

SUPPLEMENT TO SECTION 3.3

Homework: A, B, C

Section 3.3 treats max-min theory for functions of several variables. It mostly considers functions of two variables. In this supplement we discuss how to handle functions of any number of variables. We first review some terminology.

RELATIVE EXTREMA AND TAYLOR POLYNOMIALS

Consider a function $f : U \rightarrow R$ and a point $\mathbf{a} \in U$, where U is an open subset of \mathbf{R}^n .

We say that f has a *relative maximum* at \mathbf{a} if there is a number $\delta > 0$ such that $f(\mathbf{x}) \leq f(\mathbf{a})$ for all \mathbf{x} such that $\|\mathbf{x} - \mathbf{a}\| < \delta$. It has a *strict relative maximum* at \mathbf{a} if there is a number $\delta > 0$ such that $f(\mathbf{x}) < f(\mathbf{a})$ for all \mathbf{x} such that $0 < \|\mathbf{x} - \mathbf{a}\| < \delta$. Replacing $f(\mathbf{x}) \leq f(\mathbf{a})$ by $f(\mathbf{x}) \geq f(\mathbf{a})$ and $f(\mathbf{x}) < f(\mathbf{a})$ by $f(\mathbf{x}) > f(\mathbf{a})$ in these statements gives the definitions of *relative minimum* and *strict relative minimum*. The word “local” is often used in place of the term “relative.” The word “extremum” is used to mean “maximum or minimum”.

We say that \mathbf{a} is a *critical point* of f if $\nabla f(\mathbf{a}) = \mathbf{0}$. If f has a local extremum at \mathbf{a} , then \mathbf{a} must be a critical point of f . It is not true if \mathbf{a} is a critical point of f , then f must have a local extremum at \mathbf{a} . A critical point of f at which f does not have a local extremum is called a *saddle point* of f .

Recall that the second order Taylor polynomial of f at \mathbf{a} is

$$P_2(\mathbf{x}, \mathbf{a}) = f(\mathbf{a}) + \nabla f(\mathbf{a})\mathbf{h} + \frac{1}{2}\mathbf{h}^t S \mathbf{h},$$

where $\mathbf{h} = \mathbf{x} - \mathbf{a}$ and is regarded as a column vector, the gradient $\nabla f(\mathbf{a})$ is regarded as a row vector (so the matrix product $\nabla f(\mathbf{a})\mathbf{h}$ is really just the dot product of these two vectors), \mathbf{h}^t is the transpose of \mathbf{h} and so is a row vector, and S is the matrix whose i - j th entry is $\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{a})$, so that the matrix product $\mathbf{h}^t S \mathbf{h}$ is just the dot product of \mathbf{h} and the vector $S\mathbf{h}$. Note that if f is a “nice” function (the only kind we will be considering), then the mixed second partials are equal, and so S is a symmetric matrix.

Taylor’s formula says that $f(\mathbf{x}) = P_2(\mathbf{x}, \mathbf{a}) + R_2(\mathbf{x}, \mathbf{a})$, where $\lim_{\mathbf{x} \rightarrow \mathbf{a}} \frac{R_2(\mathbf{x}, \mathbf{a})}{\|\mathbf{x} - \mathbf{a}\|^2} = 0$. At a critical point $\nabla f(\mathbf{a}) = \mathbf{0}$, and so the middle term of $P_2(\mathbf{x}, \mathbf{a})$ goes away. If, in addition, $\det S \neq 0$, then \mathbf{a} is called a *non-degenerate* critical point. It can be shown in this case that f and P_2 have the same behavior at \mathbf{a} as far as relative extrema and saddle points are concerned. If one of them has a relative maximum, relative minimum, or saddle point, then the other also has, respectively, a relative maximum, relative minimum, or saddle point. Moreover, if there is a relative extremum at \mathbf{a} , it will be a strict relative extremum.

So we now have to analyze the extrema of $P_2(\mathbf{x}, \mathbf{a}) = f(\mathbf{a}) + \frac{1}{2}\mathbf{h}^t S \mathbf{h}$. Note that since $\mathbf{h} = \mathbf{x} - \mathbf{a}$, $f(\mathbf{a})$ is a constant, and $\frac{1}{2}$ is a positive constant, $P_2(\mathbf{x}, \mathbf{a})$ will have a strict local max, strict local min, or saddle point at \mathbf{a} exactly when the function $q(\mathbf{h}) = \mathbf{h}^t S \mathbf{h}$ has a strict local max, strict local min, or saddle point, respectively, at $\mathbf{0}$.

