QUADRATIC FORMS AND THE SECOND DERIVATIVE TEST

1. Quadratic Forms

Suppose \( F : \mathbb{R}^3 \to \mathbb{R} \) is given by a quadratic polynomial with no constant or linear terms. Every such function may be expressed as
\[
F(x) = x^T A x
\]
or, equivalently, as
\[
F(x) = x \cdot A x
\]
where \( A \) is a symmetric matrix whose entries depend on the coefficients of the polynomial.

For example, if \( F : \mathbb{R}^3 \to \mathbb{R} \) is given by
\[
F(x, y, z) = x^2 + 3y^2 - 7z^2 + 4xy - 6xz + 2yz
\]
then we can express \( F \) as
\[
F(x, y, z) = \begin{bmatrix} x & y & z \end{bmatrix} \begin{pmatrix} 1 & 2 & -3 \\ 2 & 3 & 1 \\ -3 & 1 & -7 \end{pmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}.
\]
Such a function is called a quadratic form. Note that if \( x = 0 \), then \( x^T A x = 0 \). In many situations, it is important to know the behavior of \( x^T A x \) for nonzero \( x \). The following theorem is very useful in such situations:

**Theorem 1.** Suppose \( A \) is a symmetric matrix. Then
\[
\lambda_{\text{MIN}} \|x\|^2 \leq x^T A x \leq \lambda_{\text{MAX}} \|x\|^2
\]
for all \( x \), where \( \lambda_{\text{MIN}} \) is the smallest eigenvalue of \( A \) and \( \lambda_{\text{MAX}} \) is the largest eigenvalue of \( A \).

Furthermore, if \( x \) is any eigenvector with eigenvalue \( \lambda \), then
\[
x^T A x = \lambda \|x\|^2.
\]

**Proof.** For simplicity we give the proof for 3 by 3 matrices. The general case is proved in exactly the same way.

Since \( A \) is symmetric, there is an orthonormal basis \( v_1, v_2, v_3 \) of \( \mathbb{R}^3 \) consisting of eigenvectors of \( A \). Let \( \lambda_1, \lambda_2, \) and \( \lambda_3 \) be the corresponding eigenvalues.

Consider a vector \( x \in \mathbb{R}^n \). Since the \( v_i \)'s form a basis, we can express \( x \) as a linear combination of the them:
\[
x = c_1 v_1 + c_2 v_2 + c_3 v_3.
\]

Note that
\[
\|x\|^2 = x \cdot x
\]
\[
= (c_1 v_1 + c_2 v_2 + c_3 v_3) \cdot (c_1 v_1 + c_2 v_2 + c_3 v_3)
\]
\[
= c_1^2 \|v_1\|^2 + c_2^2 \|v_2\|^2 + c_3^2 \|v_3\|^2
\]
\[
= c_1^2 + c_2^2 + c_3^2.
\]
(For a general basis, we would also get terms like \(c_1c_2v_1 \cdot v_2\), but here those terms vanish since the \(v_i\)'s are orthogonal to each other.)

We have shown

\[
\|x\|^2 = c_1^2 + c_2^2 + c_3^2.
\]

Now

\[
x^T A x = x \cdot A x = (c_1 v_1 + c_2 v_2 + c_3 v_3) \cdot A(c_1 v_1 + c_2 v_2 + c_3 v_3)
= (c_1 v_1 + c_2 v_2 + c_3 v_3) \cdot (c_1 A v_1 + c_2 A v_2 + c_3 A v_3)
= (c_1 v_1 + c_2 v_2 + c_3 v_3) \cdot (\lambda_1 c_1 v_1 + \lambda_2 c_2 v_2 + \lambda_3 c_3 v_3)
= \lambda_1 c_1^2 + \lambda_2 c_2^2 + \lambda_3 c_3^2.
\]

Note that replacing the \(\lambda_i\)'s by \(\lambda_{\text{MAX}}\) can only make this expression larger. Thus

\[
x^T A x \leq \lambda_{\text{MAX}}(c_1^2 + c_2^2 + c_3^2) = \lambda_{\text{MAX}} \|x\|^2
\]

since \(c_1^2 + c_2^2 + c_3^2 = \|x\|^2\) (by equation (1)).

