Convex Optimization for Machine Learning

with Mathematica Applications

Chapter 6

Multi-Variable Optimization with Equality Constraints

M. M. Hammad
Department of Mathematics
Faculty of Science
Damanhour University
Egypt

Chapter 6

Multi-Variable Optimization with Equality Constraints

6.1 Constrained Optimization Problems

Applying constraints to a problem can affect the solution, but this need not be the case as shown in Figure 6.1.

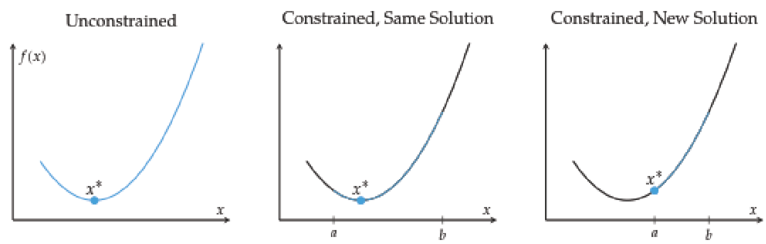


Figure 6.1.

The general nonlinear constrained optimization problems can be formulated as:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & h_1(\mathbf{x}) = 0, \quad g_1(\mathbf{x}) \le 0, \\ & \quad\vdots \qquad\qquad\quad \vdots \\ & h_m(\mathbf{x}) = 0, \quad g_p(\mathbf{x}) \le 0, \end{array} \tag{6.1.a}$$

where,

$$\mathbf{x} \in \mathbb{R}^n \tag{6.1.b}$$

$$f: \mathbb{R}^n \to \mathbb{R} \tag{6.1.c}$$

$$h_i: \mathbb{R}^n \to \mathbb{R}, \quad i = 1, \dots m \tag{6.1.d}$$

$$g_j: \mathbb{R}^n \to \mathbb{R}, \quad j = 1, \dots p \tag{6.1.e}$$

In vector notation, the problem above can be represented in the following standard form:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{h}(\mathbf{x}) = \mathbf{0} \\ & \mathbf{g}(\mathbf{x}) \le \mathbf{0}, \end{array} \tag{6.2.a}$$

where,

$$\mathbf{h}: \mathbb{R}^n \to \mathbb{R}^m, \quad \mathbf{h}(\mathbf{x}) = \begin{pmatrix} h_1(\mathbf{x}) \\ \vdots \\ h_m(\mathbf{x}) \end{pmatrix} \tag{6.2.b}$$

$$\mathbf{g}: \mathbb{R}^n \to \mathbb{R}^p, \quad \mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_p(\mathbf{x}) \end{pmatrix} \tag{6.2.c}$$

In this chapter, we consider only the optimization problems with equality constraints such that

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & h_1(\mathbf{x}) = 0 \\ & \quad\vdots \\ & h_m(\mathbf{x}) = 0 \end{array} \tag{6.3}$$

For solving this problem, there are several methods, e.g.

(1) Direct substitution method (2) Lagrange multiplier method.

6.2 Direct Substitution Method

For an optimization problem with n variables and m equality constraints, it is possible (theoretically) to express any set of m variables in terms of the remaining $(n-m)$ variables. When these expressions are substituted into the original objective function, then the reduced objective function involves only $n-m$ variables. This reduced objective function is not subjected to any constraint, so its optimum value can be found by using the unconstrained optimization techniques. Theoretically, the method of direct substitution is very simple. However, from a practical point of view, it is not convenient. In most of practical problems, the constraint equations are highly non-linear in nature. In those cases, it becomes impossible to solve them and express any m variables in terms of the remaining $(n-m)$ variables from the given constraints.

Example 6.1

Solve

$$\text{Minimize } z = 9 - 8x_1 - 6x_2 - 4x_3 + 2x_1^2 + 2x_2^2 + x_3^2 + 2x_1x_2 + 2x_1x_3$$
$$\text{subject to } x_1 + x_2 + 2x_3 = 3$$

Solution

From the given constraint, we have $x_2 = 3 - x_1 - 2x_3$, and substituting it in the objective function, we have

$$\begin{aligned} z &= 9 - 8x_1 - 6(3 - x_1 - 2x_3) - 4x_3 + 2x_1^2 + 2(3 - x_1 - 2x_3)^2 + x_3^2 + 2x_1(3 - x_1 - 2x_3) + 2x_1x_3 \\ &= 2x_1^2 + 9x_3^2 + 6x_1x_3 - 8x_1 - 16x_3 + 9 \end{aligned}$$

Now we have to optimize z with respect to the variables x_1 and x_3.

The necessary conditions for optimality of z are given by

$$\frac{\partial z}{\partial x_1} = 0 \text{ and } \frac{\partial z}{\partial x_3} = 0$$

or

$$\begin{aligned} 2x_1 + 3x_3 &= 4, \\ 3x_1 + 9x_3 &= 8 \end{aligned}$$

Solving these equations, we have $x_1 = \frac{4}{3}, x_3 = \frac{4}{9}$. Now,

$$x_2 = 3 - x_1 - 2x_3 = \frac{7}{9}$$

The Hessian matrix is given by

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 z}{\partial x_1^2} & \frac{\partial^2 z}{\partial x_1 \partial x_3} \\ \frac{\partial^2 z}{\partial x_3 \partial x_1} & \frac{\partial^2 z}{\partial x_3^2} \end{pmatrix} = \begin{pmatrix} 4 & 6 \\ 6 & 18 \end{pmatrix}$$

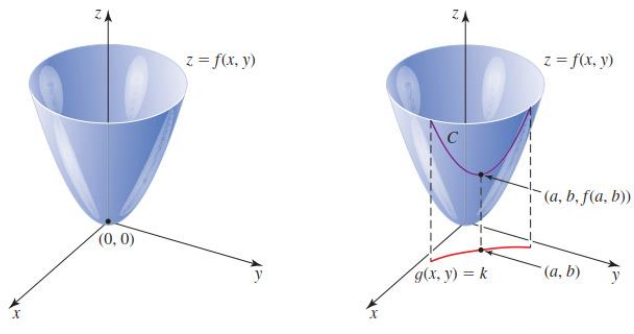
The leading principal minors are

$$H_1 = |4| = 4 \text{ and } H_2 = \begin{vmatrix} 4 & 6 \\ 6 & 18 \end{vmatrix} = 36.$$

Since $H_1 > 0$ and $H_2 > 0$, z is minimum for $x_1 = \frac{4}{3}, x_2 = \frac{7}{9}$ and $x_3 = \frac{4}{9}$ and the minimum value of z is $\frac{1}{9}$.

6.3 Lagrange Multipliers Illustration

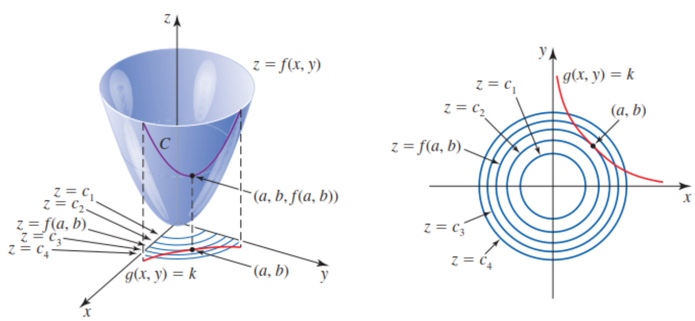
Figure 6.2a shows the graph of a function f defined by the equation $z = f(x,y)$. Observe that f has an absolute minimum at $(0,0)$ and an absolute minimum value of 0. However, if the independent variables x and y are subjected to a constraint of the form $g(x,y) = k$, then the points (x,y,z) that satisfy both $z = f(x,y)$ and $g(x,y) = k$ lie on the curve C, the intersection of the surface $z = f(x,y)$ and the $g(x,y) = k$ (Figure 6.2b). From the Figure 6.2b, you can see that the absolute minimum of f subject to the constraint $g(x,y) = k$ occurs at the point (a,b). Furthermore, f has the constrained absolute minimum value $f(a,b)$ rather than the unconstrained absolute minimum value of 0 at $(0,0)$.



(a) *f* is not subject to any constraints. **(b)** *f* is subject to a constraint.

Figure 6.2. The function has an unconstrained minimum value of 0, but it has a constrained minimum value of $f(a, b)$ when subjected to the constraint $g(x, y) = k$.

We will now consider a method, called the method of Lagrange multipliers. To see how this method works, let's reexamine the problem of finding the absolute minimum of the objective function f subject to the constraint $g(x, y) = k$. Figure 6.3a shows the level curves of f drawn in the xyz-coordinate system. These level curves are reproduced in the xy-plane in Figure 6.3b. Observe that the level curves of f with equations $f(x, y) = c$, where $c < f(a, b)$, have no points in common with the graph of the constraint equation $g(x, y) = k$ (for example, the level curves $f(x, y) = c_1$ and $f(x, y) = c_2$ shown in Figure 6.3b). Thus, points lying on these curves are not candidates for the constrained minimum of f.



(a) The level curves of *f* in the *xyz*-plane **(b)** The level curves of *f* in the *xy*-plane

Figure 6.3. The level curves of f in the xyz-plane.

On the other hand, the level curves of f with equation $f(x, y) = c$, where $c \geq f(a, b)$, do intersect the graph of the constraint equation $g(x, y) = k$ (such as the level curves of $f(x, y) = c_3$ and $f(x, y) = c_4$). These points of intersection are candidates for the constrained minimum of f.

Finally, observe that the larger c is for $c \geq f(a, b)$, the larger the value $f(x, y)$ is for (x, y) lying on the level curve $g(x, y) = k$. This observation suggests that we can find the constrained minimum of f by choosing the smallest value of c so that the level curve $f(x, y) = c$ still intersects the curve $g(x, y) = k$. At such a point (a, b) the level curve of f just touches the graph of the constraint equation $g(x, y) = k$. That is, the two curves have a common tangent at (a, b) (see Figure 6.3b). Equivalently, their normal lines at this point coincide. Putting it yet another way, the gradient vectors $\nabla f(a, b)$ and $\nabla g(a, b)$ have the same direction, so $\nabla f(a, b) = \lambda \nabla g(a, b)$ for some scalar λ. These geometric arguments suggest the following theorem.

Theorem 6.1: Let f and g have continuous first partial derivatives in some region D in the plane. If f has an extremum at a point (a, b) on the smooth constraint curve $g(x, y) = c$ lying in D and

$$\nabla g(a, b) \neq 0 \tag{6.4.a}$$

then there is a real number λ such that

$$\nabla f(a, b) = \lambda \nabla g(a, b) \tag{6.4.b}$$

The number λ is called a Lagrange multiplier.