DIAGONALIZATION OF QUADRATIC FORMS

The function $q(\mathbf{h}) = \mathbf{h}^t S \mathbf{h}$ is called a *quadratic form*. In general, working out the matrix multiplications we see that

$$q(\mathbf{h}) = \sum_{i=1}^n \sum_{j=1}^n s_{i,j} h_i h_j.$$

For example, if $S = \begin{bmatrix} 4 & -1 \\ -1 & 2 \end{bmatrix}$, then

$$\begin{aligned} q(\mathbf{h}) &= \begin{bmatrix} h_1 & h_2 \end{bmatrix} \begin{bmatrix} 4 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} h_1 & h_2 \end{bmatrix} \begin{bmatrix} 4h_1 - h_2 \\ -h_1 + 2h_2 \end{bmatrix} = \\ &= 4h_1^2 - h_1 h_2 - h_2 h_1 + 2h_2^2 = 4h_1^2 + 2h_2^2 - 2h_1 h_2. \end{aligned}$$

Note that for a quadratic form q we always have $q(\mathbf{0}) = 0$. A quadratic form q is called *positive definite* if $q(\mathbf{h}) > 0$ for all $\mathbf{h} \neq \mathbf{0}$. Note that this implies that q has a local minimum at $\mathbf{0}$. We say that q is *negative definite* if $q(\mathbf{h}) < 0$ whenever $\mathbf{h} \neq \mathbf{0}$. Note that this implies that q has a local maximum at $\mathbf{0}$. If there are vectors \mathbf{v} and \mathbf{w} such that $q(\mathbf{v}) > 0$ and $q(\mathbf{w}) < 0$, then q is said to be *indefinite*. Note that this implies that q has a saddle point at $\mathbf{0}$. As we will see, any quadratic form $q(\mathbf{h}) = \mathbf{h}^t S \mathbf{h}$ with $\det(S) \neq 0$ falls into one of these three categories, and there is a procedure for telling which one it falls into. From now on we will be assuming that $\det(S) \neq 0$.

The key observation is that if we make a change of variables by setting $\mathbf{h} = Q\mathbf{k}$, where Q is an $n \times n$ matrix, then we get a new quadratic form $p(\mathbf{k}) = (Q\mathbf{k})^t S Q\mathbf{k} = \mathbf{k}^t Q^t S Q \mathbf{k} = \mathbf{k}^t V \mathbf{k}$. We show that if $\det(Q) \neq 0$, then this new form has the same kind of behavior at $\mathbf{0}$ as the old form. First note that since $\det(Q) \neq 0$, we have that Q is invertible. This implies that $\mathbf{k} = Q^{-1}\mathbf{h}$, and so $\mathbf{h} = \mathbf{0}$ if and only if $\mathbf{k} = \mathbf{0}$. Note also that $q(\mathbf{h}) = p(\mathbf{k})$. Now suppose that q is positive definite and that $\mathbf{k} \neq \mathbf{0}$. Then $\mathbf{h} \neq \mathbf{0}$, and so $p(\mathbf{k}) = q(\mathbf{h}) > 0$. Thus p is positive definite. A similar argument shows that if p is positive definite, then so is q . We leave it as an exercise to show that q is negative definite if and only if p is negative definite, and that q is indefinite if and only if p is indefinite.

The advantage to this observation is that if we choose Q wisely the new form may be simpler and easier to classify. For example, let $Q = \begin{bmatrix} 1 & \frac{1}{4} \\ 0 & 1 \end{bmatrix}$. Then

$$Q^t S Q = \begin{bmatrix} 1 & 0 \\ \frac{1}{4} & 1 \end{bmatrix} \begin{bmatrix} 4 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{4} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 4 & -1 \\ 0 & \frac{7}{4} \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{4} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & \frac{7}{4} \end{bmatrix}$$

Our new form is

$$p(\mathbf{k}) = \begin{bmatrix} k_1 & k_2 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & \frac{7}{4} \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} = \begin{bmatrix} k_1 & k_2 \end{bmatrix} \begin{bmatrix} 4k_1 \\ \frac{7}{4}k_2 \end{bmatrix} = 4k_1^2 + \frac{7}{4}k_2^2.$$

Note that $p(\mathbf{k})$ is obviously positive whenever $\mathbf{k} \neq \mathbf{0}$ and so is positive definite. Since $\det(Q) \neq 0$, our original quadratic form will also be positive definite.

If the matrix of a quadratic form is diagonal, as with the new form we obtained above, then it is easy to classify the form. Let μ_1, \dots, μ_n be the diagonal entries. Then the form can be expressed as $\mu_1 k_1^2 + \dots + \mu_n k_n^2$. It will be positive definite if and only if all the μ_i are positive. It will be negative definite if and only if all the μ_i are negative. It will be indefinite if and only if some of the μ_i are positive and some of the μ_i are negative.

So, if we can always find a matrix Q with $\det(Q) \neq 0$ such that $Q^t S Q$ is a diagonal matrix D , then we can classify the quadratic form q and thus we can classify the critical point \mathbf{a} . To describe how to find Q we need to review row and column operations and elementary matrices. There are three elementary row operations:

$R_i \rightarrow R_i + cR_j$ Replace the i^{th} row by the sum of the i^{th} row and c times the j^{th} row.

