CHAPTER 3

Optimization of Quadratic Functions

In this chapter we study the problem

$$\text{(31)} \qquad \underset{x\in\mathbb{R}^n}{\text{minimize}}\ \tfrac{1}{2}x^THx + g^Tx + \beta,$$

where $H \in \mathbb{R}^{n\times n}$ is symmetric, $g \in \mathbb{R}^n$, and $\beta \in \mathbb{R}$. It has already been observed that we may as well assume that H is symmetric since

$$x^THx = \tfrac{1}{2}x^THx + \tfrac{1}{2}(x^THx)^T = x^T\left[\tfrac{1}{2}(H + H^T)\right]x,$$

where $\frac{1}{2}(H + H^T)$ is called the *symmetric part* of H. Therefore, in this chapter we assume that H is symmetric. In addition, we have also noted that an objective function can always be shifted by a constant value without changing the solution set to the optimization problem. Therefore, we assume that $\beta = 0$ for most of our discussion. However, just as in the case of integration theory where it is often helpful to choose a particular constant of integration, in many applications there is a "natural" choice for β that helps one interpret the problem as well as its solution.

The class of problems (31) is important for many reasons. Perhaps the most common instance of this problem is the linear least squares problem:

$$\text{(32)} \qquad \underset{x\in\mathbb{R}^n}{\text{minimize}}\ \tfrac{1}{2}\left\|Ax - b\right\|_2^2,$$

where $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$. By expanding the objective function in (32), we see that

$$\text{(33)} \qquad \tfrac{1}{2}\left\|Ax - b\right\|_2^2 = \tfrac{1}{2}x^T(A^TA)x - (A^Tb)^Tx + \tfrac{1}{2}\left\|b\right\|_2^2 = \tfrac{1}{2}x^THx + g^Tx + \beta,$$

where $H = A^TA$, $g = -A^Tb$, and $\beta = \frac{1}{2}\left\|b\right\|_2^2$. This connection to the linear least squares problem will be explored in detail later in this chapter. For the moment, we continue to exam the general problem (31). As in the case of the linear least squares problem, we begin by discussing characterizations of the solutions as well as their existence and uniqueness. In this discussion we try to follow the approach taken for the the linear least squares problem. However, in the case of (32), the matrix $H := A^TA$ and the vector $g = -A^Tb$ possess special features that allowed us to establish very strong results on optimality conditions as well as on the existence and uniqueness of solutions. In the case of a general symmetric matrix H and vector g it is possible to obtain similar results, but there are some twists. Symmetric matrices have many special properties that can be exploited to help us achieve our goal. Therefore, we begin by recalling a few of these properties, specifically those related to eigenvalue decomposition.

1. Optimality Properties of Quadratic Functions

Recall that for the linear least squares problem, we were able to establish a necessary and sufficient condition for optimality, namely the normal equations, by working backward from a known solution. We now try to apply this same approach to quadratic functions, in particular, we try to extend the derivation in (19) to the objective function in (34). Suppose $\overline{x}$ is a local solution to the quadratic optimization problem

$$\text{(34)} \qquad \underset{x\in\mathbb{R}^n}{\text{minimize}}\ \tfrac{1}{2}x^THx + g^Tx,$$

where $H \in \mathbb{R}^{n\times n}$ is symmetric and $g \in \mathbb{R}^n$, i.e., there is an $\epsilon > 0$ such that

$$\text{(35)} \qquad \tfrac{1}{2}\overline{x}^TH\overline{x} + g^T\overline{x} \le \tfrac{1}{2}x^THx + g^Tx \quad \forall\, x \in \overline{x} + \epsilon\mathbb{B}_2,$$

where $\overline{x} + \epsilon\mathbb{B}_2 := \{\overline{x} + \epsilon u \,|\, u \in \mathbb{B}_2\}$ and $\mathbb{B}_2 := \{u \,|\, \|u\|_2 \le 1\}$ (hence, $\overline{x} + \epsilon\mathbb{B}_2 = \{x \,|\, \|\overline{x} - x\|_2 \le \epsilon\}$). Note that, for all $x \in \mathbb{R}^n$,

$$\begin{aligned}
\overline{x}^T H\overline{x} &= (x + (\overline{x} - x))^T H(x + (\overline{x} - x)) \\
&= x^T Hx + 2x^T H(\overline{x} - x) + (\overline{x} - x)^T H(\overline{x} - x) \\
&= x^T Hx + 2(\overline{x} + (x - \overline{x}))^T H(\overline{x} - x) + (\overline{x} - x)^T H(\overline{x} - x) \\
&= x^T Hx + 2\overline{x}^T H(\overline{x} - x) + 2(x - \overline{x})^T H(\overline{x} - x) + (\overline{x} - x)^T H(\overline{x} - x) \\
&= x^T Hx + 2\overline{x}^T H(\overline{x} - x) - (\overline{x} - x)^T H(\overline{x} - x).
\end{aligned} \tag{36}$$

Therefore, for all $x \in \overline{x} + \epsilon\mathbb{B}_2$,

$$\begin{aligned}
\tfrac{1}{2}\overline{x}^T H\overline{x} + g^T\overline{x} &= (\tfrac{1}{2}x^T Hx + g^T x) + (H\overline{x} + g)^T(\overline{x} - x) - \tfrac{1}{2}(\overline{x} - x)^T H(\overline{x} - x) \\
&\ge (\tfrac{1}{2}\overline{x}^T H\overline{x} + g^T\overline{x}) + (H\overline{x} + g)^T(\overline{x} - x) - \tfrac{1}{2}(\overline{x} - x)^T H(\overline{x} - x)\ , \qquad \text{(since } \overline{x} \text{ is a local solution)}
\end{aligned}$$

and so

$$\tfrac{1}{2}(\overline{x} - x)^T H(\overline{x} - x) \ge (H\overline{x} + g)^T(\overline{x} - x) \qquad \forall\, x \in \overline{x} + \epsilon\mathbb{B}_2. \tag{37}$$

Let $0 \le t \le \epsilon$ and $v \in \mathbb{B}_2$ and define $x = \overline{x} + tv \in \overline{x} + \epsilon\mathbb{B}_2$. If we plug $x = \overline{x} + tv$ into (37), then

$$\frac{t^2}{2}v^T Hv \ge -t(H\overline{x} + g)^T v. \tag{38}$$

Dividing this expression by $t > 0$ and taking the limit as $t \downarrow 0$ tells us that

$$0 \le (H\overline{x} + g)^T v \quad \forall\ v \in \mathbb{B}_2\ ,$$

which implies that $H\overline{x} + g = 0$. Plugging this information back into (38) gives

$$\frac{t^2}{2}v^T Hv \ge 0 \quad \forall\, v \in \mathbb{B}_2\ .$$

Dividing by $t^2/2$ for $t \ne 0$ tells us that

$$v^T Hv \ge 0 \quad \forall\, v \in \mathbb{B}_2$$

or equivalently, that $v^T Hv \ge 0 \quad \forall\, v \in \mathbb{R}^n$. This latter condition on the matrix H plays an important role in optimization theory and practice. We codify this property in the following definition.

DEFINITION 1.1. *Let $H \in \mathbb{R}^{n\times n}$.*

(1) H is said to be positive definite if $x^T Hx > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$.
(2) H is said to be positive semi-definite if $x^T Hx \ge 0$ for all $x \in \mathbb{R}^n$.
(3) H is said to be negative definite if $x^T Hx < 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$.
(4) H is said to be positive semi-definite if $x^T Hx \le 0$ for all $x \in \mathbb{R}^n$.
(5) H is said to be indefinite if H is none of the above.

