First and second-order conditions in constrained optimisation

Anthony Horsley and Andrew J. Wrobel
Department of Economics
London School of Economics
Houghton Street
London WC2A 2AE
United Kingdom.
e-mail: LSEecon123@mac.com

1 December 2003 (revised 24 September 2004)
CDAM Research Report LSE-CDAM-2003-15

Abstract

We give a concise but complete and detailed exposition of the "classical" approach, based on directional variations, to the first-order and second-order conditions (FOC and SOCs) for finite-dimensional constrained optimisation with both equality and inequality constraints. Attention is paid to liminal constraints, which are active inequality constraints with zero Lagrange multipliers. The persistent assertion in economics texts that all active constraints can be treated like equality constraints is untrue with liminal constraints, and it gives a false "sufficient" SOC. Nor can liminal constraints be ignored like inactive constraints; this would give a false "necessary" SOC. Treating liminal constraints like equalities in the *necessary* SOC, or ignoring them in the *sufficient* SOC is not incorrect, but it gives weaker optimality criteria than the standard Kuhn-Tucker multiplier rules (although the resulting Strong Sufficient SOC does have a place in the directional-derivative results of solution-sensitivity analysis without strict complementarity). We also show that the square slack-variables method, which reduces inequalities to equalities, cannot deal properly with liminal constraints in SOCs.

Keywords: First-order conditions, second-order conditions, liminal constraints, binding constraints.

1991 *Mathematics Subject Classification*: Primary 90C30.

1991 *Journal of Economic Literature Classification*: C61.

1 Introduction

This is an exposition of the "classical" approach, based on directional variations, to the first-order and second-order conditions (FOC and SOCs) for constrained optimisation. The method consists in linearisation, and second-order expansion, of the objective and constraint functions. It is presented in the modern geometric language of tangent and normal cones, but it originated in the Euler-Lagrange calculus of variations. It was designed and developed for infinite-dimensional problems long before being systematically applied to finite-dimensional problems (which, because of their size in practical applications, became of real interest only with the advent of high-speed computers). Though only such problems, with finite numbers of variables and constraints, are considered here explicitly, the framework is general.

The cone method we use has also a more recent, and more powerful, "nonclassical" variant which consists in linearising only *after* mapping the variables to the values of the nonlinear functions in question (objective and constraints), i.e., a suitable cone is constructed in their codomain (instead of the domain as in the "classical" approach). This gives a framework for both optimal control and the calculus of variations that improves on the classical results. We intend to present this in detail elsewhere, and here the image method is only noted, as are the augmented Lagrangian and the penalty method (Section 11).

The problem is therefore linearised at the outset: after the preliminary Section 2, in which we introduce the relevant functions and their derivatives, we discuss the tangent and normal cones to the constraint set in Section 3. In Section 4 we linearise the constraint functions and define the associated linearisation cone. This equals the tangent cone if constraints are regular. Section 5 provides workable criteria for regularity in terms of the constraint gradients, viz., linear independence and the significantly less stringent Mangasarian-Fromovitz Constraint Qualification (MFCQ) a.k.a. normality (which for convex programmes reduces essentially to Slater's Condition). Other implications of linear independence and normality are also noted. Section 6 gives examples of irregularity.

After this preparation, the immediate goal is to state the Lagrange multiplier rules in readily applicable forms. This is done first for the case of equality constraints only (Section 7), and then for the case of both equality and inequality constraints (Section 8). The formulations assume regularity but, provided that the constraints are linearly independent (or just normal), no explicit reference to the tangent cone is needed in applying the multiplier rules. Being simpler, the pure equality-constrained case is presented separately and first, but it is of course subsumed in the more general results with both kinds of constraint, i.e., in the Kuhn-Tucker multiplier rules.

Although this is standard material, even otherwise excellent economics texts persistently give faulty SOCs by overlooking the possibility that inequality constraints may be liminal, i.e., may have zero Lagrange multipliers. In particular, they give a "sufficient" second-order condition that is in fact *insufficient*: see [7] for a counterexample. In addi-

tion, some seem to regard that "result", which is *false* with liminal inequalities, as such an obvious extension from the equality-constrained case as not to require a proper proof [12, p. 466]. Other texts "prove" the false "theorem" [3, Theorem II.3.4 (p. 38)]. These serve as references for later texts, such as [13, E.1.16 (ii)], which also refer to mathematical sources that could have been consulted for a correct formulation of the SOC itself, as well as for a better method of proof.

As we point out in Section 9, the failure of [3] to deal correctly with inequality constraints has its roots in the method employed, which consists in converting inequalities into equalities by introducing square slack variables. Attributed to Valentine in [5, p. 39] but actually even older, the method *is* effective with FOCs—and it "is used sparingly" in, e.g., [5, pp. 261–262]. But as a way of obtaining SOCs, this approach is inherently limited to the case of no liminal constraints. It offers no option but to exclude these by assumption, which is known as strict complementarity. Recognising this would have at least produced true, though not the best, results. If strict complementarity were indispensable for further analysis, then arguably not much would have been lost. But since the work started in [1] and completed in [8], this is no longer justified. Earlier, strict complementarity used to be the standard assumption for a sensitivity analysis of the optimal solution, and it is indeed necessary if the solution and its multipliers are to have ordinary derivatives with respect to the problem's parameters: see, e.g., [4, Theorems 2.4.4 and 3.2.2] or [8, Theorem 1]. But although the solution and its multipliers are usually nondifferentiable without strict complementarity, they are still *directionally* differentiable: see, e.g., [4, Theorem 2.4.5] or [8, Theorems 3 and 4]. This suffices for most purposes, as is noted in [8, Section 3].

Proofs of necessity and sufficiency of the standard FOC and SOCs are deferred to Section 10, where the multiplier rules of Sections 7 and 8 are derived from more general results with an abstract constraint set. The Abstract Necessary FOC is put into a multiplier form by applying Farkas' Lemma, which is a separation argument extending, to the case of inequalities, the purely algebraic Factorisation Lemma. These are given in the Appendix, along with a further extension known as Motzkin's Lemma (which is used here only to reformulate the MFCQ, but is also of interest in multi-objective optimisation).

Throughout, we follow closely the exposition of Hestenes [5, Chapter 1] and [6, Chapters 3 and 4], identifying those statements in the sources that correspond to ours. The material is selected and arranged to give a concise but complete and detailed account, and extensive explanation is added to facilitate a thorough understanding of the method, its techniques and results. Application of the theory requires little other than facility with the Jacobian and Hessian matrices (reviewed briefly in Section 2, along with the multivariate Taylor expansion to second order), and with a determinantal or eigenvalue test of definiteness for quadratic forms subject to linear restrictions.

2 Problem formulation and preliminaries

The single-objective optimisation problem in question is taken to have been oriented to maximisation over an intersection of sublevel and level sets. Here, the number of decision variables, n, is assumed to be finite. So is the number of constraints, $m + l$: there are m equality and l inequality constraints. Neither the constraint set nor the maximand need be convex. The problem, then, is one of finite-dimensional nonconvex programming.

