Quadratic Forms

Recall the Simon & Blume excerpt from an earlier lecture which said that the main task of calculus is to approximate nonlinear functions with linear functions. It's actually more accurate to say that we approximate nonlinear functions with *affine* functions: given a nonlinear function $f : \mathbb{R}^n \to \mathbb{R}$, our approximating function will be of the form $\mathbf{b} + g(\Delta\mathbf{x})$, where g is a linear function and $\Delta\mathbf{x} = \mathbf{x} - \overline{\mathbf{x}}$. For example, in the case $n = 1$, if we wish to approximate f near a point $\overline{x}$ in the domain of f, our approximating function is $f(\overline{x}) + a\Delta x$, where the coefficient a is $f'(\overline{x})$, the derivative of f at $\overline{x}$: $f(\overline{x})$ plays the role of $\mathbf{b}$ in the expression $\mathbf{b} + g(\Delta\mathbf{x})$ above, and the linear function $a\Delta x$ — *i.e.*, $f'(\overline{x})\Delta x$ — plays the role of $g(\Delta\mathbf{x})$. In words, the affine approximation of f near $\overline{x}$ is the affine function with (i) the same value as f at $\overline{x}$, and (ii) the same slope (the same derivative) as f at $\overline{x}$.

We're going to find that it's important to approximate nonlinear functions not only with linear (actually, affine) functions, but also with quadratic functions. For example, for a real function $f : \mathbb{R} \to \mathbb{R}$, our quadratic approximating function will be $f(\overline{x}) + f'(\overline{x})\Delta x + \frac{1}{2}f''(\overline{x})(\Delta x)^2$. The quadratic approximation is therefore the quadratic function that has (i) the same value as f at $\overline{x}$, (ii) the same slope (the same derivative) as f at $\overline{x}$, and (iii) the same curvature (the same second derivative) as f at $\overline{x}$.

When we generalize from functions with the one-dimensional domain $\mathbb{R}$ to multivariate functions, with domain $\mathbb{R}^n$, things get a little bit more complicated. The derivative of a function $f : \mathbb{R}^n \to \mathbb{R}$ at a point $\overline{\mathbf{x}} \in \mathbb{R}^n$ is no longer just a number, but a vector in $\mathbb{R}^n$ — specifically, the gradient of f at $\overline{\mathbf{x}}$, which we write as $\nabla f(\overline{\mathbf{x}})$. And the quadratic term in the quadratic approximation to f is a *quadratic form*, which is defined by an $n \times n$ matrix $H(\overline{\mathbf{x}})$ — the second derivative of f at $\overline{\mathbf{x}}$. In these notes we're going to study quadratic forms.

Quadratic Forms

You already know that a quadratic function (from $\mathbb{R}$ into $\mathbb{R}$) is a 2nd-degree polynomial, *i.e.*, a real function $f(x) = ax^2 + bx + c$ in which $a \neq 0$. If each of the coefficients a, b, and c is non-zero, then the function has a second-degree (quadratic) term, a first-degree (linear) term, and a zero-degree (constant) term. However, a *quadratic form* is a real-valued function on $\mathbb{R}^n$ that has *only* second-degree (quadratic) terms. So a quadratic form on $\mathbb{R}$ (*i.e.*, on $\mathbb{R}^n$, where $n = 1$) is a function of the form $f(x) = ax^2$ for some non-zero coefficient $a \in \mathbb{R}$.

What about the case $n = 2$? A quadratic form on $\mathbb{R}^2$ is a function of the form $f(x_1, x_2) = a_{11}x_1x_1 + a_{12}x_1x_2 + a_{21}x_2x_1 + a_{22}x_2x_2$, or equivalently, $f(x_1, x_2) = a_{11}x_1^2 + (a_{12} + a_{21})x_1x_2 + a_{22}x_2^2$. Note that in the second expression for f we combined the coefficients a_{12} and a_{21} into their sum, so we can also write the same function as $f(x_1, x_2) = a_{11}x_1^2 + a'_{12}x_1x_2 + a_{22}x_2^2$, where $a'_{12} = a_{12} + a_{21}$, which is a common way to write quadratic forms (but without the prime). But for now we're going to use the first expression, writing the generic quadratic form on $\mathbb{R}^2$ as

$$f(x_1, x_2) = a_{11}x_1x_1 + a_{12}x_1x_2 + a_{21}x_2x_1 + a_{22}x_2x_2.$$

Note that the generic quadratic form on $\mathbb{R}^2$ can also be written as

$$f(x_1, x_2) = [x_1\ x_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Moreover, without loss of generality we can assume that $a_{12} = a_{21}$ — for if a_{12} and a_{21} are not equal, we can write them instead as $\widetilde{a}_{12}$ and $\widetilde{a}_{21}$ and then define $a_{12} = a_{21} = \frac{1}{2}(\widetilde{a}_{12} + \widetilde{a}_{21})$. Therefore the quadratic forms on $\mathbb{R}^2$ are precisely the functions $f : \mathbb{R}^2 \to \mathbb{R}$ of the form

$$f(\mathbf{x}) = f(x_1, x_2) = [x_1\ x_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{x}A\mathbf{x},$$

where A is a symmetric matrix.