Similarly,

\[
x^T A x \geq \lambda_{\text{MIN}}(c_1^2 + c_2^2 + c_3^2) = \lambda_{\text{MIN}} \|x\|^2.
\]

Finally, if \(x\) is an eigenvector with eigenvalue \(\lambda\), then

\[
x^T A x = x \cdot A x = x \cdot \lambda x = \lambda \|x\|^2.
\]

\[\square\]

**Remark.** Suppose \(v_1, \ldots, v_k\) are \(k\) of the eigenvectors of the symmetric matrix \(A\), with eigenvalues \(\lambda_1, \ldots, \lambda_k\). Suppose we have arranged them so that \(\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k\). Let \(W\) be the subspace spanned by \(v_1, \ldots, v_k\). Then

\[
\lambda_1 \|x\|^2 \leq x^T A x \leq \lambda_k \|x\|^2
\]

for every \(x \in W\). The proof is exactly the same as the proof of Theorem 1.

**Theorem 2.** Suppose \(A\) is a symmetric \(n\) by \(n\) matrix.

1. If \(\lambda_{\text{MIN}} > 0\), then \(x^T A x > 0\) for every nonzero \(x \in \mathbb{R}^n\).
2. If \(\lambda_{\text{MIN}} \geq 0\), then \(x^T A x \geq 0\) for every \(x \in \mathbb{R}^n\).
3. If \(\lambda_{\text{MAX}} < 0\), then \(x^T A x < 0\) for every nonzero \(x \in \mathbb{R}^n\).
4. If \(\lambda_{\text{MAX}} \leq 0\), then \(x^T A x \leq 0\) for every nonzero \(x \in \mathbb{R}^n\).
5. If \(\lambda_{\text{MIN}} < 0 < \lambda_{\text{MAX}}\), then \(x^T A x\) is \(> 0\) for some \(x\)'s and is \(< 0\) for other \(x\)'s.

**Definition.** In case (1), we say that \(A\) is positive definite. In case (2), we say that \(A\) is positive semidefinite. In case (3), we say that \(A\) is negative definite. In case (4), we say that \(A\) is negative semidefinite. Finally, in case (5), we say \(A\) is indefinite.

**Proof.** Suppose \(\lambda_{\text{MIN}} > 0\). If \(x\) is a nonzero vector, then by Theorem 1,

\[
x^T A x \geq \lambda_{\text{MIN}} \|x\|^2 > 0.
\]

This proves (1). Statements (2), (3), and (4) are proved in almost exactly the same way.

To prove (5), suppose that \(\lambda_{\text{MIN}} < 0 < \lambda_{\text{MAX}}\). If \(x\) is an eigenvector with eigenvalues \(\lambda_{\text{MIN}}\), then

\[
x^T A x = \lambda_{\text{MIN}} \|x\|^2 < 0.
\]
We can summarize Theorem 2 as follows:
- Eigenvalues all positive $\implies A$ positive definite
- Eigenvalues all $\geq 0 \implies A$ positive semidefinite
- Eigenvalues all negative $\implies A$ negative definite
- Eigenvalues all $\leq 0 \implies A$ negative semidefinite
- Some eigenvalues positive, some negative $\implies A$ is indefinite.

2. THE SECOND DERIVATIVE TEST

Consider a $C^2$ function $F : \mathbb{R}^n \to \mathbb{R}$. A critical point of $F$ is a point $a$ such that $\nabla F(a) = 0$.

