18.024 SPRING OF 2008
SD. SECOND-DERIVATIVE TEST FOR EXTREMA
OF FUNCTIONS OF TWO VARIABLES

Proof of the second-derivative test. Our goal is to derive *the second-derivative test*, which determines the nature of a critical point of a function of two variables, that is, whether a critical point is a local minimum, a local maximum, or a saddle point, or none of these. In general for a function of n variables, it is determined by the algebraic sign of a certain quadratic form, which in turn is determined by eigenvalues of the Hessian matrix [Apo, Section 9.11]. This approach however relies on results on eigenvalues, and it may take several lectures to fully develop. Here we focus on the simpler setting when $n = 2$ and derive a test using the algebraic sign of the second derivative of the function.

The statement of the test is in [Apo, Theorem 9.7].

Theorem 1 (The second-derivative test)**.** *Let the scalar field $f(x_1, x_2)$ have continuous second derivatives in an open ball containing $\mathbf{a} = (a_1, a_2)$. Suppose that $D_1 f(\mathbf{a}) = D_2 f(\mathbf{a}) = 0$. Let $A = D_{11} f(\mathbf{a})$, $B = D_{12} f(\mathbf{a}) = D_{21} f(\mathbf{a})$ and $C = D_{22} f(\mathbf{a})$. Let*

$$\triangle = \det \begin{bmatrix} A & B \\ B & C \end{bmatrix} = AC - B^2.$$

Then, we have

(a) *If $\triangle < 0$, then $\mathbf{a}$ is a saddle point.*
(b) *If $\triangle > 0$ and $A > 0$, then $f(\mathbf{a})$ is a local minimum.*
(c) *If $\triangle > 0$ and $A < 0$, then $f(\mathbf{a})$ is a local maximum.*
(d) *If $\triangle = 0$, then the test is inconclusive.*

The proof uses *the second-orde Taylor formula*, which we will state for general scalar fields.

Theorem 2 (Second-order Taylor formula)**.** *Let f be a scalar field with continuous second-order partial derivatives $D_{ij} f$ in an open ball $B(\mathbf{a})$. Then, for all $\mathbf{y} \in \mathbb{R}^n$ such that $\mathbf{a} + \mathbf{y} \in B(\mathbf{a})$ we have*

$$f(\mathbf{a} + \mathbf{y}) - f(\mathbf{a}) = \nabla f(\mathbf{a}) \cdot \mathbf{y} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij} f(\mathbf{a} + \theta \mathbf{y}) y_i y_j,$$

where $\mathbf{y} = (y_1, \ldots, y_n)$ and $0 < \theta < 1$.

The statement and its proof is in [Apo, Section 9.10], and hence it is omitted. The coefficient of the quadratic form in the above expansion are the second-order partial derivatives. The $n \times n$ matrix of second-order derivatives $D_{ij} f(\mathbf{a})$ is often called the *Hessian matrix*.

In our present setting, the above Taylor expansion leads to

$$\begin{aligned} f(\mathbf{a} + t\mathbf{y}) =& f(\mathbf{a}) + t(D_1 f(\mathbf{a}) h + D_2 f(\mathbf{a}) k) \\ &+ \frac{t^2}{2}(D_{11} f(\mathbf{a}^*) h^2 + 2 D_{12} f(\mathbf{a}^*) hk + D_{22} f(\mathbf{a}^*) k^2), \end{aligned} \tag{*}$$

where $\mathbf{y} = (h, k)$ and $\mathbf{a}^* = \mathbf{a} + \theta \mathbf{y}$ for some $0 < \theta < t$. Therefore,

$$\begin{aligned} \frac{2}{t^2} f(\mathbf{a} + t\mathbf{y}) - f(\mathbf{a}) =& (Ah^2 + 2Bhk + Ck^2) \\ &+ ((D_{11} f(\mathbf{a}^*) - A) h^2 + 2(D_{12} f(\mathbf{a}^*) - B) hk + (D_{22} f(\mathbf{a}^*) - C) k^2), \end{aligned} \tag{**}$$

for t small enough. This uses that $D_1 f(\mathbf{a}) = D_2 f(\mathbf{a}) = 0$. Roughly speaking, it states that

$$\frac{2}{t^2} f(\mathbf{a} + t\mathbf{y}) - f(\mathbf{a}) \sim (Ah^2 + 2Bhk + Ck^2)$$

as the last three terms in (**) are small if $\mathbf{a}^*$ is close to $\mathbf{a}$, thanks to that the second-order partial derivatives are continuous.

We are concerned with the sing of the left side of (**) when t is small. Consider the quadratic function

$$Q(h, k) = Ah^2 + 2Bhk + Ck^2,$$

where (h, k) is a unit vector in $\mathbb{R}^2$.