Note: As in the expression $\mathbf{x}A\mathbf{x}$ above, I'm not going to indicate transposes of vectors in these Quadratic Forms notes. The expression $\mathbf{x}A\mathbf{x}$ will always mean that the vector $\mathbf{x} \in \mathbb{R}^n$ is written as a row vector if it's on the left of the matrix and as a column vector if it's on the right, so that $\mathbf{x}A\mathbf{x}$ is always well-defined and its value is always a real number.

Before moving to the general case of $\mathbb{R}^n$, let's consider the case of $\mathbb{R}^3$. In this case the generic quadratic form is

$$\begin{aligned} f(x_1, x_2, x_3) = {} & a_{11}x_1x_1 + a_{22}x_2x_2 + a_{33}x_3x_3 \\ & + a_{12}x_1x_2 + a_{21}x_2x_1 + a_{13}x_1x_3 + a_{31}x_3x_1 + a_{23}x_2x_3 + a_{32}x_3x_2, \end{aligned}$$

and we can assume, as before, that $a_{12} = a_{21}$, $a_{13} = a_{31}$, and $a_{23} = a_{32}$. Therefore we can write the quadratic form as

$$f(\mathbf{x}) = f(x_1, x_2, x_3) = [x_1\ x_2\ x_3] \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{x}A\mathbf{x},$$

where A is a symmetric 3×3 matrix.

Now it should be clear how we want to define the general quadratic form, on $\mathbb{R}^n$:

Definition: A quadratic form on $\mathbb{R}^n$ is a function $f : \mathbb{R}^n \to \mathbb{R}$ of the form $f(\mathbf{x}) = \mathbf{x}A\mathbf{x}$, where A is a symmetric $n \times n$ matrix.

One important property of quadratic forms is immediately obvious:

Remark: The value of a quadratic form at the vector $\mathbf{0} \in \mathbb{R}^n$ is zero.

Because every quadratic form corresponds to a unique symmetric matrix, we can characterize various classes of quadratic forms completely in terms of properties of symmetric matrices. For example, how can we identify which quadratic forms always have nonnegative values for every vector $\mathbf{x} \in \mathbb{R}^n$? How can we identify which ones are strictly concave functions? We answer questions like these by identifying the properties of symmetric matrices that yield quadratic forms with the desired properties.

Definiteness of Quadratic Forms and Matrices

We ended the preceding section by asking how we can identify which quadratic forms on $\mathbb{R}^n$ always have nonnegative values, or which ones are strictly concave functions, etc. The pattern for general n is foreshadowed by the simple case $n = 1$, where the quadratic forms are the functions $f(x) = ax^2$. If $a > 0$ this quadratic form is positive for all nonzero values of x, and if $a < 0$ the quadratic form is negative for all nonzero values of x. Moreover, in the $a > 0$ case the function $f(\cdot)$ is strictly convex, and when $a < 0$ the function is strictly concave.

Before trying to analyze the general case of quadratic forms on $\mathbb{R}^n$ for any n, let's spend some time studying the case $n = 2$. Here the quadratic form is

$$f(\mathbf{x}) = f(x_1, x_2) = [x_1\ x_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{x}A\mathbf{x},$$

where A is a symmetric matrix. Let's rewrite the matrix as $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$ so we won't have to deal with the subscripts. So we have

$$f(\mathbf{x}) = \mathbf{x}A\mathbf{x} = [x_1\ x_2] \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = ax_1^2 + 2bx_1x_2 + cx_2^2. \tag{1}$$

What we want to know about this quadratic form is whether its value is positive (or at least non-negative) for all vectors $\mathbf{x} \neq \mathbf{0}$; or whether it's negative (or at least non-positive) for all $\mathbf{x} \neq \mathbf{0}$; or whether neither of these is true — *i.e.*, it's positive for some vectors and negative for others. If the first ("positive") statement is true, we say the quadratic form is **positive definite**; if all we can say is that it's non-negative for all nonzero vectors, we say the quadratic form is **positive semi-definite**. If the quadratic form is negative for all $\mathbf{x} \neq \mathbf{0}$, we say it's **negative definite**; and if we can only say it's non-positive for all nonzero vectors, we say it's **negative semi-definite**. If the sign can go either way, we say the quadratic form is **indefinite**. We use the same terms — positive definite, etc. — to describe the matrix A.

