

Chapter 7

Constrained optimization: the basics

Here starts the second part of the course: **Constrained optimization**. The contents of this chapter is mostly contained in Chapters 12 and 15 of Nocedal & Wright [23].

The general problem is written as

$$\min_{\mathbf{x} \in \Omega} f(\mathbf{x}), \quad \text{where} \quad (7.1a)$$

$$\Omega = \{\mathbf{x} \in \mathbb{R}^n \mid c_i(\mathbf{x}) = 0, i \in \mathcal{E}, c_i(\mathbf{x}) \geq 0, i \in \mathcal{I}\}. \quad (7.1b)$$

Assume $c_i(\mathbf{x}) \in C^1, \forall i$.

Example 7.1

1. Consider the set defined as in Figure 7.1. Here $\mathcal{E} = \{1\}$, $\mathcal{I} = \{2, 3\}$.

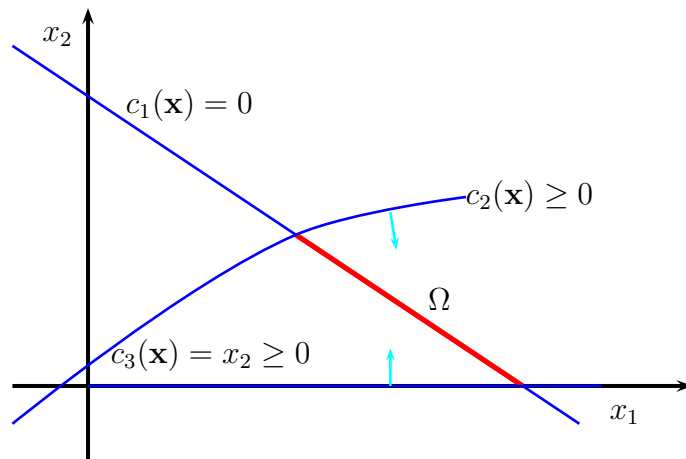


Figure 7.1: Equality and inequality constraints.

The set Ω consists of the straight line segment of $c_1 = 0$ between the curve $c_2 = 0$ and the x_1 -axis.

- Often \mathcal{E} is empty. Then Ω can have a nonempty interior as in Figure 7.2.

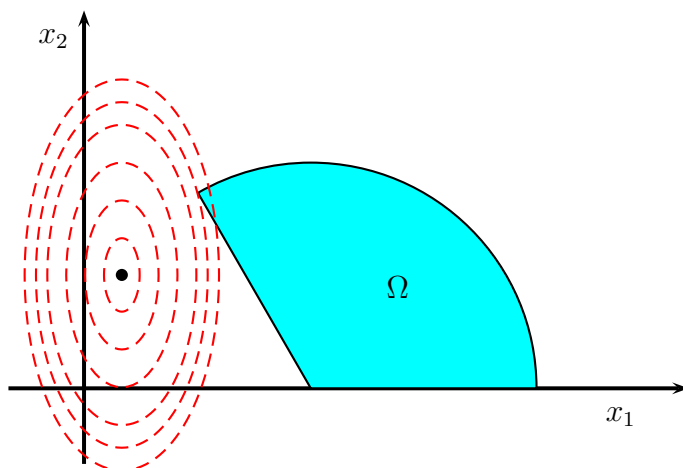


Figure 7.2: A feasible set Ω with a nonempty interior, and level sets of f (the larger the ellipse the larger the value of f).



If the unconstrained minimum of f is in Ω then the problem is essentially that of unconstrained optimization. For instance, in the distributed parameter estimation problems such as Example 4.5 there is the constraint based on the physical requirement that $\sigma > 0$, but it is often satisfied automatically in an unconstrained solution process. We will consider later on such constraints in case they are not automatically satisfied.

To make things interesting assume that the unconstrained minimizer of f is not inside Ω , as in Figure 7.2. Define at each point $\mathbf{x} \in \mathbb{R}^n$ the *active set*

$$\mathcal{A}(\mathbf{x}) = \mathcal{E} \cup \{i \in \mathcal{I} \mid c_i(\mathbf{x}) = 0\}. \quad (7.2)$$

We are looking, then, at cases where $\mathcal{A}(\mathbf{x}^*)$ is nonempty.

Before moving on let us remark that the treatment above and below relates to finding a *local* isolated minimum. Globally, constraints may turn an unconstrained problem with many local minima into a constrained one with a unique minimum. (This is in a sense the goal of branch and bound algorithms, or other search space methods.) But also, constraints may turn an unconstrained problem with a unique minimum into one with many local minima – see examples in Nocedal & Wright [23], pp. 316–317.

7.1 Necessary and sufficient conditions for a local minimum

Of course, there are first order necessary conditions, and there are second order necessary and sufficient conditions for a local minimum. They are all significantly more complicated than in the unconstrained case.

7.1.1 First order conditions

Let us consider a motivating example first.

