The idea of the accelerated gradient method, Algorithm 4.24, is to not fully accept the current iterate proposed by steepest descent (with step size $1/L$) but combine it with the old iterate. The parameter γ_k determining this convex combination arises from optimality considerations that can be found in [N].

Algorithm 4.24 Accelerated gradient descent for convex functions I

Input: *L-smooth convex function f and starting vector $\mathbf{x}_0$.*

Output: *Vector $\mathbf{y}_k$ approximating minimum: $f(\mathbf{y}_k) \approx f(\mathbf{x}^*)$.*

```
Set $\lambda_0 = 1$ and $\mathbf{y}_0 = \mathbf{x}_0$.
for $k = 0, 1, 2, \dots$ do
    Set $\lambda_{k+1} = \frac{1+\sqrt{1+4\lambda_k^2}}{2}$.
    Set $\gamma_k = \frac{1-\lambda_k}{\lambda_{k+1}}$.
    Set $\mathbf{y}_{k+1} = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$.
    Set $\mathbf{x}_{k+1} = (1-\gamma_k)\mathbf{y}_{k+1} + \gamma_k \mathbf{y}_k$.
end for
```

For analysing the convergence of Algorithm 4.24, we need the following important result on smooth convex functions.

Lemma 4.25 *For an L-smooth convex function f, we have*

$$|f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x})| \le \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$$

for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Proof. By Taylor expansion with an integral representation of the remainder term, we obtain

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(x)^T(\mathbf{y}-\mathbf{x}) + \int_0^1 \big(\nabla f(\mathbf{x} + \tau(\mathbf{y}-\mathbf{x}) - \nabla f(\mathbf{x})\big)^T(\mathbf{y}-\mathbf{x})\, \mathrm{d}\tau.$$

By the Cauchy-Schwarz inequality, we therefore get

$$\begin{aligned} &|f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(x)^T(\mathbf{y}-\mathbf{x})| \\ \le &\int_0^1 \big\|\nabla f(\mathbf{x} + \tau(\mathbf{y}-\mathbf{x}) - \nabla f(\mathbf{x})\big\|_2 \|\mathbf{y}-\mathbf{x}\|_2 \,\mathrm{d}\tau \le \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2. \end{aligned}$$

□

Theorem 4.26 *For an L-smooth convex function f, the iterates produced by Algorithm 4.24 satisfy*

$$f(\mathbf{y}_{k+1}) - f(\mathbf{x}^*) \le \frac{2L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{(k+2)^2}.$$

Proof. By Lemma 4.25,

$$\begin{aligned} f(\mathbf{y}_{k+1}) - f(\mathbf{x}_k) &= f\Big(\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)\Big) - f(\mathbf{x}_k) \\ &\le \nabla f(\mathbf{x}_k)^T\Big(- \frac{1}{L}\nabla f(\mathbf{x}_k)\Big) + \frac{L}{2}\Big\|\frac{1}{L}\nabla f(\mathbf{x}_k)\Big\|_2^2 \\ &\le -\frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|_2^2. \end{aligned}$$

This implies, using Lemma 4.19.1,

$$\begin{aligned} f(\mathbf{y}_{k+1}) - f(\mathbf{y}_k) &\le f(\mathbf{y}_{k+1}) - f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x}_k - \mathbf{y}_k) \\ &\le -\frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|_2^2 + \nabla f(\mathbf{x}_k)^T(\mathbf{x}_k - \mathbf{y}_k) \end{aligned}$$

and, similarly,

$$f(\mathbf{y}_{k+1}) - f(\mathbf{x}^*) \le -\frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|_2^2 + \nabla f(\mathbf{x}_k)^T(\mathbf{x}_k - \mathbf{x}^*).$$

Setting $\delta_k = f(\mathbf{y}_k) - f(\mathbf{x}^*)$, a convex combination of these two inequalities gives

$$\begin{aligned} \lambda_k\delta_{k+1} - (\lambda_k - 1)\delta_k &= (\lambda_k - 1)(f(\mathbf{y}_{k+1}) - f(\mathbf{y}_k)) + f(\mathbf{y}_{k+1}) - f(\mathbf{x}^*) \\ &\le -\frac{\lambda_k}{2L}\|\nabla f(\mathbf{x}_k)\|_2^2 + \nabla f(\mathbf{x}_k)^T\big(\lambda_k\mathbf{x}_k - (\lambda_k - 1)\mathbf{y}_k - \mathbf{x}^*\big). \end{aligned}$$

By definition, λ_k is defined from λ_{k-1} by satisfying the quadratic equation $\lambda_{k-1}^2 = \lambda_k^2 - \lambda_k$. Multiplying the last inequality with λ_k, this gives

$$\begin{aligned} &\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k \\ &\le -\frac{\lambda_k^2}{2L}\|\nabla f(\mathbf{x}_k)\|_2^2 + \lambda_k\nabla f(\mathbf{x}_k)^T\big(\lambda_k\mathbf{x}_k - (\lambda_k - 1)\mathbf{y}_k - \mathbf{x}^*\big) \\ &= -\frac{L}{2}\Big(\|\lambda_k(\mathbf{y}_{k+1} - \mathbf{x}_k)\|_2^2 + 2\lambda_k(\mathbf{y}_{k+1} - \mathbf{x}_k)^T(\lambda_k\mathbf{x}_k - (\lambda_k - 1)\mathbf{y}_k - \mathbf{x}^*)\Big). \end{aligned}$$