**Definition.** We say that
- $F$ has a global minimum at $a$ provided $F(a) \leq F(x)$ for all $x$.
- $F$ has a local minimum at $a$ provided $F(a) \leq F(x)$ for all $x$ in some ball around $a$.
- $F$ has a strict local minimum at $a$ provided $F(a) < F(x)$ for all $x \neq a$ in some ball around $a$.
- $F$ has a global maximum at $a$ provided $F(a) \leq F(x)$ for all $x$.
- $F$ has a local maximum at $a$ provided $F(a) \leq F(x)$ for all $x$ in some ball around $a$.
- $F$ has a strict local maximum at $a$ provided $F(a) < F(x)$ for all $x \neq a$ in some ball around $a$.

A critical point that is neither a local minimum nor a local maximum is called a saddle point.

Every global or local minimum or maximum must occur at a critical point. In general, it is difficult to determine whether a particular critical point is a global maximum or a global minimum. However, there is a test that will determine, in most cases, whether it is a local maximum, a local minimum, or a saddle point.

**Theorem 3. [Second Derivative Test]** Suppose $a$ is a critical point of a $C^2$ function $F : \mathbb{R}^n \to \mathbb{R}$.

1. If all of the eigenvalues of the Hessian $HF(a)$ are positive, then $F$ has a strict local minimum at $a$.
2. If all of the eigenvalues of the Hessian are negative, then $F$ has a strict local maximum at $a$.
3. If some eigenvalues are positive and others are negative, then $a$ is a saddle point of $F$.

**Idea of proof.** Suppose all the eigenvalues are positive. Then $\lambda_{\text{MIN}} > 0$, where $\lambda_{\text{MIN}}$ is the smallest eigenvalue.

Recall that
\[ F(a + v) \simeq F(a) + \nabla F(a) \cdot v + \frac{1}{2} v^T HF(a)v \]
4 QUADRATIC FORMS AND THE SECOND DERIVATIVE TEST

for small \( v \). More precisely,

\[
F(a + v) = F(a) + \nabla F(a) \cdot v + \frac{1}{2} v^T H F(a) v + \text{error}
\]

where the error term is small compared to \( \|v\|^2 \) (if \( v \) is small): 

\[
\lim_{v \to 0} \frac{\text{error}}{\|v\|^2} = 0.
\]

Now \( \nabla F(a) = 0 \) since \( a \) is a critical point, so equation (2) becomes

\[
F(a + v) = F(a) + \frac{1}{2} v^T H F(a) v + \text{error}
\]

Thus by Theorem 2,

\[
F(a + v) \geq F(a) + \lambda_{\text{MIN}} \|v\|^2 + \text{error}.
\]

Since \( \lambda_{\text{MIN}} > 0 \), the term \( \lambda_{\text{MIN}} \|v\|^2 \) is \( > 0 \) if \( v \) is nonzero. If \( v \) is also small, then the error term will be very small compared to \( \lambda_{\text{MIN}} \|v\|^2 \), so

\[
\lambda_{\text{MIN}} \|v\|^2 + \text{error} > 0
\]

and therefore

\[
F(a + v) > F(a).
\]

Statements (2) and (3) can be proved in a similar manner. \( \square \)

If neither the smallest nor the largest eigenvalue of the Hessian is 0, then we can use Theorem 3 to determine what kind of critical points \( a \) is. However, if \( \lambda_{\text{MIN}} = 0 \) or if \( \lambda_{\text{MAX}} = 0 \), then it is impossible to determine from the Hessian whether \( a \) is a local maximum, a local minimum, or neither: one would need to consider higher-order partial derivatives.

3. THE SIGNS OF EIGENVALUES

Note that to use the second derivative test above (Theorem 3), we don’t really need to know the eigenvalues of the Hessian: we need only know the signs of the eigenvalues. Fortunately, it is possible to get the necessary information about the signs without actually calculating the eigenvalues.