Proof:

Suppose that the smooth curve C described by $g(x, y) = c$ is represented by the vector function

$$\mathbf{r}(t) = x(t)\,\mathbf{i} + y(t)\,\mathbf{j}, \qquad \mathbf{r}'(t) = x'(t)\,\mathbf{i} + y'(t)\,\mathbf{j} \neq 0,$$

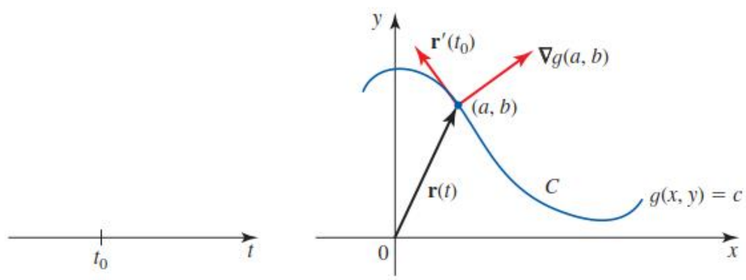
where x' and y' are continuous on an open interval I (Figure 6.4). Then the values assumed by f on C are given by

$$h(t) = f\big(x(t), y(t)\big).$$

Suppose that f has an extreme value at (a, b). If t_0 is the point in I corresponding to the point (a, b), then h has an extreme value at t_0. Therefore, $h'(t_0) = 0$. Using the Chain Rule, we have

$$\begin{aligned}\frac{d}{dt}h(t_0) &= h'(t_0)\\ &= \frac{\partial}{\partial x}f\big(x(t_0), y(t_0)\big)\frac{d}{dt}x(t_0) + \frac{\partial}{\partial y}f\big(x(t_0), y(t_0)\big)\frac{d}{dt}y(t_0)\\ &= f_x\big(x(t_0), y(t_0)\big)x'(t_0) + f_y\big(x(t_0), y(t_0)\big)y'(t_0)\\ &= f_x(a, b)x'(t_0) + f_y(a, b)y'(t_0)\\ &= \big(f_x(a, b)\,\mathbf{i} + f_y(a, b)\,\mathbf{j}\big)\cdot\big(x'(t_0)\,\mathbf{i} + y'(t_0)\,\mathbf{j}\big)\\ &= \langle \boldsymbol{\nabla} f(a, b) | \mathbf{r}'(t_0)\rangle = 0\end{aligned}$$

This shows that $\boldsymbol{\nabla} f(a, b)$ is orthogonal to $\mathbf{r}'(t_0)$. But, $\boldsymbol{\nabla} g(a, b)$ is orthogonal to $\mathbf{r}'(t_0)$. Therefore, the gradient vectors $\boldsymbol{\nabla} f(a, b)$ and $\boldsymbol{\nabla} g(a, b)$ are parallel, so there is a number λ such that $\boldsymbol{\nabla} f(a, b) = \lambda \boldsymbol{\nabla} g(a, b)$.



(a) The parameter interval I **(b)** The smooth curve C is represented by the vector function $\mathbf{r}(t)$.

Figure 6.4.

■

The Method of Lagrange Multipliers

Suppose f and g have continuous first partial derivatives. To find the maximum and minimum values of f subject to the constraint g (assuming that these extreme values exist and that $\boldsymbol{\nabla} g \neq 0$ on $g(x, y) = k$):

1. Solve the equations $\boldsymbol{\nabla} f(x, y) = \lambda \boldsymbol{\nabla} g(x, y)$ and $g(x, y) = k$ for x, y, and λ.
2. Evaluate f at each solution point found in Step 1. The largest value yields the constrained maximum of f, and the smallest value yields the constrained minimum of f.

Note that. Since

$$\boldsymbol{\nabla} f(x, y) = f_x(x, y)\,\mathbf{i} + f_y(x, y)\,\mathbf{j}, \qquad \boldsymbol{\nabla} g(x, y) = g_x(x, y)\,\mathbf{i} + g_y(x, y)\,\mathbf{j} \tag{6.5}$$

we see, by equating like components, that the vector equation

$$\boldsymbol{\nabla} f(x, y) = \lambda \boldsymbol{\nabla} g(x, y) \tag{6.6}$$

is equivalent to the two scalar equations

$$f_x(x, y) = \lambda g_x(x, y), \qquad f_y(x, y) = \lambda g_y(x, y) \tag{6.7}$$

These scalar equations together with the constraint equation $g(x, y) = k$ give a system of three equations to be solved for the three unknowns x, y, and λ.

Example 6.2

Find the maximum and minimum values of the function $f(x, y) = x^2 - 2y$ subject to $x^2 + y^2 = 9$.

Solution

The constraint equation is $g(x, y) = x^2 + y^2 - 9 = 0$. Since

$$\nabla f(x, y) = 2x\,\mathbf{i} - 2\,\mathbf{j}, \qquad \nabla g(x, y) = 2x\,\mathbf{i} + 2y\,\mathbf{j}$$

The equation $\nabla f(x, y) = \lambda \nabla g(x, y)$ becomes

$$2x\,\mathbf{i} - 2\,\mathbf{j} = \lambda(2x\,\mathbf{i} + 2y\,\mathbf{j}) = 2\lambda x\,\mathbf{i} + 2\lambda y\,\mathbf{j}$$

Equating like components and rewriting the constraint equation lead to the following system of three equations in the three variables x, y, and λ:

$$2x = 2\lambda x, \quad -2 = 2\lambda y, \quad x^2 + y^2 = 9$$

From the first equation we have

$$2x(1 - \lambda) = 0$$

so $x = 0$, or $\lambda = 1$. If $x = 0$, then the third equation gives $y = \pm 3$. If $\lambda = 1$, then the second equation gives $y = -1$, which upon substitution into the third equation yields $x = \pm 2\sqrt{2}$. Therefore, f has possible extreme values at the points $(0, -3)$, $(0, 3)$, $(-2\sqrt{2}, -1)$, and $(2\sqrt{2}, -1)$. Evaluating f at each of these points gives

$$f(0, -3) = 6, \qquad f(0,3) = -6, \qquad f\left(-2\sqrt{2}, -1\right) = 10 \quad \text{and} \quad f\left(2\sqrt{2}, -1\right) = 10$$

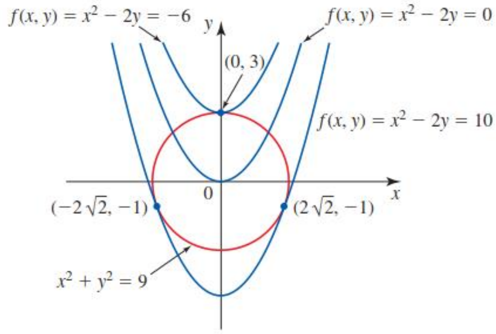


Figure 6.5. The extreme values of occur at the points where the level curves of are tangent to the graph of the constraint equation (the circle).

We conclude that the maximum value of f on the circle $x^2 + y^2 = 9$ is 10, attained at the points $\left(-2\sqrt{2}, -1\right)$ and $\left(2\sqrt{2}, -1\right)$, and that the minimum value of f on the circle is -6, attained at the point $(0,3)$. Figure 6.5 shows the graph of the constraint equation $x^2 + y^2 = 9$ and some level curves of the objective function f. Observe that the extreme values of f are attained at the points where the level curves of f are tangent to the graph of the constraint equation.

Optimizing a Function of three variables Subject to one Constraints

The proof of Lagrange's Theorem for functions of three variables is similar to that for functions of two variables.

Note Since

$$\nabla f(x, y, z) = f_x(x, y, z)\,\mathbf{i} + f_y(x, y, z)\,\mathbf{j} + f_z(x, y, z)\,\mathbf{k} \tag{6.8}$$

and

$$\nabla g(x, y, z) = g_x(x, y, z)\,\mathbf{i} + g_y(x, y, z)\,\mathbf{j} + g_z(x, y, z)\,\mathbf{k} \tag{6.9}$$

we see, by equating like components, that the vector equation

$$\nabla f(x, y, z) = \lambda \nabla g(x, y, z) \tag{6.10}$$

is equivalent to the three scalar equations

$$\begin{aligned} f_x(x, y, z) &= \lambda g_x(x, y, z) \\ f_y(x, y, z) &= \lambda g_y(x, y, z) \\ f_z(x, y, z) &= \lambda g_z(x, y, z) \end{aligned} \tag{6.11}$$

These scalar equations together with the constraint equation give a system of four equations to be solved for the four unknowns x, y, z, and λ.

Optimizing a Function Subject to Two Constraints

Some applications involve maximizing or minimizing an objective function f subject to two or more constraints. Consider, for example, the problem of finding the extreme values of $f(x,y,z)$ subject to the two constraints

$$g(x,y,z)=k \quad \text{and} \quad h(x,y,z)=l \tag{6.12}$$

It can be shown that if f has an extremum at (a,b,c) subject to these constraints, then there are real numbers (Lagrange multipliers) λ and μ such that

$$\nabla f(a,b,c)=\lambda\nabla g(a,b,c)+\mu\nabla h(a,b,c) \tag{6.13}$$

Geometrically, we are looking for the extreme values of $f(x,y,z)$ on the curve of intersection of the level surfaces $g(x,y,z)=k$ and $h(x,y,z)=l$. Condition (6.13) is a statement that at an extremum point (a,b,c), the gradient of f must lie in the plane determined by the gradient of g and the gradient of h. (See Figure 6.6.) The vector (6.13) is equivalent to three scalar equations. When combined with the two constraint equations, this leads to a system of five equations that can be solved for the five unknowns x, y, z, λ, and μ.

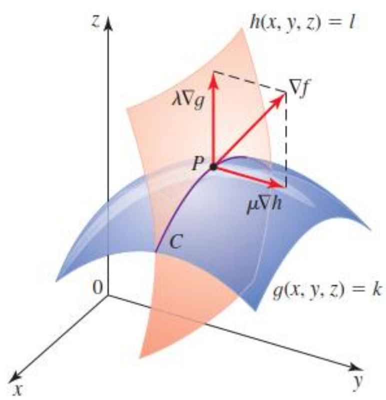


Figure 6.6. If f has an extreme value at $P(a,b,c)$, then $\nabla f(a,b,c)=\lambda\nabla g(a,b,c)+\mu\nabla h(a,b,c)$.

Example 6.3

Find the maximum and minimum values of the function $f(x,y,z)=3x+2y+4z$ subject to the constraints $x-y+2z=1$ an $x^2+y^2=4$.

Solution

Write the constraint equations in the form

$$g(x,y,z)=x-y+2z-1=0$$
$$h(x,y,z)=x^2+y^2-4=0$$

Then the equation $\nabla f(x,y,z)=\lambda\nabla g(x,y,z)+\mu\nabla h(x,y,z)$ becomes

$$3\,\mathbf{i}+2\,\mathbf{j}+4\,\mathbf{k}=\lambda(\,\mathbf{i}-\ \mathbf{j}+2\,\mathbf{k})+\mu(2x\,\mathbf{i}+2y\,\mathbf{j})=(\lambda+2\mu x)\,\mathbf{i}-(\lambda-2\mu y)\,\mathbf{j}+2\lambda\,\mathbf{k}$$

Equating like components and rewriting the constraint equations lead to the following system of five equations in the five variables, x, y, z, λ, and μ:

$$\begin{aligned}\lambda+2\mu x&=3\\-\lambda+2\mu y&=2\\2\lambda&=4\\x-y+2z&=1\\x^2+y^2&=4\end{aligned}$$

From the third equation, we have $\lambda=2$. Then

$$\begin{aligned}2+2\mu x&=3 &&\Rightarrow 2\mu x=1\\-2+2\mu y&=2 &&\Rightarrow 2\mu y=4\end{aligned}$$

The solution of last two equations is

$$x=\frac{1}{2\mu},\ \ y=\frac{2}{\mu}$$

Substituting these values of x and y into $x^2 + y^2 = 4$ yields

$$\left(\frac{1}{2\mu}\right)^2 + \left(\frac{2}{\mu}\right)^2 = 4 \quad \Rightarrow \mu^2 = \frac{17}{16}$$

Therefore,

$$\mu = \pm\frac{\sqrt{17}}{4}$$

So that

$$x = \pm\frac{2}{\sqrt{17}}, \qquad y = \pm\frac{8}{\sqrt{17}}$$

and the values of z becomes $z = \frac{1}{2}\left(1 \pm \frac{6}{\sqrt{17}}\right)$. The value of f at the point $\left(\frac{2}{\sqrt{17}}, \frac{8}{\sqrt{17}}, \frac{1}{2}\left(1 + \frac{6}{\sqrt{17}}\right)\right)$ is $2(1+\sqrt{17})$, and the value of f at the point $\left(-\frac{2}{\sqrt{17}}, -\frac{8}{\sqrt{17}}, \frac{1}{2}\left(1 - \frac{6}{\sqrt{17}}\right)\right)$ is $2(1-\sqrt{17})$. Therefore, the maximum value of f is $2(1+\sqrt{17})$ and the minimum value of f is $2(1-\sqrt{17})$.