In addition, if $\mathcal{S}^n$ is the linear space of real $n \times n$ symmetric matrices, we denote the set of real symmetric positive semi-definite and positive definite matrices by $\mathcal{S}^n_+$ and $\mathcal{S}^n_{++}$, respectively. Similarly, we denote by $\mathcal{S}^n_-$ and $\mathcal{S}^n_{--}$, the sets of real symmetric negative semi-definite and negative definite matrices, respectively.

These observations motivate the following theorem.

THEOREM 1.1. *[Existence and Uniqueness in Quadratic Optimization] Let $H \in \mathbb{R}^{n\times n}$ and $g \in \mathbb{R}^n$ be as in (34).*

(1) A local solution to the problem (34) exists if and only if $H \in \mathcal{S}^n_+$ and there exists a solution $\overline{x}$ to the equation $Hx + g = 0$ in which case $\overline{x}$ is a local solution to (34).
(2) If $\overline{x}$ is a local solution to (34), then it is a global solution to (34).
(3) The problem (34) has a unique global solution if and only if H is positive definite in which case this solution is given by $\overline{x} = -H^{-1}g$.
(4) If either H is not positive semi-definite or there is no solution to the equation $Hx + g = 0$ (or both), then

$$-\infty = \inf_{x\in\mathbb{R}^n} \tfrac{1}{2}x^T Hx + g^T x\ .$$

PROOF. (1) We have already shown that if a local solution $\overline{x}$ to (34) exists, then $H\overline{x}+g=0$ and H is positive semi-definite. On the other hand, suppose that H is positive semi-definite and $\overline{x}$ is a solution to $Hx+g=0$. Then, for all $x\in\mathbb{R}^n$, we can interchange the roles of x and $\overline{x}$ in the second line of (36) to obtain

$$x^THx = \overline{x}^TH\overline{x} + 2\overline{x}^TH(x-\overline{x}) + (x-\overline{x})^TH(x-\overline{x}). \tag{39}$$

Hence, for all $x\in\mathbb{R}^n$,

$$\tfrac{1}{2}x^THx + g^Tx = \tfrac{1}{2}\overline{x}^TH\overline{x} + g^T\overline{x} + (H\overline{x}+g)^T(x-\overline{x}) + \tfrac{1}{2}(x-\overline{x})^TH(x-\overline{x}) \geq \tfrac{1}{2}\overline{x}^TH\overline{x} + g^T\overline{x}\ ,$$

since $H\overline{x}+g=0$ and H is positive semi-definite. That is, $\overline{x}$ is a global solution to (34) and hence a local solution.

(2) The proof of part (1) shows that if H is positive semi-definite and $H\overline{x}+g=0$, then $\overline{x}$ is a global solution to (34). Hence the result follows from part (1).

(3) If (34) has a unique global solution $\overline{x}$, then $\overline{x}$ must be the unique solution to the equation $Hx+g=0$. This can only happen if H is invertible. Hence, H is invertible and positive semi-definite which implies that H is positive definite. On the other hand, if H is positive definite, then, in particular, it is positive semi-definite and there is a unique solution to the equation $Hx+g=0$, i.e., (36) has a unique global solution.

(4) The result follows if we can show that $f(x):=\frac{1}{2}x^THx+g^Tx$ is unbounded below when either H is not positive semi-definite or there is no solution to the equation $Hx+g=0$ (or both). Let us first suppose that H is not positive semi-definite, or equivalently, there is a $\hat{x}\in\mathbb{R}^n$ such that $\hat{x}^TH\hat{x}<0$. Then, for every $t\in\mathbb{R}$, $f(t\hat{x})=\frac{t^2}{2}\hat{x}^TH\hat{x}+tg^t\hat{x}$. Therefore, $f(t\hat{x})\downarrow-\infty$ as $t\uparrow+\infty$, and so f is unbounded below.

Next assume that H is positive semi-definite but $g\notin\operatorname{Ran}(H)$, i.e. there is no solution to $Hx+g=0$. Let P be the orthogonal projection onto $\operatorname{Ran}(H)$ so that $I-P$ is the orthogonal projection onto $\operatorname{Ran}(H)^\perp=\operatorname{Nul}\left(H^T\right)=\operatorname{Nul}(H)$. Set $g_1:=Pg$ and $g_2:=(I-P)g$ so that $g=g_1+g_2$. Then, for every $t\in\mathbb{R}$, $f(tg_2)=\frac{t^2}{2}g_2^THg_2+tg^tg_)2=t\left\|g_2\right\|_2^2$. Therefore, $f(tg_2)\downarrow-\infty$ as $t\downarrow-\infty$, and so again f is unbounded below. □

The identity (39) is a very powerful tool in the analysis of quadratic functions. It was the key tool in showing that every local solution to (34) is necessarily a global solution. We now show how these results can be extended to problems with linear equality constraints.

2. Minimization of a Quadratic Function on an Affine Set

In this section we consider the problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}x^THx+g^Tx \\ \text{subject to} & Ax=b, \end{array} \tag{40}$$

where $H\in\mathbb{R}^{n\times n}$ is symmetric, $g\in\mathbb{R}^n$, $A\in\mathbb{R}^{m\times n}$, and $b\in\mathbb{R}^m$. We assume that the system $Ax=b$ is consistent. That is, there exists $\hat{x}\in\mathbb{R}^n$ such that $A\hat{x}=b$ in which case

$$\{x\,|\,Ax=b\}=\hat{x}+\operatorname{Null}(A).$$

Consequently, the problem (40) is of the form

$$\underset{x\in\hat{x}+S}{\text{minimize}}\ \tfrac{1}{2}x^THx+g^Tx\ , \tag{41}$$

where S is a subspace of $\mathbb{R}^n$. This representation of the problem shows that the problem (40) is trivial if $\operatorname{Null}(A)=\{0\}$ since then the unique solution $\hat{x}$ to $Ax=b$ is the unique solution to (40). Hence, when considering the problem (40) it is always assumed that $\operatorname{Null}(A)\neq\{0\}$, and furthermore, that $m<n$.

DEFINITION 2.1. *[Affine Sets] A subset K of $\mathbb{R}^n$ is said to be an affine set if there exists a point $\hat{x}\in\mathbb{R}^n$ and a subspace $S\subset\mathbb{R}^n$ such that $K=\hat{x}+S=\{\hat{x}+u\,|\,u\in S\}$.*

We now develop necessary and sufficient optimality conditions for the problem (41), that is, for the minimization of a quadratic function over an affine set. For this we assume that we have a basis $v^1,v^2,\ldots,v^k$ for S so that $\dim(S)=k$. Let $V\in\mathbb{R}^{n\times k}$ be the matrix whose columns are the vectors $v^1,v^2,\ldots,v^k$ so that $S=\operatorname{Ran}(V)$. Then $\hat{x}+S=\left\{\hat{x}+Vz\,\middle|\,z\in\mathbb{R}^k\right\}$. This allows us to rewrite the problem (41) as

$$\underset{z\in\mathbb{R}^k}{\text{minimize}}\ \tfrac{1}{2}(\hat{x}+Vz)^TH(\hat{x}+Vz)+g^T(\hat{x}+Vz)\ . \tag{42}$$

