MSE2122 - Nonlinear Optimization
Lecture Notes VII

Fernando Dias (based on previous version by Fabricio Oliveira)

October 18, 2023

Abstract

In this lecture, we discuss the optimality conditions for constrained optimisation problems. We show how geometrical optimality can be converted into an algebraic representation using the Fritz-John optimality conditions. Next, we discuss the Karush-Kuhn-Tucker optimality condition, which require further regularity conditions on the constraints to hold as necessary conditions for optimality. We show that these regularity conditions can be translated into constraint qualification conditions, and discuss the main constraint qualification conditions one could use in practice.

Contents

1 Optimality for constrained problems

We now investigate how to derive optimality conditions for the problem:

$$(P): \text{ min. } f(x) : x \in S.$$

In particular, we are interested in understanding the role of the feasibility set S on the optimality conditions of constrained optimisation problems in the form of P. Let us first define two geometric elements that we will use to derive the optimality conditions for P.

Definition 1.1. Cone of feasible directions

et $S \subseteq \mathbb{R}^n$ be a nonempty set, and let $\overline{x} \in \mathbf{clo}(S)$. The *cone of feasible directions* D at $\overline{x} \in S$ is given by:

$$D = \{d : d \neq 0, \text{ and } \overline{x} + \lambda d \in S \text{ for all } \lambda \in (0, \delta) \text{ for some } \delta > 0\}.$$

Definition 1.2. Cone of descent directions

et $S \subseteq \mathbb{R}^n$ be a nonempty set, $f : \mathbb{R}^n \to \mathbb{R}$, and $\overline{x} \in \mathbf{clo}(S)$. The *cone of improving (i.e., descent) directions* F at $\overline{x} \in S$ is:

$$F = \{d : f(\overline{x} + \lambda d) < f(\overline{x}) \text{ for all } \lambda \in (0, \delta) \text{ for some } \delta > 0\}.$$

These cones are geometrical descriptions of the regions that, from a given point $\overline{x}$, one can obtain feasible (D) and improving (F) solutions. This is useful because it allows us to express the optimality conditions for $\overline{x}$ as it observes that $F \cap D = \emptyset$ holds. In other words, $\overline{x}$ is optimal if no feasible direction can improve the objective function value.

Although having a geometrical representation of such sets can be useful in solidifying the conditions for which a feasible solution is also optimal, we need to derive an *algebraic* representation of such sets that can be used in computations. Let us define an algebraic representation for F to reach that objective. Let us assume that $f : S \subset \mathbb{R}^n \mapsto \mathbb{R}$ is differentiable. Recall that d is a descent direction at $\overline{x}$ if $\nabla f(\overline{x})^\top d < 0$. Thus, we can define the set F_0.

$$F_0 = \{d : \nabla f(\overline{x})^\top d < 0\}$$

As an algebraic representation for F, notice that F_0 is an open half-space formed by the hyperplane with normal $\nabla f(\overline{x})$. Figure 1 illustrates the condition $F_0 \cap D = \emptyset$.

Theorem 1.3 establishes that the condition $F_0 \cap D = \emptyset$ is necessary for optimality in constrained optimisation problems.

Theorem 1.3. Geometric necessary condition

Let $S \subseteq \mathbb{R}^n$ be a nonempty set, and let $f : S \to \mathbb{R}$ be differentiable at $\overline{x} \in S$. If $\overline{x}$ is a local optimal solution to

$$(P): \text{ min. } \{f(x) : x \in S\},$$

then $F_0 \cap D = \emptyset$, where $F_0 = \{d : \nabla f(\overline{x})^\top d < 0\}$ and D is the cone of feasible directions.

The proof for this theorem consists of using the separation theorem to show that $F_0 \cap D = \emptyset$ implies that the first-order optimality condition $\nabla f(\overline{x})^\top d \geq 0$ holds.

As discussed earlier (in Lecture 4), these conditions become sufficient for optimality in the presence of convex. Moreover, if f is strictly convex, then $F = F_0$. If f is linear, it might be worth considering $F_0' = \{d \neq 0 : \nabla f(\overline{x})^\top d \leq 0\}$ to allow for considering orthogonal directions.