6.4 Lagrange Multipliers General Case

The optimization problems with equality constraints can be formulated as:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & h_1(\mathbf{x}) = 0 \\ & \quad\vdots \\ & h_m(\mathbf{x}) = 0 \end{array} \tag{6.14}$$

This set of constraints defines a hypersurface in $\mathbb{R}^n$. Each function h_i defines a surface $S_i = \{\mathbf{x} | h_i(\mathbf{x}) = 0\}$ of generally $n-1$ dimensions in the space $\mathbb{R}^n$. This surface is smooth provided $h_i(\mathbf{x}) \in C^1$. The m constraints together defines a surface S, which is the intersection of the surfaces $S_1, \ldots, S_m$; namely,

$$S = \{\mathbf{x} | h_1(\mathbf{x}) = 0\} \cap \cdots \cap \{\mathbf{x} | h_m(\mathbf{x}) = 0\}. \tag{6.15}$$

Using vector notation, we have

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{h}(\mathbf{x}) = \mathbf{0}, \end{array} \tag{6.16}$$

The function $\mathbf{h}$ takes a point $\mathbf{x} \in \mathbb{R}^n$ as input and produces the vector $\mathbf{h}(\mathbf{x}) \in \mathbb{R}^m$ as output. We assume that the function $\mathbf{h}$ is continuously differentiable, that is, $\mathbf{h} \in C^1$. Hence, we have

$$S = \{\mathbf{x} | \mathbf{h}(\mathbf{x}) = \mathbf{0}\} \tag{6.17}$$

Definition (feasible set): Any point satisfying the constraints is called a feasible point. The set of all feasible points

$$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{h}(\mathbf{x}) = \mathbf{0}\} \tag{6.18}$$

is called the feasible set.

The Jacobian matrix of a vector-valued function of several variables is the matrix of all its first-order partial derivatives. Then the Jacobian matrix of $\mathbf{h}$ is defined to be an $m \times n$ matrix, denoted by $\mathbf{J}$, whose (i,j)th entry is $J_{ij} = \frac{\partial h_i}{\partial x_j}$, or explicitly

$$\mathbf{J} = \begin{pmatrix} \frac{\partial \mathbf{h}}{\partial x_1} & \cdots & \frac{\partial \mathbf{h}}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \frac{\partial h_1}{\partial x_1} & \cdots & \frac{\partial h_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_m}{\partial x_1} & \cdots & \frac{\partial h_m}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \nabla h_1{}^T \\ \vdots \\ \nabla h_m{}^T \end{pmatrix} \tag{6.19}$$

where $\nabla h_i{}^T$ is the transpose (row vector) of the gradient of the i component. The Jacobian matrix, whose entries are functions of $\mathbf{x}$, is denoted in various ways; common notations include $D\mathbf{h}$, $\mathbf{J}_\mathbf{h}$, and $\nabla\mathbf{h}$. Some authors define the Jacobian as the transpose of the form given above.

The Jacobian of a vector-valued function in several variables generalizes the gradient of a scalar-valued function in several variables, which in turn generalizes the derivative of a scalar-valued function of a single variable.

If $\mathbf{h}$ is differentiable at a point $\mathbf{p}$ in $\mathbb{R}^n$, then its differential is represented by $\mathbf{J_h}(\mathbf{p})$. In this case, the linear transformation represented by $\mathbf{J_h}(\mathbf{p})$ is the best linear approximation of $\mathbf{h}$ near the point $\mathbf{p}$, in the sense that

$$\mathbf{h}(\mathbf{x}) \approx \mathbf{h}(\mathbf{p}) + \mathbf{J_h}(\mathbf{p})(\mathbf{x}-\mathbf{p}) + \boldsymbol{O}(\|\mathbf{x}-\mathbf{p}\|) \tag{6.20}$$

where $\boldsymbol{O}(\|\mathbf{x}-\mathbf{p}\|)$ is a quantity that approaches zero much faster than the distance between $\mathbf{x}$ and $\mathbf{p}$ does as $\mathbf{x}$ approaches $\mathbf{p}$. This approximation specializes to the approximation of a scalar function of a single variable by its Taylor polynomial of degree one, namely

$$h(x) \approx h(p) + h'(p)(x-p) + O(x-p) \tag{6.21}$$

In this sense, the Jacobian may be regarded as a kind of "first-order derivative" of a vector-valued function of several variables. In particular, this means that the gradient of a scalar-valued function of several variables may too be regarded as its "first-order derivative".

For one constraint problem, consider the case where $\nabla h(\mathbf{x}^*) = 0$, or in other words, the point which minimizes $f(\mathbf{x})$ is also a critical point of $h(\mathbf{x})$. Remember our necessary condition for a minimum is $\nabla f(\mathbf{x}^*) = \lambda \nabla h(\mathbf{x}^*)$. Since $\nabla h(\mathbf{x}^*) = 0$, this implies that $\nabla f(\mathbf{x}^*) = 0$. However, this the necessary condition for an unconstrained optimization problem, not a constrained one! In effect, when $\nabla h(\mathbf{x}^*) = 0$, the constraint is no longer considered in the problem. Hence, if any of the $\nabla h_i(\mathbf{x}^*)$ is zero, then that constraint will not be considered in the analysis. Also, there will be a row of zeros in the Jacobian, so the Jacobian will not be full rank. The generalization of the condition that $\nabla h(\mathbf{x}^*) \neq 0$ for the case when $m = 1$ is that the Jacobian matrix must be full rank.

Definition (regular point): A point $\mathbf{x}^*$ satisfying the constraints $h_1(\mathbf{x}^*) = 0, \ldots, h_m(\mathbf{x}^*) = 0$ is said to be a regular point of the constraints if the gradient vectors $\nabla h_1(\mathbf{x}^*), \ldots, \nabla h_m(\mathbf{x}^*)$ are linearly independent.

Definition (non-degenerate constraint qualification): Let $D\mathbf{h}(\mathbf{x}^*)$ be the Jacobian matrix of $\mathbf{h} = [h_1, \ldots, h_m]^T$ at $\mathbf{x}^*$, given by

$$D\mathbf{h}(\mathbf{x}^*) = \begin{pmatrix} Dh_1(\mathbf{x}^*) \\ \vdots \\ Dh_m(\mathbf{x}^*) \end{pmatrix} = \begin{pmatrix} \nabla h_1(\mathbf{x}^*)^T \\ \vdots \\ \nabla h_m(\mathbf{x}^*)^T \end{pmatrix}. \tag{6.22}$$

Then, $\mathbf{x}^*$ is regular if and only if rank $D\mathbf{h}(\mathbf{x}^*) = m$, that is, the Jacobian matrix is of full rank.

- The definition states, in effect, that $\mathbf{x}^*$ is a regular point of the constraints if it is a solution of $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ and the Jacobian $D\mathbf{h}(\mathbf{x}^*)$ has full row rank. The importance of a point $\mathbf{x}^*$ being regular for a given set of equality constraints lies in the fact that a tangent plane of the hypersurface determined by the constraints at a regular point $\mathbf{x}^*$ is well defined. Later in this chapter, the term 'tangent plane' will be used to express and describe important necessary as well as sufficient conditions for constrained optimization problems.
- Since $D\mathbf{h}(\mathbf{x})$ is a $m \times n$ matrix, it would not be possible for $\mathbf{x}$ to be a regular point of the constraints if $m > n$. This leads to an upper bound for the number of independent equality constraints, i.e., $m \leq n$. Furthermore, if $m = n$, in many cases the number of vectors $\mathbf{x}$ that satisfy $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ is finite and the optimization problem becomes a trivial one. For these reasons, we shall assume that $m < n$ throughout the rest of the book.

From linear algebra remember the following facts:

- Given a linear map $L: V \to W$ between two vector spaces V and W, the kernel of L (null space) is the vector subspace of all elements $\mathbf{v}$ of V such that $L(\mathbf{v}) = \mathbf{0}$, where $\mathbf{0}$ denotes the zero vector in W.
- Consider a linear map represented as a $m \times n$ matrix $\mathbf{A}$ with coefficients in $\mathbb{R}$, that is operating on column vectors $\mathbf{x}$ with n components over $\mathbb{R}$. The kernel of this linear map is the set of solutions to the equation $\mathbf{Ax} = \mathbf{0}$, where $\mathbf{0}$ is understood as the zero vector. The dimension of the kernel of $\mathbf{A}$ is called the nullity of $\mathbf{A}$.
$$\mathrm{Null}(\mathbf{A}) = \mathrm{Ker}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{Ax} = \mathbf{0}\}$$
- The product $\mathbf{Ax}$ can be written in terms of the dot product of vectors as follows:
$$\mathbf{Ax} = \begin{pmatrix} \langle \mathbf{a}_1 | \mathbf{x} \rangle \\ \vdots \\ \langle \mathbf{a}_m | \mathbf{x} \rangle \end{pmatrix}$$

Here, $\mathbf{a}_1, \ldots, \mathbf{a}_m$ denote the rows of the matrix $\mathbf{A}$. It follows that $\mathbf{x}$ is in the kernel of $\mathbf{A}$, if and only if $\mathbf{x}$ is orthogonal to each of the row vectors of $\mathbf{A}$.

- The row space of a matrix $\mathbf{A}$ is the span of the row vectors of $\mathbf{A}$. The dimension of the row space of $\mathbf{A}$ is called the rank of $\mathbf{A}$, and the dimension of the kernel of $\mathbf{A}$ is called the nullity of $\mathbf{A}$. These quantities are related by the rank–nullity theorem

$$\text{nullity}(\mathbf{A}) + \text{rank}(\mathbf{A}) = n$$

- So that the rank–nullity theorem can be restated as (See Figure 6.7)

$$\dim(\text{Ker } L) + \dim(\text{Im } L) = \dim(V)$$

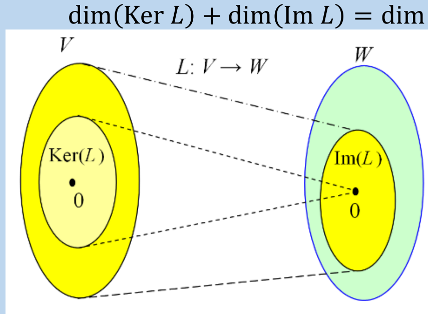


Figure 6.7.

Definition (dimension of the surface S): The set of equality constraints $h_1(\mathbf{x}) = 0, \ldots, h_m(\mathbf{x}) = 0$, $h_i: \mathbb{R}^n \to \mathbb{R}$, describes a surface

$$S = \{\mathbf{x} \in \mathbb{R}^n : h_1(\mathbf{x}) = 0, \ldots, h_m(\mathbf{x}) = 0\}. \tag{6.23}$$

Assuming the points in S are regular, the dimension of the surface S is $n - m$.

Example 6.4

Let $n = 3$ and $m = 1$ (i.e., we are operating in $\mathbb{R}^3$). Assuming that all points in S are regular, the set S is a two-dimensional surface.

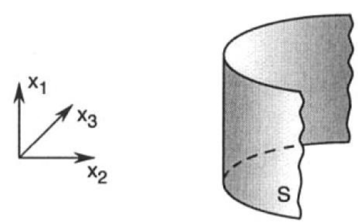
For example, let

$$h_1(\mathbf{x}) = x_2 - x_3^2 = 0.$$

Note that $\nabla h_1(\mathbf{x}) = (0,1,-2x_3)^T$, and hence for any $\mathbf{x} \in \mathbb{R}^3$, $\nabla h_1(\mathbf{x}) \neq \mathbf{0}$. In this case,

$$\dim S = \dim\{\mathbf{x}: h_1(\mathbf{x}) = 0\} = n - m = 2.$$

See Figure 6.8 for a graphical illustration.



$S=\{[x_1,x_2,x_3]^T : x_2 - x_3^2 = 0\}$

Figure 6.8. A two-dimensional surface in $\mathbb{R}^3$.

Example 6.5

Let $n = 3$ and $m = 2$. Assuming regularity, the feasible set S is a one-dimensional object (i.e., a curve in $\mathbb{R}^3$).

