First-order necessary conditions for equality-constrained optimization: Introduction to Lagrange multipliers

Mark S. Gockenbach

Both the theory of and algorithms for constrained optimization are more complicated than in the case of unconstrained optimization. I will begin with the equality-constrained nonlinear program

$$\begin{array}{ll} \min & f(x) \\ s.t. & g(x) = 0, \end{array} \tag{1}$$

where $f : \mathrm{R}^n \to \mathrm{R}$ and $g : \mathrm{R}^n \to \mathrm{R}^m$. The set of all $x \in \mathrm{R}^n$ such that $g(x) = 0$ is called the *feasible set* or *feasible region*. Geometrically, it is a lower-dimensional manifold in R^n. For example, the single constraint

$$x_1^2 + x_2^2 \Leftrightarrow 1 = 0$$

defines a circle (a one-dimensional set) in R^2, while the similar constraint

$$x_1^2 + x_2^2 + x_3^2 \Leftrightarrow 1 = 0$$

defines a sphere (a two-dimensional set) in R^3. The pair of constraints

$$\begin{array}{rcl} x_1^2 + x_2^2 + x_3^2 \Leftrightarrow 1 & = & 0 \\ x_1 + x_2 + x_3 & = & 1 \end{array}$$

defines a circle (a one-dimensional set) in R^3, the intersection of a sphere and a plane. These two constraints can be represented as $g(x) = 0$, where $g : \mathrm{R}^3 \to \mathrm{R}^2$ is defined by

$$g(x) = \left[\begin{array}{c} x_1^2 + x_2^2 + x_3^2 \Leftrightarrow 1 \\ x_1 + x_2 + x_3 \end{array} \right]. \tag{2}$$

Generally it might be expected that $g : \mathrm{R}^n \to \mathrm{R}^m$ defines a manifold of dimension $n \Leftrightarrow m$; however, while this is the typical case, it is not always true. In the preceding example, it is easy to change the second constraint, which defines a plane, so that the plane intersects the sphere in a single point or not at all.

1 Optimality conditions

Minimizers of (1) are defined in the obvious way. A global minimizer of (1) is a feasible point x^* with the property that $f(x^*) \le f(x)$ for every other feasible point x. A local minimizer is a feasible point x^* with the property that there exists $\delta > 0$ such that

$$\|x^* \Leftrightarrow x\| < \delta,\ g(x) = 0 \ \Rightarrow\ f(x^*) \le f(x).$$

For a strict local minimizer, the condition is

$$0 < \|x^* \Leftrightarrow x\| < \delta,\ g(x) = 0 \ \Rightarrow\ f(x^*) < f(x).$$

As in the case of unconstrained optimization, the algorithms that I present attempt to locate a local minimizer.

To derive an optimality condition for (1), I will relate the constrained problem to an unconstrained problem (actually, a family of unconstrained problems). A function $x : [-a, a] \to \mathrm{R}^n$ defines a *path* in R^n; if x satisfies

$$g(x(t)) = 0 \text{ for all } t \in [-a, a],$$

then x defines a *feasible path.* Suppose now that x^* is a local minimizer of (1) and x is a feasible path with $x(0) = x^*$. Then the function $\phi : [-a, a] \to \mathrm{R}$ defined by

$$\phi(t) = f(x(t))$$

has a local minimizer at $t = 0$, and hence $\phi'(0) = 0$ must hold. By the chain rule,

$$\phi'(t) = \nabla f(x(t)) \cdot \dot{x}(t),$$

where

$$\dot{x}(t) = \begin{bmatrix} \frac{dx_1}{dt}(t) \\ \frac{dx_2}{dt}(t) \\ \vdots \\ \frac{dx_n}{dt}(t) \end{bmatrix}.$$

It follows that a necessary condition for x^* to be a local minimizer of (1) is that

$$\nabla f(x^*) \cdot \dot{x}(0) = 0 \tag{3}$$

for all feasible paths x with $x(0) = x^*$. Similar to the case of unconstrained minimization, this first-order condition is also satisfied at a local maximizer.