Proposition 2.1. *Consider the two problems* (41) *and* (42)*, where the columns of the matrix V form a basis for the subspace S. The set of optimal solution to these problems are related as follows:*

$$\{\overline{x}\,|\,\overline{x} \text{ solves } (41)\} = \{\hat{x} + V\overline{z}\,|\,\overline{z} \text{ solves } (42)\}.$$

By expanding the objective function in (42), we obtain

$$\tfrac{1}{2}(\hat{x}+Vz)^T H(\hat{x}+Vz) + g^T(\hat{x}+Vz) = \tfrac{1}{2}z^T V^T HVz + (V^T(H\hat{x}+g))^T v + f(\hat{x}),$$

where $f(x) := \frac{1}{2}x^T Hx + g^T x$. If we now set $\widehat{H} := V^T HV$, $\hat{g} := V^T(H\hat{x}+g)$, and $\beta := f(\hat{x})$, then problem (42) has the form of (31):

$$\underset{z\in\mathbb{R}^k}{\text{minimize}}\ \tfrac{1}{2}z^T \widehat{H} z + \hat{g}^T z\ , \tag{43}$$

where, as usual, we have dropped the constant term $\beta = f(\hat{x})$. Since we have already developed necessary and sufficient conditions for optimality in this problem, we can use them to state similar conditions for the problem (41).

Theorem 2.1. *[Optimization of Quadratics on Affine Sets]*
Consider the problem (41).

1. *A local solution to the problem* (41) *exists if and only if $u^T Hu \ge 0$ for all $u \in S$ and there exists a vector $\overline{x} \in \hat{x} + S$ such that $H\overline{x} + g \in S^\perp$, in which case $\overline{x}$ is a local solution to* (41).
2. *If $\overline{x}$ is a local solution to* (41)*, then it is a global solution.*
3. *The problem* (41) *has a unique global solution if and only if $u^T Hu > 0$ for all $u \in S\setminus\{0\}$. Moreover, if $V \in \mathbb{R}^{n\times k}$ is any matrix such that $Ran(V) = S$ where $k = \dim(S)$, then a unique solution to* (41) *exists if and only if the matrix $V^T HV$ is positive definite in which case the unique solution $\overline{x}$ is given by*
$$\overline{x} = [I - V(V^T HV)^{-1}V^T H]\hat{x} - V(V^T HV)^{-1}V^T g\ .$$
4. *If either there exists $\overline{u} \in S$ such that $\overline{u}^T H\overline{u} < 0$ or there does not exist $\overline{x} \in \hat{x}+S$ such that $H\overline{x} + g \in S^\perp$ (or both), then*
$$-\infty = \inf_{x\in\hat{x}+S} \tfrac{1}{2}x^T Hx + g^T x\ .$$

Proof. (1) By Proposition 2.1, a solution to (41) exists if and only if a solution to (42) exists. By Theorem 1.1, a solution to (42) exists if and only if $V^T HV$ is positive semi-definite and there is a solution $\overline{z}$ to the equation $V^T(H(\hat{x}+Vz)+g) = 0$ in which case $\overline{z}$ solves (42), or equivalently, by Proposition 2.1, $\overline{x} = \hat{x} + V\overline{z}$ solves (41). The condition that $V^T HV$ is positive semi-definite is equivalent to the statement that $z^T V^T HVz \ge 0$ for all $z \in \mathbb{R}^k$, or equivalently, $u^T Hu \ge 0$ for all $u \in S$. The condition, $V^T(H(\hat{x}+V\overline{z})+g) = 0$ is equivalent to $H\overline{x} + g \in \mathrm{Null}(V^T) = \mathrm{Ran}(V)^\perp = S^\perp$.

(2) This is an immediate consequence of Proposition 2.1 and Part (2) of Theorem 1.1.

(3) By Theorem 1.1, the problem (42) has a unique solution if and only if $V^T HV$ is positive definite in which case the solution is given by $\overline{z} = (V^T HV)^{-1}V^T(H\hat{x}+g)$. Note that $V^T HV$ is positive definite if and only if $u^T Hu > 0$ for all $u \in S\setminus\{0\}$ which proves that this condition is necessary and sufficient. In addition, by Proposition 2.1, $\overline{x} = \hat{x} + V\overline{z} = [I - V(V^T HV)^{-1}V^T H]\hat{x} - V(V^T HV)^{-1}V^T g$ is the unique solution to (41).

(4) This follows the same pattern of proof using Part (4) of Theorem 1.1. ☐

Theorem 2.2. *[Optimization of Quadratics Subject to Linear Equality Constraints]*
Consider the problem (40).

1. *A local solution to the problem* (40) *exists if and only if $u^T Hu \ge 0$ for all $u \in Null(A)$ and there exists a vector pair $(\overline{x}, \overline{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that $H\overline{x} + A^T\overline{y} + g = 0$, in which case $\overline{x}$ is a local solution to* (41).
2. *If $\overline{x}$ is a local solution to* (41)*, then it is a global solution.*
3. *The problem* (41) *has a unique global solution if and only if $u^T Hu > 0$ for all $u \in Null(A)\setminus\{0\}$.*
4. *If $u^T Hu > 0$ for all $u \in Null(A)\setminus\{0\}$ and* $\mathrm{rank}\,(A) = m$*, the matrix*
$$M := \begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix} \quad \text{is invertible, and the vector} \quad \begin{bmatrix} \overline{x} \\ \overline{y} \end{bmatrix} = M^{-1}\begin{bmatrix} -g \\ b \end{bmatrix}$$
has $\overline{x}$ as the unique global solution to (41).

(5) If either there exists $\overline{u} \in Null(A)$ *such that* $\overline{u}^T H\overline{u} < 0$ *or there does not exist a vector pair* $(\overline{x}, \overline{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ *such that* $H\overline{x} + A^T\overline{y} + g = 0$ *(or both), then*

$$-\infty = \inf_{x \in \hat{x}+S} \tfrac{1}{2} x^T Hx + g^T x \ .$$

REMARK 2.1. *The condition that* rank $(A) = m$ *in Part (4) of the theorem can always be satisfied by replacing* A *by first row reducing* A *to echelon form.*

PROOF. (1) Recall that $\text{Null}(A)^\perp = \text{Ran}(A^T)$. Hence, $w \in \text{Null}(A)$ if and only if there exists $y \in \mathbb{R}^m$ such that $w = A^T y$. By setting $w = H\overline{x} + g$ the result follows from Part (1) of Theorem 2.1.

(2) Again, this is an immediate consequence of Proposition 2.1 and Part (2) of Theorem 1.1.

(3) This is just Part (3) of Theorem 2.1.

(4) Suppose $M \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, then $Hx + A^T y = 0$ and $Ax = 0$. If we multiply $Hx + A^T y$ on the left by x^T, we obtain $0 = x^T Hx + x^T A^T y = x^T Hx$ which implies that $x = 0$ since $x \in \text{Null}(A)$. But then $A^T y = 0$, so that $y = 0$ since rank $(A) = m$. Consequently, $\text{Null}(M) = \{0\}$, i.e., M is invertible. The result now follows from Part (1).

(5) By Part (1), this is just a restatement of Theorem 2.1 Part (4). □

The vector $\overline{y}$ appearing in this Theorem is call a *Lagrange multiplier* vector. Lagrange multiplier vectors play an essential role in constrained optimization and lie at the heart of what is called *duality theory*. This theory is more fully developed in Chapter **??**.

We now study how one might check when H is positive semi-definite as well as solving the equation $Hx + g = 0$ when H is positive semi-definite.