For example, let

$$h_1(\mathbf{x}) = x_1 = 0,$$
$$h_2(\mathbf{x}) = x_2 - x_3^2 = 0.$$

In this case, $\nabla h_1(\mathbf{x}) = (1,0,0)^T$, and $\nabla h_2(\mathbf{x}) = (0,1,-2x_3)^T$. Hence, the vectors $\nabla h_1(\mathbf{x})$ and $\nabla h_2(\mathbf{x})$ are linearly independent in $\mathbb{R}^3$. Thus,

$$\dim S = \dim\{\mathbf{x}: h_1(\mathbf{x}) = 0, h_2(\mathbf{x}) = 0\} = n - m = 1.$$

See Figure 6.9 for a graphical illustration.

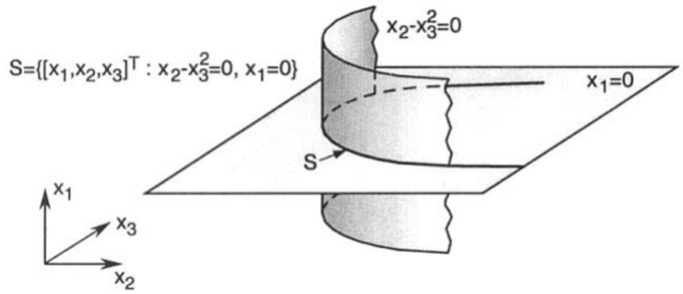


Figure 6.9. A one-dimensional surface in $\mathbb{R}^3$.

Example 6.6

Discuss and sketch the feasible region described by the equality constraints

$$h_1(\mathbf{x}) = -x_1 + x_3 - 1 = 0$$
$$h_2(\mathbf{x}) = x_1^2 + x_2^2 - 2x_1 = 0$$

Solution

The Jacobian of the constraints is given by

$$\mathbf{J_h} = \begin{pmatrix} -1 & 0 & 1 \\ 2x_1 - 2 & 2x_2 & 0 \end{pmatrix}$$

which has rank 2 except at $\mathbf{x} = (1 \quad 0 \quad x_3)^T$, where $\nabla h_2(1 \quad 0 \quad x_3)^T = (0,0,0)^T$. Since $\mathbf{x} = (1 \quad 0 \quad x_3)^T$ does not satisfy the constraint $h_2(\mathbf{x})$, any point $\mathbf{x}$ satisfying $h_1(\mathbf{x}) = 0$ and $h_2(\mathbf{x}) = 0$ is regular. The constraints $h_1(\mathbf{x}) = 0$ and $h_2(\mathbf{x}) = 0$ describe a curve which is the intersection between the cylinder, $h_2(\mathbf{x}) = 0$, and the plane, $h_1(\mathbf{x}) = 0$. To display the curve, we derive a parametric representation for the curve as follows. $h_2(\mathbf{x}) = 0$ can be written as

$$(x_1 - 1)^2 + x_2^2 = 1$$

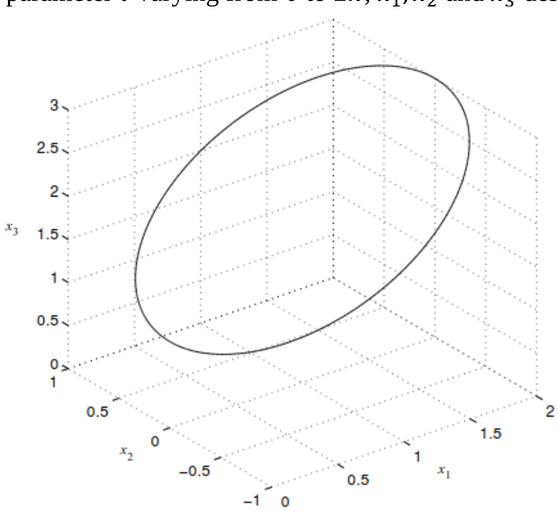
which suggests the parametric expressions

$$x_1 = 1 + \cos t$$
$$x_2 = \sin t$$

for x_1 and x_2. Now x_1 and x_2 in conjunction with $h_1(\mathbf{x}) = 0$ give

$$x_3 = 2 + \cos t$$

With parameter t varying from 0 to 2π, x_1, x_2 and x_3 describe the curve shown in Figure 6.10.



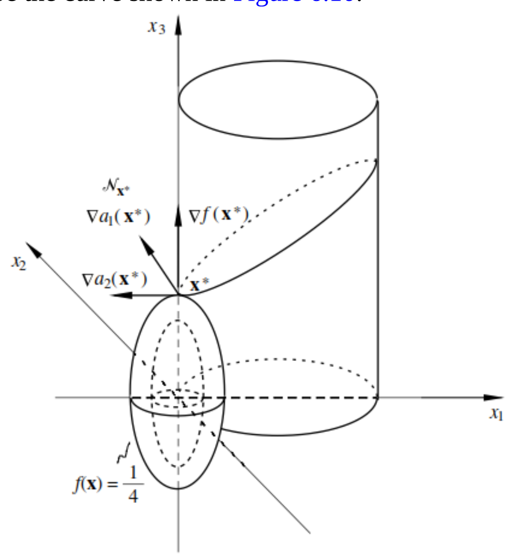


Figure 6.10.

Definition (curve C on a surface S): A curve C on a surface S is a set of points $\{\mathbf{x}(t) \in S: t \in (a,b)\}$, continuously parameterized by $t \in (a,b)$; that is, $\mathbf{x}: (a,b) \to S$ is a continuous function.

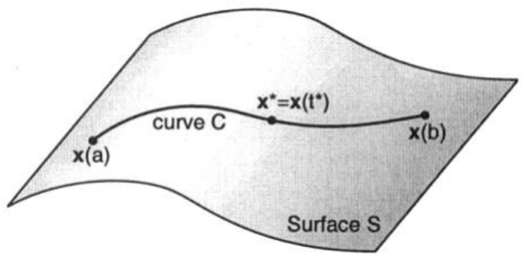


Figure 6.11. A curve on a surface.

A graphical illustration of the definition of a curve is given in Figure 6.11. The definition of a curve implies that all the points on the curve satisfy the equation describing the surface. The curve C passes through a point $\mathbf{x}^*$ if there exists $t^* \in (a,b)$ such that $\mathbf{x}(t^*) = \mathbf{x}^*$. Intuitively, we can think of a curve $C = \{\mathbf{x}(t): t \in (a,b)\}$ as the path traversed by a point $\mathbf{x}$ traveling on the surface S. The position of the point at time t is given by $\mathbf{x}(t)$.

Definition (differentiable and twice differentiable curve): The curve $C = \{\mathbf{x}(t): t \in (a,b)\}$ is differentiable if

$$\dot{\mathbf{x}}(t) = \frac{d\mathbf{x}}{dt}(t) = \begin{pmatrix} \dot{x}_1(t) \\ \vdots \\ \dot{x}_n(t) \end{pmatrix} \tag{6.24}$$

exists for all $t \in (a,b)$.
The curve $C = \{\mathbf{x}(t): t \in (a,b)\}$ is twice differentiable if

$$\ddot{\mathbf{x}}(t) = \frac{d^2\mathbf{x}}{dt^2}(t) = \begin{pmatrix} \ddot{x}_1(t) \\ \vdots \\ \ddot{x}_n(t) \end{pmatrix} \tag{6.25}$$

exists for all $t \in (a,b)$.

Note that both $\dot{\mathbf{x}}(t)$ and $\ddot{\mathbf{x}}(t)$ are n-dimensional vectors. We can think of $\dot{\mathbf{x}}(t)$ and $\ddot{\mathbf{x}}(t)$ as the "velocity" and "acceleration," respectively, of a point traversing the curve C with position $\mathbf{x}(t)$ at time t. The vector $\dot{\mathbf{x}}(t)$ points in the direction of the instantaneous motion of $\mathbf{x}(t)$. Therefore, the vector $\dot{\mathbf{x}}(t^*)$ is tangent to the curve C at $\mathbf{x}^*$ (see Figure 6.12).

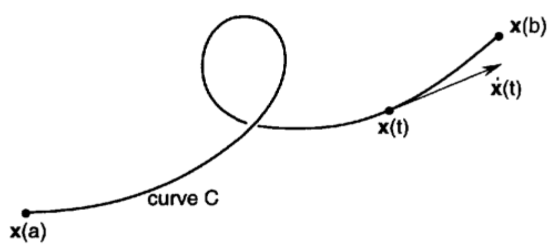


Figure 6.12. Geometric interpretation of the differentiability of a curve.

We are now ready to introduce the notions of a tangent space. For this, recall the set

$$S = \{\mathbf{x} \in \mathbb{R}^n: \mathbf{h}(\mathbf{x}) = \mathbf{0}\}, \tag{6.26}$$

where $\mathbf{h} \in C^1$. We think of S as a surface in $\mathbb{R}^n$. Now consider all differentiable curves on S passing through a point $\mathbf{x}^*$. The tangent plane at $\mathbf{x}^*$ is defined as the collection of the derivatives at $\mathbf{x}^*$ of all these differentiable curves. Figure 6.13 illustrates the notion of a tangent plane.

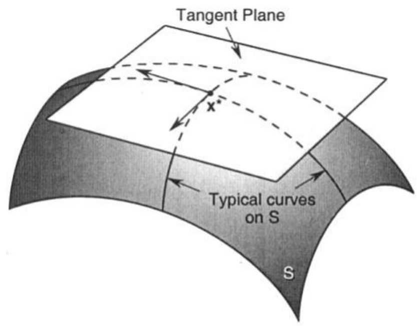


Figure 6.13. The tangent plane to the surface S at the point $\mathbf{x}^*$.

The tangent plane of a smooth function $f(\mathbf{x})$ at a given point $\mathbf{x}^*$ can be defined as a hyperplane that passes through point $\mathbf{x}^*$ with $\nabla f(\mathbf{x}^*)$ as the normal. For example, for $n = 2$ the contours, tangent plane, and gradient of a smooth function are related to each other as illustrated in Figure 6.14. Following this idea, the tangent plane at point $\mathbf{x}^*$ can be defined analytically as the set

$$T(\mathbf{x}^*) = \{\mathbf{x}: \langle \nabla f(\mathbf{x}^*) | \mathbf{x} - \mathbf{x}^* \rangle = \mathbf{0}\}. \tag{6.27}$$

In other words, a point $\mathbf{x}$ lies on the tangent plane if the vector that connects $\mathbf{x}^*$ to $\mathbf{x}$ is orthogonal to the gradient $\nabla f(\mathbf{x}^*)$, as can be seen in Figure 6.14.

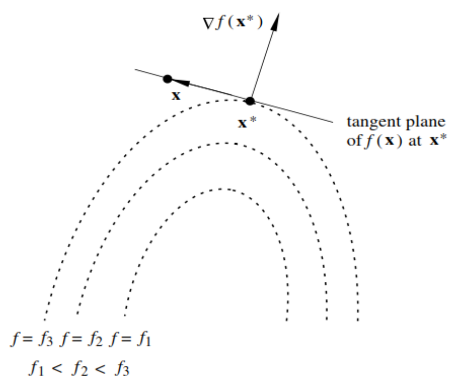


Figure 6.14.

Ideally, we would like to express this tangent plane in terms of derivatives of constraint functions h_i that define the surface. By using the Taylor series of constraint function $h_i(\mathbf{x})$ at the feasible point $\mathbf{x}^*$, we can write

$$\begin{aligned} h_i(\mathbf{x}^* + \mathbf{s}) &= h_i(\mathbf{x}^*) + \langle \nabla h_i\, (\mathbf{x}^*) | \mathbf{s} \rangle + O(\|\mathbf{s}\|) \\ &= \langle \nabla h_i\, (\mathbf{x}^*) | \mathbf{s} \rangle + O(\|\mathbf{s}\|) \end{aligned} \tag{6.28}$$

since $h_i(\mathbf{x}^*) = 0$. It follows that $h_i(\mathbf{x}^* + \mathbf{s}) = 0$ is equivalent to

$$\langle \nabla h_i\, (\mathbf{x}^*) | \mathbf{s} \rangle = 0 \quad \text{for } \; i = 1, \ldots, m \tag{6.29}$$

In other words, $\mathbf{s}$ is feasible if and only if it is orthogonal to the gradients of the constraint functions.

