Mathematics for AI

Adams Wai Kin Kong

School of Computer Science and Engineering

Nanyang Technological University, Singapore

adamskong@ntu.edu.sg

Maximum and Minimum Values

In this section we see how to use partial derivatives to locate maxima and minima of functions of two variables.

Look at the hills and valleys in the graph of *f* shown in Figure 1.

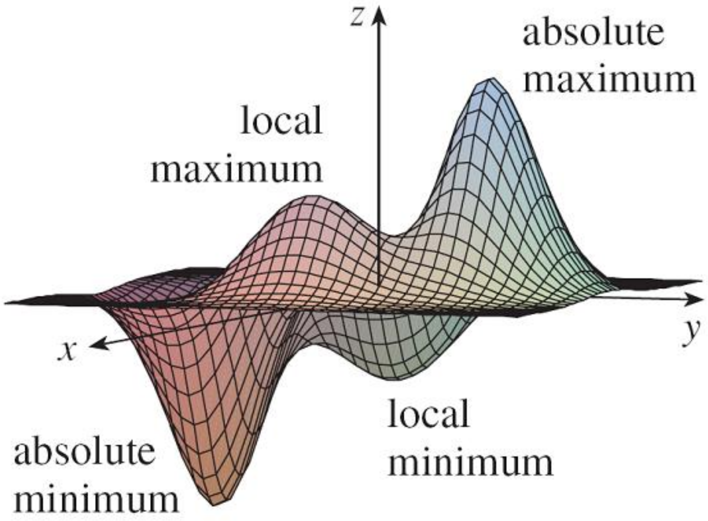


Figure 1

Maximum and Minimum Values

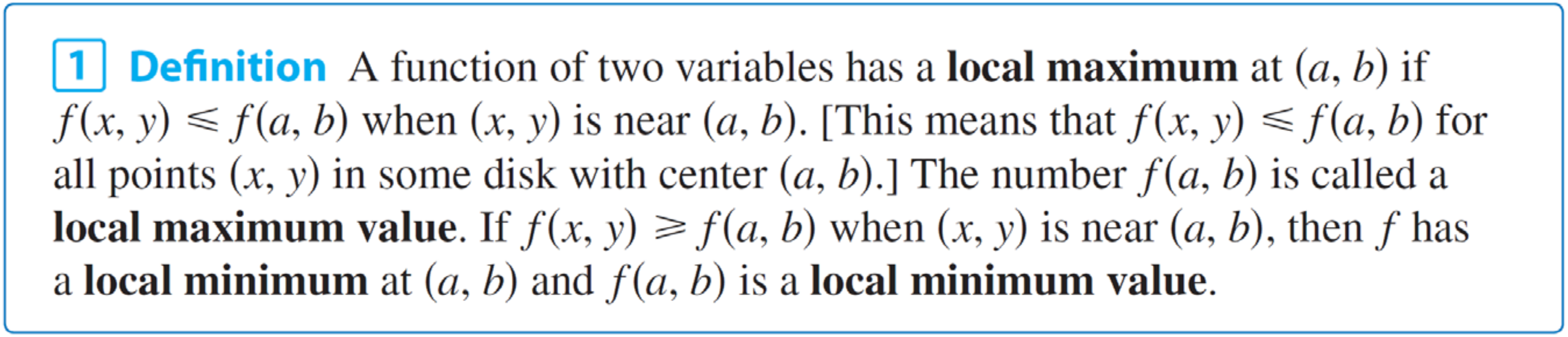
There are two points (a, b) where f has a *local maximum*, that is, where $f(a, b)$ is larger than nearby values of $f(x, y)$.

The larger of these two values is the *absolute maximum*.

Likewise, f has two *local minima*, where $f(a, b)$ is smaller than nearby values.

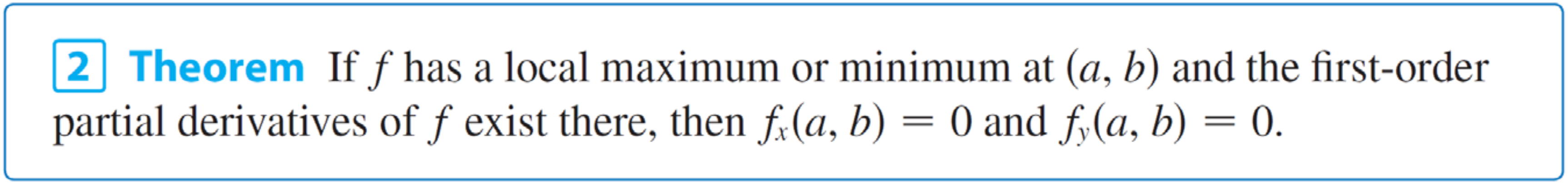
The smaller of these two values is the *absolute minimum*.

Maximum and Minimum Values



1 **Definition** A function of two variables has a **local maximum** at (a, b) if $f(x, y) \leqslant f(a, b)$ when (x, y) is near (a, b). [This means that $f(x, y) \leqslant f(a, b)$ for all points (x, y) in some disk with center (a, b).] The number $f(a, b)$ is called a **local maximum value**. If $f(x, y) \geqslant f(a, b)$ when (x, y) is near (a, b), then f has a **local minimum** at (a, b) and $f(a, b)$ is a **local minimum value**.

If the inequalities in Definition 1 hold for *all* points (x, y) in the domain of f, then f has an **absolute maximum** (or **absolute minimum**) at (a, b).



2 **Theorem** If f has a local maximum or minimum at (a, b) and the first-order partial derivatives of f exist there, then $f_x(a, b) = 0$ and $f_y(a, b) = 0$.

Maximum and Minimum Values

A point (a, b) is called a **critical point** (or *stationary point*) of f if $f_x(a, b) = 0$ and $f_y(a, b) = 0$, or if one of these partial derivatives does not exist.

Theorem 2 says that if f has a local maximum or minimum at (a, b), then (a, b) is a critical point of f.

However, as in single-variable calculus, not all critical points give rise to maxima or minima.

At a critical point, a function could have a local maximum or a local minimum or neither.

Example 1

Let $f(x, y) = x^2 + y^2 - 2x - 6y + 14$.

Then

$$f_x(x, y) = 2x - 2 \qquad f_y(x, y) = 2y - 6$$

These partial derivatives are equal to 0 when $x = 1$ and $y = 3$, so the only critical point is $(1, 3)$.

By completing the square, we find that

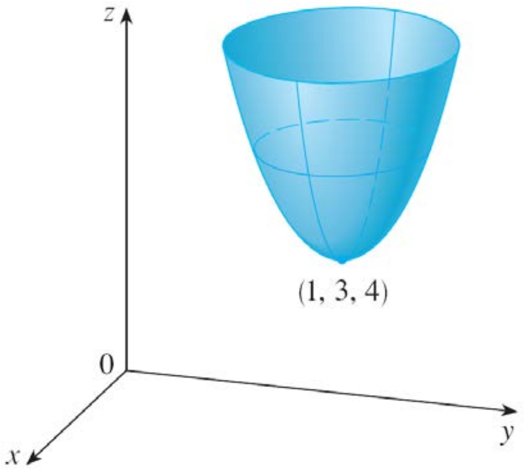
$$f(x, y) = 4 + (x - 1)^2 + (y - 3)^2$$

Example 1

Since $(x - 1)^2 \geq 0$ and $(y - 3)^2 \geq 0$, we have $f(x, y) \geq 4$ for all values of x and y.