The equality-constraint functions are h_e for $e = 1, 2, \ldots, m \geq 0$ (there may be no equality constraints at all). The inequality-constraint functions are g_i for $i = 1, \ldots, l$ (where $l \geq 0$). Each of these functions is assumed to be defined and twice continuously differentiable on an open set $D \subseteq \mathbb{R}^n$. So the constraint set is

$$C = \{x \in D : h(x) = 0,\ g(x) \leq 0\}. \tag{1}$$

The maximand, f, is also taken to be of class C^2 (twice continuously differentiable) on D.

In matrix multiplication, the n-tuple of *decision variables* $x = (x_1, \ldots, x_n)$ is regarded as a column; its transpose is a row $x^{\mathrm{T}} = [x_1, \ldots, x_n]$. In other words, the "default" arrangement of any tuple is as a column. This applies equally to the vector-valued constraint maps $h = (h_1, \ldots, h_m)$ and $g = (g_1, \ldots, g_l)$, which map $\mathbb{R}^n$ into $\mathbb{R}^m$ and $\mathbb{R}^l$, respectively. (The term "function" is used to mean a scalar-valued map.)

The *Jacobian matrix* of a map $g\colon \mathbb{R}^n \to \mathbb{R}^l$ is the $l \times n$ matrix of partial derivatives

$$\mathrm{D}g(x) = \left[\frac{\partial g_i}{\partial x_j}\right]_{i=1\ \ j=1}^{\ \ l \quad n}$$

i.e., its i-th row is the transposed gradient vector $\nabla g_i(x)^{\mathrm{T}}$.

For an l-row λ^{T} and an $l \times n$ matrix A, the matrix product $\lambda^{\mathrm{T}} A$ is a linear combination of the rows of A. So $\lambda^{\mathrm{T}} \mathrm{D}g(x)$ is a linear combination of the gradients $(\nabla g_i)_{i=1}^{\ l}$.

The matrix product Av, where v is an n-column, is similarly a linear combination of the columns of A. But its entries can also be viewed as the scalar products of v and the rows of A; for example, $\mathrm{D}g(x)\, v$ has $\nabla g_i(x)^{\mathrm{T}} v$ as its i-th entry. The scalar product $p^{\mathrm{T}} v$ is denoted also by $p \cdot v$.

The *right kernel* of A is the null space of the linear operation $v \mapsto Av$, i.e.,

$$\ker(A\cdot) = \{v \in \mathbb{R}^n : Av = 0\}.$$

Likewise, the left kernel is $\ker(\cdot A) = \left\{\lambda \in \mathbb{R}^l : \lambda^{\mathrm{T}} A = 0\right\}$.

The *Hessian matrix* of a C^2-function $L\colon \mathbb{R}^n \to \mathbb{R}$ is the symmetric square $(n \times n)$ matrix of second partial derivatives

$$\mathrm{D}^2 L(x) = \left(\frac{\partial^2 L}{\partial x_r \partial x_s}(x)\right)_{r=1\ \ s=1}^{\ \ n \quad n}.$$

When H is an $n \times n$ matrix, the function $v \mapsto v^{\mathrm{T}} H v$ is called a *quadratic form* on $\mathbb{R}^n$.

The Euclidean norm (the length) of a $v \in \mathbb{R}^n$ is[1]

$$\|v\| = (v \cdot v)^{1/2} = \left(\sum_{j=1}^{n} (v_j)^2 \right)^{1/2}.$$

When a vector is thought of as an increment to x, it may be denoted by Δx (which must be understood as a single symbol).

By definition, a function L is twice differentiable at a point x if[2]

$$\lim_{v \to 0} \frac{L(x+v) - L(x) - \nabla L(x) \cdot v}{\|v\|} = 0$$

$$\lim_{v \to 0} \frac{L(x+v) - L(x) - \nabla L(x) \cdot v - \frac{1}{2} v^{\mathrm{T}} \mathrm{D}^2 L(x)\, v}{\|v\|^2} = 0.$$

A sequence $(x(k))_{k=1}^{\infty}$ in $\mathbb{R}^n$ *converges to x from a direction Δx* if there is a sequence of positive scalars $\delta(k)$ such that

$$\delta(k) \to 0 \text{ and } \frac{x(k) - x}{\delta(k)} \to \Delta x \text{ as } k \to \infty.$$

For $\Delta x = 0$, this merely means that $x(k) \to x$ as $k \to \infty$. For $\Delta x \neq 0$, $x(k) \to x$ from the direction Δx if and only if

$$x(k) \to x \text{ and } \frac{x(k) - x}{\|x(k) - x\|} \to \frac{\Delta x}{\|\Delta x\|} \text{ as } k \to \infty. \tag{2}$$

(See, e.g., [6, p. 204].) If $x(k) \to x$ from Δx then (by substituting $x(k) - x$ for the v above)

$$\frac{\nabla L(x) \cdot \Delta x}{\|\Delta x\|} = \lim_{k \to \infty} \frac{L(x(k)) - L(x)}{\|x(k) - x\|} \tag{3}$$

$$\frac{1}{2} \frac{\Delta x^{\mathrm{T}} \mathrm{D}^2 L(x)\, \Delta x}{\|\Delta x\|^2} = \lim_{k \to \infty} \frac{L(x(k)) - L(x) - \nabla L(x) \cdot (x(k) - x)}{\|x(k) - x\|^2}. \tag{4}$$

3 Tangent and normal cones

A *cone* at the origin 0 of a vector space is a nonempty subset that is closed under scaling[3]—i.e., it is a set $K \neq \emptyset$ such that $\alpha K \subseteq K$ for each $\alpha \in \mathbb{R}_+$.[4] The *cone generated*

[1] All norms on $\mathbb{R}^n$ are equivalent, i.e., define the same convergence concept.

[2] This means twice Fréchet differentiability at x. It follows from the existence and continuity of second partial derivatives on a neighbourhood of x.

[3] By this definition, a cone need not be convex or pointed. In some literature, a *convex* set closed under scaling is called a wedge, and a pointed wedge is called a cone.

[4] Then $x + K$ is a cone at x.

by a set X (a.k.a. the conical hull of X) is the smallest cone containing X. It is the union of all rays from the origin through the points of X; i.e., $\operatorname{cone} X = \bigcup_{\alpha\in\mathbb{R}_+} \alpha X$ if $X \neq \emptyset$. Additionally $\operatorname{cone}\emptyset := \{0\}$, the zero cone. The symbol $:=$ means that the left-hand side (l.h.s.) is by definition equal to the right-hand side (r.h.s.).

The *convex hull* of X is the smallest convex set containing X; it is denoted by $\operatorname{conv} X$. The *convex cone generated* by a set X is the smallest convex cone containing X; it equals $\operatorname{cone}\operatorname{conv} X = \operatorname{conv}\operatorname{cone} X$.