Definition (tangent space): The tangent space at a point $\mathbf{x}^*$ on the surface $S = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$ is the set

$$T(\mathbf{x}^*) = \{\mathbf{y}: D\mathbf{h}(\mathbf{x}^*)\mathbf{y} = \mathbf{0}\}. \tag{6.30}$$

Notes:

1- The tangent space $T(\mathbf{x}^*)$ is the null space of the matrix $D\mathbf{h}(\mathbf{x}^*)$, that is,

$$T(\mathbf{x}^*) = \mathcal{N}\big(D\mathbf{h}(\mathbf{x}^*)\big). \tag{6.31}$$

The tangent space is therefore a subspace of $\mathbb{R}^n$.

2- Assuming $\mathbf{x}^*$ is regular (the Jacobian matrix is of full rank), the dimension of the tangent space is $n - m$, where m is the number of equality constraints $h_i(\mathbf{x}^*) = 0$.

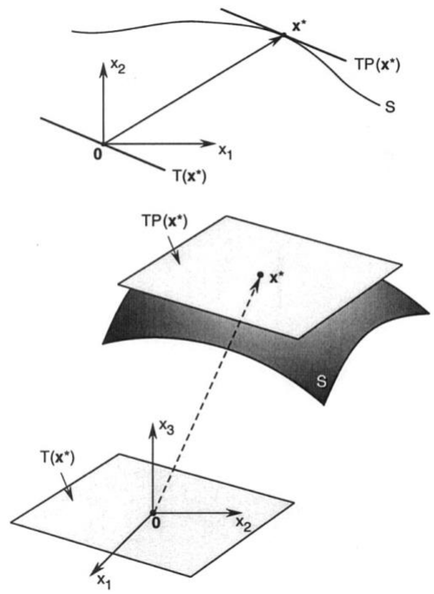


Figure 6.15. Tangent spaces and planes in $\mathbb{R}^2$ and $\mathbb{R}^3$.

3- The tangent space passes through the origin. However, it is often convenient to picture the tangent space as a plane that passes through the point $\mathbf{x}^*$. For this, we define the tangent plane at $\mathbf{x}^*$ to be the set

$$\mathrm{TP}(\mathbf{x}^*) = T(\mathbf{x}^*) + \mathbf{x}^* = \{\mathbf{x} + \mathbf{x}^* : \mathbf{x} \in T(\mathbf{x}^*)\}. \tag{6.32}$$

Figure 6.15 illustrates the relationship between the tangent plane and the tangent space.

Example 6.7

Let

$$S = \{\mathbf{x} \in \mathbb{R}^3 : h_1(\mathbf{x}) = x_1 = 0, h_2(\mathbf{x}) = x_1 - x_2 = 0\}.$$

Solution

Then, S is the x_3-axis in $\mathbb{R}^3$ (see Figure 6.16). We have

$$D\mathbf{h}(\mathbf{x}) = \begin{pmatrix} \nabla h_1(\mathbf{x})^T \\ \nabla h_2(\mathbf{x})^T \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & -1 & 0 \end{pmatrix}.$$

Because ∇h_1 and ∇h_2 are linearly independent when evaluated at any $\mathbf{x} \in S$, all the points of S are regular. The tangent space at an arbitrary point of S is

$$\begin{aligned} T(\mathbf{x}) &= \{\mathbf{y} : \langle \nabla h_1(\mathbf{x}), \mathbf{y} \rangle = 0, \langle \nabla h_2(\mathbf{x}), \mathbf{y} \rangle = 0\} \\ &= \left\{ \mathbf{y} : \begin{pmatrix} 1 & 0 & 0 \\ 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \mathbf{0} \right\} \\ &= \{(0,0,\alpha)^T : \alpha \in \mathbb{R}\} \\ &= \text{the } x_3\text{-axis in } \mathbb{R}^3. \end{aligned}$$

In this example, the tangent space $T(\mathbf{x})$ at any point $\mathbf{x} \in S$ is a one-dimensional subspace of $\mathbb{R}^3$.

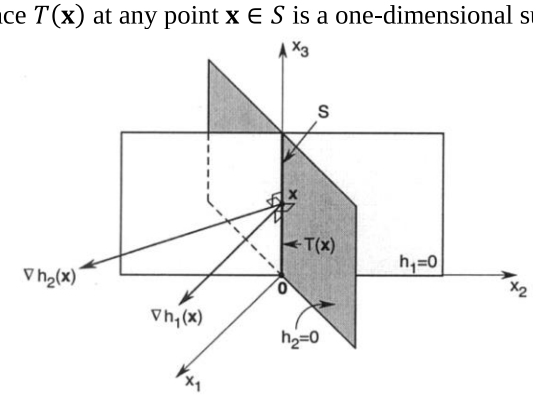


Figure 6.16. $S = \{\mathbf{x} \in \mathbb{R}^3 : h_1(\mathbf{x}) = x_1 = 0, h_2(\mathbf{x}) = x_1 - x_2 = 0\}$.

Intuitively, we would expect the definition of the tangent space at a point on a surface to be the collection of all "tangent vectors" to the surface at that point. We have seen that the derivative of a curve on a surface at a point is a tangent vector to the curve, and hence to the surface. The above intuition agrees with our definition whenever $\mathbf{x}^*$ is regular, as stated in the theorem below.

Theorem 6.2: Suppose $\mathbf{x}^* \in S$ is a regular point, and $T(\mathbf{x}^*)$ is the tangent space at $\mathbf{x}^*$. Then, $\mathbf{y} \in T(\mathbf{x}^*)$ if and only if there exists a differentiable curve in S passing through $\mathbf{x}^*$ with derivative $\mathbf{y}$ at $\mathbf{x}^*$.

Proof:

$\Leftarrow$: Suppose there exists a curve $\{\mathbf{x}(t): t \in (a,b)\}$ in S such that $\mathbf{x}(t^*) = \mathbf{x}^*$ and $\dot{\mathbf{x}}(t^*) = \mathbf{y}$ for some $t^* \in (a,b)$. Then,

$$\mathbf{h}\big(\mathbf{x}(t)\big) = \mathbf{0}$$

for all $t \in (a,b)$. If we differentiate the function $\mathbf{h}\big(\mathbf{x}(t)\big)$ with respect to t using the chain rule, we obtain

$$\frac{d}{dt}\mathbf{h}\big(\mathbf{x}(t)\big) = D\mathbf{h}\big(\mathbf{x}(t)\big)\dot{\mathbf{x}}(t) = \mathbf{0}$$

for all $t \in (a,b)$. Therefore, at t^*, we get

$$D\mathbf{h}(\mathbf{x}^*)\mathbf{y} = \mathbf{0},$$

and hence $\mathbf{y} \in T(\mathbf{x}^*)$.

$\Rightarrow$: we leave the opposite direction of the theorem as exercise.

■

We now introduce the notion of a normal space.

Definition (normal space): The normal space $N(\mathbf{x}^*)$ at a point $\mathbf{x}^*$ on the surface $S = \{\mathbf{x} \in \mathbb{R}^n: \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$ is the subspace of $\mathbb{R}^n$ spanned by the vectors $\nabla h_1(\mathbf{x}^*), \dots, \nabla h_m(\mathbf{x}^*)$, that is,

$$\begin{aligned} N(\mathbf{x}^*) &= \operatorname{span}\{\nabla h_1(\mathbf{x}^*), \dots, \nabla h_m(\mathbf{x}^*)\} \\ &= \{\mathbf{x} \in \mathbb{R}^n: \mathbf{x} = \alpha_1 \nabla h_1(\mathbf{x}^*) + \cdots + \alpha_m \nabla h_m(\mathbf{x}^*), \alpha_1, \dots, \alpha_m \in \mathbb{R}\}. \end{aligned} \quad (6.33)$$

or

$$N(\mathbf{x}^*) = \{\mathbf{x} \in \mathbb{R}^n: \mathbf{x} = D\mathbf{h}(\mathbf{x}^*)^T\mathbf{z}, \mathbf{z} \in \mathbb{R}^m\}. \quad (6.34)$$

Notes:

- We can express the normal space $N(\mathbf{x}^*)$ as
$$N(\mathbf{x}^*) = \mathcal{R}(D\mathbf{h}(\mathbf{x}^*)^T), \quad (6.35)$$
that is, the range of the matrix $D\mathbf{h}(\mathbf{x}^*)^T$.
- The normal space contains the zero vector. Assuming $\mathbf{x}^*$ is regular, the dimension of the normal space $N(\mathbf{x}^*)$ is m.
- As in the case of the tangent space, it is often convenient to picture the normal space $N(\mathbf{x}^*)$ as passing through the point $\mathbf{x}^*$ (rather than through the origin of $\mathbb{R}^n$). For this, we define the normal plane at $\mathbf{x}^*$ as the set
$$NP(\mathbf{x}^*) = N(\mathbf{x}^*) + \mathbf{x}^* = \{\mathbf{x} + \mathbf{x}^* \in \mathbb{R}^n: \mathbf{x} \in N(\mathbf{x}^*)\}. \quad (6.36)$$
Figure 6.17 illustrates the normal space and plane in $\mathbb{R}^3$ (i.e., $n = 3$ and $m = 1$).

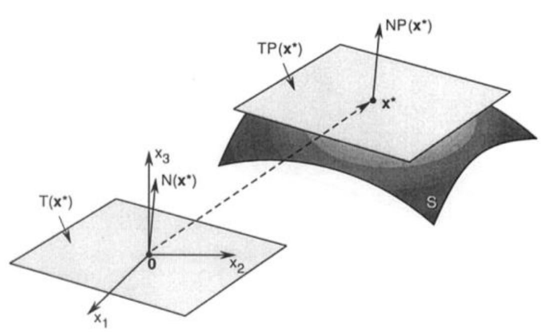


Figure 6.17. Normal space in $\mathbb{R}^3$.

We now show that the tangent space and normal space are orthogonal complements of each other.

Theorem 6.3: We have

$$T(\mathbf{x}^*) = N(\mathbf{x}^*)^{\perp} \text{ and } T(\mathbf{x}^*)^{\perp} = N(\mathbf{x}^*). \tag{6.37}$$

Proof:

By definition of $T(\mathbf{x}^*)$, we may write

$$T(\mathbf{x}^*) = \{\mathbf{y} \in \mathbb{R}^n : \langle \mathbf{x}, \mathbf{y} \rangle = 0 \text{ for all } \mathbf{x} \in N(\mathbf{x}^*)\}.$$

Hence, by definition of $N(\mathbf{x}^*)$, we have $T(\mathbf{x}^*) = N(\mathbf{x}^*)^{\perp}$. It is easy, to prove that $T(\mathbf{x}^*)^{\perp} = N(\mathbf{x}^*)$.