A feasible path x has the following property, which follows from the chain rule:

$$g(x(t)) = 0 \;\Rightarrow\; J(x(t))\dot{x}(t) = 0.$$

Here I write J for the Jacobian of g. In particular, if x is a feasible path and $x(0) = x^*$, then

$$J(x^*)\dot{x}(0) = 0,$$

which shows that $\dot{x}(0)$ belongs to $\mathcal{N}(J(x^*))$, the null space of the matrix $J(x^*)$.

I will now adopt the following notation, which is common in the optimization literature: the transpose of the Jacobian of g will be denoted ∇g. Thus $\nabla g(x) = J(x)^T$, where $J(x)$ is the Jacobian of g at x. The function g is vector-valued,

$$g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{bmatrix},$$

and each component g_i has a gradient ∇g_i. Since

$$J(x)_{ij} = \frac{\partial g_i}{\partial x_j}(x),$$

the rows of $J(x)$ are the gradients $\nabla g_1(x), \nabla g_2(x), \ldots, \nabla g_m(x)$. It follows that $\nabla g_j(x)$ is the jth column of $\nabla g(x)$, which explains the notation.[1]

[1] However, the reader should be aware that this notation is not necessarily accepted outside of the optimization literature. In continuum mechanics, for instance, $\nabla g(x)$ denotes the Jacobian of g at x!

In order to derive useful necessary conditions for nonlinear programs, it is necessary to make some assumption about the constraint g at x^*. An assumption which allows us to state the necessary conditions in terms of *Lagrange multipliers*, is called a *constraint qualification*. (I define Lagrange multipliers below.)

Constraint qualification 1.1 *For each $z \in \mathcal{N}(\nabla g(x^*)^T)$, there exists $a > 0$ and a feasible path $x : [\Leftrightarrow a, a] \to \mathrm{R}^n$ such that $x(0) = x^*$ and $\dot{x}(0) = z$.*

This constraint qualification states that $\mathcal{N}(\nabla g(x^*)^T)$ is precisely the set of feasible directions at x^*.

Assuming constraint qualification 1.1 holds at x^*, the following condition is a necessary condition for x^* to be a local minimizer of (1):

$$z \in \mathcal{N}(\nabla g(x^*)^T) \;\Rightarrow\; \nabla f(x^*) \cdot z = 0. \tag{4}$$

In order to express (4) in a form that is useful for both theory and computation, I must digress to review some basic results from linear algebra.

2 The Fundamental Theorem of Linear Algebra

The following theorem, which I present without proof, is one of the most important results from linear algebra.

Theorem 2.1 (The projection theorem) *Suppose V is any inner product space (that is, vector space with an inner product) and W is a finite-dimensional subspace of V. Given any $v \in V$, there exists a unique vector $w \in W$ closest to v. In other words, there is a unique solution to*

$$\min_{z \in W} \|v \Leftrightarrow z\|.$$

Moreover, this closest vector w is characterized by the following orthogonality condition:

$$(v \Leftrightarrow w, z) = 0 \text{ for all } z \in W. \tag{5}$$

In the above theorem, (x, y) denotes the inner product of two vectors $x, y \in V$. The vector w is called the *orthogonal projection* of v onto W, or the *best approximation* to v from W. It is sometimes denoted by $w = \mathrm{proj}_W v$.

Given any matrix $A \in \mathrm{R}^{m \times n}$, the *range* (or *column space*) of A is defined by

$$\mathcal{R}(A) = \{Ax \;:\; x \in \mathrm{R}^n\}$$

and the *null space* (or *kernel*) of A is

$$\mathcal{N}(A) = \{x \in \mathrm{R}^n \;:\; Ax = 0\}.$$

I wish to discuss the relationships that exist between the ranges and null spaces of A and A^T. The following concept is crucial.

Definition 2.2 *Suppose S is a nonempty subset of R^n. The* orthogonal complement *of S is the set*

$$S^\perp = \{y \in \mathrm{R}^n \;:\; x \cdot y = 0 \;\forall\; x \in S\}.$$

The basic properties of orthogonal complements are summarized in the following theorem.