Therefore $f(1, 3) = 4$ is a local minimum, and in fact it is the absolute minimum of f.

This can be confirmed geometrically from the graph of f, which is the elliptic paraboloid with vertex (1, 3, 4) shown in Figure 2.

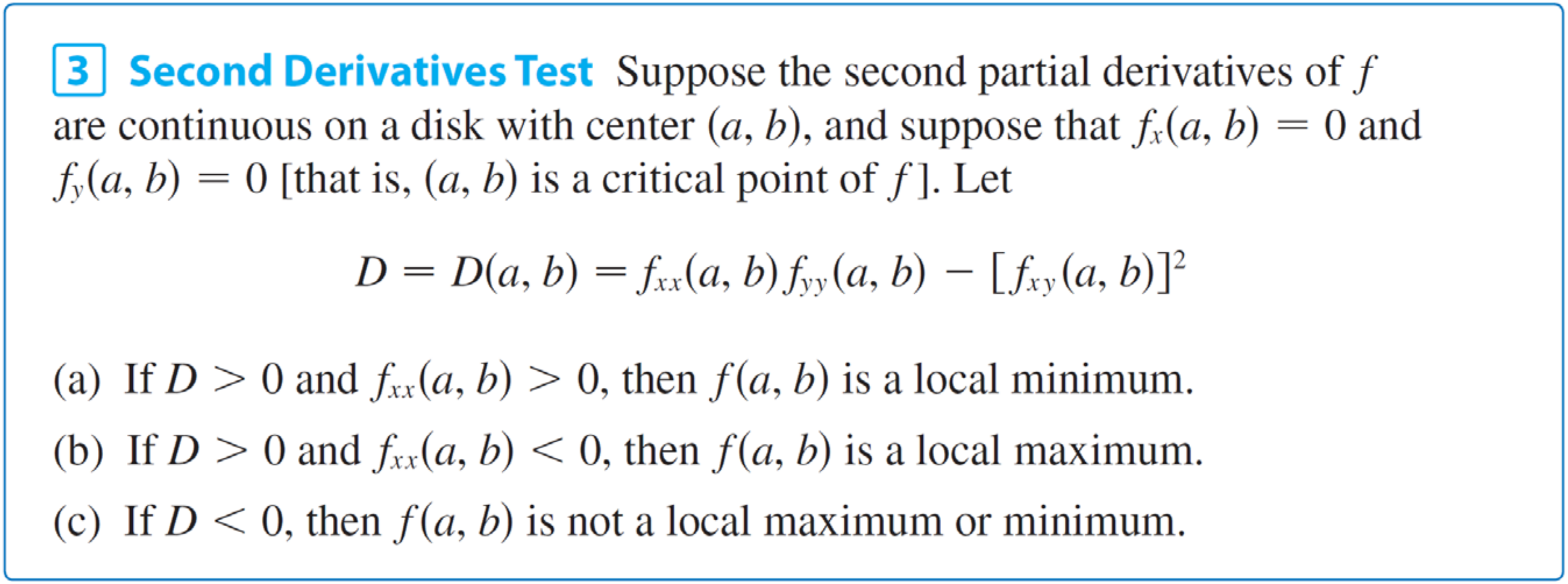


$z = x^2 + y^2 - 2x - 6y + 14$

Figure 2

Maximum and Minimum Values

The following test, is analogous to the Second Derivative Test for functions of one variable.



3 Second Derivatives Test Suppose the second partial derivatives of f are continuous on a disk with center (a, b), and suppose that $f_x(a, b) = 0$ and $f_y(a, b) = 0$ [that is, (a, b) is a critical point of f]. Let

$$D = D(a, b) = f_{xx}(a, b) f_{yy}(a, b) - [f_{xy}(a, b)]^2$$

(a) If $D > 0$ and $f_{xx}(a, b) > 0$, then $f(a, b)$ is a local minimum.

(b) If $D > 0$ and $f_{xx}(a, b) < 0$, then $f(a, b)$ is a local maximum.

(c) If $D < 0$, then $f(a, b)$ is not a local maximum or minimum.

In case (c) the point (a, b) is called a **saddle point** of f and the graph of f crosses its tangent plane at (a, b).

Second Derivative Test for Multivariable Functions

Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a function taking as input a vector $\mathbf{x} \in \mathbb{R}^n$ and outputting a scalar $f(\mathbf{x}) \in \mathbb{R}$.

Critical points of *f* are solutions of

$$\begin{cases} f_{x_1} = 0 \\ f_{x_2} = 0 \\ \quad \vdots \\ f_{x_n} = 0 \end{cases}$$

or one or more f_{x_i} does not exist.

Second Derivative Test for Multivariable Functions

If all second partial derivatives of *f* exist and are continuous over the domain of the function, then the Hessian matrix **H** of *f* is a square *n*×*n* matrix, usually defined and arranged as follows:

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

A critical point *a* is called non-degenerate if

$$\det(H(a)) \neq 0$$

Second Derivative Test for Multivariable Functions

Suppose a is a non-degenerate critical point of a function $f : \mathbb{R}^n \to \mathbb{R}$, then

- If $H_f(a)$is positive definite, then $x = a$ is a local minimum of f.
- If $H_f(a)$ is negative definite, then $x = a$ is a local maximum of f.
- If $H_f(a)$ is neither positive nor negative definite, then it is called a saddle point of f.

https://en.wikipedia.org/wiki/Definiteness_of_a_matrix

Absolute Maximum and Minimum Values

For a function *f* of one variable, the Extreme Value Theorem says that if *f* is continuous on a closed interval [*a*, *b*], then *f* has an absolute minimum value and an absolute maximum value.

According to the Closed Interval Method, we found these by evaluating *f* not only at the critical numbers but also at the endpoints *a* and *b*.

There is a similar situation for functions of two variables. Just as a closed interval contains its endpoints, a **closed set** in $\mathbb{R}^2$ is one that contains all its boundary points.

Absolute Maximum and Minimum Values

[A boundary point of D is a point (a, b) such that every disk with center (a, b) contains points in D and also points not in D.]

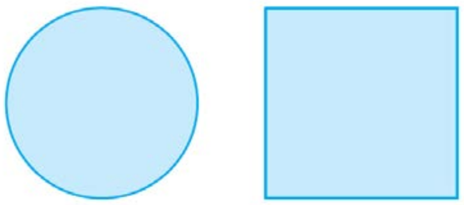
For instance, the disk

$$D = \{(x, y) | \ x^2 + y^2 \leq 1\}$$

which consists of all points on or inside the circle $x^2 + y^2 = 1$, is a closed set because it contains all of its boundary points (which are the points on the circle $x^2 + y^2 = 1$).

