SECTION C: CONTINUOUS OPTIMISATION
LECTURE 9: FIRST ORDER OPTIMALITY CONDITIONS FOR CONSTRAINED NONLINEAR PROGRAMMING

HONOUR SCHOOL OF MATHEMATICS, OXFORD UNIVERSITY
HILARY TERM 2006, DR RAPHAEL HAUSER
WITH A FEW ADDITIONS FROM DR. NICK GOULD

1. Optimality Conditions: What We Know So Far. In Lecture 2 we showed that $\nabla f(x) = 0$ and $D^2 f(x) \succeq 0$ are necessary optimality conditions for unconstrained optimisation, and we found that the stronger conditions $\nabla f(x) = 0$, $D^2 f(x) \succ 0$ are sufficient in guaranteeing that x is a strict local minimiser. In effect, sufficiency occurs because $D^2 f(x) \succ 0$ guarantees that f is locally strictly convex. Indeed, if convexity of f is a given, we can neglect second derivatives altogether, as $\nabla f(x^*) = 0$ is then a necessary and sufficient condition, see Lecture 1.

In the exercises of Problem Set 4 we used the fundamental theorem of linear inequalities to derive the LP duality theorem, yielding necessary and sufficient optimality conditions for linear programming, the simplest case of a constrained optimisation problem in which the objective and constraint functions are all linear. Note that the LP duality theorem only involved first order derivatives, and that linear programming is a special case of a *convex optimisation problem*, that is, the minimisation of a convex function over a convex domain.

It is thus natural to ask if optimality conditions of constrained optimisation problems

$$\begin{aligned} \text{(NLP)} \qquad & \min_{x\in\mathbb{R}^n} f(x) \\ & \text{s.t.} \quad g_i(x) = 0, \quad (i \in \mathcal{E}), \\ & \qquad g_j(x) \geq 0 \quad (j \in \mathcal{I}) \end{aligned}$$

mirror the situation in unconstrained optimisation, that is,

- first order conditions are necessary and sufficient for convex problems,
- second order conditions rely on strict local convexity.

In the next two lectures we will see that both points need much further refinement but hold under proper regularity assumptions.

2. First Order Necessary Optimality Conditions.

2.1. Mechanistic Interpretation. A useful picture in unconstrained optimisation is to imagine a point mass m or an infinitesimally small ball that moves on a hard surface

$$F := \big\{(x, f(x)) : x \in \mathbb{R}^n\big\}$$

without friction, see Figure 2.1. The external forces acting on the point mass are the gravity force $m\vec{g} = \left(\begin{smallmatrix} 0 \\ -mg \end{smallmatrix}\right)$ and the reaction force

$$\vec{N}_f = \frac{mg}{1+\|\nabla f(x)\|^2}\left(\begin{smallmatrix} -\nabla f(x) \\ 1 \end{smallmatrix}\right)$$

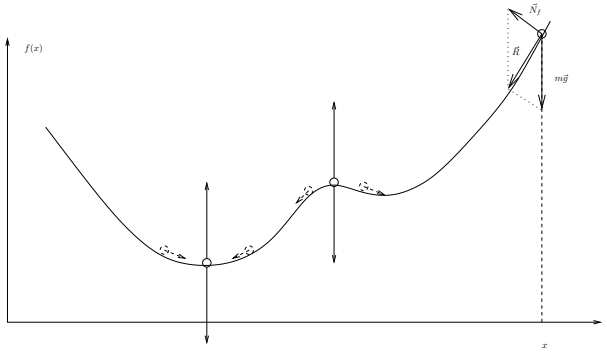

FIG. 2.1. *Mechanistic interpretation of unconstrained optimality conditions*

acting in normal direction away from the surface and countering the normal gravity component, so that the resulting total external force is

$$\vec{R} = m\vec{g} + \vec{N}_f = \frac{mg}{1+\|\nabla f(x)\|^2}\begin{bmatrix} -\nabla f(x) \\ -\|\nabla f(x)\|^2 \end{bmatrix} \perp \vec{N}_f.$$

This total force equals zero if and only if $\nabla f(x) = 0$, and if the test mass is placed at such a point then it will not move away, which is why we call such points *stationary points* of f. When the test mass is slightly moved from a local maximiser, then the external forces will pull it further away, whereas in a neighbourhood of a local minimiser they will restore the point mass to its former position. This is expressed by the second order optimality conditions: an equilibrium position is *stable* if $D^2 f(x) \succ 0$ and instable if $D^2 f(x) \prec 0$.