$R_i \rightarrow cR_i$ Replace the i^{th} row by c times the i^{th} row, where $c \neq 0$.

$R_i \leftrightarrow R_j$ Interchange the i^{th} and j^{th} rows.

If you do *one* of these elementary operations on the identity matrix I you get a matrix E which is called an elementary matrix. The basic fact about elementary matrices is that if you do this row operation on a matrix A and get the matrix B as a result, then $B = EA$. Note that E multiplies A on the *left*. Note that elementary matrices always have non-zero determinants.

There are, similarly, three column operations $C_i \rightarrow C_i + cC_j$, $C_i \rightarrow cC_i$ for $c \neq 0$, and $C_i \leftrightarrow C_j$. If you do one of these column operations on I you obtain a matrix F . The basic fact here is that if you do the column operation on a matrix M and get the result N , then $N = MF$. Note that F multiplies M on the *right*. It turns out that F is an elementary matrix, and in fact it is the transpose of the elementary matrix associated with the corresponding row operation.

Look back at our example, where we started with $S = \begin{bmatrix} 4 & -1 \\ -1 & 2 \end{bmatrix}$. The row operation $R_2 \rightarrow R_2 + \frac{1}{4}R_1$ transforms S into $T = \begin{bmatrix} 4 & -1 \\ 0 & \frac{7}{4} \end{bmatrix}$. It also transforms

$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ into $E = \begin{bmatrix} 1 & 0 \\ \frac{1}{4} & 1 \end{bmatrix}$. You can check that $ES = T$. If we now do the

corresponding column operation $C_2 \rightarrow C_2 + \frac{1}{4}C_1$ on T we get the matrix $D = \begin{bmatrix} 4 & 0 \\ 0 & \frac{7}{4} \end{bmatrix}$.

Doing this column operation on $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ gives the matrix $F = \begin{bmatrix} 1 & \frac{1}{4} \\ 0 & 1 \end{bmatrix}$. You can check that $TF = D$. Note that E and F are transposes of each other, so we have $F^t S F = D$.

In general we will need several row and column operations to get a diagonal matrix. Here is the general pattern. Start in the upper left hand corner of the given symmetric matrix S . Suppose that entry is non-zero. Call it our pivot. Do a sequence of row operations of the form $R_i \rightarrow R_i + cR_1$ to turn all the entries below the pivot into zeros. Then do the corresponding column operations $C_i \rightarrow C_i + cC_1$ to turn all the entries to the right of the pivot into zeros. This creates the first row and column of our diagonal matrix D . Next move on to the matrix obtained by deleting the first

row and column and continue in this fashion until you get D . Suppose the upper left entry of our original matrix had been zero. If some entry on the main diagonal, say $s_{i,i}$, is non-zero, then do $R_1 \leftrightarrow R_i$ followed by $C_1 \leftrightarrow C_i$. This will interchange our zero entry with the non-zero entry $s_{i,i}$. Then proceed as above. If there are no non-zero entries on the main diagonal, then look for a non-zero entry in the first column, say $s_{i,1}$. Do $R_1 \rightarrow R_1 + R_i$ followed by $C_1 \rightarrow C_1 + C_i$. This will replace our zero entry by $2s_{i,1}$. Then proceed as above. If there are no non-zero entries in the first column, then the first row and column are already those of a diagonal matrix and we move on to the next iteration. (Note that if $\det(S) \neq 0$, then this scenario cannot happen.) Now note that our row operations give us elementary matrices E_1, \dots, E_m and our column operations give us elementary matrices F_1, \dots, F_m , where E_j and F_j are transposes of each other. Thus we have

$$D = E_m \cdots E_1 S F_1 \cdots F_m = F_m^t \cdots F_1^t S F_1 \cdots F_m = (F_1 \cdots F_m)^t S (F_1 \cdots F_m) = Q^t S Q.$$

Let's try this on the matrix $S = \begin{bmatrix} 0 & 2 & -1 \\ 2 & 1 & 3 \\ -1 & 3 & 5 \end{bmatrix}$.

First we do $R_1 \leftrightarrow R_2$ to get $E_1 S = \begin{bmatrix} 2 & 1 & 3 \\ 0 & 2 & -1 \\ -1 & 3 & 5 \end{bmatrix}$.

Then $C_1 \leftrightarrow C_2$ to get $E_1 S F_1 = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & -1 \\ 3 & -1 & 5 \end{bmatrix}$.

Then $R_2 \rightarrow R_2 - 2R_1$ and $R_3 \rightarrow R_3 - 3R_1$ to get $E_3 E_2 E_1 S F_1 = \begin{bmatrix} 1 & 2 & 3 \\ 0 & -4 & -7 \\ 0 & -7 & -4 \end{bmatrix}$.

Then $C_2 \rightarrow C_2 - 2C_1$ and $C_3 \rightarrow C_3 - 3C_1$ to get $E_3 E_2 E_1 S F_1 F_2 F_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -4 & -7 \\ 0 & -7 & -4 \end{bmatrix}$.