■

By the above theorem, we can write $\mathbb{R}^n$ as the direct sum decomposition:

$$\mathbb{R}^n = N(\mathbf{x}^*) \oplus T(\mathbf{x}^*), \tag{6.38}$$

that is, given any vectors $\mathbf{v} \in \mathbb{R}^n$, there are unique vectors $\mathbf{w} \in N(\mathbf{x}^*)$ and $\mathbf{y} \in T(\mathbf{x}^*)$, such that

$$\mathbf{v} = \mathbf{w} + \mathbf{y}. \tag{6.39}$$

Example 6.8

Construct the geometrical interpretation for the three-variable problem

$$\text{Minimize } f(x) = x_1^2 + x_2^2 + \frac{1}{4}x_3^2$$

$$\text{subject to: } \quad h_1(\mathbf{x}) = -x_1 + x_3 - 1 = 0$$

$$h_2(\mathbf{x}) = x_1^2 + x_2^2 - 2x_1 = 0$$

Solution

As was discussed in Example 6.6, the above constraints describe the curve obtained as the intersection of the cylinder $h_2(\mathbf{x}) = 0$ with the plane $h_1(\mathbf{x}) = 0$. Figure 6.18 shows that the constrained problem has a global minimizer $\mathbf{x}^* = (0 \quad 0 \quad 1)^T$. At $\mathbf{x}^*$, the tangent plane becomes a line that passes through $\mathbf{x}^*$ and is parallel with the x_2 axis while the normal plane $N(\mathbf{x}^*)$ is the plane spanned by

$$\nabla h_1(\mathbf{x}^*) = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \quad \text{and} \quad \nabla h_2(\mathbf{x}^*) = \begin{pmatrix} -2 \\ 0 \\ 0 \end{pmatrix}$$

which is identical to plane $x_2 = 0$. Note that at $\mathbf{x}^*$

$$\nabla f(\mathbf{x}^*) = \begin{pmatrix} 0 \\ 0 \\ \frac{1}{2} \end{pmatrix}$$

As is expected, $\nabla f(\mathbf{x}^*)$ lies in the normal plane $N(\mathbf{x}^*)$ (see Figure. 6.18) and can be expressed as

$$\nabla f(\mathbf{x}^*) = \frac{1}{2}\nabla h_1(\mathbf{x}^*) - \frac{1}{4}\nabla h_2(\mathbf{x}^*)$$

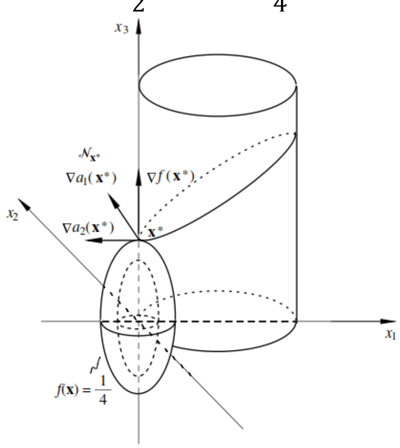


Figure 6.18.

6.6 First-Order Necessary Conditions

The necessary conditions for a point $\mathbf{x}^*$ to be a local minimizer are useful in two situations:
(a) They can be used to exclude those points that do not satisfy the necessary conditions from the candidate points;
(b) they become sufficient conditions when the objective function in question is convex.
We now generalize Lagrange theorem for the case when $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ and $\mathbf{h}: \mathbb{R}^n \longrightarrow \mathbb{R}^m, m \leq n$.

Theorem 6.4: Lagrange Theorem (First-order necessary conditions). Let $\mathbf{x}^*$ be a local minimizer (or maximizer) of $f: \mathbb{R}^n \longrightarrow \mathbb{R}$, subject to $\mathbf{h}(\mathbf{x}) = \mathbf{0}$, $\mathbf{h}: \mathbb{R}^n \longrightarrow \mathbb{R}^m, m \leq n$. Assume that $\mathbf{x}^*$ is a regular point. Then, there exists $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ such that

$$Df(\mathbf{x}^*) + \boldsymbol{\lambda}^{*T} D\mathbf{h}(\mathbf{x}^*) = \mathbf{0}^T \tag{6.40}$$

Proof:

We need to prove that

$$\nabla f(\mathbf{x}^*) = -D\mathbf{h}(\mathbf{x}^*)^T \boldsymbol{\lambda}^*$$

for some $\boldsymbol{\lambda}^* \in \mathbb{R}^m$, that is, $\nabla f(\mathbf{x}^*) \in \mathcal{R}(D\mathbf{h}(\mathbf{x}^*)^T) = N(\mathbf{x}^*)$. But, by Theorem 6.3, $N(\mathbf{x}^*) = T(\mathbf{x}^*)^{\perp}$. Therefore, it remains to show that $\nabla f(\mathbf{x}^*) \in T(\mathbf{x}^*)^{\perp}$. We proceed as follows. Suppose

$$\mathbf{y} \in T(\mathbf{x}^*).$$

Then, by Theorem 6.2, there exists a differentiable curve $\{\mathbf{x}(t): t \in (a, b)\}$ such that for all $t \in (a, b)$,

$$\mathbf{h}\big(\mathbf{x}(t)\big) = \mathbf{0}.$$

and there exists $t^* \in (a, b)$ satisfying

$$\mathbf{x}(t^*) = \mathbf{x}^*, \qquad \dot{\mathbf{x}}(t^*) = \mathbf{y}.$$

Consider now the composite function $\phi(t) = f\big(\mathbf{x}(t)\big)$. Note that t^* is a local minimizer of this function. By the first-order necessary condition for unconstrained local minimizers,

$$\frac{d\phi}{dt}(t^*) = 0$$

Applying the chain rule yields

$$\frac{d\phi}{dt}(t^*) = Df(\mathbf{x}^*)\dot{\mathbf{x}}(t^*) = Df(\mathbf{x}^*)\mathbf{y} = \langle \nabla f(\mathbf{x}^*), \mathbf{y} \rangle = 0.$$

So, all $\mathbf{y} \in T(\mathbf{x}^*)$ satisfy

$$\langle \nabla f(\mathbf{x}^*), \mathbf{y} \rangle = 0.$$

Hence, $\nabla f(\mathbf{x}^*)$ is orthogonal to the tangent plane. that is

$$\nabla f(\mathbf{x}^*) \in T(\mathbf{x}^*)^{\perp} = N(\mathbf{x}^*).$$

This implies that $\nabla f(\mathbf{x}^*)$ is a linear combination of the gradients of $\mathbf{h}$ at $\mathbf{x}^*$**.** This completes the proof.

■

- Lagrange's theorem states that if $\mathbf{x}^*$ is an extremizer, then the gradient of the objective function f can be expressed as a linear combination of the gradients of the constraints. We refer to the vector $\boldsymbol{\lambda}^*$ in the above theorem as the *Lagrange multiplier vector*, and its components the *Lagrange multipliers.*
- It should be noted that the first-order necessary conditions $Df(\mathbf{x}^*) + \boldsymbol{\lambda}^{*T} D\mathbf{h}(\mathbf{x}^*) = \mathbf{0}^T$ together with the constraints $\mathbf{h}(\mathbf{x}^*) = \mathbf{0}$ give a total of $n + m$ (generally nonlinear) equations in the $n + m$ variables comprising $\mathbf{x}^*, \boldsymbol{\lambda}^*$.
- If $\mathbf{x}^*$ is a local minimizer and $\boldsymbol{\lambda}^*$ is the associated vector of Lagrange multipliers, the set $\{\mathbf{x}^*, \boldsymbol{\lambda}^*\}$ may be referred to as the minimizer set or minimizer for short.
- Keep in mind that the Lagrange condition is only necessary, but not sufficient; that is, a point $\mathbf{x}^*$ satisfying the above equations need not be an extremizer.
- Theorem 6.4 can be related to the first-order necessary conditions for a minimum for the case of unconstrained minimization as follows. If function $f(\mathbf{x})$ is minimized without constraints, we can consider the problem as the special case of (6.6) where the number of constraints is reduced to zero. In such a case, condition of Theorem 6.4 becomes $\nabla f(\mathbf{x}^*) = \mathbf{0}$.
- Observe that $\mathbf{x}^*$ cannot be an extremizer if $\nabla f(\mathbf{x}^*) \notin N(\mathbf{x}^*)$. This situation is illustrated in Figure 6.19.

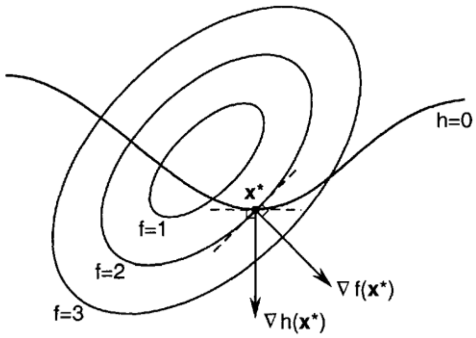


Figure 6.19. An example where the Lagrange condition does not hold.

It is convenient to introduce the so-called *Lagrangian* function $\mathcal{L}: \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}$, given by

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) \triangleq f(\mathbf{x}) + \langle \boldsymbol{\lambda}, \mathbf{h}(\mathbf{x}) \rangle. \tag{6.41}$$

The Lagrange condition for a local minimizer $\mathbf{x}^*$ can be represented using the Lagrangian function as

$$D\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0}^T \tag{6.42}$$

for some $\boldsymbol{\lambda}^*$, where the derivative operation D is with respect to the entire argument $(\mathbf{x}^T, \boldsymbol{\lambda}^T)^T$. In other words, the necessary condition in Lagrange's theorem is equivalent to the first-order necessary condition for unconstrained optimization applied to the Lagrangian function.

To see the above, denote the derivative of $\mathcal{L}$ with respect to $\mathbf{x}$ as $D_x\mathcal{L}$, and the derivative of $\mathcal{L}$ with respect to $\boldsymbol{\lambda}$ as $D_\lambda\mathcal{L}$. Then,

$$D\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = [D_x\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}), D_\lambda\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})]. \tag{6.43}$$

Note that $D_x\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = Df(\mathbf{x}) + \langle \boldsymbol{\lambda}, D\mathbf{h}(\mathbf{x}) \rangle$ and $D_\lambda\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{h}(\mathbf{x})^T$. Therefore, the Lagrange's theorem for a local minimizer $\mathbf{x}^*$can be stated as

$$\begin{aligned} D_x\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) &= \mathbf{0}^T \\ D_\lambda\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) &= \mathbf{0}^T \end{aligned} \tag{6.44}$$

for some $\boldsymbol{\lambda}^*$, which is equivalent to

$$D\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0}^T \tag{6.45}$$

In other words, the Lagrange condition can be expressed as $D\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0}^T$.

The Lagrange condition is used to find possible extremizers. This entails solving the equations:

$$\begin{aligned} D_x\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) &= \mathbf{0}^T \\ D_\lambda\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) &= \mathbf{0}^T. \end{aligned} \tag{6.46}$$

Example 6.9

$$\begin{aligned} \text{minimize: } & z = x_1 + x_2 + x_3 \\ \text{subject to: } & x_1^2 + x_2 = 3 \\ & x_1 + 3x_2 + 2x_3 = 7 \end{aligned}$$

Solution

The given program is equivalent to the unconstrained minimization of

$$z = \frac{1}{2}(x_1^2 + x_1 + 4)$$

which obviously has a solution. We may therefore apply the method of Lagrange multipliers to the original program standardized as

$$\begin{aligned} \text{minimize: } & z = x_1 + x_2 + x_3 \\ \text{subject to: } & x_1^2 + x_2 - 3 = 0 \\ & x_1 + 3x_2 + 2x_3 - 7 = 0 \end{aligned}$$

Here, $f(x_1, x_2, x_3) = x_1 + x_2 + x_3$, $n = 3$ (variables), $m = 2$ (constraints),

$$\begin{aligned} h_1(x_1, x_2, x_3) &= x_1^2 + x_2 - 3 \\ h_2(x_1, x_2, x_3) &= x_1 + 3x_2 + 2x_3 - 7 \end{aligned}$$

The Lagrangian function is then

$$\mathcal{L} = (x_1 + x_2 + x_3) - \lambda_1(x_1^2 + x_2 - 3) - \lambda_2(x_1 + 3x_2 + 2x_3 - 7)$$

and

$$\frac{\partial}{\partial x_1}\mathcal{L} = 1 - 2\lambda_1 x_1 - \lambda_2 = 0$$
$$\frac{\partial}{\partial x_2}\mathcal{L} = 1 - \lambda_1 - 3\lambda_2 = 0$$
$$\frac{\partial}{\partial x_3}\mathcal{L} = 1 - 2\lambda_2 = 0$$
$$\frac{\partial}{\partial \lambda_1}\mathcal{L} = -(x_1^2 + x_2 - 3) = 0$$
$$\frac{\partial}{\partial \lambda_2}\mathcal{L} = -(x_1 + 3x_2 + 2x_3 - 7) = 0$$

Successively solving equations, we obtain the unique solution $\lambda_2 = 0.5$, $\lambda_1 = 0.5$, $x_1 = 0.5$, $x_2 = 2.75$, and $x_3 = -0.875$, with

$$z = x_1 + x_2 + x_3 = 0.5 + 2.75 - 0.875 = 2.375$$

Since the first partial derivatives of $f(x_1, x_2, x_3)$, $h_1(x_1, x_2, x_3)$, and $h_2(x_1, x_2, x_3)$ are all continuous, and since

$$\mathbf{J} = \begin{pmatrix} \frac{\partial h_1}{\partial x_1} & \frac{\partial h_1}{\partial x_2} & \frac{\partial h_1}{\partial x_3} \\ \frac{\partial h_2}{\partial x_1} & \frac{\partial h_2}{\partial x_2} & \frac{\partial h_2}{\partial x_3} \end{pmatrix} = \begin{pmatrix} 2x_1 & 1 & 0 \\ 1 & 3 & 2 \end{pmatrix}$$

is of rank 2 everywhere (the last two columns are linearly independent everywhere), either $x_1 = 0.5$, $x_2 = 2.75$, and $x_3 = -0.875$ is the optimal solution to the program or no optimal solution exists. Checking feasible points in the region around $(0.5, 2.75, -0.875)$, we find that this point is indeed the location of a (global) minimum for the problem.