This mechanistic interpretation extends to constrained optimisation: we can interpret an inequality constraint $g(x) \geq 0$ as a hard smooth surface

$$G := \big\{(x, z) \in \mathbb{R}^n \times \mathbb{R} : g(x) = 0\big\}$$

which is parallel to the z-axis everywhere and keeps the point mass from rolling into the domain where $g(x) < 0$. Such a surface can exert only a normal force that points towards the domain $\{x : g_j(x) > 0\}$. Therefore, the reaction force must be of the form $\vec{N}_g = \mu_g\left(\begin{smallmatrix} \nabla g(x) \\ 0 \end{smallmatrix}\right)$, where $\mu_g \geq 0$. In the case depicted in Figure 2.2 where there is only one inequality constraint, the point mass is at rest and does not roll to lower terrain if the sum of external forces is zero, that is, $\vec{N}_f + \vec{N}_g + m\vec{g} = 0$. Since $\vec{N}_f = \mu_f\left(\begin{smallmatrix} -\nabla f(x) \\ 1 \end{smallmatrix}\right)$ for some $\mu_f \geq 0$, we find

$$\mu_f \begin{bmatrix} -\nabla f(x) \\ 1 \end{bmatrix} + \mu_g \begin{bmatrix} \nabla g(x) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -mg \end{bmatrix} = 0,$$

from where it follows that $\mu_f = mg$ and

$$\nabla f(x) = \lambda \nabla g(x) \tag{2.1}$$

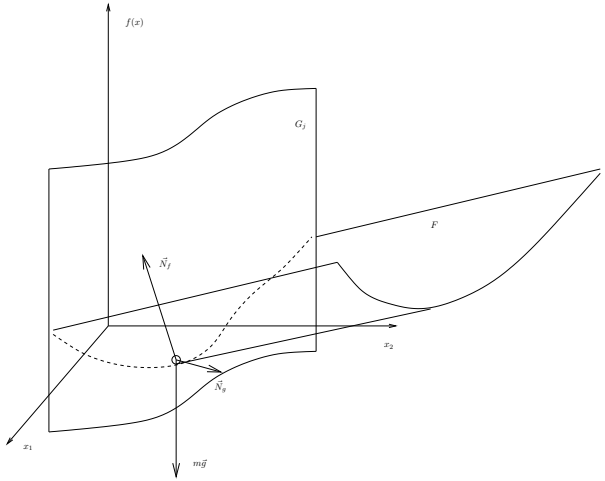

FIG. 2.2. *Mechanistic interpretation of constrained first order optimality conditions: the sum of external forces has to be zero.*

with $\lambda = \mu/mg \geq 0$. In other words, $\vec{N}_f$ is determined by the condition that its vertical component counter-balances the force of gravity, and $\vec{N}_g$ by the condition that it counter-balances the horizontal component of $\vec{N}_f$. This second condition is expressed in the balance equation (2.1).

When multiple inequality constraints are present, then the horizontal component of $\vec{N}_f$ must be counter-balanced by the *sum* of the reaction forces excerted by the constraint manifolds that touch the test mass. The balance equation (2.1) must thus be replaced with

$$\nabla f(x) = \sum_{j \in \mathcal{I}} \lambda_j \nabla g_j(x)$$

for some $\lambda_j \geq 0$, and since constraints for which $g_j(x) > 0$ cannot excert a force on the test mass, we must set $\lambda_j > 0$ for these indices, or equivalently, the equation $\lambda_j g_j(x) = 0$ must hold for all $j \in \mathcal{I}$.