Next we do $R_3 \rightarrow R_3 - \frac{7}{4}R_2$ to get $E_4 E_3 E_2 E_1 S F_1 F_2 F_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -4 & -7 \\ 0 & 0 & \frac{33}{4} \end{bmatrix}$.

Finally we do $C_3 \rightarrow C_3 - \frac{7}{4}C_2$ to get

$$D = Q^t S Q = E_4 E_3 E_2 E_1 S F_1 F_2 F_3 F_4 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & \frac{33}{4} \end{bmatrix}.$$

Thus our new form is $k_1^2 - 4k_2^2 + \frac{33}{4}k_3^2$, which is clearly indefinite. If we want to find $Q = F_1 F_2 F_3 F_4$, then we can avoid doing matrix multiplications or column operations as follows. First do the sequence of *row* operations above on I . (Do *not* do any of the column operations.) This will give the product $E_4 E_3 E_2 E_1$.

We start with $I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

First we do $R_1 \leftrightarrow R_2$ to get $E_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

Then $R_2 \rightarrow R_2 - 2R_1$ and $R_3 \rightarrow R_3 - 3R_1$ to get $E_3E_2E_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -2 & 0 \\ 0 & -3 & 1 \end{bmatrix}$.

Finally we do $R_3 \rightarrow R_3 - \frac{7}{4}R_2$ to get $E_4E_3E_2E_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -2 & 0 \\ -\frac{7}{4} & \frac{1}{2} & 1 \end{bmatrix}$.

This is Q^t . To get Q we just take the transpose to get $Q = \begin{bmatrix} 0 & 1 & -\frac{7}{4} \\ 1 & -2 & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix}$.

Note that if all you want to do is classify the quadratic form, then you do *not* have to find Q .

What would have happened if we had done a different sequence of row operations? You can check that if we had done $R_1 \leftrightarrow R_3$, $R_2 \rightarrow R_2 - \frac{3}{5}R_1$, $R_3 \rightarrow R_3 + \frac{1}{5}R_1$, $R_3 \rightarrow R_3 + \frac{13}{4}R_2$ and the corresponding column operations, we would have gotten

$$D = \begin{bmatrix} 5 & 0 & 0 \\ 0 & -\frac{4}{5} & 0 \\ 0 & 0 & \frac{33}{4} \end{bmatrix} \text{ and } Q = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & \frac{13}{4} \\ 1 & -\frac{3}{5} & \frac{7}{4} \end{bmatrix}.$$

Thus D and Q are not unique. However, it can be shown that in D we always have the same number of positive entries, the same number of negative entries, and the same number of zero entries.

QUADRATIC FORMS AND DETERMINANTS

There is another method for classifying quadratic forms which for small values of n may sometimes be easier than diagonalizing the form (for large values of n it is usually *not* easier). Given a symmetric matrix S , let S_k be the submatrix obtained by deleting the last $n - k$ rows and columns of S ; thus S is the $k \times k$ submatrix in the upper left corner of S . Let $d_k = \det(S_k)$. Note that d_1 is just the entry in the upper left hand corner of S , while since $S_n = S$ we have that $d_n = \det(S)$. We consider the signs of the d_k .

First consider the case of a diagonal matrix D with diagonal entries μ_1, \dots, μ_n . Then each D_k is a diagonal matrix with diagonal entries μ_1, \dots, μ_k . Now the determinant of a diagonal matrix is just the product of the entries on the diagonal. Thus in this case $d_1 = \mu_1$, $d_2 = \mu_1\mu_2$, $d_3 = \mu_1\mu_2\mu_3$, \dots , $d_n = \mu_1\mu_2 \cdots \mu_n$.

If all of the μ_i are positive, then clearly all of the d_i are positive. A little thought shows that the converse is also true: if all the d_i are positive, then all the μ_i must be positive. The equation $d_1 = \mu_1$ shows that $\mu_1 > 0$. Then the equation $d_2 = \mu_1\mu_2$ shows that since $d_2 > 0$ and $\mu_1 > 0$, we must have $\mu_2 > 0$, etc. Thus our form is positive definite if and only if all the d_i are positive.

Now suppose that all of the μ_i are negative. Then $d_1 = \mu_1 < 0$, $d_2 = \mu_1\mu_2 > 0$, $d_3 = \mu_1\mu_2\mu_3 < 0$, etc. So the d_i alternate in sign, beginning with the negative sign. Again, a little thought shows that if we have $d_1 < 0$, $d_2 > 0$, $d_3 < 0$, etc. then all of the μ_i are negative. Thus our form is negative definite if and only if we have this pattern for the d_k .