The *linear span* (a.k.a. linear hull) of X is the smallest linear space containing X; it is denoted by $\operatorname{span} X$.

A cone K is *line-free* if $K \cap (-K) = \{0\}$, i.e., if it contains no straight line through the origin. When K is a *convex* cone, $K \cap (-K)$ is the largest linear space contained in K; it is called the lineality space of K. When it is zero, K is called *pointed* (a.k.a. salient).

Definition 1 *A convex cone K is pointed if $K \cap (-K) = \{0\}$.*

The usual concept of a tangent to a curve is adequate for most applications. It has, however, a generalisation that is useful when C contains sequences convergent to x but no curve issuing from x.

Definition 2 *A nonzero vector Δx is a sequential tangent to a set C, at a point $x \in C$, if there exists a sequence $(x(k))_{k=1}^{\infty}$ in $C \setminus \{x\}$ that converges to x from the direction Δx, i.e., satisfies (2). "Sequential tangent" is henceforth abbreviated to "tangent".*

Definition 3 *A vector Δx is a curvilinear tangent to a set C, at a point $x \in C$, if there exists a parameterised curve $\widetilde{x}$, defined on $[0, \overline{\epsilon}]$ for some $\overline{\epsilon} > 0$, such that: $\widetilde{x}(0) = x$, $\widetilde{x}(\epsilon) \in C$ for every $\epsilon \in [0, \overline{\epsilon}]$, and $(\mathrm{d}\widetilde{x}/\mathrm{d}\epsilon)(0) = \Delta x$.*

The *tangent cone*, to C at an $x \in C$, consists of all the tangent directions and the zero vector, i.e.,

$$\mathrm{T}_x C := \operatorname{cone}\{\Delta x : \Delta x \text{ is a unit vector tangent to } C \text{ at } x\}.$$

In other words, a tangent vector to C at x is any limit of the directions of arbitrarily small displacements from x within C. A curvilinear tangent is a vector tangent to a curve in C that issues from the point in question, x.

Remark 4 *Every curvilinear tangent vector is a (sequential) tangent, i.e., $\Delta x \in \mathrm{T}_x C$ if there is a curve $\widetilde{x}$ in C that meets the conditions of Definition 3.*

The tangent cone is always closed: see, e.g., [5, Lemma 1.8.1] or [6, Lemma 4.4.2]. In general, it may be nonconvex; but it *is* convex if C is. In the convex case there is, in a sense, little need for "proper" curves: the straight-line segments lying in C, and their limits, give all the tangent directions.

Definition 5 *A vector Δx is a feasible direction (a.k.a. a linear tangent) at a point $x \in C$ if there exists a scalar $\epsilon > 0$ such that the segment joining x and $x + \epsilon\Delta x$ lies in C.*

Feasible directions form a subcone of $\mathrm{T}_x C$, denoted by

$$\mathrm{F}_x C := \{\Delta x : [x, x + \epsilon\Delta x] \subseteq C \text{ for some } \epsilon > 0\}.$$

Remark 6 *If C is convex then so is $\mathrm{F}_x C$. Furthermore, $\mathrm{T}_x C$ equals the closure of $\mathrm{F}_x C$. So $\mathrm{T}_x C$ is convex, too. Also, $C \subseteq x + \mathrm{F}_x C \subseteq x + \mathrm{T}_x C$ (a cone at x).*[5]

An *outward normal* (a.k.a. exterior normal) is a vector whose scalar product with every tangent vector is zero or less.[6]

Definition 7 *A vector p is normal to a set C at a point $x \in C$ if $p \cdot \Delta x \leq 0$ for every $\Delta x \in \mathrm{T}_x C$.*

The *normal cone*, to C at x, is

$$\mathrm{N}_x C := \{p : p \text{ is normal to } C \text{ at } x\} = \{p : p \cdot \Delta x \leq 0 \text{ for every } \Delta x \in \mathrm{T}_x C\}.$$

This is a case of polarity for cones.

Definition 8 *For any cone $K \subseteq \mathbb{R}^n$, its polar cone is*[7]

$$K^\circ := \{p \in \mathbb{R}^n : p \cdot \Delta x \leq 0 \text{ for every } \Delta x \in K\}.$$

When K is a linear subspace, K° is equal to the orthogonal complement $K^\perp$.

In these terms, the normal cone is polar to the tangent cone, i.e., $\mathrm{N}_x C := (\mathrm{T}_x C)^\circ$. Like any polar cone, $\mathrm{N}_x C$ is always convex and closed. Finally, note that the normal cone to $\mathrm{T}_x C$ at 0 is also $\mathrm{N}_x C$ (since $\mathrm{T}_0 \mathrm{T}_x C = \mathrm{T}_x C$).

[5] This is in [6, Lemma 4.4.3].

[6] Geometrically, an outward normal forms an obtuse angle (an unoriented angle of at least $\pi/2$ radians) to every tangent vector.

[7] Its negative, $-K^\circ$, is also known as the the dual of K.

4 Linearisation of constraints and their regularity

Linearisation of a set C, around a point $\overline{x} \in C$, consists in replacing C by $\overline{x} + \mathrm{T}_{\overline{x}}C$, i.e., the condition $x \in C$ is replaced by: $x - \overline{x} \in \mathrm{T}_{\overline{x}}C$. Here, "replacing" means that local optimality of $\overline{x}$ on the constraint set C is to be characterised by a necessary or sufficient condition which is to hold for every vector in $\mathrm{T}_{\overline{x}}C$.

If such an abstract condition is to expand into a multiplier rule, the tangent cone $\mathrm{T}_{\overline{x}}C$ must be described in terms of the gradients of the constraint functions. To this end, the constraints $h(x) = 0$ and $g(x) \leq 0$ are linearised to

$$\mathrm{D}h(x)(x - \overline{x}) = 0 \quad \text{and} \quad \nabla g_i(\overline{x}) \cdot (x - \overline{x}) \leq 0 \text{ for each } i \text{ with } g_i(\overline{x}) = 0 \tag{5}$$

(any inactive inequality constraint, $g_i(\overline{x}) < 0$, is irrelevant to local linearisation around $\overline{x}$). The linearised a.k.a. tangential constraints (5) are met whenever the increment $\Delta x = x - \overline{x}$ is a tangent vector, but in general they can also be met by some nontangent vectors. In other words, with

$$\mathrm{A}(\overline{x}) := \{i : g_i(\overline{x}) = 0\} \tag{6}$$

denoting the set of all the *active* inequality constraints, $\mathrm{T}_{\overline{x}}C$ is always contained in the convex cone

$$\mathrm{L}_{\overline{x}}(h, g) := \{\Delta x : \mathrm{D}h(\overline{x})\,\Delta x = 0,\ \nabla g_i(\overline{x}) \cdot \Delta x \leq 0 \text{ for every } i \in \mathrm{A}(\overline{x})\}. \tag{7}$$

In the case of no inequality constraints, $g = \emptyset$ formally, and the cone $\mathrm{L}_{\overline{x}}(h, \emptyset)$ is actually a linear space.