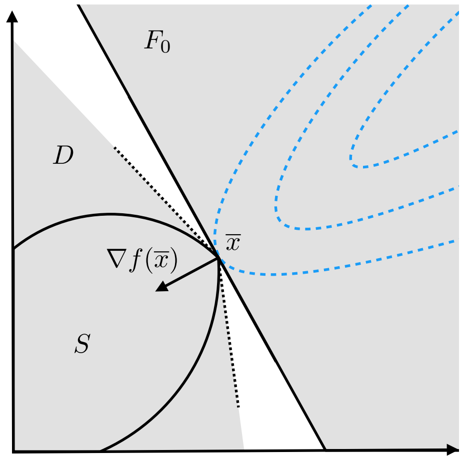

Figure 1: Illustration of the cones F_0 and D for the optimal point $\overline{x}$. Notice that D is an open set.

1.1 Inequality constrained problems

In mathematical programming applications, a set of inequalities typically expresses the feasibility set S. Let us redefine P as:

$$\begin{aligned} (P): \ \text{min.} \ & f(x) \\ \text{subject to:} \ & g_i(x) \leq 0, \ i = 1, \dots, m \\ & x \in X, \end{aligned}$$

where $g_i : \mathbb{R}^n \mapsto \mathbb{R}$, are differentiable functions for $i = 1, \dots, m$ and $X \subset \mathbb{R}$ is an nonempty open set. The differentiability of g_i, $i = 1, \dots, m$, allows for the definition of a proxy for D using the gradients of the binding constraints $i \in I = \{i : g_i(\overline{x}) = 0\}$ at $\overline{x}$. This set, denoted by G_0, is defined as:

$$G_0 = \{d : \nabla g_i(\overline{x})^\top d < 0, i \in I\}.$$

The use of G_0 is a convenient algebraic representation since it can be shown that $G_0 \subseteq D$, which is stated in Lemma 1.4. As $F_0 \cap D = \emptyset$ must hold for a locally optimal solution $\overline{x} \in S$, $F_0 \cap G_0 = \emptyset$ must also hold.

Lemma 1.4 Let $S = \{x \in X : g_i(x) \leq 0 \text{ for all } i = 1, \dots, m\}$, where $X \subset \mathbb{R}^n$ is a nonempty open set and $g_i : \mathbb{R}^n \to \mathbb{R}$ a differentiable function for all $i = 1, \dots, m$. For a feasible point $\overline{x} \in S$, let $I = \{i : g_i(\overline{x}) = 0\}$ be the index set of the binding (or active) constraints. Let

$$G_0 = \{d : \nabla g_i(\overline{x})^\top d < 0, i \in I\}$$

Then $G_0 \subseteq D$, where D is the cone of feasible directions.

In settings in which g_i is affine for some $i \in I$, it might be worth considering l$G'_0 = \{d \neq 0 : \nabla g_i(\overline{x})^\top d \leq 0, i \in I\}$ so those orthogonal feasible directions can also be represented. Notice that in this case, $D \subseteq G'_0$.

2 Fritz-John conditions

The Fritz-John conditions are the algebraic conditions that must be met for $F_0 \cap G_0 = \emptyset$ to hold. These algebraic conditions are convenient as they only involve the gradients of the binding constraints, and they can be verified computationally.

Theorem 2.1. Fritz-John necessary conditions

Let $X \subseteq \mathbb{R}^n$ be a nonempty open set, and let $f : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}$ be differentiable for all $i = 1, \dots, m$. Additionally, let $\overline{x}$ be feasible and $I = \{i : g_i(\overline{x}) = 0\}$. If $\overline{x}$ solves P locally, there exist scalars u_i, $i \in \{0\} \cup I$, such that:

$$\begin{aligned} & u_0 \nabla f(\overline{x}) + \sum_{i=1}^{m} u_i \nabla g_i(\overline{x}) = 0 \\ & u_i g_i(\overline{x}) = 0, \ i = 1, \dots, m \\ & u_i \geq 0, \ i = 0, \dots, m \\ & u = (u_0, \dots, u_m) \neq 0 \end{aligned}$$

Proof. Since $\overline{x}$ solves P locally, Theorem 1.3 guarantees that there is no d such that $\nabla f(\overline{x})^\top d < 0$ and $\nabla g_i(x)^\top d < 0$ for each $i \in I$. Let A be the matrix whose rows are $\nabla f(\overline{x})^\top$ and $\nabla g_i(\overline{x})^\top$ for $i \in I$.