Theorem 2.3 *Suppose S is a nonempty subset of R^n. Then*

1. *$S^\perp$ is a subspace of R^n.*
2. *$\left(S^\perp\right)^\perp = span(S)$, that is, $\left(S^\perp\right)^\perp$ is the smallest subspace containing S. In particular, if S is a subspace of R^n, then $\left(S^\perp\right)^\perp = S$.*
3. *If S is a subspace of R^n, then $\mathrm{R}^n = S \oplus S^\perp$, that is, every $x \in \mathrm{R}^n$ can be written uniquely as $x = y + z$, where $y \in S$ and $z \in S^\perp$.*

Proof:

1. The set $S^\perp$ is nonempty, since $0 \in S^\perp$ clearly holds. If $y, z \in S^\perp$ and $\alpha, \beta \in \mathrm{R}$, then, for any $x \in S$,
$$x \cdot (\alpha y + \beta z) = \alpha(x \cdot y) + \beta(x \cdot z) = \alpha 0 + \beta 0 = 0.$$
This shows that $\alpha y + \beta z \in S^\perp$, and hence that $S^\perp$ is a subspace of R^n.

2. If $x \in \mathrm{span}(S)$, then there exist $x^{(1)}, x^{(2)}, \ldots, x^{(k)} \in S$ and $\alpha_1, \alpha_2, \ldots, \alpha_k \in \mathrm{R}$ such that
$$x = \sum_{i=1}^{k} \alpha_i x^{(i)}.$$
I need to show that x is orthogonal to every $y \in S^\perp$. But $y \in S^\perp$ implies that $y \cdot x^{(i)} = 0$ for $i = 1, 2, \ldots, k$ (y is orthogonal to every vector in S), and so
$$y \in S^\perp \;\Rightarrow\; y \cdot x = y \cdot \left(\sum_{i=1}^{k} \alpha_i x^{(i)} \right) = \sum_{i=1}^{k} \alpha_i (y \cdot x^{(i)}) = 0.$$
Therefore $x \in \left(S^\perp\right)^\perp$.

3. Given any $x \in \mathrm{R}^n$, there exists a unique best approximation y to x in S, that is, a unique solution to
$$\min_{u \in S} \|x \Leftrightarrow u\|.$$
This best approximation y is characterized by the condition
$$(x \Leftrightarrow y) \cdot u = 0 \text{ for all } u \in S. \tag{6}$$
But if $z = x \Leftrightarrow y$, then $x = y + z$, $y \in S$, and (6) shows that $z \in S^\perp$. QED

I can now state the main theorem of this section.

Theorem 2.4 *(The Fundamental Theorem of Linear Algebra) Suppose $A \in \mathrm{R}^{m \times n}$. Then*

1. $\mathcal{N}(A)^\perp = \mathcal{R}(A^T)$ and $\mathcal{R}(A^T)^\perp = \mathcal{N}(A)$;

2. $\mathcal{N}(A^T)^\perp = \mathcal{R}(A)$ and $\mathcal{R}(A)^\perp = \mathcal{N}(A^T)$.

Proof: Since $\left(S^\perp\right)^\perp = S$ for every subspace S, the first part of 1 follows immediately from the second part. Moreover, 2 follows from applying 1 to the matrix A^T in place of A. Therefore, it suffices to prove
$$\mathcal{R}(A^T)^\perp = \mathcal{N}(A).$$
Suppose first that $x \in \mathcal{N}(A)$. If $z \in \mathcal{R}(A^T)$, there exists $y \in \mathrm{R}^m$ such that $z = A^T y$. Therefore,
$$z \cdot x = (A^T y) \cdot x = y \cdot (Ax) = 0,$$
since $Ax = 0$. Therefore $x \in \mathcal{R}(A^T)^\perp$.

On the other hand, suppose $x \in \mathcal{R}(A^T)^\perp$. Then
$$Ax = 0 \;\Leftrightarrow (Ax) \cdot y = 0 \text{ for all } y \in \mathrm{R}^m.$$
But, for any $y \in \mathrm{R}^m$,
$$(Ax) \cdot y = x \cdot (A^T y) = 0$$
since $x \in \mathcal{R}(A^T)^\perp$. This shows that $x \in \mathcal{N}(A)$. QED

3 Optimality conditions and Lagrange multipliers

Above I derived the necessary condition (4), which I repeat here for convenience:

$$z \in \mathcal{N}(\nabla g(x^*)^T) \;\Rightarrow\; \nabla f(x^*) \cdot z = 0. \tag{7}$$