Since $\mathrm{L}_{\overline{x}}(h, g)$ consists of the increments $\Delta x = x - \overline{x}$ satisfying the linearised constraints (5), it is called the *linearisation cone*. It can be larger than the tangent cone, even when the latter is also convex. For example, in the pure equality-constrained case, L_x is a always a linear space, but T_x can be a "proper" cone, i.e., not a linear space (Example 19 below). However, the two cones should be equal if multiplier rules are to hold—and this condition is known as regularity.

Lemma 9 $\mathrm{T}_x C \subseteq \mathrm{L}_x(h, g)$.[8]

Proof. Take any nonzero $\Delta x \in \mathrm{T}_x C$, scale it to unit length and take a sequence $x(k)$ in C converging to x from the direction Δx. Then

$$\mathrm{D}h(x)\,\Delta x = \lim_k \frac{h(x(k)) - h(x)}{\|x(k) - x\|} = 0$$

[8] This is in, e.g., [5, p.35, lines 12–14], [6, pp. 221–222] and [10, 5.2.12].

since $h(x(k)) = h(x)$ for each k. Similarly, if $g_i(x) = 0$ then

$$\nabla g_i(x) \cdot \Delta x = \lim_k \frac{g_i(x(k)) - g_i(x)}{\|x(k) - x\|} \leq 0$$

since $g_i(x(k)) \leq 0$ for each k. ■

Definition 10 *A point $x \in C$ is regular for the representation of the constraint set C by the constraint functions h and g (abbreviated to: x is regular for h and g) if* $\mathrm{T}_x C = \mathrm{L}_x(h, g)$.

Regularity is also known as Abadie's Constraint Qualification (ACQ) and as the Basic CQ. The somewhat stronger requirement that every vector from $\mathrm{L}_x(h, g)$ be a *curvilinear* tangent is known as Kuhn-Tucker regularity or KTCQ; it holds for linearly independent constraints (Lemma 13 below). A constraint with a vanishing gradient may fail the KTCQ and still be regular.[9]

For the first-order multiplier rule, all that matters is that the normal cone $\mathrm{N}_x C$ be expressed in terms of the constraint gradients. Regularity ensures this because the polar of $\mathrm{L}_x(h, g)$ is always equal to the convex cone generated by the constraint gradients (without regard to sign in the case of equalities), i.e.,

$$\begin{aligned} \mathrm{L}_x(h, g)^\circ &= \operatorname{cone}\operatorname{conv}\left(\{\pm\nabla h_e(x) : e = 1, \ldots, m\} \cup \{\nabla g_i(x) : i \in \mathrm{A}(x)\}\right) \\ &= \operatorname{span}\{\nabla h_e(x) : e = 1, \ldots, m\} + \operatorname{cone}\operatorname{conv}\{\nabla g_i(x) : i \in \mathrm{A}(x)\} \end{aligned} \tag{8}$$

by Farkas' Lemma (Lemma 34). And if x is regular then

$$\mathrm{N}_x C := (\mathrm{T}_x C)^\circ = \mathrm{L}_x(h, g)^\circ . \tag{9}$$

Known as *quasi-regularity*, this is exactly what is needed for the Kuhn-Tucker first-order multiplier rule to hold. The condition is somewhat weaker than regularity because the two polars in (9) are equal if and only if $\mathrm{L}_x(h, g)$ equals the closed convex hull of $\mathrm{T}_x C$, and this can be larger than $\mathrm{T}_x C$ itself. (It is larger if, and only if, $\mathrm{T}_x C$ is nonconvex. For such an example, see Section 18.)

The simplest nonlinear example of (Kuhn-Tucker) regularity is a single constraint with a nonzero gradient; this generalises to regularity of linearly independent constraints and further to regularity of so-called normal constraints (Lemmas 13 and 16 below).

[9] For a one-variable example with a C^2-constraint, take either h or g equal to $x^5 \sin(1/x)$ for $x \neq 0$, and to 0 for $x = 0$. Its derivative at $x = 0$ vanishes, and this point fails the KTCQ because the constraint set contains no interval $[-\epsilon, 0]$ or $[0, \epsilon]$ with $\epsilon > 0$. But the constraint is nevertheless regular, since the (sequential) tangent cone is $\mathrm{T}_0 = \mathbb{R} = \mathrm{L}_0$.

Example 11 (Single regular equality constraint) *The tangent cone to a hypersurface C with the locus equation $h(x) = 0$, at a point $\overline{x}$ with $\nabla h(\overline{x}) \neq 0$, is the hyperplane*

$$\mathrm{T}_{\overline{x}} C = \{\Delta x : \nabla h(\overline{x}) \cdot \Delta x = 0\}. \tag{10}$$

The normal cone is the straight line

$$\mathrm{N}_{\overline{x}} C = \operatorname{span}\{\nabla h(\overline{x})\}. \tag{11}$$

Example 12 (Single regular inequality constraint) *The tangent cone to a sublevel set $C = \{x : g(x) \leq 0\}$, at a point $\overline{x}$ with $g(\overline{x}) = 0$ and $\nabla g(\overline{x}) \neq 0$, is the half-space*

$$\mathrm{T}_{\overline{x}} C = \{\Delta x : \nabla g(\overline{x}) \cdot \Delta x \leq 0\}. \tag{12}$$

The normal cone is the ray (half-line)

$$\mathrm{N}_{\overline{x}} C = \operatorname{cone}\{\nabla g(\overline{x})\}. \tag{13}$$

So, with nonzero constraint gradients, regularity means that the tangent and normal cones can be expressed in terms of the tangent and normal cones to the individual constraints. Specifically, the tangent cone equals the intersection (7) of the hyperplanes (10) and half-spaces (12) which are tangent to the individual constraints (Examples 11 and 12). And the normal cone is the (algebraic) sum (8) of the straight lines (11) and outward rays (13) which are normal to the individual constraints.

Comment: Regularity can be thought of as commutativity of two operations, viz., the linearisation and the definition of a constraint set in terms of constraint functions: starting from the functions (h, g), one can first define the set C and then linearise it to the cone $\mathrm{T}_x C$. The alternative is to linearise h and g first, and then define the cone $\mathrm{L}_x(h, g)$. Either way, the result is the same—in the regular case.