Example 7.2

1. We start with one equality constraint in \mathbb{R}^2 : $c(x_1, x_2) = 0$. See Figure 7.3. At any point \mathbf{x} the gradient $\nabla f(\mathbf{x})$ is orthogonal to the tangent

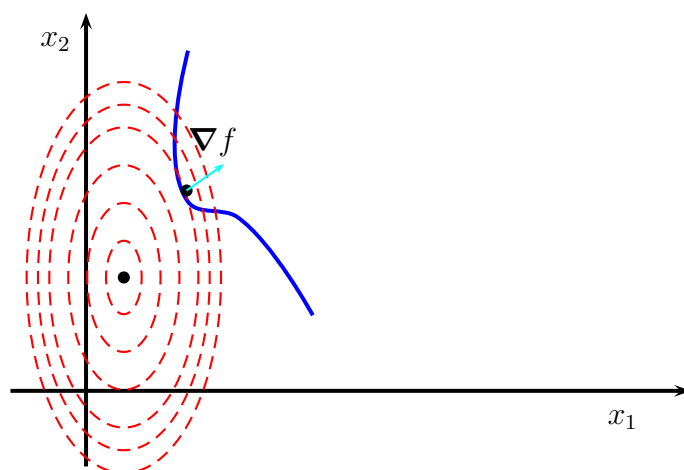


Figure 7.3: One equality constraint and the level sets of f . At \mathbf{x}^* the gradient is orthogonal to the tangent of the constraint.

of the level set at that point. At the minimum point \mathbf{x}^* obviously the constraint and the level set for $f(\mathbf{x}^*)$ have the same tangent direction, hence $\nabla f(\mathbf{x}^*)$ is parallel to $\nabla c(\mathbf{x}^*)$. This means that there is a constant of proportionality, λ^* , such that

$$\nabla f(\mathbf{x}^*) = \lambda^* \nabla c(\mathbf{x}^*).$$

2. Next, suppose that there is only one inequality constraint, $c(x_1, x_2) \geq 0$. The feasible region Ω is the domain to the right of the solid blue curve in Figure 7.3. Thus, not only is $\nabla f(\mathbf{x}^*)$ parallel to $\nabla c(\mathbf{x}^*)$, it must also point into the interior of the feasible set, which means that the constant of proportionality λ^* cannot be negative:

$$\nabla f(\mathbf{x}^*) = \lambda^* \nabla c(\mathbf{x}^*), \quad \lambda^* \geq 0.$$

This λ^* is called a *Lagrange multiplier*. ◆

Let us generalize Example 7.2 for the case of several equality constraints, $\mathcal{E} = \{1, \dots, m\}$, in \mathbb{R}^n . Recall that, in general, for \mathbf{x}^* to be a critical point we must have that

$$\nabla f(\mathbf{x}^*)^T \mathbf{p} \geq 0$$

for all feasible directions \mathbf{p} . (This comes straight from a Taylor expansion, considering direction vectors with $\|\mathbf{p}\|$ very small.) Here we understand the term “feasible direction” in the limit sense, i.e., a direction tangential to the (nonlinear) constraint manifold. So, we require $\nabla f(\mathbf{x}^*)$ to be orthogonal to the *tangent space*

$$M = \{\mathbf{y} \mid \mathbf{y}^T \nabla c_i(\mathbf{x}^*) = 0, \quad i = 1, \dots, m\}. \quad (7.3)$$

Denote

$$A_*^T = [\nabla c_1(\mathbf{x}^*), \nabla c_2(\mathbf{x}^*), \dots, \nabla c_m(\mathbf{x}^*)].$$

This is an $n \times m$ matrix. Then the tangent space M of (7.3) is equal to the nullspace of A_* , because we can write

$$M = \{\mathbf{y} \mid \mathbf{y}^T A_*^T = \mathbf{0}^T\} = \{\mathbf{y} \mid A_* \mathbf{y} = \mathbf{0}\} = \ker(A_*).$$

We shall make the all-important assumption of the *linear independence constraint qualification (LICQ)*, that A_*^T has a full column rank. But now, requiring that $\nabla f(\mathbf{x}^*)$ be orthogonal to $\ker(A_*)$ is equivalent to requiring that $\nabla f(\mathbf{x}^*) \in \text{Range}(A_*^T)$. The latter is equivalent to requiring that there be coefficients (called, of course, Lagrange multipliers) $\lambda_1^*, \dots, \lambda_m^*$ such that

$$\nabla f(\mathbf{x}^*) = \sum_{i=1}^m \lambda_i^* \nabla c_i(\mathbf{x}^*). \quad (7.4)$$