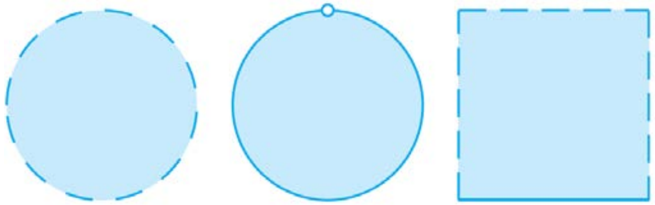
Absolute Maximum and Minimum Values

But if even one point on the boundary curve were omitted, the set would not be closed. (See Figure 11.)



Closed sets

Figure 11(a)



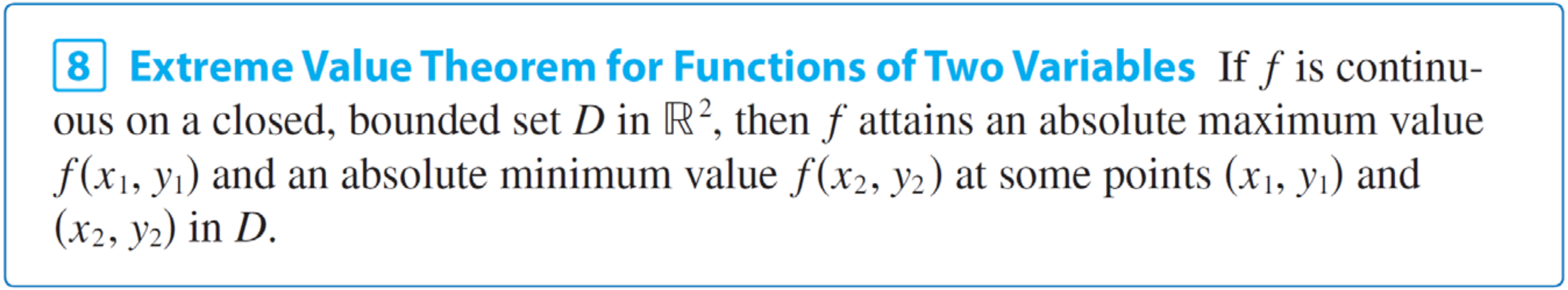
Sets that are not closed

Figure 11(b)

A **bounded set** in $\mathbb{R}^2$ is one that is contained within some disk.

Absolute Maximum and Minimum Values

Then, in terms of closed and bounded sets, we can state the following counterpart of the Extreme Value Theorem in two dimensions.

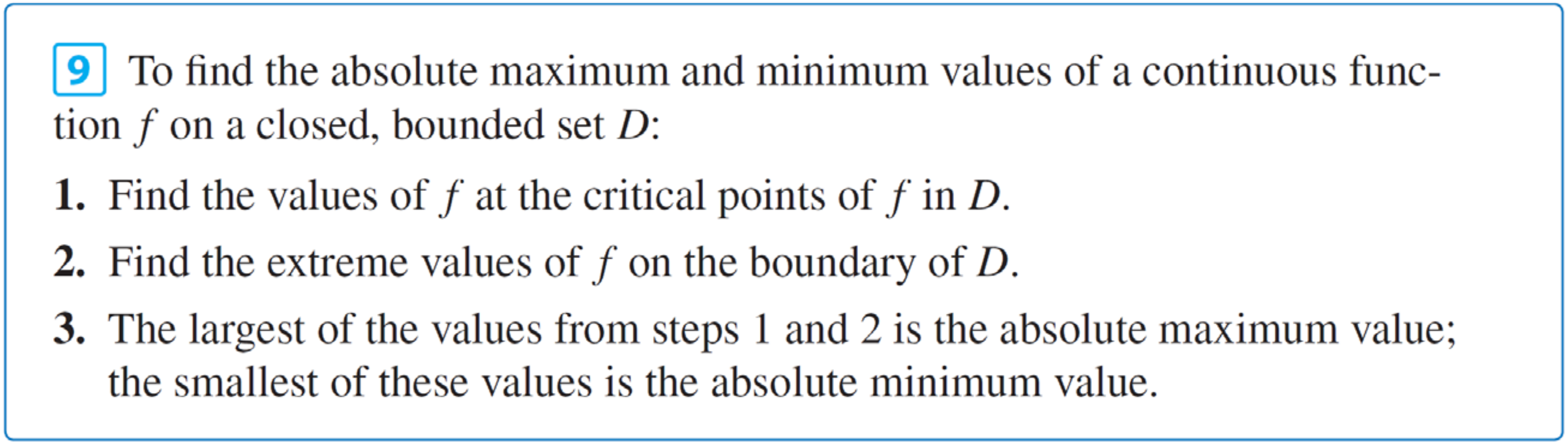


$\boxed{8}$ **Extreme Value Theorem for Functions of Two Variables** If f is continuous on a closed, bounded set D in $\mathbb{R}^2$, then f attains an absolute maximum value $f(x_1, y_1)$ and an absolute minimum value $f(x_2, y_2)$ at some points (x_1, y_1) and (x_2, y_2) in D.

Absolute Maximum and Minimum Values

To find the extreme values guaranteed by Theorem 8, we note that, by Theorem 2, if *f* has an extreme value at (x_1, y_1), then (x_1, y_1) is either a critical point of *f* or a boundary point of *D*.

Thus we have the following extension of the Closed Interval Method.



$\boxed{9}$ To find the absolute maximum and minimum values of a continuous function *f* on a closed, bounded set *D*:

1. Find the values of *f* at the critical points of *f* in *D*.
2. Find the extreme values of *f* on the boundary of *D*.
3. The largest of the values from steps 1 and 2 is the absolute maximum value; the smallest of these values is the absolute minimum value.

Example 7

Find the absolute maximum and minimum values of the function $f(x, y) = x^2 - 2xy + 2y$ on the rectangle $D = \{(x, y) \mid 0 \leq x \leq 3, 0 \leq y \leq 2\}$.

Solution:

Since f is a polynomial, it is continuous on the closed, bounded rectangle D, so Theorem 8 tells us there is both an absolute maximum and an absolute minimum.

According to step 1 in (9), we first find the critical points. These occur when

$$f_x = 2x - 2y = 0 \qquad\qquad f_y = -2x + 2 = 0$$

Example 7 – *Solution*

So the only critical point is $(1, 1)$, and the value of f there is $f(1, 1) = 1$.

In step 2 we look at the values of f on the boundary of D, which consists of the four line segments L_1, L_2, L_3, L_4 shown in Figure 12.