We now consider the case of an arbitrary symmetric matrix S . Suppose we diagonalize it, so we have $D = Q^tSQ$. Then $\det(D) = \det(Q^t)\det(S)\det(Q) =$

$\det(Q) \det(S) \det(Q) = (\det(Q))^2 \det(S)$. Since $\det(Q) \neq 0$ we have $(\det(Q))^2 > 0$, and so $\det(D)$ and $\det(S)$ have the same sign.

Now suppose that the form given by S is positive definite. Then D is positive definite, so all its diagonal entries μ_i are positive, hence $\det(D) > 0$, hence $\det(S) > 0$. We now observe that the matrix S_k gives a quadratic form q_k on \mathbf{R}^k ; if we take a vector (h_1, h_2, \dots, h_k) in \mathbf{R}^k , then the value of q_k on this vector is the same as that of q on the vector $(h_1, h_2, \dots, h_k, 0, \dots, 0)$ in \mathbf{R}^n , where we have padded out the original vector with $n - k$ zeroes. It follows that q_k is also positive definite. By the argument given above we must have that the determinant d_k of S_k is positive. So we have shown that all the d_k are positive.

Conversely, suppose that all the d_k are positive. Think about how we do row and column operations to obtain D . Since $s_{1,1} = d_1 \neq 0$, we will be doing operations of the form $R_i \rightarrow R_i + cR_1$ and $C_i \rightarrow C_i + cC_1$ to produce zeroes below and to the right of $s_{1,1}$. Note that the first diagonal entry of D is $\mu_1 = s_{1,1} = d_1$. Row and column operations of this type do not change the determinants of any of the S_k , so all of our d_k will be the same as before. Now our new S_2 has the form $\begin{bmatrix} d_1 & 0 \\ 0 & s \end{bmatrix}$, so we have $d_2 = d_1 s$. Since $d_2 \neq 0$ we must have $s \neq 0$. This means that when we go to work on our next round of row and column operations they will be of the form $R_i \rightarrow R_i + cR_2$ and $C_i \rightarrow C_i + cC_2$. The second diagonal entry of D is $\mu_2 = s$. Since d_1 and d_2 are positive and $d_2 = d_1 s$, we must have that μ_2 is positive. Continuing in this fashion we get that all of the μ_i are positive, and so the form is positive definite.

Thus we have shown that the form is positive definite if and only if all of the d_i are positive. We now show that the form is negative definite if and only if we have $d_1 < 0$, $d_2 > 0$, $d_3 < 0$, etc. The basic observation here is that the form q is negative definite if and only if the form $-q$ is positive definite. If q has matrix S , then $-q$ has matrix $-S$. The k^{th} submatrix of $-S$ is $-S_k$, where S_k is the k^{th} submatrix of S . Let $e_k = \det(-S_k)$. Well, $\det(-S_k) = (-1)^k \det(S_k)$. (You pull a -1 out of each of the k rows of $-S_k$.) Thus $e_k = (-1)^k d_k$; we have $e_1 = -d_1$, $e_2 = d_2$, $e_3 = -d_3$, etc. So, if q is negative definite, then $-q$ is positive definite, so all the e_k are positive, so $d_1 = -e_1 < 0$, $d_2 = e_2 > 0$, $d_3 = -e_3 < 0$, etc. This line of reasoning can be reversed, so if the d_i alternate in sign with d_1 negative, then q is negative definite.

So, we now know exactly when q is positive definite or negative definite. So if the d_i follow any other pattern, then q cannot be positive definite or negative definite, and so must be indefinite.

Let's look back at the example $S = \begin{bmatrix} 0 & 2 & -1 \\ 2 & 1 & 3 \\ -1 & 3 & 5 \end{bmatrix}$. We have $d_1 = 0$, $d_2 = \det \begin{bmatrix} 0 & 2 \\ 2 & 1 \end{bmatrix} = -4$, $d_3 = \det(S) = -33$. This does not fit the pattern $+++$ or $-+-$, so the form is indefinite.

EXAMPLES

Now let's go back to the problem of classifying the critical points of a function $f : \mathbf{R}^n \rightarrow \mathbf{R}$. Suppose $\nabla f(\mathbf{a}) = \mathbf{0}$. Then $P_2 = f(\mathbf{a}) + \frac{1}{2} \mathbf{h}^t \mathbf{S} \mathbf{h}$, where S is the matrix whose i - j^{th} entry is $f_{x_i x_j}(\mathbf{a})$. Recall that the critical point \mathbf{a} is non-degenerate if and

only if $\det S \neq 0$. We have seen that classifying a non-degenerate critical point is equivalent to classifying the quadratic form $q(\mathbf{h}) = \mathbf{h}^t S \mathbf{h}$. Note that since $\frac{1}{2} > 0$ we can work either with q and its matrix S or with the form $\frac{1}{2}q$ and its matrix $M = \frac{1}{2}S$. Which one we choose depends on which way the arithmetic is easier. In this section we illustrate both the diagonalization and determinant methods for classifying the quadratic forms and thus classifying non-degenerate critical points.