We can use Farkas' theorem to show that if $Ad < 0$ is inconsistent, there is nonzero $p \geq 0$ such that $A^\top p = 0$. Letting $p = (u_0, u_{i_1}, \dots, u_{i_{|I|}})$ for $I = \{i_1, \dots, i_{|I|}\}$ and making $u_i = 0$ for $i \neq I$, the result follows.

😎

The proof considers that, if $\overline{x}$ is optimal, then $f(\overline{x})^\top d \geq 0$ holds and a matrix A formed by:

$$A = \begin{bmatrix} \nabla f(\overline{x}) \\ \nabla g_{i_1}(\overline{x}) \\ \vdots \\ \nabla g_{i_{|I|}}(\overline{x}) \end{bmatrix}$$

with $I = \{i_1, \dots, i_{|I|}\}$, will violate $Ad < 0$. This is used with a variant of Farkas' theorem (known as Gordan's theorem) to show that the alternative system $A^\top p = 0$, with $p \geq 0$ holds, which, by setting $p = [u_0, u_{i_1}, \dots, u_{i_{|I|}}]$ and enforcing that the remainder of the gradients $\nabla g_i(\overline{x})$, for $i \notin I$, are removed by setting $u_i = 0$, which leads precisely to the Fritz-John conditions.

The multipliers u_i, for $i = 0, \dots, m$, are named Lagrangian multipliers due to the connection with Lagrangian duality, as we will see later. Also, notice that for nonbinding constraints ($g_i(\overline{x}) < 0$ for $i \notin I$), u_i must be zero to form the Fritz-John conditions. This condition is called complementary slackness.

Unfortunately, The Fritz-John conditions are too weak, which is problematic in some rather common settings. A point $\overline{x}$ satisfies the Fritz-John conditions only if $F_0 \cap G_0 = \emptyset$, which is trivially satisfied when $G_0 = \emptyset$.

For example, the Fritz-John conditions are trivially satisfied for points where some of the gradient vanishes (i.e., $\nabla f(\overline{x}) = 0$ or $\nabla g_i(\overline{x}) = 0$ for some $i = 1, \dots, m$). Sets with no relative interior near $\overline{x}$ also satisfy Fritz-John conditions.

An interesting case is for problems with equality constraints, as illustrated in Figure 2. In general, if the additional regularity condition that the gradients $\nabla g_i(\overline{x})$ are linearly independent does not hold, $\overline{x}$ trivially satisfies the Fritz-John conditions.

3 Karush-Kuhn-Tucker conditions

The Karush-Kuhn-Tucker (KKT) conditions are the Frizt-John conditions with an extra regularity requirement for $\overline{x} \in S$. This regularity requirement is called *constraint qualification* and, in a general sense, is meant to prevent the trivial case $G_0 = \emptyset$, thus making the optimality conditions stronger (i.e., more stringent).

This is achieved by making $u_0 = 1$ in Theorem 2.1, which ultimately implies that the gradients $\nabla g_i(\overline{x})$ for $i \in I$ must be linearly independent. This condition is called *linearly independent constraint qualification* (LICQ) and is one of several known constraint qualifications that can be used to guarantee regularity of $\overline{x} \in S$.

Theorem 3.1 establishes the KKT conditions as necessary for local optimality of $\overline{x}$ assuming that

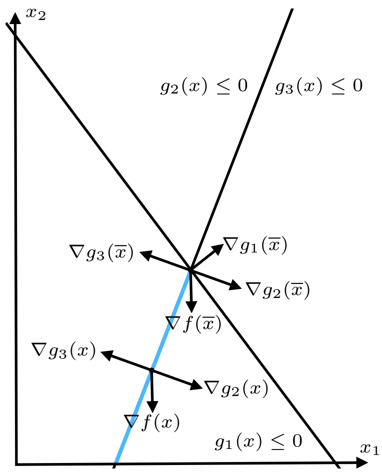

Figure 2: All points in the blue segment satisfy FJ conditions, including the minimum $\overline{x}$.