Next, we generalize Example 7.2 for the case of several inequality constraints in \mathbb{R}^n . At a candidate point \mathbf{x}^* only the active constraints, i.e. those in $\mathcal{A}(\mathbf{x}^*)$, matter. These look like equality constraints at \mathbf{x}^* , and it follows that an expression such as (7.4) must hold for the active constraints. Moreover, just like in Example 7.2 the coefficients of the nonlinear combination must be nonnegative, $\lambda_i^* \geq 0$, $i \in \mathcal{A}(\mathbf{x}^*)$. (A leap of faith is required to obtain this result: consult [23] if you are a non-believer.) We can add zero-multipliers for the inactive constraints and write this in general as

$$\nabla f(\mathbf{x}^*) = \sum_{i \in \mathcal{I}} \lambda_i^* \nabla c_i(\mathbf{x}^*), \quad \lambda_i^* \geq 0, \quad \lambda_i^* c_i(\mathbf{x}^*) = 0, \quad \forall i \in \mathcal{I}. \quad (7.5)$$

The condition that $\lambda_i^* c_i(\mathbf{x}^*) = 0 \forall i$ is called the *complementarity condition*.

The general case is now combined from the two cases just considered. We first define the *Lagrangian*

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(\mathbf{x}). \quad (7.6)$$

Next, recall the definition of active set (7.2) and define A_*^T to be the matrix whose columns are the gradients $\nabla c_i(\mathbf{x}^*)$ of all active constraints, that is to say those belonging to $\mathcal{A}(\mathbf{x}^*)$. We make the constraint qualification assumption

- **(LICQ)**: The matrix A_*^T has a full column rank.

Then, the following Karush-Kuhn-Tucker (KKT) first order conditions are necessary for a minimum: There is a vector of Lagrange multipliers $\boldsymbol{\lambda}^*$ such that

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0}, \quad (7.7a)$$

$$c_i(\mathbf{x}^*) = 0, \quad \forall i \in \mathcal{E}, \quad (7.7b)$$

$$c_i(\mathbf{x}^*) \geq 0, \quad \forall i \in \mathcal{I}, \quad (7.7c)$$

$$\lambda_i^* \geq 0, \quad \forall i \in \mathcal{I}, \quad (7.7d)$$

$$\lambda_i^* c_i(\mathbf{x}^*) = 0, \quad \forall i \in \mathcal{E} \cup \mathcal{I}. \quad (7.7e)$$

It can be shown that without constraint qualification these conditions are not quite necessary. But with this constraint qualification there exists a unique $\boldsymbol{\lambda}^*$ which satisfies the KKT conditions.

If f and the c_i are all convex functions then the KKT conditions are sufficient for a minimum as well. In general they are not.

Example 7.3

Consider minimizing a quadratic function under linear equality constraints,

$$\begin{aligned} \min \quad & f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T H \mathbf{x} - \mathbf{d}^T \mathbf{x} \\ \text{s.t.} \quad & A \mathbf{x} = \mathbf{b}, \end{aligned} \quad (7.8)$$

where A is $m \times n$, $m \leq n$. The LICQ condition is that A have a full row rank.

This is a *quadratic programming* problem with *equality constraints*. We have

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{2} \mathbf{x}^T H \mathbf{x} - \mathbf{d}^T \mathbf{x} - \boldsymbol{\lambda}^T (\mathbf{b} - A \mathbf{x}),$$

and the KKT conditions read

$$\begin{aligned} H\mathbf{x} - \mathbf{d} + A^T\boldsymbol{\lambda} &= \mathbf{0}, \\ \mathbf{b} - A\mathbf{x} &= \mathbf{0}, \end{aligned}$$

which can be arranged as

$$\begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{d} \\ \mathbf{b} \end{pmatrix}. \quad (7.9)$$

The KKT matrix $K = \begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix}$ is symmetric, but it is *indefinite* even if H is s.p.d. Since A has full row rank the KKT matrix is nonsingular if $\mathbf{y}^T H \mathbf{y} \neq 0$ for all $\mathbf{y} \in \ker(A)$, $\mathbf{y} \neq \mathbf{0}$. (Exercise: show this.) This is true of course especially if H is nonsingular.

If the KKT matrix is nonsingular then there is exactly one critical point \mathbf{x}^* satisfying the necessary conditions. ◆

The situation is more complicated, even for quadratic programming problems, when there are inequality constraints.

Example 7.4

Consider the problem

$$\begin{aligned} \min \quad & (x_1 - \frac{3}{2})^2 + (x_2 - \frac{1}{8})^2 \\ \text{s.t.} \quad & |x_1| + |x_2| \leq 1. \end{aligned}$$

Note that the objective function can be written as

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T(2I)\mathbf{x} - (3, 1/4)^T\mathbf{x}$$

plus a meaningless constant. The unconstrained minimum is at $\hat{\mathbf{x}} = (3/2, 1/8)$, which does not satisfy the constraints. Moreover, the constraints can be written in the form

$$\mathbf{b} - A\mathbf{x} \geq 0$$

where $A \in \mathbb{R}^{4 \times 2}$ as follows,

$$\begin{aligned} 1 - x_1 - x_2 &\geq 0, \\ 1 - x_1 + x_2 &\geq 0, \\ 1 + x_1 - x_2 &\geq 0, \\ 1 + x_1 + x_2 &\geq 0. \end{aligned}$$

So, this is a quadratic programming problem with inequality constraints.