By basic manipulation, one can see that

$$\begin{aligned} &\|\lambda_k(\mathbf{y}_{k+1} - \mathbf{x}_k)\|_2^2 + 2\lambda_k(\mathbf{y}_{k+1} - \mathbf{x}_k)^T(\lambda_k\mathbf{x}_k - (\lambda_k - 1)\mathbf{y}_k - \mathbf{x}^* \\ &= \|\lambda_k\mathbf{y}_{k+1} - (\lambda_k - 1)\mathbf{y}_k - \mathbf{x}^*\|_2^2 - \|\lambda_k\mathbf{x}_k - (\lambda_k - 1)\mathbf{y}_k - \mathbf{x}^*\|_2^2 \\ &= \|\lambda_{k+1}\mathbf{x}_{k+1} - (\lambda_{k+1} - 1)\mathbf{y}_{k+1} - \mathbf{x}^*\|_2^2 - \|\lambda_k\mathbf{x}_k - (\lambda_k - 1)\mathbf{y}_k - \mathbf{x}^*\|_2^2 \\ &= \|\mathbf{u}_{k+1}\|_2^2 - \|\mathbf{u}_k\|_2^2, \end{aligned}$$

where we set $\mathbf{u}_k := \lambda_k\mathbf{x}_k - (\lambda_k - 1)\mathbf{y}_k - \mathbf{x}^*$. In summary, we have

$$\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k \le \frac{L}{2}\big(\|\mathbf{u}_k\|_2^2 - \|\mathbf{u}_{k+1}\|_2^2\big).$$

Formally setting $\lambda_{-1} = 0$ and summing up this inequality for $0, \dots, k$ yields

$$\lambda_k^2\delta_{k+1} \le \frac{L}{2}\big(\|\mathbf{u}_0\|_2^2 - \|\mathbf{u}_{k+1}\|_2^2\big) \le \frac{L}{2}\|\mathbf{u}_0\|_2^2 = \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2.$$

We will now show by induction that $\lambda_k \geq (k+2)/2$. The ineqality is obviously satisfied for $k = 0$. The induction step follows from

$$\lambda_{k+1} = \frac{1 + \sqrt{1 + 4\lambda_k^2}}{2} \geq \frac{1 + \sqrt{1 + (k+2)^2}}{2} \geq \frac{k+3}{2}.$$

This concludes the proof. □

Remark 4.27 When implementing Algorithm 4.24, it is important to supply a (reasonably tight) upper bound for the Lipschitz constant L. In specific cases, such a bound may be available, but in general it needs to be estimated. We refer to [P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. 2008] for a backtracking procedure for estimating L. ◇

Theorem 4.26 constitutes a significant improvement compared to the expected convergence of standard gradient descent applied to a smooth convex function, see Part 1 of Theorem 4.20. It would be quite satisfactory if we also achieved an improvement for *μ-strongly* convex functions compared to Part 2 of Theorem 4.20. It turns out that Algorithm 4.24 does *not* achieve this goal. For this purpose, one needs to incoporate (a reasonably tight lower bound of) μ. We refer to [B. O'Donoghue and E. Candès. Adaptive Restart for Accelerated Gradient Schemes. 2012] for a more detailed discussion, which also discusses the estimation of μ. Algorithm 4.28 summarizes the resulting procedure.

Algorithm 4.28 Accelerated gradient descent for convex functions II

Input: *L-smooth μ-strongly convex function f and starting vector $\mathbf{x}_0$.*

Output: *Vector $\mathbf{y}_k$ approximating minimum: $f(\mathbf{y}_k) \approx f(\mathbf{x}^*)$.*

```
Set $\mathbf{y}_0 = \mathbf{x}_0$.
for $k = 0, 1, 2, \dots$ do
    Set $\mathbf{y}_{k+1} = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$.
    Set $\mathbf{x}_{k+1} = \mathbf{y}_{k+1} + \frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}(\mathbf{y}_{k+1} - \mathbf{y}_k)$.
end for
```

Theorem 4.29 *For an L-smooth μ-strongly convex function f, the iterates produced by Algorithm 4.24 satisfy*

$$\begin{aligned} f(\mathbf{y}_k) - f(\mathbf{x}^*) &\leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \left(f(\mathbf{x}_0) - f^* + \frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2\right) \\ &\leq L\left(1 - \sqrt{\frac{\mu}{L}}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2. \end{aligned}$$

Proof. The first inequality is Theorem 2.2.3 in [N]. The second inequality follows from the first using Lemma 4.25. □

Chapter 5

Constrained optimization

A **constrained optimization problem** takes the form

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \quad \text{subject to} \quad g(\mathbf{x}) \geq 0,\ h(\mathbf{x}) = 0, \tag{5.1}$$

with $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$, $h : \mathbb{R}^n \to \mathbb{R}^p$.

The conditions $g(\mathbf{x}) \geq 0$ and $h(\mathbf{x}) = 0$ impose constraints on $\mathbf{x}$, which are called **inequality constraints** and **equality constraints**, respectively. By defining the **feasible set**

$$\Omega = \{\mathbf{x} \in \mathbb{R}^n \mid g(\mathbf{x}) \geq 0,\ h(\mathbf{x}) = 0\} \tag{5.2}$$

we can rewrite (5.1) as

$$\min_{\mathbf{x}\in\Omega} f(\mathbf{x}),$$

which makes it visually closer to an unconstrained optimization problem.