It remains to discuss the influence of equality constraints when they are present. Replacing $g_i(x) = 0$ by the two inequality constraints $g_i(x) \geq 0$ and $-g_i(x) \geq 0$, our mechanistic interpretation yields two parallel surfaces G_i^+ and G_i^-, leaving an infinitesimally thin space between them within which our point mass is constrained to move, see Figure 2.3. The net reaction force of the two surfaces is of the form

$$\lambda_i^+ \nabla g_i(x) + \lambda_i^- \nabla(-g_i)(x) = \lambda_i \nabla g_i(x),$$

where we replaced the difference $\lambda_i^+ - \lambda_i^-$ of the bound-constrained variables $\lambda_i^+, \lambda_i^- \geq 0$ by a single unconstrained variable $\lambda_i = \lambda_i^+ - \lambda_i^-$. Note that in this case the conditions $\lambda_i^+ g_i(x) = 0$, $\lambda_i^-(-g_i(x)) = 0$ are satisfied automatically, since $g_i(x) = 0$ if x is feasible.

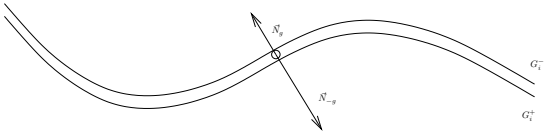

FIG. 2.3. *Mechanistic interpretation of an equality constraint: the net reaction force can point to either side of G_i. Here, we are looking down the z-axis.*

In summary, our mechanistic motivation suggests that if x is a local minimiser of (NLP), then there exist *Lagrange multipliers* $\lambda \in \mathbb{R}^{|\mathcal{I} \cup \mathcal{E}|}$ such that

$$\begin{aligned} \nabla f(x) - \sum_{i \in \mathcal{I} \cup \mathcal{E}} \lambda_i \nabla g_i(x) &= 0 \\ g_i(x) &= 0 \qquad (i \in \mathcal{E}) \\ g_j(x) &\geq 0 \qquad (j \in \mathcal{I}) \\ \lambda_j g_j(x) &= 0 \qquad (j \in \mathcal{I}) \\ \lambda_j &\geq 0 \qquad (j \in \mathcal{I}). \end{aligned}$$

These are the so-called Karush-Kuhn-Tucker (KKT) conditions. Our intuitive motivation cannot replace a rigorous proof, but the physical interpretation provides an easy explanation and provides a jog for memory.

2.2. Constraint Qualification. Before we start deriving the KKT conditions more rigorously, we introduce a few technical concepts and some notation.

DEFINITION 2.1. *Let $x^* \in \mathbb{R}^n$ be feasible for the problem (NLP). We say that the inequality constraint $g_j(x) \geq 0$ is* active *at x^* if $g(x^*) = 0$. We write $\mathcal{A}(x^*) := \{j \in \mathcal{I} : g_j(x^*) = 0\}$ for the set of indices corresponding to active inequality constraints.*

Of course, equality constraints are always active, but we will account for their indices separately. If $\mathcal{J} \subset \mathcal{E} \cup \mathcal{I}$ is a subset of indices, we will write $g_{\mathcal{J}}$ for the vector-valued map that has g_i $(i \in \mathcal{J})$ as components in some specific order. Furthermore, we write g for $g_{\mathcal{E} \cup \mathcal{I}}$.

There are situations in which our mechanical picture does not apply: if two inequality constraints have first order contact at a local minimiser, as in Figure 2.4, then they cannot annul the horizontal part of $\vec{N}_f$. In this case the mechanistic interpretation is flawed. When there are more constraints constraints, then generalisations of this situation can occur. In order to prove the KKT conditions, we must therefore make a regularity assumption on the constraints:

DEFINITION 2.2. *If $\{\nabla g_i : i \in \mathcal{E} \cup \mathcal{A}(x^*)\}$ is a linearly independent set of vectors, we say that the* linear independence constraint qualification *(LICQ) holds at x^*.*

The LICQ assumption guarantees that the linearisation of (NLP) around x^* is differential-topologically equivalent to (NLP) in a neighbourhood of x^*: the dimension of the manifold formed by the strictly feasible points in a neighbourhood of x^* must