Case 1: If $\triangle = AC - B^2 < 0$, then we claim that Q takes both positive and negative signs. Indeed, if $A \neq 0$ then the quadratic equation $Ax^2 + 2Bx + C = 0$ has two distinct real roots, and the graph of $y = Ax^2 + 2Bx + C$ represents a parabola which crosses the x-axis at two distinct points; otherwise if $A = 0$ then $B \neq 0$ must hold, and $y = Ax^2 + 2Bx + C$ represents a straight line with a non-zero slope. In either case, $Ax^2 + 2Bx + C$ takes both positive and negative values. This proves the claim.

Case 2: If $\triangle = AC - B^2 > 0$ then Q takes only one sign. Since (h, k) is a unit vector, let $(h, k) = (\cos t, \sin t)$, where $t \in [0, 2\pi]$. $Q(t) = A(\cos t, \sin t)$ is a continuous function in t. We will show that $Q(t)$ takes only one sign. First, $Q(0) = A \neq 0$ takes a definitive sign. Suppose now that is a $t_0 \in [0, 2\pi]$ such that $Q(t_0) = 0$. That is, $Q(\cos t_0, \sin t_0) = Q(h_0, k_0) = 0$ for some (h_0, k_0). If $k_0 \neq 0$ this means that h_0/k_0 is a real root of $Ax^2 + 2Bx + C = 0$. But, since $B^2 - AC < 0$, the quadratic equation can only have complex roots, and we are lead to a contradiction. Similarly, if $k_0 = 0 (and h_0 \neq 0)$, we consider the quadratic equation $Cx^2 + 2Bx + A = 0$, and, by the same argument, we get a contradiction. This proves the claim.

Our next step is to study implications of the previous results for the nature of the critical points.

First **we prove part (a)** of the theorem. Assume that $B^2 - AC > 0$. Let $\mathbf{y}_0$ be a unit vector for which $Q(\mathbf{y}_0) > 0$. Examining the formula (**), we notice that the expression $(2/t^2)(f(\mathbf{a}+t\mathbf{y}) - f(\mathbf{a})$ approaches $Q(\mathbf{y}_0)$ as $t \to 0$. Let $\mathbf{x} = \mathbf{a} + t\mathbf{y}$ and let $t \to 0$. Then, $\mathbf{x}$ approaches $\mathbf{a}$ along the straight line from $\mathbf{a}$ to $\mathbf{a} + t\mathbf{y}_0$ and the expression $f(\mathbf{x}) - f(\mathbf{a})$ approaches zero through positive values. On the other hand, if $\mathbf{y}_1$ is a point such that $Q(\mathbf{y}_1) < 0$, then by the same argument, the expression $f(\mathbf{x}) - f(\mathbf{a})$ approaches zero through negative values as $\mathbf{x}$ approaches $\mathbf{a}$ along the straight line from $\mathbf{a}$ to $\mathbf{a} + t\mathbf{y}_1$. Therefore, $\mathbf{a}$ is a saddle point of f.

Next, **we prove parts (b) and (c)** of the theorem. Examining (**) again, we know that $|Q(\mathbf{y})| > 0$ for all unit vectors $\mathbf{y}$. Then $|Q(\mathbf{y})|$ has a positive minimum, say m, while $\mathbf{y}$ exhausts all unit vectors. Now choose $\delta > 0$ small enough that

$$|D_{1,1} f(\mathbf{a}^*) - A|, |D_{1,2} f(\mathbf{a}^*) - B|, |D_{2,2} f(\mathbf{a}^*) - C| < m/3$$

whenever $|\mathbf{a}^* - \mathbf{a}| < \delta$. This uses continuity of the second-order partial derivatives of f. It $0 < t < \delta$, then $\mathbf{a}^*$ is on the line from $\mathbf{a}$ to $\mathbf{a} + \delta\mathbf{y}$. Since $\mathbf{y}$ is a unit vector, the right side of (*) has the same sign as A whenever $0 < t < \delta$. If $A > 0$ this means that $f(\mathbf{x}) - f(\mathbf{a}) > 0$ whenever $0 < |\mathbf{x} - \mathbf{a}| < \delta$, which says that f has a local minimum at $\mathbf{a}$. Similarly, $A < 0$ implies that f has a local maximum at $\mathbf{a}$.

The proof of (d) of the theorem is by an example, which is left as a pset problem.

Example 3. (1) Let $f(x, y) = x^2 + 3xy + y^2$. The origin is a critical point of f since $\bigtriangledown f(x, y) = (2x + 3y, 3x + 2y)$. Here, $A = 2, B = 3, C = 2$. Since $\triangle = -5$, which is negative, the origin is a saddle point of f.