5.1 Fundamentals

The goal of this section is to develop some theoretical understanding of the constrained optimization problem (5.1).

There may be the naive hope that imposing constraints makes it easier to solve an optimization problem, for example, because the constraints exclude local minima that are not global minima. Nearly always, the opposite is the case: imposing constraints makes optimization much harder! For example[3], consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^2} (x_2 + 100)^2 + 0.01x_1^2 \quad \text{subject to} \quad x_2 - \cos x_1 \geq 0. \tag{5.3}$$

Without the constraint, the (unique) local minimum is given by $(0, -100)^T$. With the constraints, there are local solutions near the points $\mathbf{x}^{(k)} = (k\pi, -1)^T$ for all odd integers k, see Figure 5.1.

[3] This and the following examples are taken from [NW].

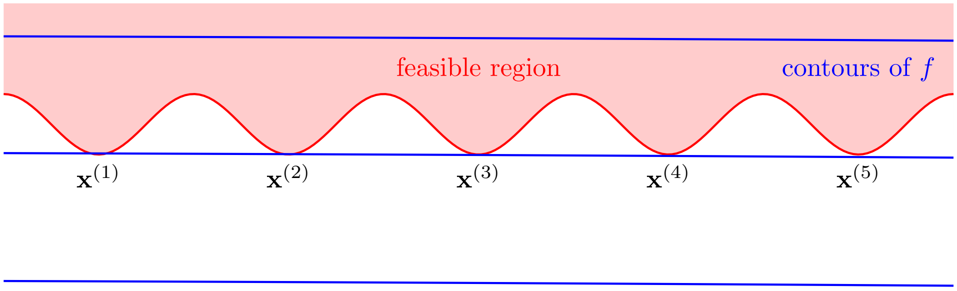

Figure 5.1. *Local solutions of* (5.3).

Note that the concept of local solutions corresponds to the notion of local minima from Chapter 4 restricted to the feasible set Ω defined in (5.2). A point $\mathbf{x}^* \in \Omega$ is called a **local solution** of (5.1) if there is a neighborhood $\mathcal{N}$ such that $f(\mathbf{x}^*) \le f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{N} \cap \Omega$. Similarly, $\mathbf{x}^*$ is called a **strict local solution** if $f(\mathbf{x}^*) < f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{N} \cap \Omega$ with $\mathbf{x} \neq \mathbf{x}^*$. The stronger concept of **isolated local solution** requires that $\mathbf{x}^*$ is the only local solution in $\mathcal{N} \cap \Omega$.

Throughout this chapter the following concepts will play a central role.

Definition 5.1 *A point* $\mathbf{x} \in \mathbb{R}^n$ *is called* **feasible** *if* $\mathbf{x} \in \Omega$. *The* **active set** $\mathcal{A}(x)$ *at a feasible* $\mathbf{x}$ *is defined as*

$$\mathcal{A}(\mathbf{x}) = \big\{ i \in \{1, \dots, m\} \mid g_i(\mathbf{x}) = 0 \big\}.$$

5.1.1 Two simple examples

Before deriving general optimality conditions, it is extremely helpful to look at some simple examples first.

Example 5.2 Consider a problem in two variables with a single equality constraint:

$$\min_{\mathbf{x} \in \mathbb{R}^2} x_1 + x_2 \quad \text{subject to} \quad x_1^2 + x_2^2 - 2 = 0. \tag{5.4}$$

The feasible set is a circle of radius $\sqrt{2}$ centered at 0. The solution of (5.4) is given by $(-1, -1)^T$. Geometrically, this can be seen by choosing a line $x_1 + x_2 \equiv \text{const}$ such that it touches the circle in the bottom left corner. Another (slightly less obvious) way to see this is to choose any other point on the circle and observe that we can always move along the circle (i.e., we stay feasible) and decrease the target function value at the same time. From Figure 5.2, we see that the constraint normal ∇h (with $h(\mathbf{x}) = x_1^2 + x_2^2 - 2$ is parallel to ∇f (with $f(\mathbf{x}) = x_1 + x_2$) at the solution $\mathbf{x}^*$. In other words, there is a scalar λ^* such that

$$\nabla f(\mathbf{x}^*) = \lambda^* \nabla h(\mathbf{x}^*). \tag{5.5}$$

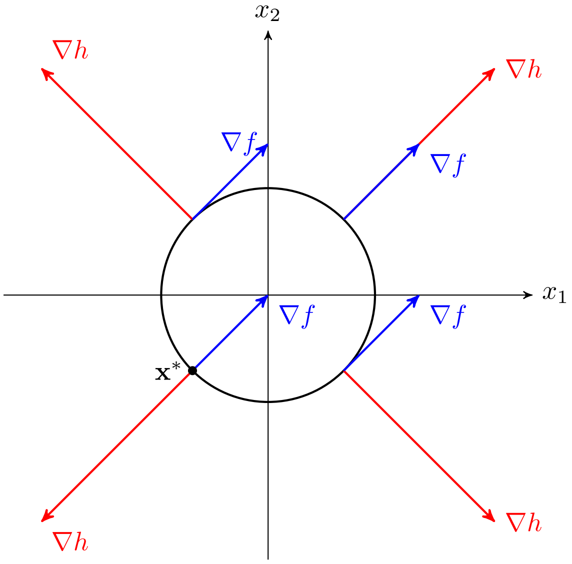

Figure 5.2. *Constraint and target function gradients for* (5.4)*.*

By introducing the *Lagrangian function*

$$\mathcal{L}(\mathbf{x},\lambda) = f(\mathbf{x}) - \lambda h(\mathbf{x}),$$

we can state (5.5) equivalently as follows: At the solution $\mathbf{x}^*$ there is a scalar λ^* such that

$$\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^*,\lambda^*) = 0. \tag{5.6}$$