LICQ holds. For notational simplicity, let us assume for now that:

$$(P): \text{ min. } \{f(x) : g_i(x) \leq 0, i = 1, \ldots, m, x \in X\}.$$

Theorem 3.1. Karush-Kuhn-Tucker necessary conditions

Let $X \subseteq \mathbb{R}^n$ be a nonempty open set, and let $f : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}$ be differentiable for all $i = 1, \ldots, m$. Additionally, for a feasible $\overline{x}$, let $I = \{i : g_i(\overline{x}) = 0\}$ and suppose that $\nabla g_i(\overline{x})$ are linearly independent for all $i \in I$. If $\overline{x}$ solves P locally, there exist scalars u_i for $i \in I$ such that:

$$\begin{aligned}
&\nabla f(\overline{x}) + \sum_{i=1}^{m} u_i \nabla g_i(\overline{x}) = 0 \\
&u_i g_i(\overline{x}) = 0, \;\; i = 1, \ldots, m \\
&u_i \geq 0, \;\; i = 1, \ldots, m
\end{aligned}$$

Proof. By Theorem 2.1, there exists nonzero $(\hat{u}_i)$ for $i \in \{0\} \cup I$ such that:

$$\begin{aligned}
&\hat{u}_0 \nabla f(\overline{x}) + \sum_{i=1}^{m} \hat{u}_i \nabla g_i(\overline{x}) = 0 \\
&\hat{u}_i \geq 0, \;\; i = 0, \ldots, m
\end{aligned}$$

Note that $\hat{u}_0 > 0$, as the linear independence of $\nabla g_i(\overline{x})$ for all $i \in I$ implies that $\sum_{i=1}^{m} \hat{u}_i \nabla g_i(\overline{x}) \neq 0$. Now, let $u_i = \hat{u}_i / u_0$ for each $i \in I$ and $u_i = 0$ for all $i \notin I$.

😎

The proof builds upon the Fritz-John conditions, which under the assumption that the gradients of the active constraints $\nabla g_i(\overline{x})$ for $i \in I$ are independent, the multipliers $\hat{u}_i$ can be rescaled so that $u_0 = 1$.

The general conditions, including inequality and equality constraints, are posed as follows. Notice that the Lagrange multipliers v_i associated with the equality constraints $h(\overline{x}) = 0$ for $i = 1, \ldots, l$ are not restricted in sign, and the complementary slackness condition is not explicitly stated since it holds redundantly. These can be obtained by replacing equality constraints $h(x) = 0$ with two equivalent inequalities $h_-(x) \leq 0$ and $-h_+(x) \leq 0$ and writing the conditions in Theorem 3.1. Also, notice that, without constraints, the KKT conditions reduce to the unconstrained first-order condition $\nabla f(\overline{x}) = 0$.

$$\begin{aligned}
&\nabla f(\overline{x}) + \sum_{i=1}^{m} u_i \nabla g_i(\overline{x}) + \sum_{i=1}^{l} v_i \nabla h_i(\overline{x}) = 0 && \text{(dual feasibility 1)} \\
&u_i g_i(\overline{x}) = 0, \qquad i = 1, \ldots, m && \text{(complementary slackness)} \\
&\overline{x} \in X, \;\; g_i(\overline{x}) \leq 0, \;\; i = 1, \ldots, m && \text{(primal feasibility)} \\
&h_i(x) = 0, \qquad i = 1, \ldots, l && \\
&u_i \geq 0, \qquad i = 1, \ldots, m && \text{(dual feasibility 2)}
\end{aligned}$$

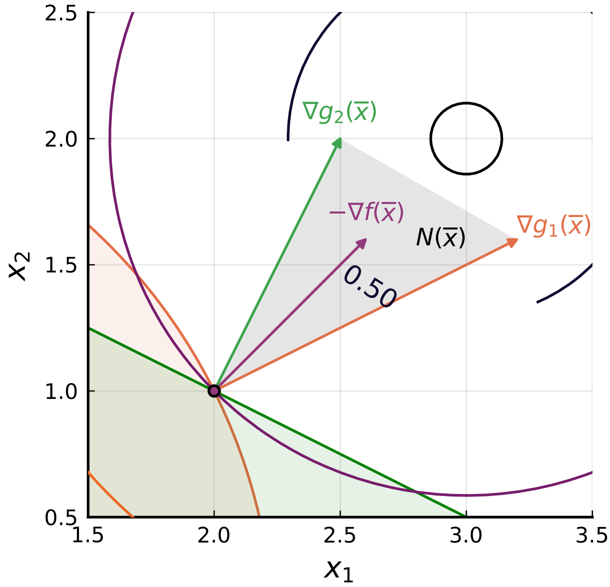

Figure 3: Graphical illustration of the KKT conditions at the optimal point $\overline{x}$