5 Linear independence and Mangasarian-Fromovitz normality as conditions for regularity

Constraints are regular if their gradients are linearly independent. Known as LICQ, this condition can be restated as a rank condition on the constraints' Jacobian—viz., that the Jacobian matrix of the equality and active inequality constraint functions be of full row-rank. For brevity, denote the number of active inequality constraints by

$$\mathrm{a}(x) := \operatorname{card} \mathrm{A}(x)$$

and let $g_{\mathrm{A}(x)}$ mean the (finite) sequence of the active constraint functions; then $\mathrm{D}g_{\mathrm{A}(x)}(x)$ is the corresponding $\mathrm{a}(x) \times n$ submatrix of $\mathrm{D}g(x)$.[10]

[10] Formally, $g_{\mathrm{A}(x)} = (g_{i_k})_{k=1}^{\mathrm{a}(x)}$, and the k-th row of $\mathrm{D}g_{\mathrm{A}(x)}(x)$ is $\nabla g_{i_k}(x)^{\mathrm{T}}$, where i_k increases with k and $\mathrm{A}(x) = \{i_k : k = 1, 2, \ldots, \mathrm{a}(x)\}$.

Lemma 13 (Linearisation under LICQ) *If $h(x) = 0$ and $g(x) \leq 0$, and the vectors $(\nabla h_e(x))_{e=1}^m$ and $(\nabla g_i(x))_{i \in \mathrm{A}(x)}$ are linearly independent or, equivalently,*

$$\operatorname{rank} \begin{bmatrix} \mathrm{D}h(x) \\ \mathrm{D}g_{\mathrm{A}(x)}(x) \end{bmatrix} = m + \mathrm{a}(x) \tag{14}$$

(i.e., this matrix has a full row rank), then:[11]

1. *Every $\Delta x \in \mathrm{L}_x(h, g)$ is a curvilinear tangent vector at x to the set C given by (1), i.e., the Kuhn-Tucker Constraint Qualification holds at x.*[12]

   In other words, there is a curve $\widetilde{x}\colon [0, \overline{\epsilon}] \to C$ (for some $\overline{\epsilon} > 0$) with $\widetilde{x}(0) = x$ and $(\mathrm{d}\widetilde{x}/\mathrm{d}\epsilon)(0) = \Delta x$. This implies that $h_e(\widetilde{x}(\epsilon)) = 0$ for each e and $\epsilon \leq \overline{\epsilon}$, i.e., $h \circ \widetilde{x} = 0$. In addition, $\widetilde{x}$ can be chosen so that $g_i(\widetilde{x}(\epsilon)) = \epsilon \nabla g_i(x) \cdot \Delta x$ for each $i \in \mathrm{A}(x)$ and each $\epsilon \leq \overline{\epsilon}$, i.e., $g_{\mathrm{A}(x)}(\widetilde{x}(\epsilon)) = \epsilon \mathrm{D}\left(g_{\mathrm{A}(x)}\right)(x) \Delta x$.[13] *If h and g are of class C^k (for an integer $k \geq 1$), then $\widetilde{x}$ can be chosen to be of class C^k also.*

2. *So $\mathrm{L}_x(h, g) = \mathrm{T}_x C$ (i.e., x is regular for the representation of C by h and g).*

Proof. Part 2 follows from Part 1 by Remark 4.

For Part 1, $\widetilde{x}(\epsilon)$ can be obtained from $x + \epsilon\Delta x$ by adding a correction term $\widetilde{c}(\epsilon)$, which is generally needed because $h(x + \epsilon\Delta x)$ is not exactly 0 and $g(x + \epsilon\Delta x)$ is not linear in ϵ.[14] (Nor is $g_i(x + \epsilon\Delta x)$ always nonpositive if $g_i(x) = 0$ and $\nabla g_i(x) \cdot \Delta x = 0$; although if $\nabla g_i(x) \cdot \Delta x < 0$ then $g_i(x + \epsilon\Delta x) < 0$ for small enough $\epsilon > 0$.) Fix any basis for $\mathbb{R}^n$; like any vector, $\widetilde{c}(\epsilon)$ is a linear combination of the basis vectors. To minimise the number of nonzero coefficients, choose a subsequence $E(1), \ldots, E(m + \mathrm{a}(x))$ of as many basis vectors as there are active inequality and equality constraints in such a way that the $(m + \mathrm{a}(x))$-square matrix

$$J = \begin{bmatrix} \left[\ \nabla h_e(x) \cdot E(k)\ \right]_{e=1 \quad k=1}^{m \qquad m+\mathrm{a}(x)} \\ \left[\ \nabla g_i(x) \cdot E(k)\ \right]_{i \in \mathrm{A}(x) \quad k=1}^{\qquad\quad m+\mathrm{a}(x)} \end{bmatrix} = \begin{bmatrix} \mathrm{D}h(x) \\ \mathrm{D}g_{\mathrm{A}(x)}(x) \end{bmatrix} E$$

is nonsingular; then seek a correction of the form

$$\widetilde{c}(\epsilon) = \sum_{k=1}^{m+\mathrm{a}(x)} E(k)\, \widetilde{r}_k(\epsilon) = E\widetilde{r}(\epsilon)$$

[11] This is in [5, Lemma 1.10.1]; it is also a part of [6, Theorem 4.10.3]. For equality constraints only, it is also in [6, Theorem 3.5.1] and [10, 5.1.7]; this is known as the Tangent Space Theorem.

[12] The KTCQ is further discussed in, e.g., [13, 1.D].

[13] In other words, any active constraint function g_i decreases along the curve $\widetilde{x}$ at a constant nonnegative rate, viz., $-\nabla g_i(x) \cdot \Delta x$ per unit of the curve's parameter.

[14] The correction is *not* needed if the constraints are locally linear, i.e., of the form $h(x) = Bx + \text{const.}$, etc.—so linear constraints are always regular, even when they are linearly dependent.

where the columns $E(1)$, $E(2)$, ... are concatenated in the $n \times (m + \mathrm{a}(x))$ matrix E with entries $E_{jk} := E(k)_j$. (If the unit coordinate vectors of $\mathbb{R}^n$ are chosen as the basis, then J is simply any maximal nonsingular submatrix of the Jacobian of $\left(h, g_{\mathrm{A}(x)}\right)$, the matrix E consists of the corresponding columns of the $n \times n$ unit matrix $\mathrm{I}(n)$, and $c = Er$ is just r but with $n - m - \mathrm{a}(x)$ zeros inserted at the right places.)