3. The Principal Minor Test for Positive Definiteness

Let $H \in \mathcal{S}^n$. We wish to obtain a test of when H is positive definite. First note that $H_{ii} = e_i^T He_i$, so that H can be positive definite only if $H_{ii} > 0,\ i = 1, \ldots, n$. This is only a "sanity check" for whether a matrix is positive definite. That is, if any diagonal element of H is not positive, then H cannot be positive definite. In this section we develop a necessary and sufficient condition for H to be positive definite based on the determinant. We begin with the following lemma.

LEMMA 3.1. *Let* $H \in \mathcal{S}^n$, $u \in \mathbb{R}^n$, *and* $\alpha \in \mathbb{R}$, *and consider the block matrix*

$$\hat{H} := \begin{bmatrix} H & u \\ u^T & \alpha \end{bmatrix} \in \mathcal{S}^{(n+1)} .$$

(1) The matrix $\hat{H}$ *is positive semi-definite if and only if* H *is positive semi-definite and there exists a vector* $z \in \mathbb{R}^n$ *such that* $u = Hz$ *and* $\alpha \geq z^T Hz$.

(2) The matrix $\hat{H}$ *is positive definite if and only if* H *is positive definite and* $\alpha > u^T H^{-1} u$.

PROOF. (1) Suppose H is positive semi-definite, and there exists z such that $u = Hz$ and $\alpha \geq z^T Hz$. Then for any $\hat{x} = \begin{bmatrix} x \\ x_n \end{bmatrix}$ where $x_n \in \mathbb{R}$ and $x \in \mathbb{R}^n$, we have

$$\begin{aligned} \hat{x}^T \hat{H} \hat{x} &= x^T Hx + 2x^T H x_n z + x_n^2 \alpha \\ &= (x + x_n z)^T H (x + x_n z) + x_n^2(\alpha - z^T Hz) \geq 0. \end{aligned}$$

Hence, $\hat{H}$ is positive semi-definite.

Conversely, suppose that $\hat{H}$ is positive semi-definite. Write $u = u_1 + u_2$ where $u_1 \in \text{Ran}(H)$ and $u_2 \in \text{Ran}(H)^\perp = \text{Null}(H)$, so that there is a $z \in \mathbb{R}^n$ such that $u_1 = Hz$. Then, for all $\hat{x} = \begin{pmatrix} x \\ x_n \end{pmatrix} \in \mathbb{R}^{(n+1)}$,

$$\begin{aligned} 0 \leq \hat{x}^T \hat{H} \hat{x} &= x^T Hx + 2x_n u^T x + \alpha x_n^2 \\ &= x^T Hx + 2x_n (u_1 + u_2)^T x + \alpha x_n^2 \\ &= x^T Hx + 2x_n z^T Hx + x_n^2 z^T Hz + x_n^2(\alpha - z^T Hz) + 2x_n u_2^T x \\ &= (x + x_n z)^T H (x + x_n z) + x_n^2(\alpha - z^T Hz) + 2x_n u_2^T x. \end{aligned}$$

Taking $x_n = 0$ tells us that H is positive semi-definite, and taking $\hat{x} = \begin{pmatrix} -tu_2 \\ 1 \end{pmatrix}$ for $t \in \mathbb{R}$ gives

$$\alpha - 2t \|u_2\|_2^2 \geq 0 \quad \text{for all } t \in \mathbb{R},$$

which implies that $u_2 = 0$. Finally, taking $\hat{x} = \begin{pmatrix} -z \\ 1 \end{pmatrix}$, tells us that $z^T H z \leq \alpha$ which proves the result.

(2) The proof follows the pattern of Part (1) but now we can take $z = H^{-1}u$. □

If the matrix H is invertible, we can apply a block Gaussian elimination to the matrix $\hat{H}$ in the lemma to obtain a matrix with block upper triangular structure:

$$\begin{bmatrix} I & 0 \\ (-H^{-1}u)^T & 1 \end{bmatrix} \begin{bmatrix} H & u \\ u^T & \alpha \end{bmatrix} = \begin{bmatrix} H & u \\ 0 & (\alpha - u^T H^{-1} u) \end{bmatrix}.$$

This factorization tells us that

$$\begin{aligned} \det \begin{bmatrix} H & u \\ u^T & \alpha \end{bmatrix} &= \det \begin{bmatrix} I & 0 \\ (-H^{-1}u)^T & 1 \end{bmatrix} \det \begin{bmatrix} H & u \\ u^T & \alpha \end{bmatrix} \\ &= \det \left(\begin{bmatrix} I & 0 \\ (-H^{-1}u)^T & 1 \end{bmatrix} \begin{bmatrix} H & u \\ u^T & \alpha \end{bmatrix} \right) \\ &= \det \begin{bmatrix} H & u \\ 0 & (\alpha - u^T H^{-1} u) \end{bmatrix} \\ &= \det(H)(\alpha - u^T H^{-1} u). \end{aligned} \tag{44}$$

We use this determinant identity in conjunction with the previous lemma to establish a test for whether a matrix is positive definite based on determinants. The test requires us to introduce the following elementary definition.

Definition 3.1. *[Principal Minors] The kth principal minor of a matrix $B \in \mathbb{R}^{n \times n}$ is the determinant of the upper left–hand corner $k \times k$–submatrix of B for $1 \leq k \leq n$.*

Proposition 3.1. *[The Principal Minor Test] Let $H \in \mathcal{S}^n$. Then H is positive definite if and only if each of its principal minors is positive.*

Proof. The proof proceeds by induction on the dimension n of H. The result is clearly true for $n = 1$. We now assume the result is true for $1 \leq k \leq n$ and show it is true for dimension $n + 1$. Write

$$H := \begin{bmatrix} \hat{H} & u \\ u^T & \alpha \end{bmatrix}.$$

Then Lemma 3.1 tells us that H is positive definite if and only if $\hat{H}$ is positive definite and $\alpha > u^T \hat{H}^{-1} u$. By the induction hypothesis, $\hat{H}$ is positive definite if and only if all of its principal minors are positive. If we now combine this with the expression (44), we get that H is positive definite if and only if all principal minors of $\hat{H}$ are positive and, by (44), $\det(H) = \det(\hat{H})(\alpha - u^T \hat{H}^{-1} u) > 0$, or equivalently, all principal minors of H are positive. □

This result only applies to positive definite matrices, and does not provide insight into how to solve linear equations involving H such as $Hx + g = 0$. These two issues can be addressed through the Cholesky factorization.

4. The Cholesky Factorizations

We now consider how one might solve a quadratic optimization problem. Recall that a solution only exists when H is positive semi-definite and there is a solution to the equation $Hx + g = 0$. Let us first consider solving the equation when H is positive definite. We use a procedure similar to the LU factorization but which also takes advantage of symmetry.