This suggests that we can solve equality-constrained optimization problem by adapting the methods from Chapter 4 to find stationary points of $\mathcal{L}$. Of course, having a point $\mathbf{x}$ that satisfies (5.6) does in general *not* imply that $\mathbf{x}$ is a local solution. For example, (5.6) is also satisfied at $\mathbf{x} = (1,1)^T$ with $\lambda = 1/2$. ◇

Example 5.3 In this example, we turn the equality constraint in (5.4) into an inequality constraint:

$$\min_{\mathbf{x}\in\mathbb{R}^2} x_1 + x_2 \quad \text{subject to} \quad 2 - x_1^2 - x_2^2 \geq 0. \tag{5.7}$$

Again, the solution is given by $(-1,-1)^T$. The feasible set is now the circle we had before *and* its interiour. Note that the constraint normal ∇g (with $g(\mathbf{x}) = 2 - x_1^2 - x_2^2$) now points into the interior of the feasible set at every point on the boundary, see Figure 5.3. Suppose now that $\mathbf{x}$ is feasible and we are looking for a step $\mathbf{x} \mapsto \mathbf{x} + \mathbf{s}$ such that f is decreased and the inequality constraint is still met. In first order, these two conditions on $\mathbf{s}$ amount to

$$\nabla f(\mathbf{x})^T\mathbf{s} < 0, \qquad 0 \leq g(\mathbf{x}+\mathbf{s}) \approx g(\mathbf{x}) + \nabla g(\mathbf{x})^T\mathbf{s}. \tag{5.8}$$

In order to find such a step we have to discriminate between two cases.

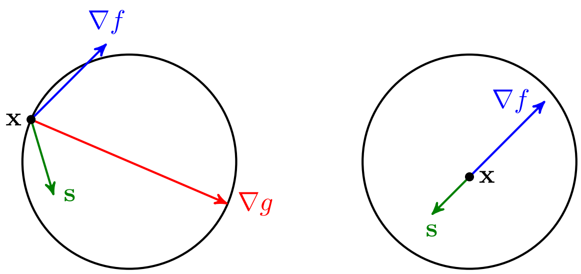

Figure 5.3. *Two feasible points* $\mathbf{x}$ *for* (5.7) *with possible improvement directions* $\mathbf{s}$.

Case of inactive g**.** If $\mathbf{x}$ is in the interior of the circle, we have $g(\mathbf{x}) > 0$. Consequently, any sufficiently small step $\mathbf{s}$ will guarantee the feasibility of $\mathbf{x}+\mathbf{s}$. Provided that $\nabla f(\mathbf{x}) \neq 0$, *any* step

$$\mathbf{s} = -\alpha \nabla f(\mathbf{x})$$

will satisfy (5.8) with $\alpha > 0$ sufficiently small.

Case of active g**.** If $\mathbf{x}$ is on the boundary of the circle, we have $g(\mathbf{x}) = 0$. The conditions (5.8) then become

$$\nabla f(\mathbf{x})^T \mathbf{s} < 0, \qquad \nabla g(\mathbf{x})^T \mathbf{s} \geq 0. \tag{5.9}$$

Both conditions define open/closed half-spaces of $\mathbb{R}^2$, see Figure 5.4. Their intersection is non-empty, unless $\nabla f(\mathbf{x})$ and $\nabla g(\mathbf{x})$ point in the same direction. The latter happens when

$$\nabla f(\mathbf{x}) = \alpha \nabla g(\mathbf{x}), \qquad \alpha \geq 0. \tag{5.10}$$

The optimality conditions for both cases, g inactive and g active, can again be summarized by using the Lagrangian

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x}).$$

At some point $\mathbf{x}^*$ there is *no* feasible step in the sense of conditions (5.8) if

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \lambda^*) = 0 \tag{5.11}$$

for some $\lambda^* \geq 0$ satisfying

$$\lambda^* \cdot g(\mathbf{x}^*) = 0. \tag{5.12}$$

A condition of the form (5.12) is called *complementarity* condition. If $g(\mathbf{x}^*) \neq 0$ (inactive g), it implies that $\lambda^* = 0$ and hence (5.11) comes down to the definition $\nabla f(\mathbf{x}^*) = 0$ of a stationary point. If $g(\mathbf{x}^*) = 0$ (active g), (5.12) is satisfied for any $\lambda^* \geq 0$. In this case, (5.11) is identical with (5.10). ◇

In Example 5.3, we have seen for a single inequality constraint that it is important to discriminate between the two situations when the constraint is active and when it is inactive. This illustrates the importance of active sets from Definition 5.1.