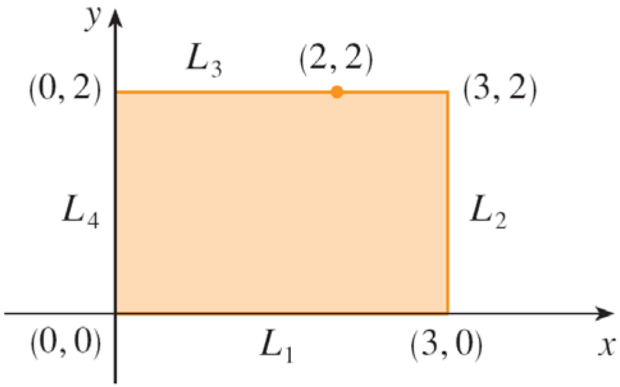


Figure 12

Example 7 – *Solution*

On L_1 we have $y = 0$ and

$$f(x, 0) = x^2 \qquad 0 \leq x \leq 3$$

This is an increasing function of x, so its minimum value is $f(0, 0) = 0$ and its maximum value is $f(3, 0) = 9$.

On L_2 we have $x = 3$ and

$$f(3, y) = 9 - 4y \qquad 0 \leq y \leq 2$$

This is a decreasing function of y, so its maximum value is $f(3, 0) = 9$ and its minimum value is $f(3, 2) = 1$.

Example 7 – *Solution*

On L_3 we have $y = 2$ and

$$f(x, 2) = x^2 - 4x + 4 \qquad 0 \leq x \leq 3$$

Simply by observing that $f(x, 2) = (x - 2)^2$, we see that the minimum value of this function is $f(2, 2) = 0$ and the maximum value is $f(0, 2) = 4$.

Example 7 – *Solution*

Finally, on L_4 we have $x = 0$ and

$$f(0, y) = 2y \qquad 0 \le y \le 2$$

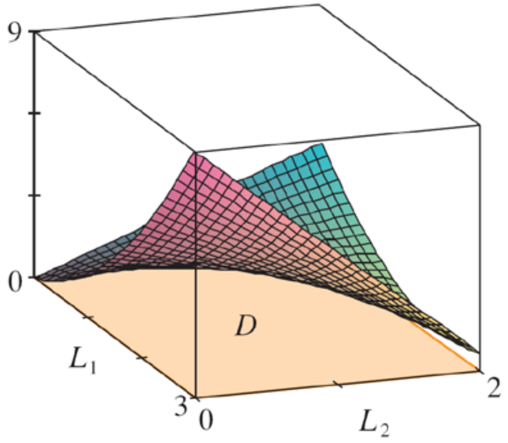
with maximum value $f(0, 2) = 4$ and minimum value $f(0, 0) = 0$.

Thus, on the boundary, the minimum value of f is 0 and the maximum is 9.

Example 7 – *Solution*

In step 3 we compare these values with the value $f(1, 1) = 1$ at the critical point and conclude that the absolute maximum value of f on D is $f(3, 0) = 9$ and the absolute minimum value is $f(0, 0) = f(2, 2) = 0$.

Figure 13 shows the graph of f.



$f(x, y) = x^2 - 2xy + 2y$

Figure 13

Lagrange Multipliers

In this section we present Lagrange's method for maximizing or minimizing a general function $f(x, y, z)$ subject to a constraint (or side condition) of the form $g(x, y, z) = k$.

It's easier to explain the geometric basis of Lagrange's method for functions of two variables. So we start by trying to find the extreme values of $f(x, y)$ subject to a constraint of the form $g(x, y) = k$.

In other words, we seek the extreme values of $f(x, y)$ when the point (x, y) is restricted to lie on the level curve $g(x, y) = k$.

Lagrange Multipliers

Figure 1 shows this curve together with several level curves of f.

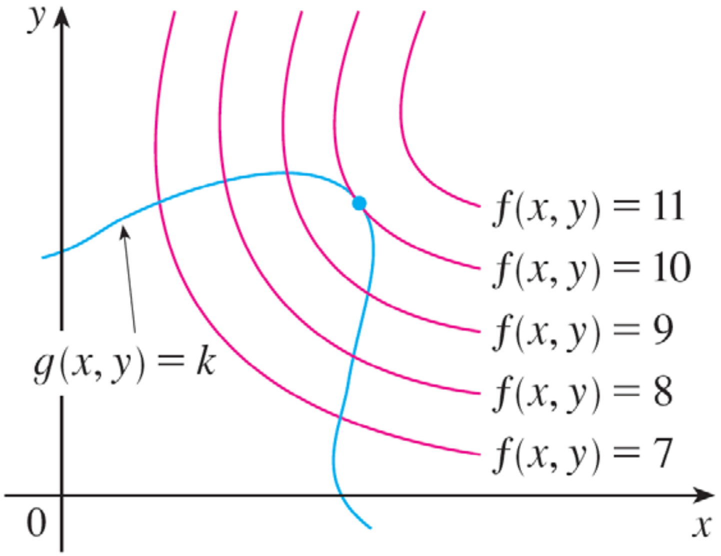


Figure 1

These have the equations $f(x, y) = c$, where $c = 7, 8, 9, 10, 11$.

Lagrange Multipliers

To maximize $f(x, y)$ subject to $g(x, y) = k$ is to find the largest value of c such that the level curve $f(x, y) = c$ intersects $g(x, y) = k$.

It appears from Figure 1 that this happens when these curves just touch each other, that is, when they have a common tangent line. (Otherwise, the value of c could be increased further.)

Lagrange Multipliers

This means that the normal lines at the point (x_0, y_0) where they touch are identical. So the gradient vectors are parallel; that is, $\nabla f(x_0, y_0) = \lambda \nabla g(x_0, y_0)$ for some scalar λ.

This kind of argument also applies to the problem of finding the extreme values of $f(x, y, z)$ subject to the constraint $g(x, y, z) = k$.

Thus the point (x, y, z) is restricted to lie on the level surface S with equation $g(x, y, z) = k$.

Lagrange Multipliers

Instead of the level curves in Figure 1, we consider the level surfaces $f(x, y, z) = c$ and argue that if the maximum value of f is $f(x_0, y_0, z_0) = c$, then the level surface $f(x, y, z) = c$ is tangent to the level surface $g(x, y, z) = k$ and so the corresponding gradient vectors are parallel.