Example 6.10

Consider the problem of extremizing the objective function

$$f(\mathbf{x}) = x_1^2 + x_2^2$$

on the ellipse

$$\{(x_1, x_2)^T : h(\mathbf{x}) = x_1^2 + 2x_2^2 - 1 = 0\}.$$

Solution

We have

$$\nabla f(\mathbf{x}) = (2x_1, 2x_2)^T,$$
$$\nabla h(\mathbf{x}) = (2x_1, 4x_2)^T.$$

$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda h(\mathbf{x})$. Thus,

$$\begin{aligned} D_x\mathcal{L}(\mathbf{x}, \lambda) &= D_x\big(f(\mathbf{x}) + \lambda h(\mathbf{x})\big) \\ &= D_x\big(x_1^2 + x_2^2 + \lambda(x_1^2 + 2x_2^2 - 1)\big) \\ &= D_x(x_1^2 + \lambda x_1^2 + x_2^2 + 2\lambda x_2^2 - \lambda) \\ &= (2x_1 + 2\lambda x_1, 2x_2 + 4\lambda x_2) \end{aligned}$$

and

$$D_\lambda\mathcal{L}(\mathbf{x}, \lambda) = h(\mathbf{x}) = x_1^2 + 2x_2^2 - 1.$$

Setting $D_x\mathcal{L}(\mathbf{x}, \lambda) = \mathbf{0}^T$ and $D_\lambda\mathcal{L}(\mathbf{x}, \lambda) = 0$ we obtain three equations in three unknowns

$$2x_1 + 2\lambda x_1 = 0$$
$$2x_2 + 4\lambda x_2 = 0$$
$$x_1^2 + 2x_2^2 = 1.$$

All feasible points in this problem are regular. From the first of the above equations, we get either $x_1 = 0$ or $\lambda = -1$. For the case where $x_1 = 0$, the second and third equations imply that $\lambda = -1/2$ and $x_2 = \pm 1/\sqrt{2}$. For the case where $\lambda = -1$, the second and third equations imply that $x_1 = \pm 1$ and $x_2 = 0$. Thus, the points that satisfy the Lagrange condition for extrema are

$$\mathbf{x}^{(1)} = \begin{pmatrix} 0 \\ 1/\sqrt{2} \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} 0 \\ -1/\sqrt{2} \end{pmatrix}, \quad \mathbf{x}^{(3)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{x}^{(4)} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

Because

$$f\big(\mathbf{x}^{(1)}\big) = f\big(\mathbf{x}^{(2)}\big) = \frac{1}{2} \quad \text{and} \quad f\big(\mathbf{x}^{(3)}\big) = f\big(\mathbf{x}^{(4)}\big) = 1$$

we conclude that if there are minimizers, then they are located at $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ and if there are maximizers, then they are located at $\mathbf{x}^{(3)}$ and $\mathbf{x}^{(4)}$. It turns out that, indeed, $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are minimizers and $\mathbf{x}^{(3)}$ and $\mathbf{x}^{(4)}$are maximizers. This problem can be solved graphically, as illustrated in Figure 6.20.

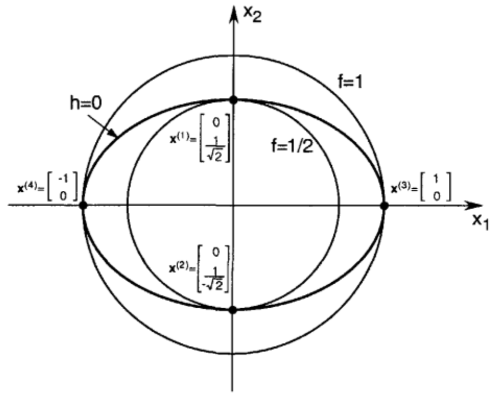


Figure 6.20. Graphical solution of the problem

Example 6.11

Find the points that satisfy the necessary conditions for a minimum for the problem in Example 6.8.

Solution

We have

$$
\begin{aligned}
Df(\mathbf{x}) &= \begin{pmatrix} 2x_1 & 2x_2 & \frac{1}{2}x_3 \end{pmatrix}, \\
D\mathbf{h}(\mathbf{x}) &= \begin{pmatrix} -1 & 0 & 1 \\ 2x_1 - 2 & 2x_2 & 0 \end{pmatrix}, \\
\boldsymbol{\lambda} &= \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}
\end{aligned}
$$

Thus, $Df(\mathbf{x}) + \boldsymbol{\lambda}^T D\mathbf{h}(\mathbf{x}) = \mathbf{0}^T$ together with the constraints $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ give a total of 5 equations in the 5 variables comprising $\mathbf{x}^*$, $\boldsymbol{\lambda}^*$. We have

$$
\begin{aligned}
Df(\mathbf{x}) + \boldsymbol{\lambda}^T D\mathbf{h}(\mathbf{x}) &= \begin{pmatrix} 2x_1 & 2x_2 & \frac{1}{2}x_3 \end{pmatrix} + (\lambda_1 \quad \lambda_2)\begin{pmatrix} -1 & 0 & 1 \\ 2x_1 - 2 & 2x_2 & 0 \end{pmatrix} \\
&= \begin{pmatrix} 2x_1 & 2x_2 & \frac{1}{2}x_3 \end{pmatrix} + \begin{pmatrix} -\lambda_1 + \lambda_2(2x_1 - 2) & 2\lambda_2 x_2 & \lambda_1 \end{pmatrix} \\
&= \begin{pmatrix} 2x_1 - \lambda_1 + \lambda_2(2x_1 - 2) & 2x_2 + 2\lambda_2 x_2 & \frac{1}{2}x_3 + \lambda_1 \end{pmatrix} \\
&= \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}
\end{aligned}
$$

and $\mathbf{h}(\mathbf{x}) = \mathbf{0}$

$$
\begin{pmatrix} h_1(\mathbf{x}) \\ h_2(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} -x_1 + x_3 - 1 \\ x_1^2 + x_2^2 - 2x_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}
$$

Hence, we have

$$
\begin{aligned}
2x_1 - \lambda_1 + \lambda_2(2x_1 - 2) &= 0 \\
2x_2 + 2\lambda_2 x_2 &= 0 \\
\frac{1}{2}x_3 + \lambda_1 &= 0 \\
-x_1 + x_3 - 1 &= 0 \\
x_1^2 + x_2^2 - 2x_1 &= 0
\end{aligned}
$$

Solving the above system of equations, we obtain two solutions, i.e.,

$$
\mathbf{x}_1^* = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\lambda}_1^* = \begin{pmatrix} -\frac{1}{2} \\ \frac{1}{4} \end{pmatrix}
$$

and

$$\mathbf{x}_2^* = \begin{pmatrix} 2 \\ 0 \\ 3 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\lambda}_2^* = \begin{pmatrix} -\frac{3}{2} \\ -\frac{11}{4} \end{pmatrix}$$

6.7 Second-Order Conditions

By an argument analogous to that used for the unconstrained case, we can also derive the corresponding second-order conditions for equality constrained problems. We assume that $f: \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{h}: \mathbb{R}^n \to \mathbb{R}^m$ are twice continuously differentiable, that is, $f, \mathbf{h} \in C^2$. Let

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \langle \boldsymbol{\lambda}, \mathbf{h}(\mathbf{x}) \rangle = f(\mathbf{x}) + \lambda_1 h_1(\mathbf{x}) + \cdots + \lambda_m h_m(\mathbf{x}) \tag{6.47}$$

be the Lagrangian function. Let $\boldsymbol{L}(\mathbf{x}, \boldsymbol{\lambda})$ be the Hessian matrix of $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ with respect to $\mathbf{x}$, that is,

$$\boldsymbol{L}(\mathbf{x}, \boldsymbol{\lambda}) = \boldsymbol{F}(\mathbf{x}) + \lambda_1 \boldsymbol{H}_1(\mathbf{x}) + \cdots + \lambda_m \boldsymbol{H}_m(\mathbf{x}), \tag{6.48}$$

where $\boldsymbol{F}(\mathbf{x})$ is the Hessian matrix of f at $\mathbf{x}$, and $\boldsymbol{H}_k(\mathbf{x})$ is the Hessian matrix of h_k at $\mathbf{x}$, $k = 1, \dots, m$, given by

$$\boldsymbol{H}_k(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 h_k}{\partial x_1^2}(\mathbf{x}) & \cdots & \frac{\partial^2 h_k}{\partial x_n \partial x_1}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 h_k}{\partial x_1 \partial x_n}(\mathbf{x}) & \cdots & \frac{\partial^2 h_k}{\partial x_n^2}(\mathbf{x}) \end{pmatrix}. \tag{6.49}$$

We introduce the notation $[\boldsymbol{\lambda H}(\mathbf{x})]$:

$$[\boldsymbol{\lambda H}(\mathbf{x})] = \lambda_1 \boldsymbol{H}_1(\mathbf{x}) + \cdots + \lambda_m \boldsymbol{H}_m(\mathbf{x}). \tag{6.50}$$

Using the above notation, we can write

$$\boldsymbol{L}(\mathbf{x}, \boldsymbol{\lambda}) = \boldsymbol{F}(\mathbf{x}) + [\boldsymbol{\lambda H}(\mathbf{x})]. \tag{6.51}$$

Theorem 6.5 (Second-Order Necessary Conditions): Let $\mathbf{x}^*$ be a local minimizer of $f: \mathbb{R}^n \to \mathbb{R}$ subject to $\mathbf{h}(\mathbf{x}) = \mathbf{0}$, $\mathbf{h}: \mathbb{R}^n \to \mathbb{R}^m$, $m \le n$, and $f, \mathbf{h} \in C^2$. Suppose $\mathbf{x}^*$ is regular. Then, there exists $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ such that

1. $Df(\mathbf{x}^*) + \boldsymbol{\lambda}^{*T} D\mathbf{h}(\mathbf{x}^*) = \mathbf{0}^T$; and
2. for all $\mathbf{y} \in T(\mathbf{x}^*)$, we have $\langle \mathbf{y} | \boldsymbol{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) | \mathbf{y} \rangle \ge 0$.