The KKT conditions can be interpreted geometrically as follows. Consider the cone spanned by the active constraints at $\overline{x}$, defined as $N(\overline{x}) = \{\sum_{i \in I} u_i \nabla g_i(\overline{x}) : u_i \geq 0\}$. A solution $\overline{x}$ will then satisfy the KKT conditions if $-\nabla f(\overline{x}) \in N(\overline{x})$, which is equivalent to $-\nabla f(\overline{x}) = \sum_{i=1}^{m} u_i \nabla g_i(\overline{x})$. Figure 3 illustrates this condition.

4 Constraint qualification

Constraint qualification is a technical condition that needs to be assessed in the context of nonlinear optimisation problems. As we rely on an algebraic description of the set of directions G_0 that serves as a proxy for D, it is important to be sure that the former is a reliable description of the latter.

Specifically, constraint qualification can be seen as a certification that the geometry of the feasible region and gradient information obtained from the constraints that form it are related to an optimal solution. Remind that gradients can only provide a *first-order* approximation of the feasible region, which might lead to mismatches. This is typically when the feasible region has cusps or a single feasible point.

Constraint qualification can be seen as certificates for proper relationships between the set of feasible directions:

$$G'_0 = \{d \neq 0 : \nabla g_i(\overline{x})^\top d \leq 0, i \in I\}$$

and the cone of tangents (or tangent cone):

$$T = \{d : \ d = \lim_{k \to \infty} \lambda_k (x_k - \overline{x}), \lim_{k \to \infty} x_k = \overline{x}, x_k \in S, \lambda_k > 0, \forall k\} \tag{1}$$

with $S = \{g_i(x) \leq 0, \ i = 1, \dots, m; h(x) = 0, \ i = 1, \dots, l; x \in X\}$.

The cone of tangents represents all directions in which the feasible region allows for an arbitrarily small movement from the point $\overline{x}$ while retaining feasibility. As the name suggests, it is normally formed by the tangent lines to S at $\overline{x}$. However, if the point is in the interior of $S \subseteq \mathbb{R}^n$, then $T = \mathbb{R}^n$.

One way of interpreting the cone of tangents as defined in (1) is the following: consider a sequence of feasible points $x \in S$ in any trajectory you like, but in a way that the sequence converges to $\overline{x}$. Then, take the last (in a limit sense, since $k \to \infty$) x_k and consider this direction from which x_k came onto $\overline{x}$. The collection of all these directions from all possible trajectories is what forms the cone of tangents.

Constraint qualification holds when $T = G'_0$ holds for $\overline{x}$, a condition named *Abadie's constraint qual-*

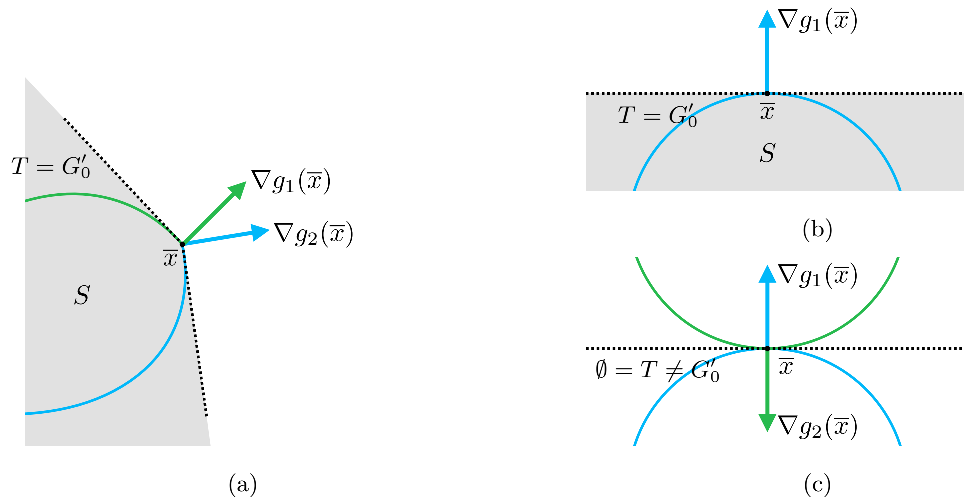

Figure 4: CQ holds for 4a and 4b, since the tangent cone T and the cone of feasible directions G'_0 (denoted by the dashed black lines and a grey area) match; for 4c, they do not match, as $T = \emptyset$