Now let's see if we can figure out some conditions on the matrix A that will tell us which definiteness property it has — and therefore which property the quadratic form $\mathbf{x}A\mathbf{x}$ has. Certainly if $a = c = 0$ then the quadratic form in (1) is indefinite: $x_1x_2 > 0$ for some vectors $\mathbf{x}$, and $x_1x_2 < 0$ for other vectors $\mathbf{x}$. In fact, if just *one* of the coefficients a or c is zero, the quadratic form is indefinite: for example, if $a = 0$ then $\mathbf{x}A\mathbf{x} = 2bx_1x_2 + cx_2^2 = (2bx_1 + cx_2)x_2$, so the sign of $\mathbf{x}A\mathbf{x}$ depends on the sign of $2bx_1 + cx_2$, which will clearly be positive for some vectors $\mathbf{x}$ and negative for others.

So let's assume that both $a \neq 0$ and $c \neq 0$. Now we can use the trick of "completing the square" to change this expression into a sum of two squares, as follows:

$$\begin{aligned}
\mathbf{x}A\mathbf{x} &= ax_1^2 + 2bx_1x_2 + cx_2^2 \\
&= a\big(x_1^2 + 2\frac{b}{a}x_1x_2\big) + cx_2^2 + \frac{b^2}{a}x_2^2 - \frac{b^2}{a}x_2^2 \\
&= a\big(x_1^2 + 2\frac{b}{a}x_1x_2 + \frac{b^2}{a^2}x_2^2\big) + \big(c - \frac{b^2}{a}\big)x_2^2 \\
&= a\big(x_1 + \frac{b}{a}x_2\big)^2 + \frac{1}{a}(ac - b^2)x_2^2 \\
&= a\big(x_1 + \frac{b}{a}x_2\big)^2 + \frac{1}{a}|A|x_2^2.
\end{aligned}$$

Now it's clear that if $a > 0$ and $|A| > 0$, then $\mathbf{x}A\mathbf{x}$ is positive definite: these two inequalities ensure that the only way both terms in the sum can be zero is if $x_2 = 0$ (to make the second term zero), in which case x_1 has to be zero as well in order to make the first term zero. Of course, $a > 0$ and $|A| > 0$ also guarantee that $\mathbf{x}A\mathbf{x}$ can't be negative for any vector $\mathbf{x}$, so $\mathbf{x}A\mathbf{x}$ is indeed positive definite. A parallel argument shows that $\mathbf{x}A\mathbf{x}$ is negative definite if $a < 0$ and $|A| > 0$: in this case the coefficient on the second term is negative if and only if a and $|A|$ have opposite signs.

Notice that although we assumed that a and c are both nonzero, we didn't actually use the fact that $c \neq 0$. However, one of the conditions for definiteness (positive or negative) is that $|A| > 0$, and this requires that $ac > 0$ — *i.e.*, that a and c have the same sign. Also note that we could have carried out the above argument with the roles of a and c reversed. Therefore, we have the following theorem, where we revert to the notation a_{ij} for the entries in the matrix A:

Theorem: A 2×2 symmetric matrix A is
- positive definite if and only if $|A| > 0$ and either $a_{11} > 0$ or $a_{22} > 0$,
  which is equivalent to $|A| > 0$ and *both* $a_{11} > 0$ and $a_{22} > 0$;
- negative definite if and only if $|A| > 0$ and either $a_{11} < 0$ or $a_{22} < 0$,
  which is equivalent to $|A| > 0$ and *both* $a_{11} < 0$ and $a_{22} < 0$.

What are necessary and sufficient conditions for the matrix A and the associated quadratic form $\mathbf{x}A\mathbf{x}$ to be positive or negative *semidefinite*? If A is positive semidefinite — *i.e.*, $\mathbf{x}A\mathbf{x} \geqq 0$ for all $\mathbf{x} \neq \mathbf{0}$ — we clearly must have $a \geqq 0$ and $c \geqq 0$. If either $a > 0$ or $c > 0$, then we must have $|A| \geqq 0$; and if $a = c = 0$, then we must have $b = 0$ as well, in which case $|A| = 0$. Therefore the conditions $a \geqq 0, c \geqq 0$, and $|A| \geqq 0$ together must all hold if A is positive semidefinite. Are these conditions also sufficient? Suppose that $|A| \geqq 0$. If $a = c = 0$, then $|A| \geqq 0$ implies that $b = 0$, so that A is the zero matrix, and $\mathbf{x}A\mathbf{x} = 0$ for all $\mathbf{x} \in \mathbb{R}^2$. And it's clear in the above expression for $\mathbf{x}A\mathbf{x}$ that if $a > 0$ (or, symmetrically, $c > 0$) and also $|A| \geqq 0$, then $\mathbf{x}A\mathbf{x} \geqq 0$ for all $\mathbf{x} \neq \mathbf{0}$. Therefore the conditions $a \geqq 0, c \geqq 0$, and $|A| \geqq 0$ are sufficient as well as necessary for A to be positive semidefinite. A parallel argument provides the conditions for A to be negative semidefinite, and we have the following theorem:

Theorem: A 2×2 symmetric matrix A is
- positive semidefinite if and only if $a_{11} \geqq 0, a_{22} \geqq 0$, and $|A| \geqq 0$;
- negative semidefinite if and only if $a_{11} \leqq 0, a_{22} \leqq 0$, and $|A| \geqq 0$.