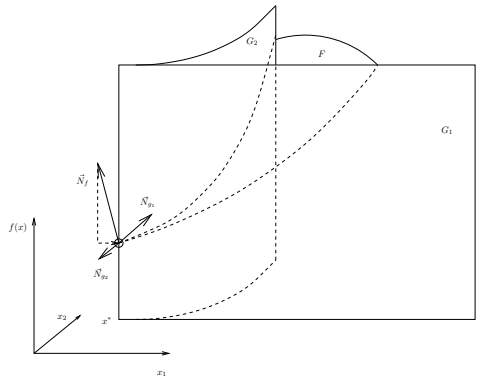

FIG. 2.4. *The mechanistic interpretation breaks down because the surfaces G_1 and G_2 are tangential to one another at x^*.*

remain the same after replacing each of the constraint surfaces by their tangent plane at x^*, see Figure 2.6. We illustrate two situations in which the dimension of the strictly feasible set changes in the linearised problem:

- Some of the active inequality constraint surfaces may not intersect properly: Figure 2.5 shows that when two constraint surfaces have first order contact, then the dimension of the strictly feasible set in the linearised problem collapses. Figure 2.6 shows that this doesn't happen when the constraint surfaces intersect properly.
- Some of the equality constraints may not intersect properly: Figure 2.7 shows that when two equality constraint surfaces have first order contact at x^*, then the set of points that satisfy the equality constraints has higher dimension in the linearised problem. At $\bar{x}$ however the equality constraints intersect properly and the dimension remains the same.

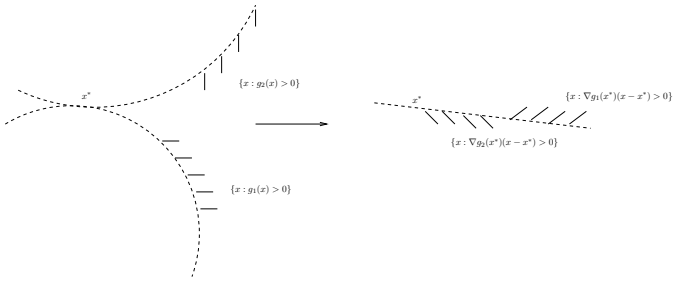

FIG. 2.5. *Two active inequality constraints do not intersect properly*

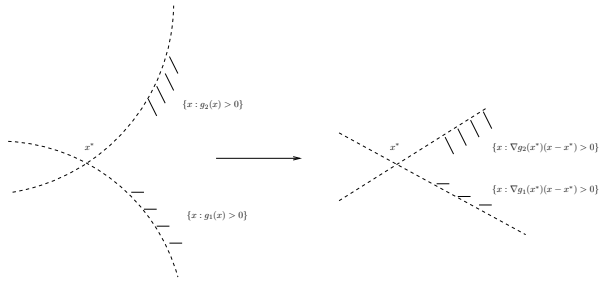

FIG. 2.6. *Two active inequality constraints intersect properly*

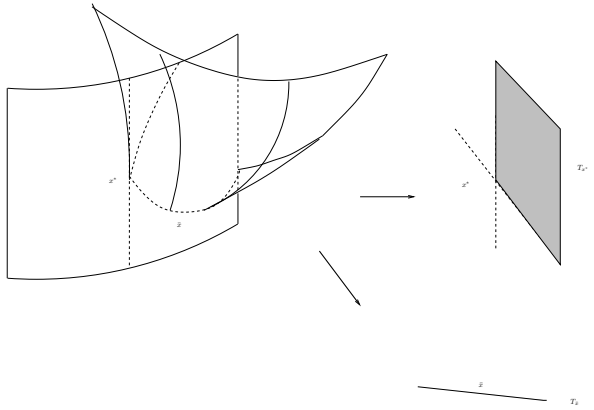

FIG. 2.7. *Two equality constraints have first order contact at x^* but intersect properly at $\bar{x}$. T_x is defined as $T_x := \{y : \nabla g_1(x)(y - x) = 0, \nabla g_2(x)(y - x) = 0\}$.*