ification. In the presence of equality constraints, the condition becomes $T = G'_0 \cap H_0$, with:

$$H_0 = \{d : \nabla h_i(\overline{x})^\top d = 0, i = 1, \dots, l\}.$$

Figure 4 illustrates the tangent cone T and the cone of feasible directions (G'_0) for cases when constraint qualification holds (Figures 4a and 4b) for which case $T = G'_0$, and a case for when it does not (Figure 4c, where $T = \emptyset$ and G'_0 is given by the dashed black line).

The importance of Abadie constraint qualification is that it allows for generalising the KKT conditions by replacing the condition with the linear independence of the gradients $\nabla g_i(\overline{x})$ for $i \in I$. This allows us to state the KKT conditions as presented in Theorem 4.1.

Theorem 4.1. Karush-Kuhn-Tucker necessary conditions II

Consider the problem

$$(P): \text{ min. } \{f(x) : g_i(x) \leq 0, i = 1, \dots, m, x \in X\}.$$

Let $X \subseteq \mathbb{R}^n$ be a nonempty open set, and let $f : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}$ be differentiable for all $i = 1, \dots, m$. Additionally, for a feasible $\overline{x}$, let $I = \{i : g_i(\overline{x}) = 0\}$ and *suppose that Abadie CQ holds* at $\overline{x}$. If $\overline{x}$ solves P locally, there exist scalars u_i for $i \in I$ such that:

$$\begin{aligned}
&\nabla f(\overline{x}) + \sum_{i=1}^{m} u_i \nabla g_i(\overline{x}) = 0 \\
&u_i g_i(\overline{x}) = 0, \ \ i = 1, \dots, m \\
&u_i \geq 0, \ \ i = 1, \dots, m.
\end{aligned}$$

Despite being a more general result, Theorem 4.1 is of little use, as Abadie's constraint qualification cannot be straightforwardly verified in practice. Alternatively, we can rely on verifiable constraint qualification conditions that imply Abadie's constraint qualification. Examples include:

1. **Linear independence (LI)CQ:** holds at $\overline{x}$ if $\nabla g_i(\overline{x})$, for $i \in I$, as well as $\nabla h_i(\overline{x})$, $i = 1, \dots, l$ are *linearly independent.*
2. **Affine CQ:** holds for all $x \in S$ if g_i, for all $i = 1, \dots, m$, and h_i, for all $i = 1, \dots, l$, are *affine.*
3. **Slater's CQ:** holds for all $x \in S$ if g_i is a *convex* function for all $i = 1, \dots, m$, h_i is an *affine* function for all $i = 1, \dots, l$, and there exists $x \in S$ such that $g_i(x) < 0$ for all $i = 1, \dots, m$.

Slater's constraint qualification is the most frequently used, particularly in convex optimisation problems. One important point to notice is the requirement of not having an empty relative interior, which can be a source of error.

Consider, for example: $P = \{\text{min. } x_1 : x_1^2 + x_2 \leq 0, x_2 \geq 0\}$. Notice that P is convex and therefore the KKT system for P is:

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = 0; u_1, u_2 \geq 0,$$

which has no solution. Thus, the KKT conditions are not necessary for the global optimality of $(0, 0)$. This is due to the lack of CQ since the feasible region is the single point $(0, 0)$ and that KKT conditions are only sufficient (not necessary) in the presence of convexity.

Corollary 4.2 summarises the setting in which one should expect the KKT conditions to be necessary and sufficient conditions for global optimality, i.e., convex optimisation.

Corollary 4.2 (Necessary and sufficient KKT conditions)**.** Suppose that Slater's CQ holds. Then, if f is convex, the conditions of Theorem 4.1 are *necessary and sufficient* for $\overline{x}$ to be a globally optimal solution.