This pattern generalizes to $\mathbb{R}^n$ for arbitrary n as follows: first think of the components a_{ij} of the 2×2 matrix A as 1×1 "submatrices" of A; because they're 1×1 we'll call them "order-1" submatrices, and we'll say that A itself, which is 2×2, is an order-2 submatrix. For an $n \times n$ matrix A we'll say that a submatrix of order k consists of k of the rows and k of the columns of A — or we could equivalently say that an order-k submatrix is formed by deleting $n-k$ rows and columns. Now note that in the 2×2 example, the conditions we developed involved only submatrices on the diagonal, a_{11} and a_{22}, as well as A. We could say each of these submatrices was formed by deleting *the same* row and column: for a_{11} we deleted the second row and the second column; for a_{22} we deleted the first row and column; for A itself we deleted no rows and columns. Finally, note that the conditions in the 2×2 case were conditions on the signs of the determinants of these submatrices. The following definition generalizes these ideas to $n \times n$ matrices.

Definition: Let A be an $n \times n$ matrix. For each $k = 1, 2, \ldots, n$ an **order-k principal submatrix** of A is a $k \times k$ matrix formed by deleting the same $n-k$ rows and columns. The order-k submatrices formed by deleting the *last* $n-k$ rows and columns are called the **leading** principal submatrices of A. The determinant of a submatrix of A is called a **minor** of A.

Therefore the leading principal submatrices of a 2×2 matrix A are the matrices a_{11} and A itself; the leading principal minors are a_{11} and $|A|$. The leading principal minors of a 3×3 matrix A are

$$a_{11} \quad \text{and} \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \quad \text{and} \quad |A|.$$

The leading principal minors of a 4×4 matrix A are

$$a_{11} \quad \text{and} \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \quad \text{and} \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \quad \text{and} \quad |A|.$$

The general versions of our 2×2 theorems are as follows:

Theorem: An $n \times n$ symmetric matrix is
- positive definite if and only if all of its leading principal minors are positive,
  or equivalently, if and only if *all* of its principal minors are positive;
- negative definite if and only if all of its order-k leading principal minors have sign $(-1)^k$;
  or equivalently, if and only if *all* of its order-k leading principal minors have sign $(-1)^k$.

Theorem: An $n \times n$ symmetric matrix is
- positive semidefinite if and only if all of its principal minors are non-negative;
- negative semidefinite if and only if all of its nonzero order-k principal minors have sign $(-1)^k$.

Corollary: An $n \times n$ symmetric matrix is indefinite if and only if it has both a negative principal minor and an order-k principal minor with sign $(-1)^{k+1}$. (Note that these could both be the same principal minor, as in the following example.)

Example: In order that the matrix A be positive definite it's necessary and sufficient that the leading principal minors all be positive. It therefore seems natural to think that the parallel result should hold for positive *semidefiniteness*: that A is positive semidefinite if and only if all of its leading principal minors are nonnegative. Here's a counterexample, which shows that merely having nonnegative *leading* principal minors is not sufficient to ensure that A is positive semidefinite: we need to consider all the principal minors.

Let $A = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 2 \end{bmatrix}$. All three leading principal minors are either positive or zero:

$$|A_1| = a_{11} = 1, \quad |A_2| = \begin{vmatrix} 1 & 0 \\ 0 & 0 \end{vmatrix} = 0, \text{ and } |A_3| = |A| = 0.$$

However, the order-2 non-leading principal minor

$$\begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} = \begin{vmatrix} 1 & 2 \\ 2 & 2 \end{vmatrix} = -2,$$

which is inconsistent with both positive and negative semidefiniteness: it's negative, which is inconsistent with positive semidefiniteness; and its order is $k = 2$ and its sign is $-1 \neq (-1)^k$, which is inconsistent with negative semidefiniteness. Note that $\mathbf{x}A\mathbf{x} = x_1^2 + 2x_3^2 + 4x_1x_3$. When $\mathbf{x} = (1, 0, 1)$, then $\mathbf{x}A\mathbf{x} = 7$; when $\mathbf{x} = (1, 0, -1)$, then $\mathbf{x}A\mathbf{x} = -1$; this verifies directly, without having to consider principal minors, that A is neither negative semidefinite nor positive semidefinite — *i.e.*, that it's indefinite.