Consider an \( n \) by \( n \) symmetric matrix \( A \). For \( k = 1, 2, \ldots, n \), let \( d_k \) be the determinant of the \( k \) by \( k \) matrix in the upper left corner of \( A \). Thus

\[
d_1 = a_{11}, \quad d_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \quad d_3 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}, \ldots, \quad d_n = \det A.
\]

Theorem 4. Suppose that \( A \) is a symmetric \( n \) by \( n \) matrix.

1. If \( d_1, d_2, \ldots, d_n \) are all positive, then all the eigenvalues of \( A \) are positive (so \( A \) is positive definite).

2. If \( d_1, d_3, d_5, \ldots \) are all negative and if \( d_2, d_4, d_6, \ldots \) are all positive, then all the eigenvalues of \( A \) are positive (so \( A \) is positive definite).

3. If \( d_n \neq 0 \) and if the \( d_k \)'s do not fit the pattern in (1) or in (2), then some eigenvalues of \( A \) are positive and others are negative (so \( A \) is indefinite).

See section 6 below for the proof.
4. Easier Second Derivative Test

If we combine Theorems 3 and 4, we get

**Theorem 5 (Easier 2nd Derivative Test).** Suppose that \( a \) is a critical point of \( C^2 \) function \( F : \mathbb{R}^n \to \mathbb{R} \), that \( A = HF(a) \) is the Hessian, and that \( d_1, d_2, \ldots \) are as in equation (3).

1. If \( d_1, d_2, \ldots, d_n \) are all positive, then \( a \) is a strict local minimum of \( F \).
2. If \( d_1, d_3, d_5, \ldots \) are all negative and if \( d_2, d_4, d_6, \ldots \) are all positive, then \( a \) is a strict local maximum of \( F \).
3. If \( d_n \neq 0 \) and if the \( d_k \)'s do not fit the pattern in (1) or in (2), then \( a \) is a saddle point of \( F \).

If \( d_n = 0 \), then \( a \) is called a *degenerate critical point* of \( F \). In that case, Theorem 5 gives no information about whether \( a \) is a local maximum, a local minimum, or a saddle point.

5. Examples

1. Suppose the Hessian at a critical point \( a \) of a function \( F : \mathbb{R}^3 \to \mathbb{R} \) is

\[
\begin{bmatrix}
0 & 0 & 1 \\
0 & -7 & 2 \\
1 & 2 & 3
\end{bmatrix}
\]

Determine what kind of critical point \( a \) is.

**Solution:**

\[
d_3 = \begin{vmatrix} 0 & 0 & 1 \\ 0 & -7 & 2 \\ 1 & 2 & 3 \end{vmatrix} = -1(-7)1 = 7.
\]

Also, \( d_1 = 0 \). Thus \( d_3 \neq 0 \), but the \( d_k \)'s do not fit pattern (1) or (2) of Theorem 5, so \( a \) is a saddle point.

2. Same problem, but where the Hessian is

\[
\begin{bmatrix}
-2 & 0 & 1 \\
0 & -3 & 0 \\
1 & 0 & -4
\end{bmatrix}
\]

**Solution:** \( d_1 = -2, d_2 = \begin{vmatrix} -2 & 0 \\ 0 & -3 \end{vmatrix} = 6 \), and

\[
d_3 = \begin{vmatrix} -2 & 0 & 1 \\ 0 & -3 & 0 \\ 1 & 0 & -4 \end{vmatrix} = (-3)\begin{vmatrix} -2 & 1 \\ 1 & -4 \end{vmatrix} = (-3)((-2)(-4) - 1^2) = (-3)(7) = -21.
\]

Now \( d_1 \) and \( d_3 \) are both negative and \( d_2 \) is positive, so \( a \) is a strict local maximum.