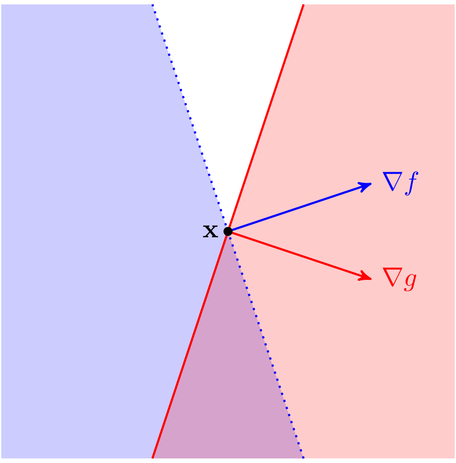

Figure 5.4. *Geometric illustration of* (5.9).

5.1.2 First-order necessary conditions

To develop optimality conditions for constrained optimization, we need to take the geometry of Ω induced by the constraints into account.

Definition 5.4 *The* **tangent cone** *of a set* $\Omega \subset \mathbb{R}^n$ *at* $\mathbf{x} \in \Omega$ *is defined as*

$$T_\Omega(\mathbf{x}) = \left\{ \mathbf{d} \in \mathbb{R}^n \mid \exists \eta_k > 0, \mathbf{x}_k \in \Omega : \ \eta_k \overset{k\to\infty}{\to} 0, \ \mathbf{x}_k \overset{k\to\infty}{\to} \mathbf{x}, \ \frac{1}{\eta_k}(\mathbf{x}_k - \mathbf{x}) \overset{k\to\infty}{\to} \mathbf{d} \right\}.$$

Put in words, Definition 5.4 means that $\mathbf{d}$ is in the tangent cone if it can be written as the limit of finite difference quotients $\frac{\mathbf{x}_k - \mathbf{x}}{\eta_k}$ along the feasible set. It is an easy exercise to see that $T_\Omega(\mathbf{x})$ is indeed a cone.[4]

Definition 5.4 directly leads to the following necessary condition.

Theorem 5.5 *If* $\mathbf{x}^*$ *is a local solution of* (5.1) *then*

$$\nabla f(\mathbf{x}^*)^T \mathbf{d} \geq 0 \qquad \forall \mathbf{d} \in T_\Omega(\mathbf{x}^*). \tag{5.13}$$

Proof. Consider $\mathbf{d} \in T_\Omega(\mathbf{x}^*)$. Then there are sequences $\mathbf{x}_k \to \mathbf{x}^*$ and $\eta_k > 0$ such that $\mathbf{d}_k := \frac{1}{\eta_k}(\mathbf{x}_k - \mathbf{x}^*) \to \mathbf{d}$. Since $\mathbf{x}^*$ is a local solution, we have $f(\mathbf{x}_k) - f(\mathbf{x}^*) \geq 0$ for sufficiently large k. Hence, it follows from the Taylor expansion that

$$\begin{aligned} 0 &\leq \frac{1}{\eta_k}(f(\mathbf{x}_k) - f(\mathbf{x}^*)) = \frac{1}{\eta_k}\nabla f(\mathbf{x}^*)^T(\mathbf{x}_k - \mathbf{x}^*) + \frac{1}{\eta_k} o(\|\mathbf{x}_k - \mathbf{x}^*\|) \\ &= \nabla f(\mathbf{x}^*)^T \mathbf{d}_k + o(1) \overset{k\to\infty}{\to} \nabla f(\mathbf{x}^*)^T \mathbf{d}, \end{aligned}$$

which completes the proof. □

[4] Recall that a set K is called a cone if $k \in K$ implies $\lambda k \in K$ for all $\lambda > 0$.

Theorem 5.5 is mathematically very elegant: The condition (5.13) is brief and does not depend on the particular parametrization of $T_\Omega(\mathbf{x}^*)$. Unfortunately, this implicit nature also makes it very hard to work with (5.13). It would be more practical to have conditions that are explicit in g and h. For this purpose, we need to linearize Definition 5.4. For the rest of the section we assume g and h to be continuously differentiable.

Definition 5.6 *The set of* **linearized feasible directions** *of a feasible point* $\mathbf{x} \in \Omega$ *is defined as*

$$\mathcal{F}(\mathbf{x}) = \left\{ \mathbf{d} \in \mathbb{R}^n \mid \nabla g_i(\mathbf{x})^T \mathbf{d} \geq 0 \ \forall i \in \mathcal{A}(\mathbf{x}), h'(\mathbf{x})\mathbf{d} = 0 \right\}.$$

The hope is that the two cones $T_\Omega(\mathbf{x})$ and $\mathcal{F}(\mathbf{x})$ are identical. In many cases, this is reasonable to expect. However, one can also construct counterexamples quite easily.

Example 5.7 We consider the non-optimal point $\mathbf{x}_0 = (-\sqrt{2}, 0)^T$ for the equality-constrained problem (5.4) from Example 5.2, that is, $f(\mathbf{x}) = x_1 + x_2$ and $h(\mathbf{x}) = x_1^2 + x_2^2 - 2$.

By considering the two directions on the circle $\mathbf{x}_0$ can be approached, one computes $T_\Omega(\mathbf{x}_0) = \{(0, d_2)^T \mid d_2 \in \mathbb{R}\}$. On the other hand, we have $\mathbf{d} \in \mathcal{F}(\mathbf{x}_0)$ if

$$0 = h'(\mathbf{x})\mathbf{d} = -2\sqrt{2} d_1.$$

Hence, $\mathcal{F}(\mathbf{x}_0)$ is identical with $T_\Omega(\mathbf{x}_0)$.