(2) Let $f(x,y) = 3x^2 - 5xy + 3y^2$. Again, the origin is a critical point since $\bigtriangledown f(x,y) = (6x - 5y, -5x + 6y)$. Here, $A = 6, B = -5, C = 6$. Since $\triangle = 11$, which is positive, and since $A > 0$ the origin is a local minimum point.

Example 4. Find the critical points of $f(x,y) = x \sin y$ and determine whether they are local minima, local maxima or saddle points.
solution. It is straightforward that

$$\bigtriangledown f(x,y) = (\sin y, x \cos y).$$

If $\sin y = 0$ then $y = n\pi$, where n is an integer. Since $\cos y \neq 0$ at these points, $x \cos y = 0$ implies that $x = 0$. Thus, all critical points of f are of the form $(0, n\pi)$, where n is an integer. Next, since

$$A = 0, \quad B = \cos y = \pm 1, \quad C = -x \sin y = 0$$

at critical points, the discriminant is calculated as $\delta = -1 < 0$. Therefore, all critical points are saddle.

Exercise. Find the maxima, minima and saddle points of $z = (x^2 - y^2)e^{(-x^2-y^2)/2}$.

Exercise. Let $z = (x^2 + y^2)\cos(x + 2y)$. Show that $(0,0)$ is a critical point. Is it a local minimum or a local maximum?

Examples for use of Lagrange's multiplier method. Assuming that among all rectangular boxes with fixed surface area of 10 square meters there is a box of largest possible volume. Find its dimension.
solution. Let the lengths of the sides of the cube be $x, y, z \geqslant 0$, respectively. The volume is $f(x,y,z) = xyz$. The constraint is $2(xy + yz + xz) = 10$. Lagrange multiplier conditions are

$$yz = \lambda(y + z); \quad xz = \lambda(x + z); \quad xy = \lambda(x + y)$$

and

$$xy + yz + xz = 5.$$

First of all, $x \neq 0$; for $x = 0$ implies $yz = 5$ and $\lambda z = 0$ so that $\lambda = 0$ and this leads to a contradiction since $yz = 0$. Similarly, $y \neq 0$, $z \neq 0$ and $x + y \neq 0$ and so on. Thus, we get

$$\lambda = \frac{yz}{y+z} = \frac{xz}{x+z} = \frac{xy}{x+y},$$

whence $x = y = z$. Substituting these values into the constraint equation, we obtain $x = y = z = \sqrt{5/3}$, This (cubical) shape must therefore maximize the volume, assuming there is a box of maximum volume.

Now we make a couple of remarks.

1. The solution in the example does *not* demonstrate that the cube is the rectangular box of largest volume with a given fixed surface area; it proves that the cube is the only possible candidate for a maximum. The distinction between showing that *there is only one possible solution* to a problem and that, in fact, *a solution exists* is quite subtle.

Indeed, Queen Dido (ca. 900 B.C.) realized that among all planar regions with fixed circumference the disc is the region of maximum area. It is not terribly hard to prove this fact assuming that there does exists a region of maximum area. However, proving that such a region of maximum area exists is quite another (and difficult) matter. A complete proof was not given until the second half of the nineteenth century by a German mathematician Weierstrass (1815–1897).

2. The problem in showing that $f(x,y,z) = xyz$ has a maximum lies in the fact that f is a continuous function which is defined on the unbounded surface $xy + yz + zx = 5$ and not on a bounded set which includes its boundary. The way to show the existence of a maximum of

$f(x, y, z) = xyz$ subject to $xy + yz + zx = 5$ is to show that if either x, y or z tend to infinity, then $f(x, y, z) \to 0$. We may then conclude that the maximum of f on the surface $xy + yz + zx = 5$. Indeed, multiplying the equation of the surface by z we obtain the equation $xyz + xz_y^2 z^2 = 5z \to 0$ s $x \to \infty$. Since $x, y, z \geqslant 0$ it follows that $xyz = f(x, y, z) \to 0$. Similarly, $f(x, y, z) \to 0$ if either y or z tend to infinity. Therefore, a box of maximum volume must exist.

3. There are also second derivative tests available for constrained extrema. Please consult [MaTr].

Exercise. Show that the volume of the largest rectangular parallelepiped that can be inscribed in the ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

is $8abc/3\sqrt{3}$.

REFERENCES

[Apo] T. Apostol, *Calculus*, vol. II, Second edition, Wiley, 1967.
[MaTr] J. Marsde and A. Tromba, *Vector Calculus*, Third edition, W.H. Freeman, N.Y., 1988.

E-mail address: verahur@math.mit.edu