Suppose

$$H = \begin{bmatrix} \alpha_1 & h_1^T \\ h_1 & \widetilde{H}_1 \end{bmatrix}, \quad \text{where } \widetilde{H}_1 \in \mathcal{S}^n.$$

Note that $\alpha_1 = e_1^T H e_1 > 0$ since H is positive definite (if $\alpha_1 \leq 0$, then H cannot be positive definite), so there is no need to apply a permutation. Multiply H on the left by the Gaussian elimination matrix for the first column, we obtain

$$L_1^{-1} H = \begin{bmatrix} 1 & 0 \\ -\frac{h_1}{\alpha_1} & I \end{bmatrix} \begin{bmatrix} \alpha_1 & h_1^T \\ h_1 & \widetilde{H}_1 \end{bmatrix} = \begin{bmatrix} \alpha_1 & h_1^T \\ 0 & \widetilde{H}_1 - \alpha_1^{-1} h_1 h_1^T \end{bmatrix}.$$

By symmetry, we have

$$L_1^{-1} H L_1^{-T} = \begin{bmatrix} \alpha_1 & h_1^T \\ 0 & \widetilde{H}_1 - \alpha_1^{-1} h_1 h_1^T \end{bmatrix} \begin{bmatrix} 1 & -\frac{h_1^T}{\alpha_1} \\ 0 & I \end{bmatrix} = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \widetilde{H}_1 - \alpha_1^{-1} h_1 h_1^T \end{bmatrix}.$$

Set $H_1 = \widetilde{H}_1 - \alpha^{-1} h_1 h_1^T$. Observe that for every non-zero vector $v \in \mathbb{R}^{(n-1)}$,

$$v^T H_1 v = \begin{pmatrix} 0 \\ v \end{pmatrix}^T \begin{bmatrix} \alpha_1 & 0 \\ 0 & H_1 \end{bmatrix} \begin{pmatrix} 0 \\ v \end{pmatrix} = \left(L_1^{-T} \begin{pmatrix} 0 \\ v \end{pmatrix} \right)^T H \left(L_1^{-T} \begin{pmatrix} 0 \\ v \end{pmatrix} \right) > 0,$$

which shows that H_1 is positive definite. Decomposing H_1 as we did H gives

$$H_1 = \begin{bmatrix} \alpha_2 & h_2^T \\ h_2 & \widetilde{H}_2 \end{bmatrix}, \quad \text{where} \quad \widetilde{H}_2 \in \mathcal{S}^{(n-1)}.$$

Again, $\alpha_2 > 0$ since H_1 is positive definite (if $\alpha_2 \leq 0$, then H cannot be positive definite). Hence, we can repeat the reduction process for H_1. Continuing in this way, if at any stage we discover and $\alpha_i \leq 0$, then we terminate, since H cannot be positive definite.

If we can continue this process n times, we will have constructed a lower triangular matrix

$$L := L_1 L_2 \cdots L_n \quad \text{such that} \quad L^{-1} H L^{-T} = D, \quad \text{where} \quad D := \operatorname{diag}(\alpha_1, \alpha_2, \ldots, \alpha_n)$$

is a diagonal matrix with strictly positive diagonal entries. On the other hand, if at some point in the process we discover an α_i that is not positive, then H cannot be positive definite and the process terminates. That is, this computational procedure simultaneously tests whether H is positive definite as it tries to diagonalize H. We will call this process the *Cholesky diagonalization procedure*. It is used to establish the following factorization theorem.

THEOREM 4.1. *[The Cholesky Factorization] Let $H \in \mathcal{S}^n_+$ have rank k. Then there is a lower triangular matrix $L \in \mathbb{R}^{n \times k}$ such that $H = LL^T$. Moreover, if the rank of H is n, then there is a positive diagonal matrix D and a lower triangular matrix $\widetilde{L}$ with ones on its diagonal such that $H = \widetilde{L} D \widetilde{L}^T$.*

PROOF. Let the columns of the matrix $V_1 \in \mathbb{R}^{n \times k}$ be an orthonormal basis for $\operatorname{Ran}(H)$ and the columns of $V_2 \in \mathbb{R}^{n \times (n-k)}$ be an orthonormal basis for $\operatorname{Null}(H)$ and set $V = [V_1\ V_2] \in \mathbb{R}^{n \times n}$. Then

$$\begin{aligned} V^T H V &= \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} H [V_1\ V_2] \\ &= \begin{bmatrix} V_1^T H V_1 & V_1^T H V_2 \\ V_2^T H V_1 & V_2^T H V_2 \end{bmatrix} \\ &= \begin{bmatrix} V_1^T H V_1 & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

Since $\operatorname{Ran}(H) = \operatorname{Null}(H^T)^\perp = \operatorname{Null}(H)^\perp$, $V_1 H V_1^T \in \mathbb{R}^{k \times k}$ is symmetric and positive definite. By applying the procedure described prior to the statement of the theorem, we construct a nonsingular lower triangular matrix $\widetilde{L} \in \mathbb{R}^{k \times k}$ and a diagonal matrix $D = \operatorname{diag}(\alpha_1, \alpha_2, \ldots, \alpha_k)$, with $\alpha_i > 0,\ i = 1, \ldots, k$, such that $V_1 H V_1^T = \widetilde{L} D \widetilde{L}^T$. Set $\widehat{L} = \widetilde{L} D^{1/2}$ so that $V_1 H V_1^T = \widehat{L} \widehat{L}^T$. If $k = n$, taking $V = I$ proves the theorem by setting $L = \widehat{L}$. If $k < n$,

$$H = [V_1\ V_2] \begin{bmatrix} \widehat{L} \widehat{L}^T & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = (V_1 \widehat{L})(V_1 \widehat{L})^T.$$

Let $(V_1 \widehat{L})^T \in \mathbb{R}^{k \times n}$ have reduced QR factorization $(V_1 \widehat{L})^T = QR$ (see Theorem 5.1). Since $\widehat{L}^T$ has rank k, $Q \in \mathbb{R}^{k \times k}$ is unitary and $R = [R_1\ R_2]$ with $R_1 \in \mathbb{R}^{k \times k}$ nonsingular and $R_2 \in \mathbb{R}^{k \times (n-k)}$. Therefore,

$$H = (V_1 \widehat{L})(V_1 \widehat{L})^T = R^T Q^T Q R = R^T R.$$

The theorem follows by setting $L = R^T$. □

When H is positive definite, the factorization $H = LL^T$ is called the Cholesky factorization of H, and when $\text{rank}\,(H) < n$ it is called the *generalized Cholesky factorization* of H. In the positive definite case, the Cholesky diagonalization procedure computes the Cholesky factorization of H. On the other hand, when H is only positive semi-definite, the proof of the theorem provides a guide for obtaining the generalized Cholesky factorization.

4.1. Computing the Generalized Cholesky Factorization.

Step1: Initiate the Cholesky diagonalization procedure. If the procedure successfully completes n iterations, the Cholesky factorization has been obtained. Otherwise the procedure terminates at some iteration $k+1 < n$. If $\alpha_{k+1} < 0$, proceed no further since the matrix H is not positive semi-definite. If $\alpha_{k+1} = 0$, proceed to Step 2.