3. Same problem, but where the Hessian is

\[
\begin{bmatrix}
5 & 1 & 2 \\
1 & 2 & 4 \\
1 & 4 & 8
\end{bmatrix}
\]

**Solution:** The columns are linearly dependent (columns 3 is two times column 2), so the determinant \( d_3 = 0 \). Thus \( a \) is a degenerate critical point.
6. PROOF OF THEOREM 4

The proof of Theorem 4 is based on the following fact:

**Lemma.** Suppose that $Q$ is an $n$ by $n$ matrix. Let $M$ be the $(n-1)$ by $(n-1)$ matrix in the upper left corner of $Q$, and suppose that $M$ is positive definite. Then $Q$ is positive definite if and only if $\det Q > 0$.

**Proof.** Since $Q$ is symmetric, there is an orthonormal basis $v_1, \ldots, v_n$ of eigenvectors of $Q$. Let $\lambda_1, \ldots, \lambda_n$ be the corresponding eigenvalues. We may arrange them so that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. (Thus $\lambda_1$ and $\lambda_2$ are the smallest two eigenvalues.)

We claim that there is a nonzero linear combination $w = c_1 v_1 + c_2 v_2$ of $v_1$ and $v_2$ such that the last coordinate of $w$ is 0:

$$(c_1 v_1 + c_2 v_2) \cdot e_n = 0$$

i.e., so that

(*)

$$c_1 (v_1 \cdot e_n) + c_2 (v_2 \cdot e_n) = 0.$$  

(If $v_1 \cdot e_n = 0$, we can let $c_1 = 1$ and $c_2 = 0$; otherwise let $c_1 = 1$ and use (*) to solve for $c_2$.)

Let $x$ be the vector in $\mathbb{R}^{n-1}$ whose components are the first $(n-1)$ components of $w$. Thus

$$w = \begin{bmatrix} x \\ 0 \end{bmatrix}.$$  

Note that

$$(4) \quad w \cdot Qw = \begin{bmatrix} x \\ 0 \end{bmatrix} \cdot \begin{bmatrix} x \\ 0 \end{bmatrix} = \begin{bmatrix} x \\ 0 \end{bmatrix} \cdot \begin{bmatrix} Mx \\ 0 \end{bmatrix} = x \cdot Mx > 0$$

since $M$ is positive definite. On the other hand,

$$(5) \quad w^T Q w \leq \lambda_2 \| w \|^2$$

by the remark after the proof of Theorem 1. Combining (4) and (5) shows that

$$\lambda_2 \| w \|^2 > 0$$

so $\lambda_2 > 0$. Since $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, this implies that $\lambda_2, \lambda_3, \ldots, \lambda_n$ are all positive, so

$$\lambda_2 \lambda_3 \cdots \lambda_n > 0.$$  

Since $\det Q = \lambda_1 \lambda_2 \cdots \lambda_n$, this implies that $\det Q > 0$ if and only if $\lambda_1 > 0$. But by Theorem 2, $\lambda_1 > 0$ if and only if $Q$ is positive definite. \[\square\]

Now we can see why Theorem 4 is true:

**Proof of Theorem 4.** Suppose $d_1, d_2, \ldots, d_n$ are all positive. We must prove that $A_n$ is positive definite.

Let $A(k)$ be the $k$ by $k$ matrix in the upper left corner of $A$. We claim that the matrices $A(1), A(2), \ldots, A(n)$ are all positive definite.

Now $a_{11} = d_1 > 0$, so the matrix $A(1) = [a_{11}]$ is certainly positive definite.

Once we know $A(k)$ is positive definite, then (since $d_{k+1} > 0$) we can apply the Lemma (with $Q = A(k+1)$ and $M = A(k)$) to conclude that $A(k+1)$ is positive definite.

So: $A(1)$ is positive definite, which implies that $A(2)$ is positive definite, which in turn implies that $A(3)$ is positive definite, and so on. Thus all the $A(k)$ are positive.
definite. In particular, $A = A_{(n)}$ is positive definite. This proves statement (1) of the theorem.

Statement (2) is just statement (1) applied to the matrix $-A$.

The proof of statement (3) is similar, but we’ll skip it. □