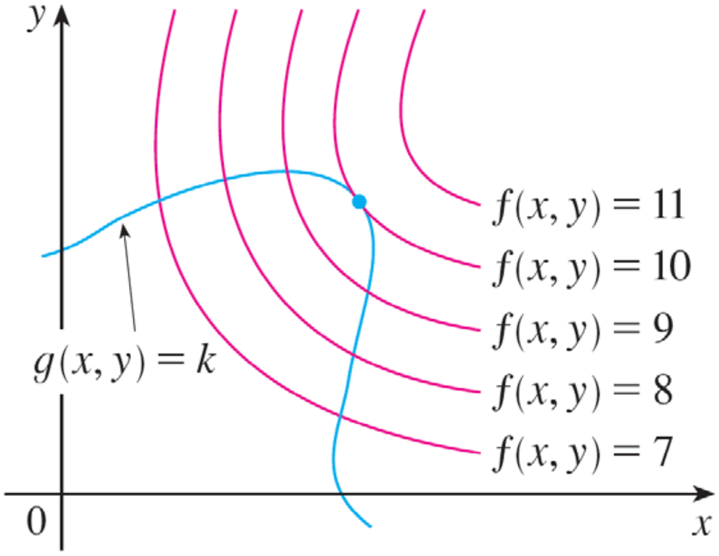


Figure 1

Lagrange Multipliers

This intuitive argument can be made precise as follows. Suppose that a function f has an extreme value at a point $P(x_0, y_0, z_0)$ on the surface S and let C be a curve with vector equation $\mathbf{r}(t) = \langle x(t), y(t), z(t) \rangle$ that lies on S and passes through P.

If t_0 is the parameter value corresponding to the point P, then $\mathbf{r}(t_0) = \langle x_0, y_0, z_0 \rangle$.

The composite function $h(t) = f(x(t), y(t), z(t))$ represents the values that f takes on the curve C.

Lagrange Multipliers

Since f has an extreme value at (x_0, y_0, z_0), it follows that h has an extreme value at t_0, so $h'(t_0) = 0$. But if f is differentiable, we can use the Chain Rule to write

$$\begin{aligned} 0 &= h'(t_0) \\ &= f_x(x_0, y_0, z_0)x'(t_0) + f_y(x_0, y_0, z_0)y'(t_0) + f_z(x_0, y_0, z_0)z'(t_0) \\ &= \nabla f(x_0, y_0, z_0) \cdot \mathbf{r}'(t_0) \end{aligned}$$

This shows that the gradient vector $\nabla f(x_0, y_0, z_0)$ is orthogonal to the tangent vector $\mathbf{r}'(t_0)$ to every such curve C. But we already know that the gradient vector of g, $\nabla g(x_0, y_0, z_0)$, is also orthogonal to $\mathbf{r}'(t_0)$ for every such curve.

Lagrange Multipliers

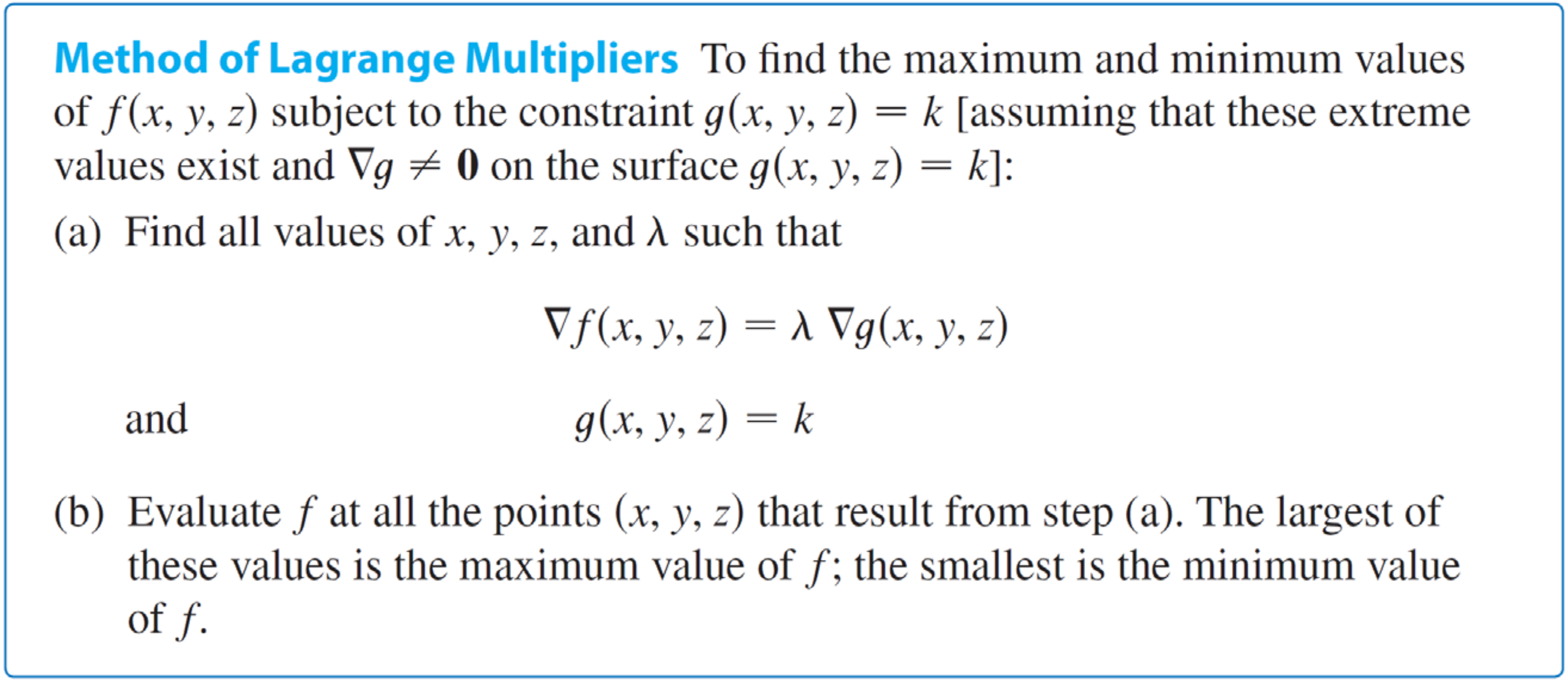
This means that the gradient vectors $\nabla f(x_0, y_0, z_0)$ and $\nabla g(x_0, y_0, z_0)$ must be parallel. Therefore, if $\nabla g(x_0, y_0, z_0) \neq \mathbf{0}$, there is a number λ such that

$$\boxed{\nabla f(x_0, y_0, z_0) = \lambda \nabla g(x_0, y_0, z_0)} \qquad (1)$$

The number λ in Equation 1 is called a **Lagrange multiplier**.

Lagrange Multipliers

The procedure based on Equation 1 is as follows.



Method of Lagrange Multipliers To find the maximum and minimum values of $f(x, y, z)$ subject to the constraint $g(x, y, z) = k$ [assuming that these extreme values exist and $\nabla g \neq \mathbf{0}$ on the surface $g(x, y, z) = k$]:

(a) Find all values of x, y, z, and λ such that

$$\nabla f(x, y, z) = \lambda \nabla g(x, y, z)$$

and

$$g(x, y, z) = k$$

(b) Evaluate f at all the points (x, y, z) that result from step (a). The largest of these values is the maximum value of f; the smallest is the minimum value of f.

https://en.wikipedia.org/wiki/Lagrange_multiplier

Lagrange Multipliers

If we write the vector equation $\nabla f = \lambda \nabla g$ in terms of components, then the equations in step (a) become

$$f_x = \lambda g_x \qquad f_y = \lambda g_y \qquad f_z = \lambda g_z \qquad g(x, y, z) = k$$

This is a system of four equations in the four unknowns x, y, z, and λ, but it is not necessary to find explicit values for λ.