By the Implicit Function Theorem (in, e.g., [5, Theorem 1.7.1] and [6, Theorem 3.7.1]), the system of $m + \mathrm{a}(x)$ nonlinear equations for r

$$h(x + \epsilon \Delta x + Er) = 0 \tag{15}$$

$$g_{\mathrm{A}(x)}(x + \epsilon \Delta x + Er) - \epsilon \mathrm{D} g_{\mathrm{A}(x)}(x) \Delta x = 0 \tag{16}$$

determines $(\widetilde{r}_1, \widetilde{r}_2, \ldots)$ as smooth functions of ϵ, on an interval $[-\overline{\epsilon}, \overline{\epsilon}]$ for some $\overline{\epsilon} > 0$, with $\widetilde{r}(0) = 0$. This is because, at $(\epsilon, r) = (0, 0)$, the Jacobian w.r.t. r of the l.h.s. of (15)–(16) is the nonsingular matrix J. It also follows that

$$\frac{\mathrm{d}\widetilde{r}}{\mathrm{d}\epsilon}(0) = -J^{-1} \begin{bmatrix} \mathrm{D}h(x) \Delta x \\ \mathrm{D}g_{\mathrm{A}(x)}(x) \Delta x - \mathrm{D}g_{\mathrm{A}(x)}(x) \Delta x \end{bmatrix} = -J^{-1} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0$$

and so the curve

$$\widetilde{x}(\epsilon) := x + \epsilon \Delta x + E\widetilde{r}(\epsilon)$$

meets all the requirements: $\widetilde{x}(0) = x$

$$\frac{\mathrm{d}\widetilde{x}}{\mathrm{d}\epsilon}(0) = \Delta x + E\frac{\mathrm{d}\widetilde{r}}{\mathrm{d}\epsilon}(0) = \Delta x$$

and $h(\widetilde{x}(\epsilon)) = 0$ as well as $g_{\mathrm{A}(x)}(\widetilde{x}(\epsilon)) = \epsilon \mathrm{D} g_{\mathrm{A}(x)}(x) \cdot \Delta x \leq 0$ by (15)–(16), for every positive $\epsilon \leq \overline{\epsilon}$. ■

For inequality constraints, the linear independence condition can be weakened—without loss of regularity—to that of *positive* independence, i.e., nonexistence of a vanishing linear combination with positive coefficients (Definition 37). This is equivalent to the existence of a vector that meets the linearised inequality constraints *strictly* (when substituted for $x - \overline{x}$ in (5)): see Lemma 36. So the weaker CQ, known as the Mangasarian-Fromovitz *normality*, has two equivalent forms.

Lemma 14 (Two forms of MFCQ a.k.a. normality) *Given a point $x \in C$, the following two conditions are equivalent to each other:*[15]

1. *If*

$$\mu^{\mathrm{T}} \mathrm{D}h(x) + \lambda^{\mathrm{T}} \mathrm{D} g_{\mathrm{A}(x)}(x) = 0 \tag{17}$$

and $\lambda \geq 0$ (where $\lambda \in \mathbb{R}^{\mathrm{a}(x)}$), then $\lambda = 0$ and $\mu = 0$.

[15] This is in [6, Theorem 4.10.4].

2. *(a) The equality constraint gradients* $(\nabla h_e(x))_{e=1}^m$ *are linearly independent; and*
(b) there is a vector $v \in \mathbb{R}^n$ *such that* $\mathrm{D}g_{\mathrm{A}(x)}(x)\, v \ll 0$ *and* $\mathrm{D}h(x)\, v = 0$*, i.e.,*

$$\nabla h_e(x) \cdot v = 0 \text{ for each } e \quad \text{and} \quad \nabla g_i(x) \cdot v < 0 \text{ for each } i \in \mathrm{A}(x). \tag{18}$$

Proof. When $\mathrm{A}(x) \neq \emptyset$ (i.e., there is an active inequality constraint), apply Lemma 35 (Motzkin's Alternative) with $A = \emptyset$—i.e., without the matrix A—to $\mathrm{D}h(x)$ and $\mathrm{D}g_{\mathrm{A}(x)}(x)$ in place of the remaining matrices B and C (renaming ν to λ). This shows that Condition 2b (i.e., solubility of (65) with the above substitutions) is equivalent to nonexistence of a μ and $\lambda > 0$ meeting (17). Nonexistence of such μ and λ follows obviously from Condition 1 (since the latter means nonexistence of a nonzero pair μ and $\lambda \geq 0$ meeting (17)).

Furthermore, when Condition 2b holds, multiplication of (17) by such a v shows that the λ in (17) must then be 0. This means that, under Condition 2b, Condition 2a is equivalent to Condition 1. So Condition 1 is equivalent to the conjunction of Conditions 2a and 2b.

This argument applies formally also when $\mathrm{A}(x) = \emptyset$. But this case is actually trivial: if no inequality constraint is active, then Condition 2b holds vacuously, and Condition 1 reduces to Condition 2a. ■

Geometrically, the MFCQ means that: (i) the equality constraint gradients (∇h_e) are linearly independent, (ii) the convex cone generated by the inequality constraint gradients (∇g_i) is pointed, and (iii) the interior of its polar $\mathrm{L}_x(\emptyset, g)$, which is nonempty by Lemma 36, contains a vector tangent to the equality-constraint manifold (i.e., a vector v tangent to each hypersurface with the locus equation $h_e = 0$).

That the MFCQ is weaker than the LICQ is obvious from its first form.

Corollary 15 (LICQ and MFCQ) *If the vectors* $(\nabla h_e(x))_{e=1}^m$ *and* $(\nabla g_i(x))_{i \in \mathrm{A}(x)}$ *are linearly independent, then the Mangasarian-Fromovitz Constraint Qualification holds at* x*.*[16]

Proof. Linear independence of all the relevant gradients (∇h_e and ∇g_i for each active i) obviously implies Condition 1 of Lemma 14. ■

But the MFCQ is strong enough to ensure regularity.

Lemma 16 (Linearisation under MFCQ) *The Mangasarian-Fromovitz Constraint Qualification, at an* $x \in C$*, implies that* $\mathrm{L}_x(h, g) = \mathrm{T}_x C$ *(i.e., that the point* x *is regular for the representation of the constraint set* C *by the functions* h *and* g*).*[17]

[16] This is a part of the first implication in [6, Theorem 4.10.3]: the conclusion "$r = p$" means there the same as Condition 2b of Remark 14 here. Therefore, like the Proof of Remark 14, the derivation of MFCQ from LICQ in [6, p. 239, lines 6–12] uses a theorem of the alternative (Stiemke's, which is Lemma 38).

[17] This is the second implication in [6, Theorem 4.10.3], where it is derived from [6, Theorem 3.5.1] and thus from the Implicit Function Theorem, like Lemma 16 here.

Proof. When there is no inequality constraint, i.e., $l = 0$, this is a special case of Lemma 13. When $l > 0$, the Proof of Lemma 13 has to be modified. To simplify the notation, one can assume that all the inequality constraints are active.