2.3. A Key Lemma. In this section we will prove a lemma that will be extremely useful in everything that follows.

LEMMA 2.3. *Consider the problem (NLP) where f and g_i $(i \in \mathcal{E} \cup \mathcal{I})$ are C^k functions with $k \geq 1$. Let x^* be a feasible point where the LICQ holds and let $d \in \mathbb{R}^n$ be a vector such that*

$$\begin{aligned} d &\neq 0, \\ d^{\mathrm{T}} \nabla g_i(x^*) &= 0, \qquad (i \in \mathcal{E}), \\ d^{\mathrm{T}} \nabla g_j(x^*) &\geq 0, \qquad (j \in \mathcal{A}(x^*)). \end{aligned} \tag{2.2}$$

Then for $\epsilon > 0$ small enough there exists a path $x \in C^k\big((-\epsilon, +\epsilon), \mathbb{R}^n\big)$ such that

$$\begin{aligned} x(0) &= x^*, \\ \frac{d}{dt} x(0) &= d, \\ g_i(x(t)) &= t d^{\mathrm{T}} \nabla g_i(x^*) \quad (i \in \mathcal{E} \cup \mathcal{A}(x^*), t \in (-\epsilon, \epsilon)), \\ g_i(x(t)) &= 0, \qquad (i \in \mathcal{E}, t \in (-\epsilon, \epsilon)), \\ g_j(x(t)) &\geq 0 \qquad (j \in \mathcal{I}, t \geq 0). \end{aligned} \tag{2.3}$$

Proof. Let $l = |\mathcal{A}(x^*) \cup \mathcal{E}|$. Since the LICQ holds, it is possible to choose $Z \in \mathbb{R}^{(n-l)\times n}$ such that $\left[\begin{smallmatrix} Dg_{\mathcal{A}(x^*)\cup\mathcal{E}}(x^*) \\ Z \end{smallmatrix}\right]$ is a nonsingular $n \times n$ matrix. Let $h : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ be defined by

$$(x, t) \mapsto \left[\begin{smallmatrix} g_{\mathcal{A}(x^*)\cup\mathcal{E}}(x) - t Dg_{\mathcal{A}(x^*)\cup\mathcal{E}}(x^*) d \\ Z(x - x^* - td) \end{smallmatrix}\right]$$

Then $h'(x^*, 0) = \left[\begin{smallmatrix} D_x h(x^*,0) & D_t h(x^*,0) \end{smallmatrix}\right]$, where where

$$\begin{aligned} D_x h(x^*, 0) &= \left[\begin{smallmatrix} Dg_{\mathcal{A}(x^*)\cup\mathcal{E}}(x^*) \\ Z \end{smallmatrix}\right] \qquad \text{and} \\ D_t h(x^*, 0) &= -\left[\begin{smallmatrix} Dg_{\mathcal{A}(x^*)\cup\mathcal{E}}(x^*) d \\ Zd \end{smallmatrix}\right] = -D_x h(x^*, 0) d \end{aligned}$$

are the leading $n \times n$ and trailing $n \times 1$ block respectively.

Since $D_x h(x^*, 0)$ is nonsingular, the Implicit Function Theorem (see Lecture 8) implies that for $\tilde{\epsilon} > 0$ small enough there exists a unique C^k function $x : (-\tilde{\epsilon}, \tilde{\epsilon}) \to \mathbb{R}^n$ and a neighbourhood $\mathfrak{V}(x^*)$ such that for $x \in \mathfrak{V}(x^*)$, $t \in (-\tilde{\epsilon}, \tilde{\epsilon})$,

$$h(x, t) = 0 \Leftrightarrow x = x(t).$$

In particular, we have $x(0) = x^*$ and $g_i(x(t)) = t d^{\mathrm{T}} \nabla g(x^*)$ for all $i \in \mathcal{A}(x^*) \cup \mathcal{E}$ and $t \in (-\tilde{\epsilon}, \tilde{\epsilon})$. (2.2) therefore implies that $g_i(x(t)) = 0$ $(i \in \mathcal{E})$ and $g_i(x(t)) \geq 0$ $(i \in \mathcal{A}(x^*), t \in [0, \tilde{\epsilon}))$.

On the other hand, since $g_i(x^*) > 0$ $(i \notin \mathcal{A}(x^*))$, the continuity of $x(t)$ implies that there exists $\epsilon \in (0, \tilde{\epsilon})$ such that $g_j(x(t)) > 0$ $(j \in \mathcal{I} \setminus \mathcal{A}(x^*), t \in (-\epsilon, \epsilon))$.