For functions of two variables the method of Lagrange multipliers is similar to the method just described.

Lagrange Multipliers

To find the extreme values of $f(x, y)$ subject to the constraint $g(x, y) = k$, we look for values of x, y, and λ such that

$$\nabla f(x, y) = \lambda \nabla g(x, y) \quad \text{and} \quad g(x, y) = k$$

This amounts to solving three equations in three unknowns:

$$f_x = \lambda g_x \qquad f_y = \lambda g_y \qquad g(x, y) = k$$

Example 1

A rectangular box without a lid is to be made from 12 m^2 of cardboard. Find the maximum volume of such a box.

Solution:

Let x, y, and z be the length, width, and height, respectively, of the box in meters.

Then we wish to maximize

$$V = xyz$$

subject to the constraint

$$g(x, y, z) = 2xz + 2yz + xy = 12$$

Example 1 – *Solution*

Using the method of Lagrange multipliers, we look for values of x, y, z, and λ such that $\nabla V = \lambda \nabla g$ and $g(x, y, z) = 12$.

This gives the equations

$$V_x = \lambda g_x$$

$$V_y = \lambda g_y$$

$$V_z = \lambda g_z$$

$$2xz + 2yz + xy = 12$$

Example 1 – *Solution*

Which become

$$\boxed{2} \qquad yz = \lambda(2z + y)$$

$$\boxed{3} \qquad xz = \lambda(2z + x)$$

$$\boxed{4} \qquad xy = \lambda(2x + 2y)$$

$$\boxed{5} \qquad 2xz + 2yz + xy = 12$$

Example 1 – *Solution*

There are no general rules for solving systems of equations. Sometimes some ingenuity is required.

In the present example you might notice that if we multiply (2) by *x*, (3) by *y*, and (4) by *z*, then the left sides of these equations will be identical.

Doing this, we have

$$\boxed{6} \qquad xyz = \lambda(2xz + xy)$$

$$\boxed{7} \qquad xyz = \lambda(2yz + xy)$$

$$\boxed{8} \qquad xyz = \lambda(2xz + 2yz)$$

Example 1 – *Solution*

We observe that $\lambda \neq 0$ because $\lambda = 0$ would imply $yz = xz = xy = 0$ from (2), (3), and (4) and this would contradict (5).

Therefore, from (6) and (7), we have

$$2xz + xy = 2yz + xy$$

which gives $xz = yz$.

But $z \neq 0$ (since $z = 0$ would give $V = 0$), so $x = y$.

Example 1 – *Solution*

From (7) and (8) we have

$$2yz + xy = 2xz + 2yz$$

which gives $2xz = xy$ and so (since $x \neq 0$) $y = 2z$.

If we now put $x = y = 2z$ in (5), we get

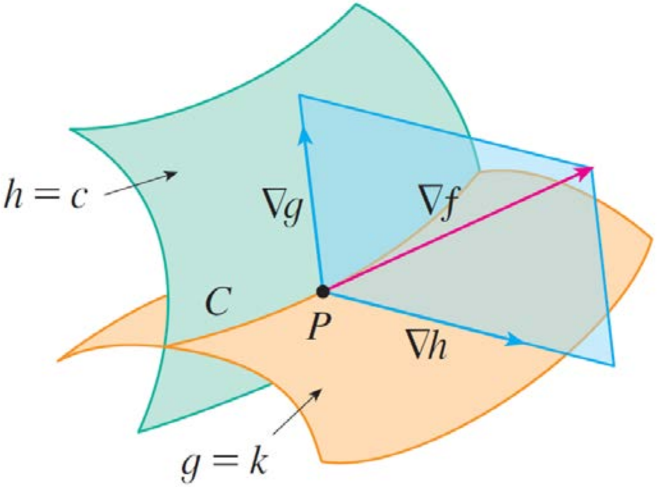
$$4z^2 + 4z^2 + 4z^2 = 12$$

Since x, y, and z are all positive, we therefore have $z = 1$ and so $x = 2$ and $y = 2$.

Two Constraints

Suppose now that we want to find the maximum and minimum values of a function $f(x, y, z)$ subject to two constraints (side conditions) of the form $g(x, y, z) = k$ and $h(x, y, z) = c$.

Geometrically, this means that we are looking for the extreme values of f when (x, y, z) is restricted to lie on the curve of intersection C of the level surfaces $g(x, y, z) = k$ and $h(x, y, z) = c$. (See Figure 5.)



Two Constraints

Suppose f has such an extreme value at a point $P(x_0, y_0, z_0)$. We know from the beginning of this section that ∇f is orthogonal to C at P.

But we also know that ∇g is orthogonal to $g(x, y, z) = k$ and ∇h is orthogonal to $h(x, y, z) = c$, so ∇g and ∇h are both orthogonal to C.

This means that the gradient vector $\nabla f(x_0, y_0, z_0)$ is in the plane determined by $\nabla g(x_0, y_0, z_0)$ and $\nabla h(x_0, y_0, z_0)$. (We assume that these gradient vectors are not zero and not parallel.)

Two Constraints

So there are numbers λ and μ (called Lagrange multipliers) such that

$$\boxed{\nabla f(x_0, y_0, z_0) = \lambda \nabla g(x_0, y_0, z_0) + \mu \nabla h(x_0, y_0, z_0)} \tag{16}$$

In this case Lagrange's method is to look for extreme values by solving five equations in the five unknowns *x*, *y*, *z*, λ, and μ.

Two Constraints

These equations are obtained by writing Equation 16 in terms of its components and using the constraint equations:

$$f_x = \lambda g_x + \mu h_x$$

$$f_y = \lambda g_y + \mu h_y$$

$$f_z = \lambda g_z + \mu h_z$$

$$g(x, y, z) = k$$

$$h(x, y, z) = c$$

Example 5

Find the maximum value of the function $f(x, y, z) = x + 2y + 3z$ on the curve of intersection of the plane $x - y + z = 1$ and the cylinder $x^2 + y^2 = 1$.

Solution:

We maximize the function $f(x, y, z) = x + 2y + 3z$ subject to the constraints $g(x, y, z) = x - y + z = 1$ and $h(x, y, z) = x^2 + y^2 = 1$.

Example 5 – *Solution*

The Lagrange condition is $\nabla f = \lambda \nabla g + \mu \nabla h$, so we solve the equations

$$\boxed{17} \qquad 1 = \lambda + 2x\mu$$

$$\boxed{18} \qquad 2 = -\lambda + 2y\mu$$

$$\boxed{19} \qquad 3 = \lambda$$

$$\boxed{20} \qquad x - y + z = 1$$

$$\boxed{21} \qquad x^2 + y^2 = 1$$

Putting $\lambda = 3$ [from (19)] in (17), we get $2x\mu = -2$, so $x = -1/\mu$. Similarly, (18) gives $y = 5/(2\mu)$.