The situation changes if we replace $h(\mathbf{x})$ by $h(\mathbf{x}) = (x_1^2 + x_2^2 - 2)^2$. This still describe the same feasibility set but now

$$0 = h'(\mathbf{x})\mathbf{d} = 0,$$

that is, no constraint is imposed on $\mathbf{d}$. Hence, $\mathcal{F}(\mathbf{x}_0) = \mathbb{R}^2$ now differs from $T_\Omega(\mathbf{x}_0)$! ◇

There are many flavors of so called *constraint qualifications* that impose a condition to guarantee $T_\Omega(\mathbf{x}) = \mathcal{F}(\mathbf{x})$ at a feasible point $\mathbf{x}$. We will work with the following, relatively strong condition.

Definition 5.8 (LICQ) *We say that the* **linear independence constraint qualification (LICQ)** *holds at* $\mathbf{x} \in \Omega$ *if the active constraint gradients*

$$\{\nabla g_i(\mathbf{x}) \mid i \in \mathcal{A}(\mathbf{x})\} \cup \{\nabla h_i(\mathbf{x}) \mid i \in \{1, \ldots, p\}\}$$

form a linearly independent set.

Lemma 5.9 *Let* $\mathbf{x}^\star \in \Omega$. *Then:*

1. $T_\Omega(\mathbf{x}^\star) \subset \mathcal{F}(\mathbf{x}^\star)$.
2. *If LICQ holds at* $x^\star$ *then* $T_\Omega(\mathbf{x}^\star) = \mathcal{F}(\mathbf{x}^\star)$.

Proof. 1. Let $\mathbf{d} \in T_\Omega(\mathbf{x}^*)$ and consider the corresponding sequences $\mathbf{x}_k \to \mathbf{x}^*$ and $\eta_k > 0$ such that $\frac{1}{\eta_k}(\mathbf{x}_k - \mathbf{x}^*) \to \mathbf{d}$. Clearly, we have

$$\mathbf{x}_k = \mathbf{x}_* + \eta\mathbf{d} + o(\eta_k).$$

Taking into account that the equality constraints $h_1, \dots, h_p$ are satisfied for $\mathbf{x}_k$, we obtain from the Taylor expansion that

$$0 = \frac{1}{\eta_k} h_i(\mathbf{x}_k) = \frac{1}{\eta_k}\big(h_i(\mathbf{x}^*) + \eta_k \nabla h_i(\mathbf{x}^*)^T\mathbf{d} + o(\eta_k)\big) = \nabla h_i(\mathbf{x}^*)^T\mathbf{d} + o(1).$$

Hence, $\nabla h_i(\mathbf{x}^*)^T\mathbf{d} = 0$ follows from taking the limit $k \to \infty$. In the same way, we obtain $\nabla g_i(\mathbf{x}^*)^T\mathbf{d} \geq 0$ for $i \in \mathcal{A}(\mathbf{x}^*)$.

2. By simply dropping inactive constraints beforehand, we may assume without loss of generality that all inequality constraints are active, that is, $\mathcal{A}(\mathbf{x}) = \{1, \dots, m\}$. Then the matrix which collects all gradients,

$$A = \begin{pmatrix} g'(\mathbf{x}^\star) \\ h'(\mathbf{x}^\star) \end{pmatrix} \in \mathbb{R}^{(m+p)\times n},$$

has full row rank.[5] Now, let $Z \in \mathbb{R}^{n\times(n-m-p)}$ be a basis for the null space of A, that is, Z has full column rank and $AZ = 0$.

We now consider the (parametrized) nonlinear system of equations

$$R(\mathbf{x}, \eta) := \begin{pmatrix} g(\mathbf{x}) - \eta g'(\mathbf{x}^\star)\mathbf{d} \\ h(\mathbf{x}) - \eta h'(\mathbf{x}^\star)\mathbf{d} \\ Z^T(\mathbf{x} - \mathbf{x}^* - \eta\mathbf{d}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{5.14}$$

with an arbitrary choice of $\mathbf{d} \in \mathcal{F}(\mathbf{x}^*)$. For $\eta = 0$, the Jacobian of R at $\mathbf{x}^\star$ is given by

$$\frac{\partial}{\partial \mathbf{x}} R(\mathbf{x}^*, 0) = \begin{pmatrix} g'(\mathbf{x}^\star) \\ h'(\mathbf{x}^\star) \\ Z^T \end{pmatrix}.$$

The construction of Z implies that this Jacobian is nonsingular. Hence, by the implicit function theorem, for all η sufficiently small the nonlinear system (5.14) has a solution $\mathbf{x}$ (which is even the unique solution in some neighborhood of $\mathbf{x}^*$) that depends continuously differentiable on η.

We now consider a sequence $\eta_k \to 0$ with $\eta_k > 0$ along with the corresponding solution $\mathbf{x}_k$ of (5.14). Then the first two equations of (5.14) imply that $\mathbf{x}_k$ is feasible. Now, by the Taylor expansion

$$0 = R(\mathbf{x}_k, \eta_k) = \left[\frac{\partial}{\partial \mathbf{x}} R(\mathbf{x}^*, 0)\right](\mathbf{x}_k - \mathbf{x}^* - \eta_k\mathbf{d}) + o(\|\mathbf{x}_k - \mathbf{x}^*\|).$$

[5] A small subtlety: this seems to imply $m + p \leq n$, which need not to be the case in (5.1). Since we have thrown away all inactive constraints, the m in the proof might be smaller than the original m, so keeping the notation is convenient but slightly abusive.