First consider any $\Delta x \in \mathrm{L}_x(h, g)$ with $\mathrm{D}g(x)\Delta x \ll 0$ (in addition to $\mathrm{D}h(x)\Delta x = 0$).[18] If actually there is no equality constraint, i.e., $m = 0$, then—to show that Δx is a tangent to C at x—it suffices to take the curve $\epsilon \mapsto x + \epsilon\Delta x$ (since $\nabla g_i(x) \cdot \Delta x < 0$ implies that $g_i(x + \epsilon\Delta x) < g_i(0) = 0$ for small $\epsilon > 0$ and each i). But if $m > 0$ then a correction term $\widetilde{c}(\epsilon)$ must be added to ensure that the curve meets the equality constraints. This is done as in the Proof of Lemma 13: from any basis for $\mathbb{R}^n$, choose m vectors $E(1), \ldots, E(m)$ such that the $m \times m$ square matrix

$$J := \mathrm{D}h(x)E = \left[\ \nabla h_e(x) \cdot E(k)\ \right]_{e=1\ \ k=1}^{m\quad m}$$

is nonsingular, and seek a correction of the form

$$\widetilde{c}(\epsilon) = \sum_{k=1}^{m} E(k)\,\widetilde{r}_k(\epsilon) = E\widetilde{r}(\epsilon)$$

where the columns $E(1), \ldots, E(m)$ are concatenated in an $n \times m$ matrix with entries $E_{jk} := E(k)_j$. By the Implicit Function Theorem, the system of m nonlinear equations for r

$$h(x + \epsilon\Delta x + Er) = 0 \tag{19}$$

determines $(\widetilde{r}_1, \ldots, \widetilde{r}_m)$ as smooth functions of ϵ, on an interval $[-\overline{\epsilon}, \overline{\epsilon}]$ for some $\overline{\epsilon} > 0$, with $\widetilde{r}(0) = 0$. This is because the Jacobian w.r.t. r of the l.h.s. of (19) is the nonsingular matrix J. It also follows that

$$\frac{\mathrm{d}\widetilde{r}}{\mathrm{d}\epsilon}(0) = -J^{-1}\mathrm{D}h(x)\Delta x = -J^{-1}0 = 0$$

and so the curve

$$\widetilde{x}(\epsilon) := x + \epsilon\Delta x + E\widetilde{r}(\epsilon)$$

meets all the requirements: $\widetilde{x}(0) = x$,

$$\frac{\mathrm{d}\widetilde{x}}{\mathrm{d}\epsilon}(0) = \Delta x + E\frac{\mathrm{d}\widetilde{r}}{\mathrm{d}\epsilon}(0) = \Delta x$$

and $\widetilde{x}(\epsilon) \in C$ for every sufficiently small $\epsilon > 0$. This is because $h(\widetilde{x}(\epsilon)) = 0$ by (19), and because for each i

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}(g_i \circ \widetilde{x})(0) = \nabla g_i(x) \cdot \left(\Delta x + \frac{\mathrm{d}\widetilde{r}}{\mathrm{d}\epsilon}(0)\right) = \nabla g_i(x) \cdot \Delta x < 0 \tag{20}$$

[18] The MFCQ means that the set of such vectors is nonempty, in which case it is the relative interior of L_x, i.e., its interior in $\ker(\mathrm{D}h(x)\cdot)$.

which implies that $g_i(\widetilde{x}(\epsilon)) < g_i(\widetilde{x}(0)) = 0$ for small enough $\epsilon > 0$.

Finally, given an arbitrary $\Delta x \in \mathrm{L}_x(h, g)$, use the MFCQ to take a v satisfying (18). Then $\Delta x + \epsilon v \in \mathrm{T}_x C$ for every $\epsilon > 0$ by the preceding argument (since $\mathrm{D}g(x)(\Delta x + \epsilon v) \ll 0$ and $\mathrm{D}h(x)(\Delta x + \epsilon v) v = 0$). As $\epsilon \searrow 0$, it follows that $\Delta x \in \mathrm{T}_x C$ because $\mathrm{T}_x C$ is closed. ■

As the Proof of Lemma 16 shows, the MFCQ implies the existence of a point satisfying all the inequality constraints strictly, i.e., a point x^{S} with

$$g_i\left(x^{\mathrm{S}}\right) < 0 \text{ for each } i \quad \text{and} \quad h_e\left(x^{\mathrm{S}}\right) = 0 \text{ for each } e. \tag{21}$$

In convex programming this is known as Slater's Condition (SCQ). It implies, conversely, that the MFCQ holds after discarding any dependent equality constraints.

Remark 17 *The Mangasarian-Fromovitz Constraint Qualification (at any $x \in C$) implies Slater's. For the converse, assume that g_i is convex and h_e is linear, for each i and e (and their domain D is a convex open set). Then:*

1. *For every $x \in C$, (21) implies that the vector $v := x - x^{\mathrm{S}}$ meets (18).*
2. *So if I is any maximal set of linearly independent equality constraints, i.e., $(\nabla h_e)_{e \in I}$ are linearly independent with the same span as the whole system $(\nabla h_e)_{e=1}^{m}$, then the MFCQ holds for the representation of C by $h_I := (\nabla h_e)_{e \in I}$ and g (at every $x \in C$).*

Proof. That the MFCQ implies SCQ is shown in proving Lemma 16, after (20): for a small $\epsilon > 0$, $\widetilde{x}(\epsilon)$ will do as x^{S} in (21).

For the converse, if $g_i(x) = 0$ then, by the gradient inequality for a convex function,

$$\nabla g_i(x) \cdot \left(x - x^{\mathrm{S}}\right) \leq g_i(x) - g_i\left(x^{\mathrm{S}}\right) = -g_i\left(x^{\mathrm{S}}\right) < 0.$$

And if $h_e(x) = 0$ then $\nabla h_e \cdot \left(x - x^{\mathrm{S}}\right) = h_e(x) - h_e\left(x^{\mathrm{S}}\right) = 0$. This proves Part 1. Part 2 follows because h_I defines the same constraint set as h. ■

Comments:

1. These qualifications have implications other than regularity. The MFCQ means that there is no degenerate multiplier system in the Fritz John formulation of the necessary first-order condition, which assigns a multiplier $\nu \geq 0$ to the objective f. Such a generalised multiplier system exists always, also for irregular constraints: see, e.g., [5, p. 181] or [6, Theorem 6.6.2 and 6.10.3]. When $\nu > 0$, it can be set equal to 1 by scaling. When $\nu = 0$, the multiplier system is called *degenerate* (or singular). Nonexistence of a degenerate multiplier system is exactly the MFCQ (as is obvious from its first form, viz., Condition 1 of Lemma 14). Regularity, being a weaker condition, ensures only that a non-degenerate multiplier system exists (Theorems 22 and 25 below), without excluding the existence of a degenerate system. LICQ guarantees that a multiplier system exists, is not degenerate, and is unique after scaling ν to 1.

2. An equality constraint can be expressed as a pair of inequality constraints; this does not affect regularity (since the linearisation cone remains unchanged). Of course, neither LICQ nor MFCQ can hold once the constraints include a pair of opposite inequalities (with the same constraint function). The conversion of an equality to a pair of inequalities by itself gives rise to a degenerate multiplier system (with just two nonzero entries); this effect can be removed by strengthening the Fritz John FOC as in, e.g., [6, Theorem 5.7.1].

3. After linearisation, equalities can be replaced by inequalities without any disadvantage; this is done in the proofs of Appendix A.