Example 5 – *Solution*

Substitution in (21) then gives

$$\frac{1}{\mu^2} + \frac{25}{4\mu^2} = 1$$

and so $\mu^2 = \frac{29}{4}$, $\mu = \pm\sqrt{29}/2$.

Then $x = \mp 2/\sqrt{29}$, $y = \pm 5/\sqrt{29}$, and, from (20),
$z = 1 - x + y = 1 \pm 7/\sqrt{29}$.

Example 5 – *Solution*

The corresponding values of *f* are

$$\mp\frac{2}{\sqrt{29}} + 2\left(\pm\frac{5}{\sqrt{29}}\right) + 3\left(1 \pm \frac{7}{\sqrt{29}}\right) = 3 \pm \sqrt{29}$$

Therefore the maximum value of *f* on the given curve is $3 + \sqrt{29}$.

Taylor Series, one-dimensional case

Taylor's theorem (one-dimensional case): Suppose $f \epsilon C^n[a,b]$, f^{n+1} exists on $[a,b]$, and $x_0 \in [a,b]$. For every $x \in [a,b]$, there exists $\xi(x)$ between x_0 and x with

$$f(x) = P_n(x) + R_n(x).$$

where

$$P_n(x) = \sum_{k=0}^{n} \frac{f^k(x_o)}{k!}(x - x_o)^k$$

and

$$R_n(x) = \frac{f^{n+1}(\xi(x))}{(n+1)!}(x - x_o)^{n+1}$$

$P_n(x)$ is called the ***n*th Taylor polynormal** for f about x_o and $R_n(x)$ is called the **remainder term (or truncation error)**. When $n \to \infty$, $P_n(x)$ is called the **Taylor series** for for f about x_o

Taylor Series, one-dimensional case

When $\triangle x = |x_0 - x|$ is small, $f(x)$ can be approximated by $P_n(x)$. In other words,

$$f(x) \approx P_n(x) = \sum_{k=0}^{n} \frac{f^k(x_o)}{k!}(x - x_o)^k = \sum_{k=0}^{n} \frac{f^k(x_o)}{k!} \triangle x^k$$

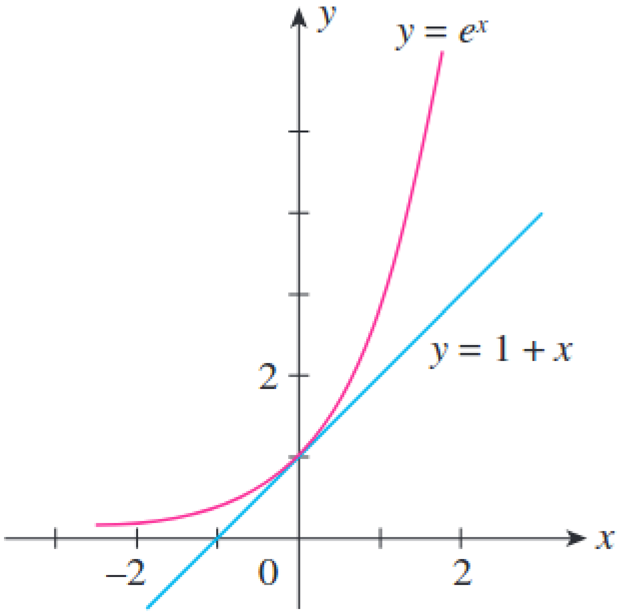
The first order (linear) approximation:

$$f(x) \approx f(x_o) + f'(x_o)(x - x_o)$$

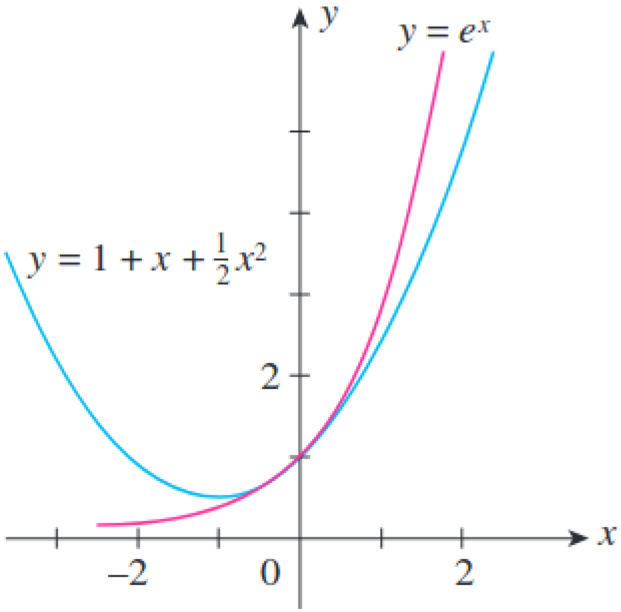
The second order (quadratic) approximation:

$$f(x) \approx f(x_o) + f'(x_o)(x - x_o) + \frac{1}{2} f''(x_o)(x - x_0)^2$$

Taylor Series, one-dimensional case



Linear approximation



Quadratic approximation

Image taken from Applied Calculus for the Managerial, Life and Social Sciences 8th ed

Taylor Series, high-dimensional case

For a multivariate function $f\colon \Re^n \rightarrow \Re$, the first order and second order **Taylor polynomials** can be expressed in terms of the gradient of f and Hessian matrix:

The first order:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x_o}) + (\boldsymbol{x} - \boldsymbol{x_o})^T \nabla f(\boldsymbol{x_o})$$

The second order:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x_o}) + (\boldsymbol{x} - \boldsymbol{x_o})^T \nabla f(\boldsymbol{x_o}) + \frac{1}{2!}(\boldsymbol{x} - \boldsymbol{x_0})^T \boldsymbol{H}(\boldsymbol{x_o})(\boldsymbol{x} - \boldsymbol{x_0})$$

where *H* is the Hessian matrix of *f and both* $\boldsymbol{x}$ and $\boldsymbol{x_0}$ are n-dimensional column vector.