1. Classify the critical points of $f(x, y, z) = x^2 - 4x + 16 + 2y^2 + 11z^2 + 22z - 2xy + 10y + 6yz$.

$f_x = 2x - 4 - 2y$, $f_y = 4y - 2x + 10 + 6z$, and $f_z = 22z + 22 + 6y$. Setting each of these equal to zero and rearranging a bit gives the system of equations

$$\begin{array}{rcl} 2x & -2y & = 4 \\ -2x & +4y & +6z = -10 \\ & 6y & +22z = -22 \end{array}$$

This system has augmented matrix

$$\left[\begin{array}{ccc|c} 2 & -2 & 0 & 4 \\ -2 & 4 & 6 & -10 \\ 0 & 6 & 22 & -22 \end{array} \right].$$

Doing the row operations $R_2 \rightarrow R_2 + R_1$ and then $R_3 \rightarrow R_3 - 3R_2$ puts it into the row echelon form

$$\left[\begin{array}{ccc|c} 2 & -2 & 0 & 4 \\ 0 & 2 & 6 & -6 \\ 0 & 0 & 4 & -4 \end{array} \right].$$

Back substitution yields $4z = -4$, $z = -1$, $2y + 6z = -6$, $2y - 6 = -6$, $y = 0$, $2x - 2y = 4$, $2x = 4$, $x = 2$. Thus we have one critical point $\mathbf{a} = (2, 0, -1)$.

We now classify this critical point. $f_{xx} = 2$, $f_{yy} = 4$, $f_{zz} = 22$, $f_{xy} = f_{yx} = -2$, $f_{xz} = f_{zx} = 0$, and $f_{yz} = f_{zy} = 6$. Note that these all happen to be constants. If they had not been constants, we would have had to plug in the values $x = 2$, $y = 0$, and $z = -1$ to obtain the values of these derivatives at \mathbf{a} before going on. We next have

$$S = \begin{bmatrix} 2 & -2 & 0 \\ -2 & 4 & 6 \\ 0 & 6 & 22 \end{bmatrix}.$$

Since all the entries are even, it will be easier to work with

$$M = \frac{1}{2}S = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & 3 \\ 0 & 3 & 11 \end{bmatrix}.$$

Here is the diagonalization method: We do $R_2 \rightarrow R_2 + R_1$ to get

$$\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 3 \\ 0 & 3 & 11 \end{bmatrix}.$$

Then we do $C_2 \rightarrow C_2 + C_1$ to get

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 3 \\ 0 & 3 & 11 \end{bmatrix}.$$

Next we do $R_3 \rightarrow R_3 - 3R_2$ to get

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 3 \\ 0 & 0 & 2 \end{bmatrix}.$$

Finally we do $C_3 \rightarrow C_3 - 3C_2$ to get

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Since all the diagonal entries are positive, the form is positive definite, and hence f has a local minimum at $(2, 0, -1)$.

Here is the determinant method: $d_1 = \det([1]) = 1 > 0$, $d_2 = \det\left(\begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}\right) =$

$1 > 0$, and $d_3 = \det\left(\begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & 3 \\ 0 & 3 & 11 \end{bmatrix}\right) = 2 > 0$. Since we have the pattern

+++ we know that the form is positive definite, and hence that f has a local minimum at $(2, 0, -1)$.

2. Classify the critical points of $f(x, y, z) = 2xy + 2xz - x^2 - 2y^2 - 3z^2$.

$f_x = 2y + 2z - 2x$, $f_y = 2x - 4y$, and $f_z = 2x - 6z$. Setting each of these equal to zero and rearranging a bit gives the system of equations

$$\begin{array}{rcl} -2x & +2y & +2z = 0 \\ 2x & -4y & = 0 \\ 2x & & -6z = 0 \end{array}$$

This system has augmented matrix

$$\left[\begin{array}{ccc|c} -2 & 2 & 2 & 0 \\ 2 & -4 & 0 & 0 \\ 2 & 0 & -6 & 0 \end{array} \right].$$

Doing the row operations $R_2 \rightarrow R_2 + R_1$ and $R_3 \rightarrow R_3 + R_1$, and then $R_3 \rightarrow R_3 + R_2$ puts it into the row echelon form

$$\left[\begin{array}{ccc|c} -2 & 2 & 2 & 0 \\ 0 & -2 & 2 & 0 \\ 0 & 0 & -2 & 0 \end{array} \right].$$

Back substitution yields $-2z = 0$, $z = 0$, $-2y + 2z = 0$, $-2y = 0$, $y = 0$, $-2x + 2y + 2z = 0$, $2x = 0$, $x = 0$. Thus we have one critical point $\mathbf{a} = (0, 0, 0)$.

We now classify this critical point. $f_{xx} = -2$, $f_{yy} = -4$, $f_{zz} = -6$, $f_{xy} = f_{yx} = 2$, $f_{xz} = f_{zx} = 2$, and $f_{yz} = f_{zy} = 0$. Note that these all happen to be constants. If they had not been constants, we would have had to plug in the values $x = 0$, $y = 0$, and $z = 0$ to obtain the values of these derivatives at \mathbf{a} before going on.