Step 2: In Step 1, the factorization

$$\widehat{L}^{-1} H \widehat{L}^{-T} = \begin{bmatrix} \widehat{D} & 0 \\ 0 & \widehat{H} \end{bmatrix},$$

where

$$\widehat{L} = \begin{bmatrix} \widehat{L}_1 & 0 \\ \widehat{L}_2 & I_{(n-k)} \end{bmatrix}$$

with $\widehat{L}_1 \in \mathbb{R}^{k\times k}$ lower triangular with ones on the diagonal, $\widehat{D} = \text{diag}(\alpha_1, \alpha_2, \ldots, \alpha_k) \in \mathbb{R}^{k\times k}$ with $\alpha_i > 0$ $i = 1, \ldots, k$, and $\widehat{H} \in \mathbb{R}^{(n-k)\times(n-k)}$ with $\widehat{H}$ symmetric has a nontrivial null space. Let the full QR factorization of $\widehat{H}$ be given by

$$\widehat{H} = [U_1\ U_2] \begin{bmatrix} R_1 & R_2 \\ 0 & 0 \end{bmatrix} = U \begin{bmatrix} R \\ 0 \end{bmatrix},$$

where
- $U = [U_1\ U_2] \in \mathbb{R}^{k\times k}$ is unitary,
- the columns of $U_1 \in \mathbb{R}^{k\times k_1}$ form an orthonormal basis for $\text{Ran}(\widehat{H})$ with $k_1 = \text{rank}\left(\widehat{H}\right) < k$,
- the columns of $U_2 \in \mathbb{R}^{k\times(k-k_1)}$ for an orthonormal basis for $\text{Null}(\widehat{H})$,
- $R_1 \in \mathbb{R}^{k_1\times k_1}$ is upper triangular and nonsingular,
- $R_2 \in \mathbb{R}^{k_1\times(k-k_1)}$, and
- $R = [R_1\ R_2] \in \mathbb{R}^{k_1\times k}$.

Consequently,

$$\begin{aligned}
\begin{bmatrix} U_1^T\widehat{H}U_1 & 0 \\ 0 & 0 \end{bmatrix} &= \begin{bmatrix} U_1^T\widehat{H}U_1 & U_1^T\widehat{H}U_2 \\ U_2^T\widehat{H}U_1 & U_2^T\widehat{H}U_2 \end{bmatrix} \\
&= U^T\widehat{H}U \\
&= \begin{bmatrix} R \\ 0 \end{bmatrix} [U_1\ U_2] \\
&= \begin{bmatrix} RU_1 & RU_2 \\ 0 & 0 \end{bmatrix},
\end{aligned}$$

and so $RU_2 = 0$ and $U_1^T\widehat{H}U_1 = RU_1 \in \mathbb{R}^{k_1\times k_1}$ is a nonsingular symmetric matrix.
Note that only the reduced QR factorization of $H = U_1R$ is required since $U_1^T\widehat{H}U_1 = RU_1$.

Step 4: Initiate the Cholesky diagonalization procedure on $U_1^T\widehat{H}U_1$. If the procedure successfully completes k_1 iterations, the Cholesky factorization

$$U_1^T\widehat{H}U_1 = \widehat{L}_3\widehat{L}_3^T$$

has been obtained. If this does not occur, the procedure terminates at some iteration $j < k_1$ with $\alpha_j < 0$ since $U_1^T\widehat{H}U_1$ is nonsingular. In this case, terminate the process since H cannot be positive semi-definfite. Otherwise proceed to Step 5.

Step 5: We now have

$$\begin{aligned}
H &= \begin{bmatrix} \widehat{L}_1 & 0 \\ \widehat{L}_2 & I_{(n-k)} \end{bmatrix} \begin{bmatrix} \widehat{D} & 0 \\ 0 & \widehat{H} \end{bmatrix} \begin{bmatrix} \widehat{L}_1^T & \widehat{L}_2^T \\ 0 & I_{(n-k)} \end{bmatrix} \\
&= \begin{bmatrix} \widehat{L}_1 & 0 \\ \widehat{L}_2 & I_{(n-k)} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & U \end{bmatrix} \begin{bmatrix} \widehat{D} & 0 \\ 0 & U^T\widehat{H}U \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & U^T \end{bmatrix} \begin{bmatrix} \widehat{L}_1^T & \widehat{L}_2^T \\ 0 & I_{(n-k)} \end{bmatrix} \\
&= \begin{bmatrix} \widehat{L}_1 & 0 & 0 \\ \widehat{L}_2 & U_1 & U_2 \end{bmatrix} \begin{bmatrix} \widehat{D} & 0 & 0 \\ 0 & \widehat{L}_3\widehat{L}_3^T & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \widehat{L}_1^T & \widehat{L}_2^T \\ 0 & U_1^T \\ 0 & U_2^T \end{bmatrix} \\
&= \begin{bmatrix} \widetilde{L}_1\widehat{D}^{1/2} & 0 & 0 \\ \widehat{L}_2\widehat{D}^{1/2} & U_1\widehat{L}_3 & 0 \end{bmatrix} \begin{bmatrix} \widehat{D}^{1/2}\widetilde{L}_1^T & \widehat{D}^{1/2}\widehat{L}_2^T \\ 0 & \widehat{L}_3^T U_1^T \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} L_1 & 0 \\ L_2 & U_1\widehat{L}_3 \end{bmatrix} \begin{bmatrix} L_1^T & L_2^T \\ 0 & \widehat{L}_3^T U_1^T \end{bmatrix},
\end{aligned}$$

where $L_1 = \widetilde{L}_1\widehat{D}^{1/2} \in \mathbb{R}^{k\times k}$ is lower triangular, $L_2 = \widehat{L}_2\widehat{D}^{1/2} \in \mathbb{R}^{(n-k)\times k}$, and $U_1\widehat{L}_3 \in \mathbb{R}^{(n-k)\times k_1}$. In particular, $k+k_1 = \text{rank}\,(H)$ since L_1 has rank k and $U_1\widehat{L}_3$ has rank k_1. Let $\widehat{L}_3^T U_1^T$ have QR factorization $\widehat{L}_3^T U_1^T = VL_3^T$, where $V \in \mathbb{R}^{k_1\times k_1}$ is unitary and $L_3 \in \mathbb{R}^{k_1\times(n-k)}$ is lower triangular. Then

$$H = \begin{bmatrix} L_1 & 0 \\ L_2 & U_1\widehat{L}_3 \end{bmatrix} \begin{bmatrix} L_1^T & L_2^T \\ 0 & \widehat{L}_3^T U_1^T \end{bmatrix} = \begin{bmatrix} L_1 & 0 \\ L_2 & L_3V^T \end{bmatrix} \begin{bmatrix} L_1^T & L_2^T \\ 0 & VL_3^T \end{bmatrix} = \begin{bmatrix} L_1 & 0 \\ L_2 & L_3 \end{bmatrix} \begin{bmatrix} L_1^T & L_2^T \\ 0 & L_3^T \end{bmatrix},$$

since $V^TV = I_{k_1}$. This is the generalized Cholesky factorization of H.

4.2. Computing Solutions to the Quadratic Optimization Problem via Cholesky Factorizations.

Step 1: Apply the procedure described in the previous section for computing the generalized Cholesky factorization of H. If it is determined that H is not positive definite, then proceed no further since the problem (31) has no solution and the optimal value is $-\infty$.

Step 2: Step 1 provides us with the generalized Cholesky factorization for $H = LL^T$ with $L^T = [L_1^T\ L_2^T]$, where $L_1 \in \mathbb{R}^{k\times k}$ and $L_2 \in \mathbb{R}^{(n-k)\times k}$ with $k = \text{rank}\,(H)$. Write

$$g = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix},$$

where $g_1 \in \mathbb{R}^k$ and $g_2 \in \mathbb{R}^{(n-k)}$. Since $\text{Ran}(H) = \text{Ran}(L)$, the system $Hx + g = 0$ is solvable if and only if $-g \in \text{Ran}(L)$. That is, there exists $w \in \mathbb{R}^k$ such that $Lw = -g$, or equivalently,

$$L_1w = -g_1 \quad \text{and} \quad L_2w = -g_2.$$

Since L_1 is invertible, the system $L_1w = -g_1$ has as its unique solution $\overline{w} = L_1^{-1}g_1$. Note that $\overline{w}$ is easy to compute by forward substitution since L_1 is lower triangular. Having $\overline{w}$ check to see if $L_2\overline{w} = -g_2$. If this is not the case, then proceed no further, since the system $Hx + g = 0$ has no solution and so the optimal value in (31) is $-\infty$. Otherwise, proceed to Step 3.