Taylor Series, high-dimensional case

Hessian matrix

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Newton's Method

Now we want to determine $\triangle\, \boldsymbol{x} = \boldsymbol{x} - \boldsymbol{x_o}$ to minizine the equation

$$f(\boldsymbol{x_o}) + (\boldsymbol{x} - \boldsymbol{x_o})^T \nabla f(\boldsymbol{x_o}) + \frac{1}{2!}(\boldsymbol{x} - \boldsymbol{x_0})^T H(\boldsymbol{x_o})(\boldsymbol{x} - \boldsymbol{x_0}).$$

Rewriting the equation in term of $\triangle\, \boldsymbol{x}$, we have

$$\min_{\triangle \boldsymbol{x}} \left(f(\boldsymbol{x_o}) + \triangle\, \boldsymbol{x}^T\, \nabla f(\boldsymbol{x_o}) + \frac{1}{2!} \triangle\, \boldsymbol{x}^T\, H(\boldsymbol{x_o}) \triangle\, \boldsymbol{x} \right)$$

Using the matrix differentiation formulas,

$$\frac{\partial}{\partial \Delta x}\left(f(\boldsymbol{x_o}) + \triangle\, \boldsymbol{x}^T\, \nabla f(\boldsymbol{x_o}) + \frac{1}{2!} \triangle\, \boldsymbol{x}^T\, H(\boldsymbol{x_o}) \triangle\, \boldsymbol{x} \right) = \nabla f^T(\boldsymbol{x_o}) + \triangle\, \boldsymbol{x}^T\, H(\boldsymbol{x_o}) = 0$$

Thus,

$$\nabla f^T(\boldsymbol{x_o}) + (\boldsymbol{x} - \boldsymbol{x_o})^T H(\boldsymbol{x_o}) = 0$$

$$\boldsymbol{x}^T = \boldsymbol{x}_o^T - \nabla f^T(\boldsymbol{x_o}) H^{-1}(\boldsymbol{x_o})$$

$$\boldsymbol{x} = \boldsymbol{x}_o - \boldsymbol{H}^{-1}(\boldsymbol{x_o}) \nabla \boldsymbol{f}\ (\boldsymbol{x_o})$$

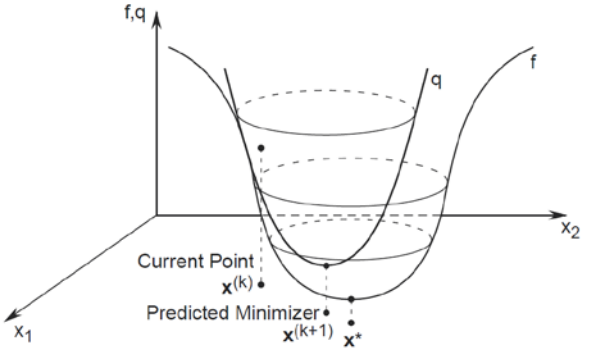


Image taken from: http://netlab.unist.ac.kr/wp-content/uploads/sites/183/2014/03/newton_method.png

Newton’s Method

In the implementation, we don’t compute H^{-1} because of computational cost. Rewriting, $\nabla f^T(\boldsymbol{x_o}) + \triangle \boldsymbol{x}^T H(\boldsymbol{x_o}) = 0$, we have

$$H(\boldsymbol{x_o}) \triangle \boldsymbol{x} = -\nabla f(\boldsymbol{x_o})$$

which is a linear system. Solving linear systems is much easier than compute the inverse.

Newton's Method

Objective: $\min f(\boldsymbol{x})$

Input: $\boldsymbol{x_0}$ and stopping criterion, e.g., max number of iterations, $f(\boldsymbol{x}) < \varepsilon$ or $\|\boldsymbol{x_t} - \boldsymbol{x_{t-1}}\| < \delta$.

Output t, $\boldsymbol{x_{t+1}}$ and $f(\boldsymbol{x_t})$

```
Step 1: Set $t = 0$
Step 2: while until the stopping criterion is fulfilled
        Step 3 Compute $H(\boldsymbol{x_t})$ and $\nabla f(\boldsymbol{x_t})$
        Step 4  Solve the $n \times n$ linear system $H(\boldsymbol{x_t}) \triangle \boldsymbol{x} = -\nabla f(\boldsymbol{x_t})$
        Step 5 Set $\boldsymbol{x_{t+1}} = \boldsymbol{x_t} + \triangle \boldsymbol{x}$
        Step 6 Step $\boldsymbol{t} = \boldsymbol{t} + \boldsymbol{1}$
Step 8: Output $t$, $\boldsymbol{x_{t+1}}$ and $f(\boldsymbol{x_t})$
```

Newton's Method

The Newton's method automatically controls the step size. It is different from the gradient descent method.

Tends to be extremely fragile unless function very smooth and starting close to minimum.

For large *n*, solving the linear system $H(\boldsymbol{x_o}) \triangle \boldsymbol{x} = -\nabla f(\boldsymbol{x_o})$ is still very costly.

Under some conditions, Newtons' method is quadratic local convergence. It means that $||\boldsymbol{x_{t+1}} - \boldsymbol{x}^*|| \leq \frac{L}{2m} ||\boldsymbol{x_t} - \boldsymbol{x}^*||^2$, where $\boldsymbol{x}^*$ is the solution of $\min f(\boldsymbol{x})$ and L and m are constant independent of $\boldsymbol{x}$ *(Theorem 5.2, page 8.4, Introduction to Nonlinear Optimization – Theory, Algorithms, and Application with MATLAB.)*

Gradient Descent

Gradient descent is based on the fact that ∇f indicates the direction of the fastest increase. Thus, $-\nabla f$ is the direction of the fastest decrease.

Gradient descent uses the equation below to search $\boldsymbol{x_t}$

$$\boldsymbol{x_{t+1}} = \boldsymbol{x_t} - \alpha \nabla f(\boldsymbol{x_t})$$

where $\alpha > 0$ is called learning rate in machine learning. As with the newton method, $\boldsymbol{x_o}, \alpha$ and stopping criterion are required input.

Gradient Descent

Consider $g(\alpha) = f(\boldsymbol{x_t} - \alpha\nabla f(\boldsymbol{x_t}))$ is a function of α. Using the first order Taylor polynomial to approximate it,

$$\mathrm{g}(\alpha) \approx \mathrm{g}(0) + \alpha \times \frac{dg(\alpha)}{d\alpha}$$

$$\mathrm{g}(\alpha) \approx f(\boldsymbol{x_t}) + \alpha \times \frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{u}} \times \frac{\partial \boldsymbol{u}}{\partial \alpha}, \quad where\ u = \boldsymbol{x_t} - \alpha\nabla f(\boldsymbol{x_t})$$

$$\mathrm{g}(\alpha) \approx f(\boldsymbol{x_t}) - \alpha\nabla f^T(\boldsymbol{x_t})\nabla f(\boldsymbol{x_t})$$

$$= f(\boldsymbol{x_t}) - \alpha\|\nabla f(\boldsymbol{x_t})\|^2 \leq f(\boldsymbol{x_t}).$$

In other words, for small α, the approximation would be accurate and $f(\boldsymbol{x_t} - \alpha\nabla f(\boldsymbol{x_t}))$ is likely smaller than $f(\boldsymbol{x_t})$. But for small α, more iterations are required to reach the optimal point.

Gradient Descent

How to select α?

- Constant learning rate α for all iterations
  - Good: simple.
  - Bad: If α is too large, it may not converge.
  - If α is too small, it would be very slow.
- Use a large α at the beginning and reduce it gradually. For example, for each k iterations, α is updated by α_{new}= $\alpha_{old} \times \beta$, where $0 < \beta < 1$. This approach uses a large α at the beginning.