n$.

Proof. Let $z = x + iy$. Then $\Phi(z,\bar{z}) = F(x,y)$, $\psi(z,\bar{z}) = G(x,y)$. By lemma 8.1 applied to each variable z_k,

$$\frac{\partial \Phi}{\partial w_k} = \frac{\partial F}{\partial x_k} + i \frac{\partial F}{\partial y_k} ,$$

$$\frac{\partial \psi}{\partial w_k} = \frac{\partial G}{\partial x_k} + i \frac{\partial G}{\partial y_k}$$

for $z = a$, $w = \bar{a}$, and $k = 1, \ldots, n$.

Then lemma (8.7) follows from lemma (8.5) .

9. References.

(9.1) P. R. Garabedian, Partial Differential Equations, Wiley, 1964; Chap. 3.

(9.2) David L. Phillips, A technique for the numerical solution of certain integral equations of the first kind, J. Assoc. Comput. Mach. vol. 9 (1962), pp. 84-97.

(9.3) S. Twomey, On the numerical solution of Fredholm integral equations of the first kind by the inversion of the linear system produced by quadrature, J. Assoc. Comput. Mach. vol. 10 (1963), pp. 97-101.