Let us examine this problem. The feasible set is in a unit diamond about the origin. From the location of the unconstrained minimum it is obvious that the only constraints of possible relevance are

$$\begin{aligned}x_1 + x_2 &\leq 1, \\x_1 - x_2 &\leq 1.\end{aligned}$$

If only the first is active then we get

$$\begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \lambda_1 \end{pmatrix} = \begin{pmatrix} 3 \\ 1/4 \\ 1 \end{pmatrix}.$$

This yields $\mathbf{x} = (1.1875, -.1875)$, which is infeasible.

Adding the second constraint we get

$$\begin{pmatrix} 2 & 0 & 1 & 1 \\ 0 & 2 & 1 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 1/4 \\ 1 \\ 1 \end{pmatrix}.$$

The solution is $\mathbf{x} = (1, 0)^T$, $\lambda_{[1:2]} = (.625, .375)^T$. This is feasible, hence a critical point. Indeed, here it is obviously the minimizer. Note that by complementarity,

$$\mathbf{x}^* = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \boldsymbol{\lambda}^* = \begin{pmatrix} .625 \\ .375 \\ 0 \\ 0 \end{pmatrix}.$$

◆

Example 7.5

For the trust region method (trying to globalize methods for unconstrained optimization) we had a quadratic objective function with one (nonlinear) constraint:

$$\begin{aligned}\min \quad & m(\mathbf{p}) = \frac{1}{2} \mathbf{p}^T B \mathbf{p} + \mathbf{g}^T \mathbf{p} \\ \text{s.t.} \quad & \frac{1}{2} \mathbf{p}^T \mathbf{p} \leq \frac{1}{2} \Delta^2.\end{aligned}$$

The Lagrangian here is $\mathcal{L} = \frac{1}{2}\mathbf{p}^T B\mathbf{p} + \mathbf{g}^T \mathbf{p} - \frac{\lambda}{2}(\Delta^2 - \mathbf{p}^T \mathbf{p})$. The KKT conditions yield

- $B\mathbf{p} + \lambda\mathbf{p} = -\mathbf{g}$,
- either $\|\mathbf{p}\| \leq \Delta$ and $\lambda = 0$, or $\|\mathbf{p}\| = \Delta$ and $\lambda \geq 0$.

This is what was claimed way back in Section 4.2. ◆

Example 7.6

The linear programming problem (LP) in *primal form* is written as

$$\begin{aligned} \min \quad & f(\mathbf{x}) = \mathbf{c}^T \mathbf{x}, \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b}, \\ & \mathbf{x} \geq \mathbf{0}. \end{aligned} \tag{7.10}$$

Thus, both objective function and constraints are linear.

Note that this notation does not fit exactly into our previous (and future) conventions: \mathbf{c} is just a constant (“cost”) vector and the inequality constraints are separated from the matrix notation. We assume that $A \in \mathbb{R}^{m \times n}$ has a full row rank m .

The Lagrangian can be written as $\mathcal{L} = \mathbf{c}^T \mathbf{x} - \sum_{i=1}^m \lambda_i (\sum_{j=1}^n a_{ij} x_j - b_i) - \sum_{j=1}^n s_j x_j$, i.e. we use both λ_i and s_j (s for ‘slack’) to denote the multipliers. The KKT conditions therefore read

$$\begin{aligned} \mathbf{c} - A^T \boldsymbol{\lambda} - \mathbf{s} &= \mathbf{0} \\ A\mathbf{x} &= \mathbf{b} \\ s_i x_i &= 0, \quad i = 1, \dots, n \\ \mathbf{x} \geq \mathbf{0}, \mathbf{s} &\geq \mathbf{0}. \end{aligned}$$

These conditions are necessary as well as sufficient for a minimum. They are called *complementarity slackness* conditions, and express the LP problem without any objective function to minimize or maximize. This is also called the *primal-dual form*. ◆

Example 7.7

All example above have the LICQ satisfied. Is that always the case?!

Unfortunately, no. In fact, it is easy to generate examples where LICQ does not hold: Consider a constrained minimization problem with just one equality constraint, $\hat{c}(\mathbf{x}) = 0$, and assume that there is a solution \mathbf{x}^* and a Lagrange multiplier λ^* such that

$$\mathbf{0} \neq \nabla f(\mathbf{x}^*) = \lambda^* \nabla \hat{c}(\mathbf{x}^*).$$

But now, consider the same minimization problem subject to the constraint

$$c(\mathbf{x}) = \hat{c}^2(\mathbf{x}) = 0.$$

We have at any feasible point, particularly at \mathbf{x}^* ,

$$\nabla c(\mathbf{x}^*) = 2\hat{c}(\mathbf{x}^*)\nabla \hat{c}(\mathbf{x}^*) = \mathbf{0}.$$

Hence, there is no Lagrange multiplier for the same problem with c replacing \hat{c} . Indeed, A_* is a row of zeros, hence its nullspace is \mathbb{R}^n , and the only vector orthogonal to all of \mathbb{R}^n is the zero vector.