We next have

$$S = \begin{bmatrix} -2 & 2 & 2 \\ 2 & -4 & 0 \\ 2 & 0 & -6 \end{bmatrix}.$$

Since all the entries are even, it will be easier to work with

$$M = \frac{1}{2}S = \begin{bmatrix} -1 & 1 & 1 \\ 1 & -2 & 0 \\ 1 & 0 & -3 \end{bmatrix}.$$

Here is the diagonalization method: We do $R_2 \rightarrow R_2 + R_1$ and $R_3 \rightarrow R_3 + R_1$ to get

$$\begin{bmatrix} -1 & 1 & 1 \\ 0 & -1 & 1 \\ 0 & 1 & -2 \end{bmatrix}.$$

Then we do $C_2 \rightarrow C_2 + C_1$ and $C_3 \rightarrow C_3 + C_1$ to get

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & -2 \end{bmatrix}.$$

Next we do $R_3 \rightarrow R_3 + R_2$ to get

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}.$$

Finally we do $C_3 \rightarrow C_3 + C_2$ to get

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

Since all the diagonal entries are negative, the form is negative definite, and hence f has a local maximum at $(0, 0, 0)$.

Here is the determinant method: $d_1 = \det([-1]) = -1 < 0$, $d_2 = \det\left(\begin{bmatrix} -1 & 1 \\ 1 & -2 \end{bmatrix}\right) =$

$1 > 0$, and $d_3 = \det\left(\begin{bmatrix} -1 & 1 & 1 \\ 1 & -2 & 0 \\ 1 & 0 & -3 \end{bmatrix}\right) = -1 < 0$. Since we have the pattern

$- + -$ we know that the form is negative definite, and hence that f has a local maximum at $(0, 0, 0)$.

3. Classify the critical points of $f(x, y, z) = x^2 + y^2 + 3z^2 + 2xz + 4yz$.

$f_x = 2x + 2z$, $f_y = 2y + 4z$, and $f_z = 6z + 2x + 4y$. Setting each of these equal to zero and rearranging a bit gives the system of equations

$$\begin{array}{rcl} 2x & +2z & = 0 \\ & 2y +4z & = 0 \\ 2x & +4y & +6z = 0 \end{array}$$

This system has augmented matrix

$$\left[\begin{array}{ccc|c} 2 & 0 & 2 & 0 \\ 0 & 2 & 4 & 0 \\ 2 & 4 & 6 & 0 \end{array} \right].$$

Doing the row operations $R_3 \rightarrow R_3 - R_1$ and then $R_3 \rightarrow R_3 - 2R_2$ puts it into the row echelon form

$$\left[\begin{array}{ccc|c} 2 & 0 & 2 & 0 \\ 0 & 2 & 4 & 0 \\ 0 & 0 & -4 & 0 \end{array} \right].$$

Back substitution yields $-4z = 0$, $z = 0$, $2y + 4z = 0$, $2y = 0$, $y = 0$, $2x + 2z = 0$, $2x = 0$, $x = 0$. Thus we have one critical point $\mathbf{a} = (0, 0, 0)$.

We now classify this critical point. $f_{xx} = 2$, $f_{yy} = 2$, $f_{zz} = 6$, $f_{xy} = f_{yx} = 0$, $f_{xz} = f_{zx} = 2$, and $f_{yz} = f_{zy} = 4$. Note that these all happen to be constants. If they had not been constants, we would have had to plug in the values $x = 0$, $y = 0$, and $z = 0$ to obtain the values of these derivatives at \mathbf{a} before going on. We next have

$$S = \begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & 4 \\ 2 & 4 & 6 \end{bmatrix}.$$

Since all the entries are even, it will be easier to work with

$$M = \frac{1}{2}S = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 1 & 2 & 3 \end{bmatrix}.$$

Here is the diagonalization method: We do $R_3 \rightarrow R_3 - R_1$ to get

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 0 & 2 & 2 \end{bmatrix}.$$

Then we do $C_3 \rightarrow C_3 - C_1$ to get

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 2 \end{bmatrix}.$$

Next we do $R_3 \rightarrow R_3 - 2R_2$ to get

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & -2 \end{bmatrix}.$$

Finally we do $C_3 \rightarrow C_3 - 2C_2$ to get

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix}.$$

Since we have both positive and negative entries on the diagonal, the form is indefinite, and hence f has a saddle point at $(0, 0, 0)$.

Here is the determinant method: $d_1 = \det([1]) = 1 > 0$, $d_2 = \det\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) =$

$1 > 0$, and $d_3 = \det\left(\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 1 & 2 & 3 \end{bmatrix}\right) = -2 < 0$. Since we do not have the pattern

$+++$ or $-+-$ we know that the form is indefinite, and hence that f has a saddle point at $(0, 0, 0)$.