Proof:

The existence of $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ such that $Df(\mathbf{x}^*) + \boldsymbol{\lambda}^{*T} D\mathbf{h}(\mathbf{x}^*) = \mathbf{0}^T$ follows from Lagrange's theorem. It remains to prove the second part of the result. Suppose $\mathbf{y} \in T(\mathbf{x}^*)$, that is, $\mathbf{y}$ belongs to the tangent space to $S = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$ at $\mathbf{x}^*$. Because $\mathbf{h} \in C^2$, following the argument of Theorem 6.2, there exists a twice differentiate curve $\{\mathbf{x}(t) : t \in (a, b)\}$ on S such that

$$\mathbf{x}(t^*) = \mathbf{x}^*, \;\; \dot{\mathbf{x}}(t^*) = \mathbf{y}$$

for some $t^* \in (a, b)$. Observe that by assumption, t^* is a local minimizer of the function $\phi(t) = f\big(\mathbf{x}(t)\big)$. From the second-order necessary condition for unconstrained minimization, we obtain

$$\frac{d^2\phi}{dt^2}(t^*) \ge 0.$$

Using the following formula

$$\frac{d}{dt}\langle \mathbf{y}(t), \mathbf{z}(t) \rangle = \langle \mathbf{z}(t), \frac{d\mathbf{y}}{dt}(t) \rangle + \langle \mathbf{y}(t), \frac{d\mathbf{z}}{dt}(t) \rangle$$

and applying the chain rule yields

$$\begin{aligned}\frac{d^2\phi}{dt^2}(t^*) &= \frac{d}{dt}\left[Df\big(\mathbf{x}(t^*)\big)\dot{\mathbf{x}}(t^*)\right]\\ &= \langle\dot{\mathbf{x}}(t^*)|\boldsymbol{F}(\mathbf{x}^*)|\dot{\mathbf{x}}(t^*)\rangle + Df(\mathbf{x}^*)\ddot{\mathbf{x}}(t^*)\\ &= \langle\mathbf{y}|\boldsymbol{F}(\mathbf{x}^*)|\mathbf{y}\rangle + Df(\mathbf{x}^*)\ddot{\mathbf{x}}(t^*) \geq 0.\end{aligned}$$

Because $\mathbf{h}\big(\mathbf{x}(t)\big) = \mathbf{0}$ for all $t \in (a, b)$, we have

$$\frac{d^2}{dt^2}\langle\boldsymbol{\lambda}^*, \mathbf{h}\big(\mathbf{x}(t)\big)\rangle = 0.$$

Thus, for all $t \in (a, b)$,

$$\begin{aligned}\frac{d^2}{dt^2}\langle\boldsymbol{\lambda}^*, \mathbf{h}\big(\mathbf{x}(t)\big)\rangle &= \frac{d}{dt}\left[\langle\boldsymbol{\lambda}^*, \frac{d}{dt}\mathbf{h}\big(\mathbf{x}(t)\big)\rangle\right]\\ &= \frac{d}{dt}\left[\sum_{k=1}^{m}\lambda_k^*\frac{d}{dt}h_k\big(\mathbf{x}(t)\big)\right]\\ &= \frac{d}{dt}\left[\sum_{k=1}^{m}\lambda_k^* Dh_k\big(\mathbf{x}(t)\big)\dot{\mathbf{x}}(t)\right]\\ &= \sum_{k=1}^{m}\lambda_k^*\frac{d}{dt}\Big(Dh_k\big(\mathbf{x}(t)\big)\dot{\mathbf{x}}(t)\Big)\\ &= \sum_{k=1}^{m}\lambda_k^*\left[\langle\dot{\mathbf{x}}(t)\big|\boldsymbol{H_k}\big(\mathbf{x}(t)\big)\big|\dot{\mathbf{x}}(t)\rangle + Dh_k\big(\mathbf{x}(t)\big)\ddot{\mathbf{x}}(t)\right]\\ &= \sum_{k=1}^{m}\left[\langle\dot{\mathbf{x}}(t)\big|\lambda_k^*\boldsymbol{H_k}\big(\mathbf{x}(t)\big)\big|\dot{\mathbf{x}}(t)\rangle + \lambda_k^* Dh_k\big(\mathbf{x}(t)\big)\ddot{\mathbf{x}}(t)\right]\\ &= \langle\dot{\mathbf{x}}(t)\big|\textstyle\sum_{k=1}^{m}\lambda_k^*\boldsymbol{H_k}\big(\mathbf{x}(t)\big)\big|\dot{\mathbf{x}}(t)\rangle + \left(\displaystyle\sum_{k=1}^{m}\lambda_k^* Dh_k\big(\mathbf{x}(t)\big)\right)\ddot{\mathbf{x}}(t)\\ &= \langle\dot{\mathbf{x}}(t)\big|\big[\boldsymbol{\lambda}^*\boldsymbol{H}\big(\mathbf{x}(t)\big)\big]\big|\dot{\mathbf{x}}(t)\rangle + \boldsymbol{\lambda}^{*T}D\mathbf{h}\big(\mathbf{x}(t)\big)\ddot{\mathbf{x}}(t) = 0.\end{aligned}$$

In particular, the above is true for $t = t^*$, that is,

$$\langle\mathbf{y}|[\boldsymbol{\lambda}^*\boldsymbol{H}(\mathbf{x}^*)]|\mathbf{y}\rangle + \boldsymbol{\lambda}^{*T}D\mathbf{h}(\mathbf{x}^*)\ddot{\mathbf{x}}(t^*) = 0.$$

Adding the above equation to the inequality

$$\langle\mathbf{y}|\boldsymbol{F}(\mathbf{x}^*)|\mathbf{y}\rangle + Df(\mathbf{x}^*)\ddot{\mathbf{x}}(t^*) \geq 0$$

yields

$$\langle\mathbf{y}|(\boldsymbol{F}(\mathbf{x}^*) + [\boldsymbol{\lambda}^*\boldsymbol{H}(\mathbf{x}^*)])|\mathbf{y}\rangle + \Big(Df(\mathbf{x}^*) + \boldsymbol{\lambda}^{*T}D\mathbf{h}(\mathbf{x}^*)\Big)\ddot{\mathbf{x}}(t^*) \geq 0,$$

But, by Lagrange's theorem, $Df(\mathbf{x}^*) + \boldsymbol{\lambda}^{*T}D\mathbf{h}(\mathbf{x}^*) = \mathbf{0}$. Therefore,

$$\langle\mathbf{y}|(\boldsymbol{F}(\mathbf{x}^*) + [\boldsymbol{\lambda}^*\boldsymbol{H}(\mathbf{x}^*)])|\mathbf{y}\rangle = \langle\mathbf{y}|\boldsymbol{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)|\mathbf{y}\rangle \geq 0,$$

which proves the result.

■

Observe that $\boldsymbol{L}(\mathbf{x}, \boldsymbol{\lambda})$ plays a similar role as the Hessian matrix $\boldsymbol{F}(\mathbf{x})$ of the objective function f did in the unconstrained minimization case. However, we now require that $\boldsymbol{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) \geq 0$ only on $T(\mathbf{x}^*)$ rather than on $\mathbb{R}^n$.

The above conditions are necessary, but not sufficient, for a point to be a local minimizer. We now present, without a proof, sufficient conditions for a point to be a strict local minimizer.

Theorem 6.6 (Second-Order Sufficient Conditions): Suppose $f, \mathbf{h} \in C^2$ and there exist a point $\mathbf{x}^* \in \mathbb{R}^n$ and $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ such that

1. $Df(\mathbf{x}^*) + \boldsymbol{\lambda}^{*T}D\mathbf{h}(\mathbf{x}^*) = \mathbf{0}^T$; and
2. for all $\mathbf{y} \in T(\mathbf{x}^*)$, $\mathbf{y} \neq \mathbf{0}$, we have $\langle\mathbf{y}|\boldsymbol{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)|\mathbf{y}\rangle > 0$.

Then, $\mathbf{x}^*$ is a strict local minimizer of f subject to $\mathbf{h}(\mathbf{x}) = \mathbf{0}$.

The above theorem states that if an $\mathbf{x}^*$ satisfies the Lagrange condition, and $\boldsymbol{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is positive definite on $T(\mathbf{x}^*)$, then $\mathbf{x}^*$ is a strict local minimizer.

Example 6.12

Consider the problem

$$\text{maximize } (x_1-1)^2 + (x_2-1)^2$$
$$\text{subject to } x_1^2 + x_2^2 - 1 = 0.$$

Solution

The Lagrangian and subsequent first-order conditions would be

$$\mathcal{L}(x_1, x_2, \lambda) = (x_1-1)^2 + (x_2-1)^2 - \lambda(x_1^2 + x_2^2 - 1)$$

$$\boldsymbol{\nabla}_{\mathbf{x}}\mathcal{L}(x_1, x_2, \lambda) = \begin{pmatrix} 2x_1(1-\lambda) - 2 \\ 2x_2(1-\lambda) - 2 \end{pmatrix} = \mathbf{0}$$

From the two equations we conclude $x_1 = x_2$, together with $x_1^2 + x_2^2 - 1 = 0$, we have two first-order stationary solutions ($x_1 = x_2 = \frac{1}{\sqrt{2}}, \lambda = 1 - \sqrt{2}$) and ($x_1 = x_2 = -\frac{1}{\sqrt{2}}, \lambda = 1 + \sqrt{2}$).

The Lagrangian Hessian matrix $\boldsymbol{L}(\mathbf{x}, \boldsymbol{\lambda})$, at two λs, becomes

$$\begin{pmatrix} 2(1-\lambda) & 0 \\ 0 & 2(1-\lambda) \end{pmatrix} \Rightarrow \begin{pmatrix} 2\sqrt{2} & 0 \\ 0 & 2\sqrt{2} \end{pmatrix} \text{ with } \left(\lambda = 1 - \sqrt{2}\right)$$

$$\Rightarrow \begin{pmatrix} -2\sqrt{2} & 0 \\ 0 & -2\sqrt{2} \end{pmatrix} \text{ with } \left(\lambda = 1 + \sqrt{2}\right)$$

where the first one is positive definite and the second negative definite, and they remain so in tangent subspace. Thus, $x_1 = x_2 = \frac{1}{\sqrt{2}}$ is a minimum and $x_1 = x_2 = -\frac{1}{\sqrt{2}}$ is a maximum.

Example 6.13

In Example 6.8 it was found that

$$\mathbf{x}_2^* = \begin{pmatrix} 2 \\ 0 \\ 3 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\lambda}_2^* = \begin{pmatrix} -\frac{3}{2} \\ -\frac{11}{4} \end{pmatrix}$$

satisfy the first-order necessary conditions for a minimum for the problem of Example 6.11. Check whether the second-order necessary conditions for a minimum are satisfied.

Solution

We can write

$$\nabla_x^2 \boldsymbol{L}(\mathbf{x}_2^*, \boldsymbol{\lambda}_2^*) = \begin{pmatrix} -\frac{7}{2} & 0 & 0 \\ 0 & -\frac{7}{2} & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}$$

and

$$D\mathbf{h}(\mathbf{x}_2^*) = \begin{pmatrix} -1 & 0 & 1 \\ 2 & 0 & 0 \end{pmatrix}$$

It can be readily verified that the null space of $D\mathbf{h}(\mathbf{x}_2^*)$ is the one-dimensional space spanned by $\mathbf{y} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$. Since

$$\langle \mathbf{y} | \nabla_x^2 \boldsymbol{L}(\mathbf{x}_2^*, \boldsymbol{\lambda}_2^*) | \mathbf{y} \rangle = -\frac{7}{2} < 0$$

we conclude that $\{\mathbf{x}_2^*, \boldsymbol{\lambda}_2^*\}$ does not satisfy the second-order necessary conditions.

Example 6.14

Check whether the second-order sufficient conditions for a minimum are satisfied in the minimization problem of Example 6.11.

Solution

We compute

$$\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}_1^*, \boldsymbol{\lambda}_1^*) = \begin{pmatrix} \frac{5}{2} & 0 & 0 \\ 0 & \frac{5}{2} & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}$$

which is positive definite. Hence Theorem 6.6 implies that x_1^* is a strong local minimizer.