One can argue that formulating the problem above with c instead of \hat{c} is stupid. Indeed, when formulating a constrained optimization problem in practice, a user should be careful not to introduce extra difficulties via a careless formulation. There is no method, let alone software package, that solves all such problems! But there are less obvious cases where LICQ does not hold.

Here is an example adopted from J. Nocedal's lecture at UBC, Sept. 22, 2003. Consider the following:

$$\begin{aligned} \min \quad & (x_1 - 1)^2 + (x_2 - 1)^2, \\ \text{s.t.} \quad & \mathbf{x} \geq \mathbf{0}, \quad x_1 x_2 = 0. \end{aligned}$$

The unconstrained minimum is at $\hat{\mathbf{x}} = (1, 1)^T$ and the level sets are circles about $\hat{\mathbf{x}}$. The constraints say that the minimum is on one of the axes, so clearly, by inspection there are two minima, at $\mathbf{x}^a = (0, 1)^T$ and at $\mathbf{x}^b = (1, 0)^T$.

The constraint gradients are

$$\nabla c_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \nabla c_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \nabla c_3 = \begin{pmatrix} x_2 \\ x_1 \end{pmatrix}.$$

Thus, at \mathbf{x}^a the active constraint matrix is

$$[\nabla c_1, \nabla c_3] = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix},$$

which is singular, and at \mathbf{x}^b it is

$$[\nabla c_2, \nabla c_3] = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix},$$

which is also singular! The LICQ condition does not hold at any of the minimum points.

(Exercise: form the KKT conditions and see what is happening there: perhaps there are still Lagrange multipliers?)

Is this just another unique, sick example in two variables? Unfortunately, there are applications that do give rise to sensible constrained minimization problems where LICQ does not hold. Usual software would not work for such applications.

◆

7.1.2 Second order conditions

Recall that in the unconstrained case we examined the Hessian $\nabla^2 f(\mathbf{x}^*)$ at critical points, i.e., points where $\nabla f(\mathbf{x}^*) = \mathbf{0}$. A necessary condition for a minimum was that the Hessian be positive semi-definite, whereas a positive definite Hessian was sufficient.

In the constrained case it appears by (7.7a) that the condition that the gradient of f vanish is replaced by the condition that the gradient of the Lagrangian \mathcal{L} with respect to \mathbf{x} (not with respect to $\boldsymbol{\lambda}$) vanish. Perhaps for the second order conditions we should be looking at the Hessian with respect to \mathbf{x} of the Lagrangian?

Indeed it turns out that

$$\mathcal{L}_{\mathbf{xx}}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \nabla^2 f(\mathbf{x}^*) - \sum_{i \in \mathcal{A}(\mathbf{x}^*)} \lambda_i^* \nabla^2 c_i(\mathbf{x}^*) \quad (7.11)$$

replaces $\nabla^2 f(\mathbf{x}^*)$ in seeking necessary and sufficient conditions for a minimum. But, unlike the unconstrained case, $\mathcal{L}_{\mathbf{xx}}(\mathbf{x}^*)$ need not be positive definite for a sufficient condition. Assume LICQ and recall the basic Taylor expansion about a candidate for a minimum: for $\|\mathbf{p}\|$ arbitrarily small we write

$$f(\mathbf{x}^* + \mathbf{p}) = f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)^T \mathbf{p} + \frac{1}{2} \mathbf{p}^T \nabla^2 f(\mathbf{x}^*) \mathbf{p} + \dots \geq f(\mathbf{x}^*).$$

The term involving the Hessian need only be considered where the previous term vanishes, i.e. for directions satisfying $\nabla f(\mathbf{x}^*)^T \mathbf{p} = 0$. By (7.4)-(7.3) these are the directions belonging to the tangent space of the constraint

manifold, $M = \ker(A_*)$. We then “guess” that $\mathcal{L}_{\mathbf{xx}}(\mathbf{x}^*)$ need only be positive definite on $\ker(A_*)$, not all of \mathbb{R}^n .

The general case turns out to be a little more complicated. Unless there is *strict complementarity*, i.e., unless for each $i \in \mathcal{A}(\mathbf{x}^*)$ either $c_i(\mathbf{x}^*) = 0$ or $\lambda_i^* = 0$ but *not both*, we must expand the subspace $\ker(A_*)$ to the set $F(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ defined by

$$\mathbf{w} \in F(\mathbf{x}^*, \boldsymbol{\lambda}^*) \text{ iff } \begin{cases} \nabla c_i(\mathbf{x}^*)^T \mathbf{w} = 0 & \text{all } i \in \mathcal{E}; \\ \nabla c_i(\mathbf{x}^*)^T \mathbf{w} = 0 & \text{all } i \in \mathcal{A}(\mathbf{x}^*) \cap \mathcal{I}, \lambda_i^* > 0; \\ \nabla c_i(\mathbf{x}^*)^T \mathbf{w} \geq 0 & \text{all } i \in \mathcal{A}(\mathbf{x}^*) \cap \mathcal{I}, \lambda_i^* = 0. \end{cases} \quad (7.12)$$

Clearly, if the third case in (7.12) does not occur then $F = \ker(A_*)$.