Step 3: Use back substitution to solve the equation $L_1^Ty = \overline{w}$ for $\overline{y} := L_1^{-T}\overline{w}$ and set

$$\overline{x} = \begin{pmatrix} \overline{y} \\ 0 \end{pmatrix}.$$

Then

$$H\overline{x} = LL^T\overline{x} = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} [L_1^T\ L_2^T] \begin{pmatrix} \overline{y} \\ 0 \end{pmatrix} = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} \overline{w} = -g\ .$$

Hence, $\overline{x}$ solves the equation $Hx + g = 0$ and so is an optimal solution to the quadratic optimization problem (31).

5. Linear Least Squares Revisited

We have already seen that the least squares problem is a special case of the problem of minimizing a quadratic function. But what about the reverse? Part (4) of Theorem 1.1 tells us that, in general, the reverse cannot be true since the linear least squares problem always has a solution. But what about the case when the quadratic optimization problem has a solution? In this case the matrix H is necessarily positive semi-definite and a solution to the system $Hx+g=0$ exists. By Theorem 4.1, there is a lower triangular matrix $L \in \mathbb{R}^{n\times k}$, where $k = \text{rank}\,(H)$, such that $H = LL^T$. Set $A := L^T$. In particular, this implies that $\text{Ran}(H) = \text{Ran}(L) = \text{Ran}(A^T)$. Since $Hx+g = 0$, we know that $-g \in \text{Ran}(H) = \text{Ran}(A^T)$, and so there is a vector $b \in \mathbb{R}^k$ such that $-g = A^T b$. Consider the linear least squares problem

$$\min_{x\in\mathbb{R}^n} \tfrac{1}{2}\left\|Ax-b\right\|_2^2 .$$

As in (33), expand the objective in this problem to obtain

$$\begin{aligned}\tfrac{1}{2}\left\|Ax-b\right\|_2^2 &= \tfrac{1}{2}x^T(A^TA)x - (A^Tb)^Tx + \tfrac{1}{2}\left\|b\right\|_2^2 \\ &= \tfrac{1}{2}x^TLL^Tx + g^Tx + \beta \\ &= \tfrac{1}{2}x^THx + g^Tx + \beta,\end{aligned}$$

where $\beta = \frac{1}{2}\left\|b\right\|_2^2$. We have just proved the following result.

Proposition 5.1. *A quadratic optimization problem of the form* (31) *has an optimal solution if and only if it is equivalent to a linear least squares problem.*

6. The Conjugate Gradient Algorithm

The Cholesky factorization is an important and useful tool for computing solutions to the quadratic optimization problem, but it is too costly to be employed in many very large scale applications. In some applications, the matrix H is too large to be stored or it is not available as a data structure. However, in these problems it is often the case that the matrix vector product Hx can be obtained for a given vector $x \in \mathbb{R}^n$. This occurs, for example, in signal processing applications. In this section, we develop an algorithm for solving the quadratic optimization problem (34) that only requires access to the matrix vector products Hx. Such an algorithm is called a *matrix free* method since knowledge the whole matrix H is not required. In such cases the Cholesky factorization is inefficient to compute. The focus of this section is the study of the matrix free method known as the *conjugate gradient algorithm*. Throughout this section we assume that the matrix H is positive definite.

6.1. Conjugate Direction Methods. Consider the problem (34) where it is known that H is symmetric and positive definite. In this case it is possible to define a notion of *orthogonality* or *congugacy* with respect to H.

Definition 6.1 (Conjugacy). *Let $H \in \mathcal{S}^n_{++}$. We say that the vectors $x, y \in \mathbb{R}^n\backslash\{0\}$ are H-conjugate (or H-orthogonal) if $x^THy = 0$.*

Proposition 6.1. *[Conjugacy implies Linear Independence]*
If $H \in \mathcal{S}^n_{++}$ and the set of nonzero vectors d^0, $d^1, \ldots, d^k$ are (pairwise) H-conjugate, then these vectors are linearly independent.

Proof. If $0 = \sum_{i=0}^{k} \mu_i d^i$, then for $\bar{i} \in \{0, 1, \ldots, k\}$

$$0 = (d^{\bar{i}})^T H[\sum_{i=0}^{k} \mu_i d^i] = \mu_{\bar{i}}(d^{\bar{i}})^T H d^{\bar{i}},$$

Hence $\mu_i = 0$ for each $i = 0, \ldots, k$. □

Let $x^0 \in \mathbb{R}^n$ and suppose that the vectors $d^0, d^1, \ldots, d^{k-1} \in \mathbb{R}^n$ are H-conjugate. Set $S = \text{Span}(d^0, d^1, \ldots, d^{k-1})$. Theorem 2.1 tells us that there is a unique optimal solution $\overline{x}$ to the problem $\min\left\{\frac{1}{2}x^THx + g^Tx \,\middle|\, x \in x^0 + S\right\}$, and that $\overline{x}$ is uniquely identified by the condition $H\overline{x} + g \in S^\perp$, or equivalently, $0 = (d^j)^T(H\overline{x} + g)$, $j = 0, 1, \ldots, k-1$. Since $\overline{x} \in x_0 + S$, there are scalars $\mu_0, \ldots, \mu_{n-1}$ such that

$$\overline{x} = x^0 + \mu_0 d^0 + \ldots + \mu_{k-1} d^{k-1}, \tag{45}$$

and so, for each $j = 0, 1, \ldots, k-1$,

$$\begin{array}{rcl} 0 & = & (d^j)^T(H\overline{x} + g) \\ & = & (d^j)^T\left(H(x^0 + \mu_0 d^0 + \ldots + \mu_{k-1}d^{k-1}) + g\right) \\ & = & (d^j)^T(Hx^0 + g) + \mu_0 (d^j)^T H d^0 + \ldots + \mu_{k-1}(d^j)^T H d^{k-1} \\ & = & (d^j)^T(Hx^0 + g) + \mu_j (d^j)^T H d^j \ . \end{array}$$

Therefore,

(46)
$$\mu_j = \frac{-(Hx^0 + g)^T(d^j)}{(d^j)^T H d^j} \quad j = 0, 1 \ldots, k-1 \ .$$

This observation motivates the following theorem.

THEOREM 6.1. *[Expanding Subspace Theorem]*
Consider the problem (34) with $H \in \mathcal{S}^n_{++}$, and set $f(x) = \frac{1}{2}x^T Hx + g^T x$. Let $\{d^i\}_{i=0}^{n-1}$ be a sequence of nonzero H-conjugate vectors in $\mathbb{R}^n$. Then, for any $x^0 \in \mathbb{R}^n$ the sequence $\{x^k\}$ generated according to

$$x^{k+1} := x^k + t_k d^k,$$

with

$$t_k := \arg\min\{f(x^k + td^k) : t \in \mathbb{R}\},$$

has the property that $f(x) = \frac{1}{2}x^T Hx + g^T x$ attains its minimum value on the affine set $x^0 +$ Span $\{d^0, \ldots, d^{k-1}\}$ at the point x^k. In particular, if $k = n$, then x^n is the unique global solution to the problem (34).