In the language of the previous section, (7) can be expressed as

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

But then, by the Fundamental Theorem of Linear Algebra,

$$\nabla f(x^*) \in \mathcal{R}(\nabla g(x^*)),$$

that is, there exists $\lambda^* \in \mathrm{R}^m$ such that

$$\nabla f(x^*) = \nabla g(x^*)\lambda^* = \sum_{j=1}^{m} \lambda_j^* \nabla g_j(x^*). \tag{8}$$

The vector λ^* is called a *Lagrange multiplier* (or a vector of Lagrange multipliers) for (1). The reader should notice that (8), being an system of equations, is a much more tractable necessary condition than (3) or (7). Indeed, taking into account the constraint itself, the necessary condition can be expressed as follows: If $x^* \in \mathrm{R}^n$ is a local minimizer of (1) and constraint qualification 1.1 holds at x^*, then there exists $\lambda^* \in \mathrm{R}^m$ such that

$$\begin{aligned} g(x^*) &= 0, && (9)\\ \nabla f(x^*) &= \nabla g(x^*)\lambda^*. && (10) \end{aligned}$$

Conditions (9–10) form a system of $n+m$ (nonlinear) equations for the $n+m$ unknowns x^* and λ^*; these conditions must hold whether x^* is a local minimizer or a local maximizer of the nonlinear program (1).

I will now give two examples, one illustrating the existence of Lagrange multipliers and the other showing the need for a constraint qualification.

Example 3.1 *I define* $f : \mathrm{R}^3 \to \mathrm{R}$ *by*

$$f(x) = (x_1 \Leftrightarrow 3)^2 + 2(x_2 \Leftrightarrow 1)^2 + 3(x_3 \Leftrightarrow 1)^2$$

and $g : \mathrm{R}^3 \to \mathrm{R}$ *by*

$$g(x) = \begin{bmatrix} x_1^2 + x_2^2 + x_3^2 \Leftrightarrow 1 \\ x_1 + x_2 + x_3 \end{bmatrix}.$$

The feasible set of the nonlinear program

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & g(x) = 0 \end{aligned}$$

is a circle, the intersection of the sphere $x_1^2 + x_2^2 + x_3^2 = 1$ *and the plane* $x_1 + x_2 + x_3 = 0$. *The gradients are given by*

$$\nabla f(x) = \begin{bmatrix} 2(x_1 \Leftrightarrow 3) \\ 4(x_2 \Leftrightarrow 1) \\ 6(x_3 \Leftrightarrow 1) \end{bmatrix}, \quad \nabla g(x) = \begin{bmatrix} 2x_1 & 1 \\ 2x_2 & 1 \\ 2x_3 & 1 \end{bmatrix}.$$

The necessary conditions for a local optimum are

$$\begin{aligned} g(x) &= 0, \\ \nabla f(x) &= \nabla g(x)\lambda, \end{aligned}$$

which reduce to

$$\begin{aligned} x_1^2 + x_2^2 + x_3^2 &= 1, \\ x_1 + x_2 + x_3 &= 0, \\ 2(x_1 - 3) &= 2\lambda_1 x_1 + \lambda_2, \\ 4(x_2 - 1) &= 2\lambda_1 x_2 + \lambda_2, \\ 6(x_3 - 1) &= 2\lambda_1 x_3 + \lambda_2. \end{aligned}$$

This is a system of five equations for the five unknowns $x_1, x_2, x_3, \lambda_1, \lambda_2$. It can be solved algebraically or numerically. I used MATLAB's Symbolic Toolbox to solve the system and then rounded the solutions to five digits. In this case, which is very simple, there are exactly two solutions,

$$x^* \doteq \begin{bmatrix} 0.058183 \\ 0.73440 \\ -0.67622 \end{bmatrix}, \ \lambda^* \doteq \begin{bmatrix} 3.1883 \\ -5.7454 \end{bmatrix}, \ f(x^*) \doteq 17.923$$

and

$$x^* \doteq \begin{bmatrix} 0.67622 \\ -0.73440 \\ 0.058183 \end{bmatrix}, \ \lambda^* \doteq \begin{bmatrix} 0.81171 \\ -5.7454 \end{bmatrix}, \ f(x^*) \doteq 14.077.$$

The values of f at the two solutions show that the first is the global maximizer and the second is the global minimizer.