Using the nonsingularity of the Jacobian and dividing by η_k gives

$$\frac{1}{\eta_k}(\mathbf{x}_k - \mathbf{x}^*) = \mathbf{d} + o\left(\frac{\|\mathbf{x}_k - \mathbf{x}^*\|}{\eta_k}\right),$$

from which it follows that $\mathbf{d} \in T_\Omega(\mathbf{x}^*)$. (Note that the limit on the left side exists since $\eta \mapsto \mathbf{x}(\eta)$ is differentiable in zero.) □

We have now collected all ingredients to carry over the machinery from linear programming to the nonlinear case. We recall the following central result from the course on discrete optimization.

Lemma 5.10 (Farkas' lemma) *Let $A_g \in \mathbb{R}^{n\times m}, A_h \in \mathbb{R}^{n\times p}, \mathbf{c} \in \mathbb{R}^n$. Then the following two statements are equivalent.*

1. *For any $\mathbf{d} \in \mathbb{R}^n$ with $A_g^T\mathbf{d} \geq 0$ and $A_h^T\mathbf{d} = 0$ we have $\mathbf{c}^T\mathbf{d} \geq 0$.*
2. *There exists $\mathbf{u} \in \mathbb{R}^m$ with $\mathbf{u} \geq 0$ and $\mathbf{v} \in \mathbb{R}^p$ such that $\mathbf{c} = A_g\mathbf{u} + A_h\mathbf{v}$.*

Proof. See Theorem 2.9 in [E]. □

Let us now consider a local solution $\mathbf{x}^*$ of (5.1) for which the LICQ holds. By Lemma 5.9, $T_\Omega(\mathbf{x}^\star) = \mathcal{F}(\mathbf{x}^\star)$. Thus, Theorem 5.5 implies

$$\nabla f(\mathbf{x}^*)^T\mathbf{d} \geq 0 \qquad \forall \mathbf{d} \in \mathcal{F}(\mathbf{x}^\star).$$

This is the first statement of Farkas' lemma where the columns of A_g contain all gradients $\nabla g_i(\mathbf{x}^*)$ with active inequality constraints, $A_h = h'(\mathbf{x}^*)^T$, and $\mathbf{c} = \nabla f(\mathbf{x}^*)$. The second, equivalent statement of Farkas' lemma then yields the existence of $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ such that

$$\nabla f(\mathbf{x}^*) - \sum_{i\in\mathcal{A}(\mathbf{x}^*)} \lambda_i^* \nabla g_i(\mathbf{x}^*) + \sum_{i=1}^{p} \mu_i^* \nabla h_i(\mathbf{x}^*) = 0. \tag{5.15}$$

Note that the values of $\lambda_i^\star$ for $i \notin \mathcal{A}(\mathbf{x}^*)$ do not enter (5.15). Choosing these "missing" values to be zero is equivalent to the complementarity condition

$$\lambda_i g_i(\mathbf{x}^*) = 0, \qquad i = 1, \ldots, m. \tag{5.16}$$

This also allows us to rewrite (5.15) more compactly as

$$\nabla f(\mathbf{x}^*) - g'(\mathbf{x}^*)^T\lambda^* - h'(\mathbf{x}^*)^T\mu^* = 0.$$

The left-hand side turns out to be the gradient with respect to $\mathbf{x}$ at $(\mathbf{x}^*, \lambda^*, \mu^*)$ of the **Lagrangian function**

$$\mathcal{L}(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) - \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) - \sum_{i=1}^{p} \mu_i h_i(\mathbf{x}). \tag{5.17}$$

The following theorem is the main result of this section and summarizes our findings.

Theorem 5.11 (First-order necessary conditions) *Suppose that $\mathbf{x}^*$ is a local solution of the constrained optimization problem* (5.1) *with continuously differentiable f, g, h, such that the LICQ holds. Then there are Lagrange multipliers λ^*, μ^* such that the following conditions are satisfied:*

(1) $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \lambda^*, \mu^*) = 0$*;*

(2) $h(\mathbf{x}^*) = 0$*;*

(3a) $g(\mathbf{x}^*) \geq 0$*;*

(3b) $\lambda^* \geq 0$*;*

(3c) $\lambda_i^* g_i(\mathbf{x}^*) = 0$ *for* $i = 1, \ldots, m$*.*

The conditions of Theorem 5.11 are known as **Karush-Kuhn-Tucker conditions** or, short, **KKT conditions**.

5.1.3 Special cases of the KKT conditions*

The constrained optimization problem (5.1) is called **convex** if the function f is convex, the functions g_i, $i = 1, \ldots, m$, are concave and h is affine linear. This implies that the feasible set is convex because

$$\begin{aligned} g_i(\beta \mathbf{x} + (1-\beta)\mathbf{y}) &\geq \beta g_i(\mathbf{x}) + (1-\beta) g_i(\mathbf{y}) \geq 0, \\ h(\beta \mathbf{x} + (1-\beta)\mathbf{y}) &= \beta h(\mathbf{x}) + (1-\beta) h(\mathbf{y}) = 0 \end{aligned}$$

holds for all $\mathbf{x}, \mathbf{y} \in \Omega$ and $\beta \in [0,1]$. More importantly, the KKT conditions turn out to be not only necessary but also *sufficient* for convex problems.