- **Second-order necessary conditions:**

Suppose that \mathbf{x}^* is a local minimum of (7.1) and that the LICQ condition is satisfied. Let $\boldsymbol{\lambda}^*$ be such that the KKT conditions (7.7) are satisfied and define F by (7.12). Then

$$\mathbf{w}^T \nabla_{\mathbf{xx}} \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) \mathbf{w} \geq 0, \quad \forall \mathbf{w} \in F(\mathbf{x}^*, \boldsymbol{\lambda}^*). \quad (7.13)$$

- **Second-order sufficient conditions:**

Suppose that \mathbf{x}^* is a feasible point of (7.1b) and that there is $\boldsymbol{\lambda}^*$ such that the KKT conditions (7.7) are satisfied. If also

$$\mathbf{w}^T \nabla_{\mathbf{xx}} \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) \mathbf{w} > 0, \quad \forall \mathbf{w} \in F(\mathbf{x}^*, \boldsymbol{\lambda}^*), \mathbf{w} \neq \mathbf{0}, \quad (7.14)$$

then \mathbf{x}^* is a strict local solution of the constrained minimization problem (7.1).

These conditions are proved in [23]. Here we proceed directly to two short examples.

Example 7.8

The very simple QP

$$\begin{aligned} \min \quad & \frac{1}{2}x_2^2 - (x_1 + x_2) \\ \text{s.t.} \quad & x_1 = 3.17 \end{aligned}$$

has the Lagrangian $\mathcal{L} = \frac{1}{2}x_2^2 - (x_1 + x_2) - \lambda_1(x_1 - 3.17)$. This is a special case of Example 7.3. The KKT conditions read

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \lambda_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 3.17 \end{pmatrix}.$$

The solution is obviously $x_1 = 3.17$, $x_2 = 1$, $\lambda_1 = 1$.

For a problem with linear constraints, $\mathcal{L}_{\mathbf{xx}} = \nabla^2 f$. For a QP, the Hessian is constant as well, and so

$$\mathcal{L}_{\mathbf{xx}} = H = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

The Hessian of the Lagrangian is thus singular, although the KKT matrix K is clearly nonsingular.

Do the sufficient conditions for a minimum hold? Here $A_* = \begin{pmatrix} 1 & 0 \end{pmatrix}$, so

$$\ker(A_*) = \left\{ \begin{pmatrix} 0 \\ \alpha \end{pmatrix}, \forall \alpha \in \mathbb{R} \right\}.$$

But for any $\alpha \neq 0$,

$$\begin{pmatrix} 0 & \alpha \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ \alpha \end{pmatrix} = \alpha^2 > 0,$$

so the sufficient condition for a minimum holds.

Note, by the way, that the eigenvalues of the KKT matrix are $1, 1, -1$. This matrix is symmetric and *indefinite*!



Example 7.9

Consider the QP with inequality constraints of Example 7.4. Here, $\mathcal{L}_{\mathbf{xx}} = H = 2I$ is positive definite. This of course is sufficient to ensure that (7.14) holds, because if the quadratic form is always positive then it is particularly positive on F . Thus, any point satisfying the KKT conditions is a minimizer for this example.



7.2 An overview of algorithms

Generally speaking, equality constraints have a more algebraic flavour to them (searching in a reduced space, on a constraint manifold, etc.), whereas inequality constraints introduce an additional combinatorial flavour to the problem (namely, which constraints are active at the solution and which are not). It is possible to distinguish between two general approaches for constrained optimization:

- **Active set**

Assuming that the unconstrained minimum of f is not in the interior of the feasibility set Ω , our solution must be on the boundary $\partial\Omega$. Active set methods thus search for the optimum along the boundary. For inequality constraints there are active set methods where we keep track of $\mathcal{A}(\mathbf{x}_k)$, moving constraints in and out of the active set as we go downhill along the boundary.

- **The rest**

The optimal solution is approached in an iterative method, either from within the feasible region Ω (these are *interior point methods*) or, more generally, by a method which may use infeasible points as well but does not move along the boundary.

Methods of the second type include those where the objective function is modified sequentially: for each such modification the corresponding unconstrained minimization problem is solved.

- **Penalty methods**

These methods, like steepest descent, are a favourite among common people due to their simplicity. Consider, for instance,

$$\min_{\mathbf{x}} \phi(\mathbf{x}, \mu) = f(\mathbf{x}) + \frac{1}{2\mu} \sum_{i \in \mathcal{E}} c_i^2(\mathbf{x})$$

where $\mu > 0$ is a parameter. This makes sense for problems with only equality constraints. One then solves a sequence of such problems for decreasing values of μ , $\mu \downarrow 0$, using $\mathbf{x}(\mu_{k-1})$ to construct the initial iterate for the unconstrained iteration for $\mathbf{x}(\mu_k)$.