Finally,

$$\frac{d}{dt} x(0) = -\big(D_x h(x^*, 0)\big)^{-1} D_t h(x^*, 0) = d$$

follows from the second part of the Implicit Function Theorem. □

2.4. KKT Conditions. We are ready to prove a theorem which shows that if x^* is a local minimiser for (NLP) and if the LICQ holds at x^* then x^* is a minimiser of the linear programming problem obtained by linearising (NLP) around x^*.

THEOREM 2.4. *If x^* is a local minimiser of (NLP) where the LICQ holds then*

$$\nabla f(x^*) \in \operatorname{cone}\left(\{\pm \nabla g_i(x^*) : i \in \mathcal{E}\} \cup \{\nabla g_j(x^*) : j \in \mathcal{A}(x^*)\}\right).$$

Proof. Suppose our claim is wrong. Then the fundamental theorem of linear inequalities implies that there exists a vector $d \in \mathbb{R}^n$ such that

$$\begin{aligned} d^{\mathrm{T}} \nabla g_j(x^*) &\geq 0, \qquad (j \in \mathcal{A}(x^*)), \\ \pm d^{\mathrm{T}} \nabla g_i(x^*) &\geq 0, \quad (\text{i.e., } d^{\mathrm{T}} \nabla g_i(x^*) = 0) \qquad (i \in \mathcal{E}), \\ d^{\mathrm{T}} \nabla f(x^*) &< 0. \end{aligned}$$

Since d satisfies (2.2), Lemma 2.3 implies that there exists a path $x : (-\epsilon, \epsilon) \to \mathbb{R}^n$ that satisfies (2.3). Taylor's theorem then implies that

$$f(x(t)) = f(x^*) + t d \nabla f(x^*) + O(t^2) < f(x^*)$$

for $0 < t \ll 1$. Since (2.3) shows that $x(t)$ is feasible for $t \in [0, \epsilon)$, this contradicts the assumption that x^* is a local minimiser. □

Note that the condition

$$\nabla f(x^*) \in \operatorname{cone}\left(\{\pm \nabla g_i(x^*) : i \in \mathcal{E}\} \cup \{\nabla g_j(x^*) : j \in \mathcal{A}(x^*)\}\right)$$

is equivalent to the existence of $\lambda \in \mathbb{R}^{|\mathcal{E}\cup\mathcal{I}|}$ such that

$$\nabla f(x^*) = \sum_{i \in \mathcal{E}\cup\mathcal{I}} \lambda_i \nabla g_i(x^*), \tag{2.4}$$

where $\lambda_j \geq 0$ $(j \in \mathcal{A}(x^*))$ and $\lambda_j = 0$ for $(j \in \mathcal{I} \setminus \mathcal{A}(x^*))$. Moreover, x^* must be feasible. Thus, Theorem 2.4 shows that when x^* is a local minimiser where the LICQ holds, then the KKT conditions must hold.

We can formulate this result in slightly more abstract form in terms of the Lagrangian associated with (NLP):

$$\begin{aligned} \mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m &\to \mathbb{R} \\ (x, \lambda) &\mapsto f(x) - \sum_{i=1}^{m} \lambda_i g_i(x). \end{aligned}$$

Equation (2.4) says that the derivative of the Lagrangian with respect to the x coordinates is zero. Putting all the pieces together, we obtain the following theorem:

THEOREM 2.5 (First Order Necessary Optimality Conditions). *If x^* is a local minimiser of (NLP) where the LICQ holds then there exists $\lambda^* \in \mathbb{R}^m$ such that (x^*, λ^*) solves the following system of inequalities,*

$$\begin{aligned} D_x\mathcal{L}(x^*, \lambda^*) &= 0, \\ \lambda_j^* &\geq 0 \quad (j \in \mathcal{I}), \\ \lambda_i^* g_i(x^*) &= 0 \quad (i \in \mathcal{E} \cup \mathcal{I}), \\ g_j(x^*) &\geq 0 \quad (j \in \mathcal{I}), \\ g_i(x^*) &= 0 \quad (i \in \mathcal{E}). \end{aligned}$$

2.5. The Method of Lagrange Multipliers. In this section we show by ways of an example how the KKT conditions can be used to solve nonlinear optimisation problems. This method of solving optimisation problems is called *the method of Lagrange multipliers*. Even though this approach can be carried through explicitly only when the number of constraints is small, numerical algorithms for nonlinear programming are actually designed to do the same in situations where the calculations cannot be done by hand.