The reader should appreciate that the previous example is special in that it is possible to solve the necessary conditions algebraically and identify all possible optima. In most problems of practical interest, an algebraic solution is not possible and numerical algorithms are needed; moreover, finding the global optima is usually not practical.

Example 3.2 *I will now consider the equality-constrained nonlinear program with objective function*

$$f(x) = x_1^2 + (x_2 - 1)^2 + x_3^2$$

and constraint function

$$g(x) = \begin{bmatrix} x_2^2 - x_3 \\ 2x_2^2 - x_3 \end{bmatrix}.$$

It is easy to see that $g(x) = 0$ is satisfied only by points on the x_1-axis, that is, by x satisfying $x_2 = x_3 = 0$. The gradients are

$$\nabla f(x) = \begin{bmatrix} 2x_1 \\ 2(x_2 - 1) \\ 2x_3 \end{bmatrix}, \ \nabla g(x) = \begin{bmatrix} 0 & 0 \\ 2x_2 & 4x_2 \\ -1 & -1 \end{bmatrix}.$$

Since $g(x) = 0$ implies that $x_2 = x_3 = 0$, the necessary conditions (9–10) imply

$$\begin{bmatrix} 2x_1 \\ -2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -\lambda_1 - \lambda_2 \end{bmatrix},$$

which has no solution. Therefore, in this case, the Lagrange multiplier does not exist. It is easy to see that $x^* = 0$ is the global minimizer for the problem.

Since the Lagrange multiplier does not exist, it must be the case that the constraint qualification fails to hold at x^*. Indeed,

$$\nabla g(0)^T z = 0 \ \Rightarrow \ z_3 = 0,$$

so the null space of $\nabla g(x^*)$ is the x_1x_2-plane. However, the only feasible paths stay on the x_1-axis. It follows that for any $z \in \mathcal{N}(\nabla g(x^*)^T)$ with $z_2 \neq 0$ (for example, $z = (0, 1, 0)$), there is no feasible path x with $x(0) = x^*$ and $\dot{x}(0) = z$.

4 A practical constraint qualification

Constraint qualification 1.1 is not easily verifiable, although it is just what is needed to prove the existence of Lagrange multipliers. I now present another constraint qualification, stronger than 1.1, that is easier to verify.

Constraint qualification 4.1 *The matrix $\nabla g(x^*)$ has full rank, that is, the columns of $\nabla g(x^*)$ are linearly independent.*

A point x^* satisfying constraint qualification 4.1 is said to be a *regular point* of the feasible set defined by $g(x) = 0$ or of the nonlinear program (1).

I will now prove that constraint qualification (1.1) is satisfied at every regular point of $g(x) = 0$. The proof is fairly simple in outline, although the notation is rather difficult to follow. I will need some results concerning linear least-squares problems.

Theorem 4.2 *Suppose $A \in \mathrm{R}^{m \times n}$, where $m \geq n$, has full rank. Then, for any $b \in \mathrm{R}^m$, the least-squares problem*

$$\min \|Ax \Leftrightarrow b\| \tag{11}$$

has a unique solution, given by

$$x = \left(A^T A\right)^{-1} A^T b.$$

Proof: The solution x to the least-squares problem (11) satisfies

$$Ax = \mathrm{proj}_{\mathcal{R}(A)} b.$$

Therefore, by the projection theorem, $b \Leftrightarrow Ax$ must be orthogonal to $\mathcal{R}(A)$, and hence

$$(Ay) \cdot (b \Leftrightarrow Ax) = 0 \text{ for all } y \in \mathrm{R}^n.$$

But

$$(Ay) \cdot (b \Leftrightarrow Ax) = y \cdot \left(A^T b \Leftrightarrow A^T Ax\right),$$

and

$$y \cdot \left(A^T b \Leftrightarrow A^T Ax\right) = 0 \text{ for all } y \in \mathrm{R}^n$$

holds if and only if $A^T b \Leftrightarrow A^T Ax = 0$. Since A has full rank, $A^T A$ is invertible, so the unique solution x is given by