- **Barrier methods**

These are interior point methods. Starting from a point in the feasible set, a sequence of unconstrained problems is solved such that at each stage the objective function is modified to ensure that the solution on the boundary is approached from within Ω . For instance, consider

$$\min_{\mathbf{x}} \phi(\mathbf{x}, \mu) = f(\mathbf{x}) - \mu \sum_{i \in \mathcal{I}} \log c_i(\mathbf{x})$$

where $\mu \downarrow 0$.

- **Augmented Lagrangian**

For a problem with only equality constraints, consider the following variation on a penalty method:

$$\min_{\mathbf{x}} \phi(\mathbf{x}, \boldsymbol{\lambda}, \mu) = f(\mathbf{x}) - \sum_{i \in \mathcal{E}} \lambda_i c_i(\mathbf{x}) + \frac{1}{2\mu} \sum_{i \in \mathcal{E}} c_i^2(\mathbf{x}).$$

Given estimates $\boldsymbol{\lambda}_k, \mu_k$, we solve the unconstrained minimization problem for $\mathbf{x} = \mathbf{x}_{k+1}$, then update the multipliers to $\boldsymbol{\lambda}_{k+1}, \mu_{k+1}$.

The most popular method in practice for general codes is (still?) **sequential quadratic programming** (SQP). At each iteration one solves a quadratic program (QP)

$$\begin{aligned} \min_{\mathbf{p}} \quad & \frac{1}{2} \mathbf{p}^T W_k \mathbf{p} + \nabla f(\mathbf{x}_k)^T \mathbf{p}, \\ & c_i(\mathbf{x}_k) + \nabla c_i(\mathbf{x}_k)^T \mathbf{p} = 0, \quad i \in \mathcal{E}, \quad c_i(\mathbf{x}_k) + \nabla c_i(\mathbf{x}_k)^T \mathbf{p} \geq 0, \quad i \in \mathcal{I}, \end{aligned}$$

yielding a direction \mathbf{p}_k at iterate $(\mathbf{x}_k, \boldsymbol{\lambda}_k)$. The objective function here approximates the Lagrangian \mathcal{L} near $(\mathbf{x}_k, \boldsymbol{\lambda}_k)$ and the linear constraints are the linearization of the original constraints at the current iterate. An active set method is used to solve the QP with inequality constraints.

7.3 Eliminating variables

Consider the general minimization problem (7.1) with equality constraints: \mathcal{I} is empty and $\mathcal{E} = \{1, 2, \dots, m\}$. It is tempting to use the constraints (7.1b) to eliminate some of the variables in terms of others, substitute in (7.1a) and obtain an unconstrained optimization problem.

7.3.1 Nonlinear equality constraints

For nonlinear constraints the above procedure can be feasible and even advantageous on occasions, but it is often both tricky and hard to generalize or automate.

Example 7.10

This cute example is due to R. Fletcher [14]. Consider

$$\begin{aligned} \min \quad & x_1^2 + x_2^2, \\ \text{s.t.} \quad & (x_1 - 1)^3 = x_2^2. \end{aligned}$$

It is not difficult to see by inspection (e.g. use MATLAB to plot the constraint curve and note that the objective function level sets are concentric circles about the origin) that there is one minimum point, $\mathbf{x}^* = (1, 0)^T$.

Now eliminate x_2^2 in the obvious fashion, i.e., consider the unconstrained problem

$$\min_{x_1} x_1^2 + (x_1 - 3)^3.$$

However, here letting $x_1 \downarrow -\infty$ yields an unbounded solution!

The trouble is that the hidden constraint $(x_1 - 1)^3 \geq 0$ has been ignored in the heat of simplification. ◆

So, in general it is better not to eliminate nonlinear equality constraints. Rather, consider iterative methods which linearize these constraints at each iteration, then *perhaps* eliminate within the iteration to solve the linearized problem.

7.3.2 Linear equality constraints

Linear equality constraints can be written as

$$A\mathbf{x} = \mathbf{b}, \tag{7.15}$$

where A is $m \times n$, $m \leq n$. The LICQ assumption is that $\text{rank}(A) = m$. The interesting case is when $m < n$, so let us consider that.