PROOF. Let us first compute the value of the t_k's. For $j = 0, \ldots, k-1$, define $\varphi_j : \mathbb{R} \to \mathbb{R}$ by

$$\begin{array}{rl} \varphi_j(t) & = f(x^j + td^j) \\ & = \frac{t^2}{2}(d^j)^T H d^j + t(g^j)^T d^j + f(x^j), \end{array}$$

where $g^j = Hx^j + g$. Then, for $j = 0, \ldots, k-1$, $\varphi_j'(t) = t(d^j)^T H d^j + (g^j)^T d^j$ and $\varphi_j''(t) = (d^j)^T H d^j > 0$. Since $\varphi_j''(t) > 0$, our one dimensional calculus tells us that φ_j attains its global minmimum value at the unique solution t_j to the equation $\varphi_j'(t) = 0$, i.e.,

$$t_j = -\frac{(g^j)^T d^j}{(d^j)^T H d^j}.$$

Therefore,

$$x^k = x^0 + t_0 d^0 + t_1 d^1 + \cdots + t_k d^k$$

with

$$t_j = -\frac{(g^j)^T d^j}{(d^j)^T H d^j}, \ j = 0, 1, \ldots, k.$$

In the discussion preceding the theorem it was shown that if $\overline{x}$ is the solution to the problem

$$\min\left\{ f(x) \,\middle|\, x \in x^0 + \mathrm{Span}(d^0, d^1, \ldots, d^k) \right\} \ ,$$

then $\overline{x}$ is given by (45) and (46). Therefore, if we can now show that $\mu_j = t_j, \ j = 0, 1, \ldots, k$, then $\overline{x} = x_k$ proving the result. For each $j \in \{0, 1, \ldots, k\}$ we have

$$\begin{array}{rcl} (g^j)^T d^j & = & (Hx^j + g)^T d^j \\ & = & \left(H(x^0 + t_0 d^0 + t_1 d^1 + \cdots + t_{j-1}d^{j-1}) + g\right)^T d^j \\ & = & (Hx^0 + g)^T d^j + t_0 (d^0)^T H d^j + t_1 (d^1)^T H d^j + \cdots + t_{j-1}(d^{j-1})^T H d^j \\ & = & (Hx^0 + g)^T d^j \\ & = & (g^0)^T d^j \ . \end{array}$$

Therefore, for each $j \in \{0, 1, \ldots, k\}$,

$$t_j = \frac{-(g^j)^T d^j}{(d^j)^T H d^j} = \frac{-(g^0)^T d^j}{(d^j)^T H d^j} = \mu_j,$$

which proves the result. □

6.2. The Conjugate Gradient Algorithm. The major drawback of the Conjugate Direction Algorithm of the previous section is that it seems to require that a set of H-conjugate directions must be obtained before the algorithm can be implemented. This is in opposition to our working assumption that H is so large that it cannot be kept in storage since any set of H-conjugate directions requires the same amount of storage as H. However, it is possible to generate the directions d^j one at a time and then discard them after each iteration of the algorithm. One example of such an algorithm is the Conjugate Gradient Algorithm.

The C-G Algorithm:

Initialization: $x^0 \in \mathbb{R}^n$, $d^0 = -g^0 = -(Hx^0 + g)$.

For $k = 0, 1, 2, \ldots$

$$\begin{array}{lll} t_k & := -(g^k)^T d^k / (d^k)^T H d^k & \\ x^{k+1} & := x^k + t_k d^k & \\ g^{k+1} & := H x^{k+1} + g & \qquad \text{(STOP if } g^{k+1} = 0) \\ \beta_k & := (g^{k+1})^T H d^k / (d^k)^T H d^k & \\ d^{k+1} & := -g^{k+1} + \beta_k d^k & \\ k & := k + 1. & \end{array}$$

Theorem 6.2. *[Conjugate Gradient Theorem]*
The C-G algorithm is a conjugate direction method. If it does not terminate at x^k (i.e. $g^k \neq 0$), then

1. *Span* $[g^0, g^1, \ldots, g^k] =$ *span* $[g^0, Hg^0, \ldots, H^k g^0]$
2. *Span* $[d^0, d^1, \ldots, d^k] =$ *span* $[g^0, Hg^0, \ldots, H^k g^0]$
3. $(d^k)^T H d^i = 0$ *for* $i \leq k - 1$
4. $t_k = (g^k)^T g^k / (d^k)^T H d^k$
5. $\beta_k = (g^{k+1})^T g^{k+1} / (g^k)^T g^k$.

Proof. We first prove (1)-(3) by induction. The results are clearly true for $k = 0$. Now suppose they are true for k, we show they are true for $k + 1$. First observe that

$$g^{k+1} = g^k + t_k H d^k$$

so that $g^{k+1} \in$ Span$[g^0, \ldots, H^{k+1} g^0]$ by the induction hypothesis on (1) and (2). Also $g^{k+1} \notin$ Span $[d^0, \ldots, d^k]$, otherwise, by Theorem 2.1 Part (1), $g^{k+1} = Hx^{k+1} + g = 0$ since the method is a conjugate direction method up to step k by the induction hypothesis. Hence $g^{k+1} \notin$ Span $[g^0, \ldots, H^k g^0]$ and so Span $[g^0, g^1, \ldots, g^{k+1}] =$ Span $[g^0, \ldots, H^{k+1} g^0]$, which proves (1).

To prove (2) write

$$d^{k+1} = -g^{k+1} + \beta_k d^k$$

so that (2) follows from (1) and the induction hypothesis on (2).

To see (3) observe that

$$(d^{k+1})^T H d^i = -(g^{k+1})^T H d^i + \beta_k (d^k)^T H d^i.$$

For $i = k$ the right hand side is zero by the definition of β_k. For $i < k$ both terms vanish. The term $(g^{k+1})^T H d^i = 0$ by Theorem 6.1 since $Hd^i \in$ Span$[d^0, \ldots, d^k]$ by (1) and (2). The term $(d^k)^T H d^i$ vanishes by the induction hypothesis on (3).

To prove (4) write

$$-(g^k)^T d^k = (g^k)^T g^k - \beta_{k-1} (g^k)^T d^{k-1}$$

where $(g^k)^T d^{k-1} = 0$ by Theorem 6.1.

To prove (5) note that $(g^{k+1})^T g^k = 0$ by Theorem 6.1 because $g^k \in$ Span$[d^0, \ldots, d^k]$. Hence

$$(g^{k+1})^T H d^k = \frac{1}{t_k} (g^{k+1})^T [g^{k+1} - g^k] = \frac{1}{t_k} (g^{k+1})^T g^{k+1}.$$

Therefore,

$$\beta_k = \frac{1}{t_k} \frac{(g^{k+1})^T g^{k+1}}{(d^k)^T H d^k} = \frac{(g^{k+1})^T g^{k+1}}{(g^k)^T g^k}.$$

□

Remarks:

(1) The C–G method is an example of a *descent method* since the values

$$f(x^0),\ f(x^1),\ \ldots, f(x^n)$$

form a decreasing sequence.

(2) It should be observed that due to the occurrence of round-off error the C-G algorithm is best implemented as an iterative method. That is, at the end of n steps, x^n may not be the global optimal solution and the intervening directions d^k may not be H-conjugate. Consequently, the algorithm is usually iterated until $\|g^k\|_2$ is sufficiently small. Due to the observations in the previous remark, this approach is guarenteed to continue to reduce the function value if possible since the overall method is a descent method. In this sense the C–G algorithm is self correcting.