$$x = \left(A^T A\right)^{-1} A^T b.$$

QED

Corollary 4.3 *Suppose $A \in \mathrm{R}^{m \times n}$, where $m \geq n$, has full rank and $b \in \mathrm{R}^m$. Then the orthogonal projection of b onto $\mathcal{R}(A)$ is*

$$A \left(A^T A\right)^{-1} A^T b$$

and the projection of b onto

$$\mathcal{R}(A)^{\perp} = \mathcal{N}(A^T)$$

is

$$b \Leftrightarrow A \left(A^T A\right)^{-1} A^T b.$$

Proof: The first conclusion follows immediately from the previous theorem. Also, by the previous theorem,

$$b \Leftrightarrow A\left(A^T A\right)^{-1} A^T b \in \mathcal{R}(A)^{\perp}.$$

Therefore, the second conclusion follows from the projection theorem if

$$b \Leftrightarrow \left(b \Leftrightarrow A\left(A^T A\right)^{-1} A^T b\right)$$

is orthogonal to $\mathcal{R}(A)^{\perp}$. But

$$b \Leftrightarrow \left(b \Leftrightarrow A\left(A^T A\right)^{-1} A^T b\right) = A\left(A^T A\right)^{-1} A^T b \in \mathrm{R}(A).$$

Every vector in $\mathcal{R}(A)$ is orthogonal to $\mathcal{R}(A)^{\perp}$ by definition. QED

In addition to the above results from linear algebra, I also need the following standard existence theorem from the theory of ordinary differential equations. This is a version of Peano's existence theorem.

Theorem 4.4 *Suppose $f : N \to \mathrm{R}^n$ is continuous and $x^{(0)} \in N$, where N is an open set in R^n. Then there exists $a > 0$ and $x : [\Leftrightarrow a, a] \to \mathrm{R}^n$ such that x solves the initial value problem (IVP)*

$$\begin{aligned} \dot{x} &= f(x), \\ x(0) &= x^{(0)}, \end{aligned}$$

that is,

$$\begin{aligned} \dot{x}(t) &= f(x(t)) \textit{ for all } t \in [\Leftrightarrow a, a], \\ x(0) &= x^{(0)}. \end{aligned}$$

I am now ready to prove the following theorem, which shows that regularity is a stronger condition than the constraint qualification (1.1). The idea[2] is the following: Given a vector $z \in \mathcal{N}(\nabla g(x^*)^T)$, a feasible path x with $x(0) = x^*$ and $\dot{x}(0) = z$ will be constructed as the solution to an initial value problem. The right-hand side of the ODE will be formed by projecting z onto

$$\mathcal{N}(\nabla g(x)^T) = \mathcal{R}(\nabla g(x))^{\perp}.$$

Applying the above results, the projection of z onto $\mathcal{N}(\nabla g(x)^T)$ is

$$z \Leftrightarrow \nabla g(x)\left(\nabla g(x)^T \nabla g(x)\right)^{-1} \nabla g(x)^T z.$$

Theorem 4.5 *Suppose $g : \mathrm{R}^n \to \mathrm{R}^m$ is continuously differentiable and x^* is a regular point of $g(x) = 0$. If $z \in \mathcal{N}(\nabla g(x^*)^T)$, then there exists $a > 0$ and a path $x : [\Leftrightarrow a, a] \to \mathrm{R}^n$ such that*

$$\begin{aligned} g(x(t)) &= 0 \textit{ for all } t \in [\Leftrightarrow a, a], \\ x(0) &= x^*, \\ \dot{x}(0) &= z. \end{aligned}$$

Proof: Since g is continuously differentiable and $\nabla g(x^*)$ has full rank, there exists a neighborhood N of x^* on which $\nabla g(x)$ has full rank and therefore $\left(\nabla g(x)^T \nabla g(x)\right)^{-1}$ exists. I can therefore define $f : N \to \mathrm{R}^n$ by

$$f(x) = z \Leftrightarrow \nabla g(x)\left(\nabla g(x)^T \nabla g(x)\right)^{-1} \nabla g(x)^T z,$$

[2] Taken from Tapia [1].