QUADRATIC FORMS AND EIGENVALUES

All this may seem vaguely reminiscent of using eigenvalues and eigenvectors to diagonalize an $n \times n$ matrix A . We briefly review the main points; see your linear algebra book for more details. Recall that the eigenvalues are the solutions of the characteristic equation $\det(A - \lambda I) = 0$. For each eigenvalue λ you find a basis B_λ for the eigenspace $E_\lambda = \{\mathbf{x} \mid (A - \lambda I)\mathbf{x} = \mathbf{0}\}$. Put all these basis vectors together. You will have at most n of them. If you have n vectors, then A is *diagonalizable*. You let P be the $n \times n$ matrix whose columns are these n vectors. Then $P^{-1}AP = D$, a diagonal matrix whose diagonal entries are the eigenvalues corresponding to the eigenvectors. If you have fewer than n vectors, then A is not diagonalizable.

Notice that $P^{-1}AP = D$ looks sort of like $Q^tSQ = D$. Note the differences: S is symmetric. In general A need not be symmetric. In general the inverse of a matrix is not equal to its transpose. A matrix R is called *orthogonal* if $R^{-1} = R^t$. This is equivalent to saying that the columns of R form an *orthonormal* set (the dot product of two different vectors is 0 while the dot product of a vector with itself is 1); it is also equivalent to saying that the rows of R form an orthonormal set. If there is an orthogonal matrix R such that $R^{-1}AR = D$, a diagonal matrix, then we say that A is *orthogonally diagonalizable*. Note that in this case $A = RDR^{-1} = RDR^t$ and $D^t = D$, so $A^t = (RDR^t)^t = (R^t)^t D^t R^t = RDR^t = A$. Thus if A is orthogonally diagonalizable, it must be symmetric. The converse of this statement is also true; if A is symmetric, then it is orthogonally diagonalizable. The theory guarantees that you will always have n vectors and that vectors in different B_λ 's will have dot product equal to zero. You apply the Gram-Schmidt process to each B_λ to get a new basis which is an orthonormal set. You then let R be the matrix whose columns are these

n vectors. Then $R^tAR = R^{-1}AR = D$. In this case the diagonal entries μ_i of D are the eigenvalues λ_i of A .

Let's look at the example $A = S = \begin{bmatrix} 4 & -1 \\ -1 & 2 \end{bmatrix}$. The characteristic equation is $\lambda^2 - 6\lambda + 7 = 0$. The eigenvalues are $\lambda_1 = 3 + \sqrt{2} \approx 4.41$ and $\lambda_2 = 3 - \sqrt{2} \approx 1.59$. The basis vectors for the two eigenspaces are, respectively, $\mathbf{v}_1 = (-1 - \sqrt{2}, 1) \approx (-2.41, 1)$ and $\mathbf{v}_2 = (-1 + \sqrt{2}, 1) \approx (0.41, 1)$. Applying Gram-Schmidt to each of these gives $\mathbf{w}_1 = \left(\frac{-1-\sqrt{2}}{\sqrt{4+2\sqrt{2}}}, \frac{1}{\sqrt{4+2\sqrt{2}}}\right) \approx (-0.92, 0.38)$ and $\mathbf{w}_2 = \left(\frac{-1+\sqrt{2}}{\sqrt{4-2\sqrt{2}}}, \frac{1}{\sqrt{4-2\sqrt{2}}}\right) \approx$

$(0.38, 0.92)$. So our matrix $R = \begin{bmatrix} \frac{-1-\sqrt{2}}{\sqrt{4+2\sqrt{2}}} & \frac{-1+\sqrt{2}}{\sqrt{4-2\sqrt{2}}} \\ \frac{1}{\sqrt{4+2\sqrt{2}}} & \frac{1}{\sqrt{4-2\sqrt{2}}} \end{bmatrix} \approx \begin{bmatrix} -0.92 & 0.38 \\ 0.38 & 0.92 \end{bmatrix}$, while

$$D = \begin{bmatrix} 3 + \sqrt{2} & 0 \\ 0 & 3 - \sqrt{2} \end{bmatrix} \approx \begin{bmatrix} 4.41 & 0 \\ 0 & 1.59 \end{bmatrix}.$$

Note that this is different from (and more complicated than) the matrices $Q = \begin{bmatrix} 1 & \frac{1}{4} \\ 0 & 1 \end{bmatrix}$ and $D = \begin{bmatrix} 4 & 0 \\ 0 & \frac{7}{4} \end{bmatrix}$ that we obtained using row and column operations.

Starting from scratch, orthogonal diagonalization is probably the *last* method you would want to use to classify the form. But it does show that if you happen to know the eigenvalues of S , then you can classify the form. If they are all positive, the form is positive definite. If they are all negative, the form is negative definite. If some are positive and some are negative, then it is indefinite.