EXAMPLE 2.6. *Solve the following nonlinear programming problem*

$$\begin{aligned} \min_{x \in \mathbb{R}^2} \; & f(x) = x_1^3 + x_2 \\ s.t. \quad & g_1(x) := x_1^2 + 2x_2^2 - 1 = 0 \\ & g_2(x) := x_1 \geq 0. \end{aligned} \tag{2.5}$$

We have

$$\nabla g_1(x) = \begin{bmatrix} 2x_1 \\ 4x_2 \end{bmatrix}, \; \nabla g_2(x) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \; \nabla f(x) = \begin{bmatrix} 3x_1^2 \\ 1 \end{bmatrix}.$$

If we want to find all points x^* that satisfy the conditions of Theorem 2.4 we have to distinguish two cases corresponding to $\mathcal{A}(x^*) = \emptyset$ and $\mathcal{A}(x^*) = \{2\}$.

If $\mathcal{A}(x^*) = \emptyset$ then $x_1^* > 0$. We must find $\lambda_1^* \in \mathbb{R}$ such that

$$\begin{bmatrix} 3x_1^{*2} \\ 1 \end{bmatrix} = \lambda_1^* \begin{bmatrix} 2x_1^* \\ 4x_2^* \end{bmatrix},$$

which is equivalent to

$$\lambda_1^* = \frac{3x_1^*}{2} \neq 0, \qquad \lambda_1^* = \frac{1}{4x_2^*}$$

and implies

$$6x_1^* x_2^* = 1. \tag{2.6}$$

In particular, $x_2^* \neq 0$, and hence $\nabla g_1(x^*), \nabla g_2(x^*)$ are linearly independent, that is, the LICQ holds at these points. We also need x^* to be feasible,

$$x_1^{*2} + 2x_2^{*2} = 1. \tag{2.7}$$

(2.6) and (2.7) together imply $18x_1^{*4} - 18x_1^{*2} + 1 = 0$, which shows that $x_1^{*2} \in \{0.941, 0.059\}$. Since we have assumed $x_1^* > 0$ this leaves the two possible solutions

$$x^{[1]} = \begin{bmatrix} 0.243 \\ 0.6859 \end{bmatrix}, \quad x^{[2]} = \begin{bmatrix} 0.97 \\ 0.1718 \end{bmatrix}$$

with function values $f(x^{[1]}) = 0.7002$, $f(x^{[2]}) = 1.0845$.

In the second case where $\mathcal{A}(x^*) = \{2\}$ we have $x_1^* = 0$. We need to find $\lambda_1^* \in \mathbb{R}$ and $\lambda_2^* \geq 0$ such that

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \lambda_1^* \begin{bmatrix} 0 \\ 4x_2^* \end{bmatrix} + \lambda_2^* \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

This implies $\lambda_2^* = 0$. Moreover, since x^* must be feasible and hence $2x_2^{*2} = 1 \neq 0$, we have $\lambda_1^* = (4x_2^*)^{-1}$ and $x_2 = \pm 1/\sqrt{2}$. This yields the two further candidate points

$$x^{[3]}, x^{[4]} = \pm \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}.$$

At these points we have $\nabla g_1(x^*) = \left(\begin{smallmatrix} 0 \\ \pm 2\sqrt{2} \end{smallmatrix}\right)$ and $\nabla g_2(x^*) = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$, and hence, the LICQ holds. The objective function values are $f(x^{[3]}) = 0.7071$, $f(x^{[4]}) = -0.7071$.

Overall we find four candidate points where the KKT conditions hold. Among these four points, $x^{[4]}$ has the smallest objective value, and this must be the global minimiser of our problem.