Theorem 5.12 *Let* (5.1) *be convex. Then the following statements hold:*

1. *Every local solution $\mathbf{x}^* \in \Omega$ is a global solution of* (5.1)*.*
2. *If $\mathbf{x}^*$ satisfies the KKT conditions of Theorem 5.11 then $\mathbf{x}^*$ is a global solution of* (5.1)*.*

Proof. 1. Let $\mathbf{x}^* \in \Omega$ be a local solution of (5.1) and consider an arbitrary $\mathbf{x} \in \Omega$. Then $\mathbf{x}^* + \beta \mathbf{d} \in \Omega$ for $\mathbf{d} = \mathbf{x} - \mathbf{x}^*$ and $\beta \in [0,1]$. By the convexity of f, it follows for sufficiently small $\beta > 0$ that

$$0 \leq f(\mathbf{x}^* + \beta \mathbf{d}) - f(\mathbf{x}^*) \leq (1-\beta) f(\mathbf{x}^*) + \beta f(\mathbf{x}) - f(\mathbf{x}^*) = \beta\big(f(\mathbf{x}) - f(\mathbf{x}^*)\big).$$

This implies $f(\mathbf{x}) - f(\mathbf{x}^*) \geq 0$ and, hence, $\mathbf{x}^*$ is a global solution.

2. Let $\mathbf{x}^* \in \Omega$ satisfy the KKT conditions, consider an arbitrary $\mathbf{x} \in \Omega$, and let $\mathbf{d} = \mathbf{x} - \mathbf{x}^*$. Then

$$\lambda_i^\star \nabla g_i(\mathbf{x})^T \mathbf{d} \geq \lambda_i^\star (g_i(\mathbf{x}) - g_i(\mathbf{x}^*)) = \lambda_i^* g_i(\mathbf{x}) \geq 0,$$

where the first inequality follows from the concavity of g_i. Moreover, $\nabla h(\mathbf{x}^*)^T \mathbf{d} = h(\mathbf{x}) - h(\mathbf{x}^*) = 0$. The convexity of f, together with the KKT conditions (1)+(3a)+(3b) yields

$$f(\mathbf{x}) - f(\mathbf{x}^*) \geq \nabla f(\mathbf{x}^*)^T \mathbf{d} = (\lambda^*)^T g'(\mathbf{x}^*)\mathbf{d} + (\mu^*)^T h'(\mathbf{x}^*)\mathbf{d} = (\lambda^*)^T g'(\mathbf{x}^*)\mathbf{d} \geq 0,$$

which completes the proof. □

In particular, Theorem 5.12 applies to the linear program

$$\min_{\mathbf{x} \in \mathbb{R}^n} \mathbf{c}^T \mathbf{x} \quad \text{subject to} \quad \mathbf{x} \geq 0,\ A\mathbf{x} = \mathbf{b}, \tag{5.18}$$

which we will call the **primal problem**.[6] The Lagrangian (5.17) is then given by

$$\mathcal{L}(\mathbf{x}, \lambda, \mu) = \mathbf{c}^T \mathbf{x} - \mathbf{x}^T \lambda - (A\mathbf{x} - \mathbf{b})^T \mu.$$

In principle, the constraint qualification LICQ requires the matrix $\left(\begin{smallmatrix} E \\ A \end{smallmatrix}\right)$ to have full row rank, where the rows of E consist of all unit vectors e_i^T with $i \in \mathcal{A}(x)$. However, it turns out that LICQ is in fact not needed. $\mathcal{F}$ arises from T_Ω by linearization, but the constraints in (5.18) are already linear and hence both cones must be identical.

For (5.18), the KKT conditions (which are equivalent and sufficient for a local solution) become

(p1) $A^T \mu^\star + \lambda^\star = \mathbf{c}$;

(p2) $A\mathbf{x}^\star = \mathbf{b}$;

(p3a) $\mathbf{x}^\star \geq 0$;

(p3b) $\lambda^\star \geq 0$;

(p3c) $\lambda_i^\star x_i^\star = 0$ for $i = 1, \dots, m$.

Given the nonnegativity of $\mathbf{x}^\star, \lambda^\star$, condition (p3c) is equivalent to

$$(\mathbf{x}^\star)^T \lambda^\star = 0.$$

With the same data as in (5.18), we can define the **dual problem**

$$\max_{\mathbf{x} \in \mathbb{R}^n} \mathbf{b}^T \lambda \quad \text{subject to} \quad A^T \lambda \leq \mathbf{c}. \tag{5.19}$$

If we apply Theorem 5.11 to this problem, then the resulting KKT conditions turn out to be identical! In particular, the optimal Lagrange multipliers in the primal problem are the optimal variables in the dual problem, while the optimal Lagrange multipliers in the dual problem are the optimal variables in the primal problem. We refer to Chapter 4 of [4] for more relations between the primal and dual problems.

[6] This linear program actually takes the form of the dual problem considered in (4.2) of [E]. This nuisance is hopefully considered minor by the reader.