and, since g is continuously differentiable, f is continuous. It follows that the IVP

$$\begin{aligned} \dot{x} &= f(x), \\ x(0) &= x^* \end{aligned}$$

has a solution x that exists on some interval $[\Leftrightarrow a, a]$. Since $z \in \mathcal{N}(\nabla g(x^*)^T)$, it follows that $f(x^*) = z$ and hence $\dot{x}(0) = z$. It remains only to show that x is feasible, that is, that

$$g(x(t)) = 0 \text{ for all } t \in [\Leftrightarrow a, a].$$

I will use the mean value theorem to show that this hold for each component of g, that is, that for each $i = 1, 2, \ldots, n$,

$$g_i(x(t)) = 0 \text{ for all } t \in [\Leftrightarrow a, a].$$

By the MVT, for each i and $t \in [\Leftrightarrow a, a]$, there exists t^* between 0 and t such that

$$g_i(x(t)) = g(x(0)) + t\nabla g(x(t^*)) \cdot \dot{x}(t^*).$$

By construction, $x(0) = x^*$, so $g(x(0)) = 0$. Moreover,

$$\nabla g_i(x(t^*)) \cdot \dot{x}(t^*) = \nabla g_i(x(t^*)) \cdot f(x(t^*)).$$

However, by construction, $f(x(t^*)) \in \mathcal{N}(\nabla g(x(t^*))^T)$ and hence

$$\nabla g_i(x(t^*)) \cdot f(x(t^*)) = \left(\nabla g(x(t^*))^T f(x(t^*))\right)_i = 0.$$

Therefore, $g(x(t)) = 0$. QED

Regularity is thus a sufficient condition for the existence of Lagrange multipliers. It is also a necessary and sufficient condition for the uniqueness of the multipliers. This is because, given a local optimum x^*, the Lagrange multiplier λ^* is determined by the linear system

$$\nabla g(x^*)\lambda^* = \nabla f(x^*),$$

and the solution to this system (if it exists) is unique provided the columns of $\nabla g(x^*)$ form a linearly independent set. (In Example 3.2, the two columns of $\nabla g(x^*)$ were identical, hence linearly dependent. Therefore, x^* was not a regular point, and no Lagrange multiplier existed.)

5 Digression: An application of Lagrange multipliers to spectral theory

The following fact about eigenvalues has been useful in previous lectures: If $A \in \mathrm{R}^{n\times n}$ is a symmetric matrix, then

$$\lambda_{min}(A)\|x\|^2 \leq (Ax)\cdot x \leq \lambda_{max}(A)\|x\|^2 \text{ for all } x \in \mathrm{R}^n.$$

Dividing through by $\|x\|^2$, this result can be written as

$$x \in \mathrm{R}^n, \|x\| = 1 \;\Rightarrow\; \lambda_{min}(A) \leq (Ax)\cdot x \leq \lambda_{max}(A). \tag{12}$$

I will now show that, in fact,

$$\lambda_{min}(A) = \min_{\|x\|=1} (Ax)\cdot x$$

and

$$\lambda_{max}(A) = \max_{\|x\|=1} (Ax)\cdot x.$$

I define $f : \mathrm{R}^n \to \mathrm{R}$ by

$$f(x) = \frac{1}{2}(Ax) \cdot x$$

and $g : \mathrm{R}^n \to \mathrm{R}$ by

$$g(x) = \frac{1}{2}(x \cdot x \Leftrightarrow 1),$$

I can then find the maximum and minimum of f subject to the constraint $g(x) = 0$. The gradients are

$$\nabla f(x) = Ax, \ \nabla g(x) = x,$$

and therefore the Lagrange multiplier condition is

$$Ax = \lambda x.$$

The stationary points are the eigenvectors $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ of A, and, with $x^{(i)} \cdot x^{(i)} = 1$,

$$f(x^{(i)}) = (Ax^{(i)}) \cdot x^{(i)} = \lambda_i (x^{(i)} \cdot x^{(i)}) = \lambda_i.$$

Therefore the maximum value of $(Ax) \cdot x$, subject to $\|x\| = 1$, is $\lambda_{max}(A)$, and the minimum value is $\lambda_{min}(A)$.

References

[1] R. A. Tapia. An introduction to the algorithms and theory of constrained optimization. Unpublished.