Since A in (7.15) has more columns than rows it is natural to consider eliminating some of the variables in terms of the others. Indeed, since A has a full row rank there must be m linearly independent columns in this matrix. We group a collection of such columns together and call the resulting square, nonsingular matrix, B . To simplify notation, suppose that B is composed of the first m columns of A and call the rest N , i.e.

$$A = (B|N).$$

Writing correspondingly

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_B \\ \mathbf{x}_N \end{pmatrix}$$

we have $\mathbf{b} = A\mathbf{x} = B\mathbf{x}_B + N\mathbf{x}_N$, so

$$\mathbf{x}_B = B^{-1}(\mathbf{b} - N\mathbf{x}_N),$$

and we obtain the unconstrained problem

$$\min_{\mathbf{x}_N \in \mathbb{R}^{n-m}} f \left(\begin{pmatrix} B^{-1}(\mathbf{b} - N\mathbf{x}_N) \\ \mathbf{x}_N \end{pmatrix} \right). \tag{7.16}$$

Just because we *can* do this does not necessarily mean it's always a good idea. But let us continue with the flow and introduce the following notation:

$$Y = \begin{pmatrix} B^{-1} \\ 0 \end{pmatrix}, \quad Z = \begin{pmatrix} -B^{-1}N \\ I \end{pmatrix}. \quad (7.17)$$

The $n \times m$ matrix Y and the $n \times (n-m)$ matrix Z obviously have full column ranks – in fact, the matrix $(Y|Z)$ is nonsingular.¹ Now, any feasible point \mathbf{x} (i.e., \mathbf{x} satisfying (7.15)) can be written as

$$\mathbf{x} = Y\mathbf{b} + Z\mathbf{x}_N,$$

i.e. \mathbf{x} is written as the sum of a particular solution and a contribution from the nullspace of A .

Indeed,

$$AZ = 0, \quad (7.18)$$

and, since the columns of Z are linearly independent, they form a *basis for the nullspace of A , $\ker(A)$* . See Figure 7.4 to get a feel of matrix dimensions. The second order sufficient conditions for a minimum at a point $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$

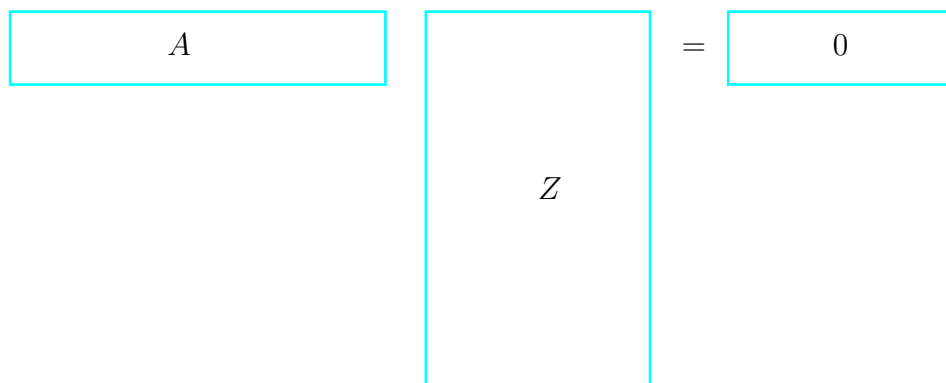


Figure 7.4: $AZ = 0$.

satisfying the KKT conditions can therefore be written as requiring that the *reduced Hessian*, or *projected Hessian*

$$H_r = Z^T \mathcal{L}_{\mathbf{xx}}(\mathbf{x}^*, \boldsymbol{\lambda}^*) Z \quad (7.19)$$

¹Note $A(Y|Z) = (I|0)$. If B is composed of columns of A other than the first m then there is an $n \times n$ permutation matrix P such that $AP = (B|N)$. Then $APP^T(Y|Z) = (I|0)$, so P^T need be applied to (7.17), corresponding to a permutation of the rows of Y and Z .

be symmetric positive definite (s.p.d.). For the QP problem (7.8) the Lagrangian is

$$\mathcal{L} = \frac{1}{2} \mathbf{x}^T H \mathbf{x} - \mathbf{d}^T \mathbf{x} - \boldsymbol{\lambda}^T (\mathbf{b} - A \mathbf{x}), \quad (7.20a)$$

so the reduced Hessian is the $(n - m) \times (n - m)$ matrix

$$H_r = Z^T H Z. \quad (7.20b)$$

The condition that H_r of (7.20b) be s.p.d. is equivalent to requiring the sufficient condition for a minimum to hold,

$$\mathbf{w}^T H \mathbf{w} > 0, \quad \forall \mathbf{w} \in \ker(A).$$

The basis formed from the columns of the matrix Z of (7.17) is not always the most stable numerically. For relatively small and dense problems, at least, other decompositions such as QR and SVD provide a more stable option. In particular, consider the singular value decomposition (SVD) of A ,

$$A = U \Sigma V^T,$$

where U is $m \times m$ and orthogonal (i.e. $U^T = U^{-1}$), V is $n \times n$ and orthogonal, and

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \ddots & \vdots & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 & \vdots & \cdots & \vdots \\ 0 & \cdots & 0 & \sigma_m & 0 & \cdots & 0 \end{pmatrix}.$$

Then we can take for Z the last $n - m$ columns of V (it is easy to verify that (7.18) holds). The columns of Z , stably obtained, now form an orthonormal basis for $\ker(A)$.