6 Examples of irregular constraint representation

As the terminology and notation of Definition 10 make clear, regularity depends on a particular choice of the functions (h and g) representing the constraint set C, and not only on C itself. This is because the linearisation cone L_x does depend on h and g, although the tangent cone T_x depends only on C. The widely used abbreviation "x is a regular point of C" can mislead, since x can be regular for one representation of C but irregular for another. For example, a regular description of C (with a nonzero Jacobian matrix) can always be made irregular by squaring the equalities and cubing the inequalities.[19] This produces constraint functions with zero gradients; such examples exist of course even with just one variable.

With two-variables, another zero-gradient example of irregularity can be constructed from a curve with a cusp: take the single constraint $0 = h(x_1, x_2) := x_1^3 + x_2^2$. Then

$$\mathrm{T}_{(0,0)}\left\{(x_1, x_2) : x_1^3 + x_2^2 = 0\right\} = \mathbb{R}_- \times \{0\} \neq \mathbb{R}^2 = \mathrm{L}_{(0,0)}(h, \emptyset)$$

because $\nabla h(0) = 0$ (where $g = \emptyset$ means absence of inequality constraints)—i.e., the tangent cone is a half-line, but the linearisation cone is the whole plane. (So the constraint is not quasi-regular either: polars of the two cones are different, too.) The same is true when the same function is used in the single inequality constraint $0 \geq g(x_1, x_2) := x_1^3 + x_2^2$, i.e.,

$$\mathrm{T}_{(0,0)}\left\{(x_1, x_2) : x_1^3 + x_2^2 \leq 0\right\} = \mathbb{R}_- \times \{0\} \neq \mathbb{R}^2 = \mathrm{L}_{(0,0)}(\emptyset, g).$$

But the *reverse* inequality constraint, $g \geq 0$ or $-g \leq 0$, *is* regular because *its* tangent cone is

$$\mathrm{T}_{(0,0)}\left\{(x_1, x_2) : x_1^3 + x_2^2 \geq 0\right\} = \mathbb{R}^2 = \mathrm{L}_{(0,0)}(\emptyset, -g).$$

[19] The set C does not change when h and g are replaced by h^2 and g^3. But at an x with $h(x) = 0$ and $g(x) = 0$, one has $\mathrm{D}\left(h^2\right)(x) = 0$ and $\mathrm{D}\left(g^3\right)(x) = 0$ by the Chain Rule. So $\mathrm{L}_x\left(h^2, g^3\right)$ is the whole space $\mathbb{R}^n$, which cannot equal $\mathrm{T}_x C$ because $\mathrm{T}_x C \subseteq \mathrm{L}_x(h, g) \neq \mathbb{R}^n$ (unless both $\mathrm{D}h(x) = 0$ and $\mathrm{D}g(x) = 0$).

This shows also that regularity does depend on whether the constraints functions serve as equality or inequality constraints, and on the sense of any inequality. By contrast, the LICQ does not distinguish between equality and active inequality constraints.

Another constraint function with a zero gradient gives an example of quasi-regularity without regularity: take the single constraint $0 = h(x_1, x_2) := x_1^2 - x_2^2$. Then $\mathrm{T}_0 C$ is C itself (i.e., two intersecting straight lines), but $\mathrm{L}_0(h, \emptyset)$ is the whole plane (since $\nabla h(0) = 0$). However, their polars are both equal to the zero cone. (A three-dimensional variant gives an example of quasi-regularity, without regularity, in which the tangent cone is convex but not polyhedral like the linearisation cone: take the single constraint $0 \geq g(x_1, x_2, x_3) := x_1^2 + x_2^2 - x_3^2$; then $\mathrm{T}_0 C$, equal to C, is the "ice-cream cone".)

Examples of irregularity with *non*-zero constraint gradients require two variables and can be constructed from two curves tangent to each other.

Example 18 (Irregular inequality constraints) *Define* $g_1(x_1, x_2) = x_1^3 - x_2$ *and* $g_2(x_1, x_2) = x_1^3 + x_2$. *(The constraint set* $C = \{g \leq 0\}$ *has a cusp.) At* $(0, 0)$, *the tangent cone is a half-line:*

$$\mathrm{T}_{(0,0)} C = \{(\Delta x_1, 0) : \Delta x_1 \leq 0\} = \mathbb{R}_- \times \{0\}.$$

But $\nabla g_2(0, 0) = (0, 1) = -\nabla g_1(0, 0)$, *so the linearisation cone is a whole line:*

$$\mathrm{L}_{(0,0)}(\emptyset, g) = \mathbb{R} \times \{0\} \neq \mathrm{T}_{(0,0)} C.$$

So the point $(0, 0)$ *is not regular. (Nor is it quasi-regular: the normal cone* $\mathrm{N} = \mathrm{T}^\circ$ *is the half-space* $\mathbb{R}_+ \times \mathbb{R}$, *but the polar* L° *is the line* $\{0\} \times \mathbb{R}$.*)*

In Example 18, both cones stay the same when just one of the inequalities is changed to an equality. Changing both inequalities into equalities gives an irregularity example with a zero tangent cone: if $h_1(x_1, x_2) = x_1^3 - x_2$ and $h_2(x_1, x_2) = x_1^3 + x_2$ then C is the single point $\{(0, 0)\}$, and $\mathrm{T}_{(0,0)} C = \{(0, 0)\}$. But $\mathrm{L}_{(0,0)}(h, \emptyset)$ is the line $\mathbb{R} \times \{0\}$, as before. So the point $(0, 0)$ is not regular or quasi-regular. (The normal cone $\mathrm{N} = \mathrm{T}^\circ$ is the half-space $\mathbb{R}_+ \times \mathbb{R}$, but the polar L° is the line $\{0\} \times \mathbb{R}$.)

A variant of Example 18 gives irregular equality-constraints with a half-line as the tangent cone (which is therefore a "proper" cone, not a linear space).

Example 19 (Irregular equality constraints) *Define* $h_1(x_1, x_2) = \left(x_1^+\right)^3 - x_2$ *and* $h_2(x_1, x_2) = \left(x_1^+\right)^3 + x_2$, *where* $x^+ := \max\{x, 0\}$. *Then* h *is of class* C^2, *and the constraint set is the half-line*

$$C = \{(x_1, 0) : x_1 \leq 0\} = \mathbb{R}_- \times \{0\}.$$

The tangent and linearisation cones are the same as in Example 18: $\mathrm{T}_{(0,0)} C$ *is the half-line* $\mathbb{R}_- \times \{0\}$, *but* $\mathrm{L}_{(0,0)}(h, \emptyset)$ *is the whole line* $\mathbb{R} \times \{0\}$.

Adjoining a redundant constraint (a constraint implied by the other constraints) spoils linear independence. But it can never spoil regularity, and it can even regularise the constraints. Indeed, this is what happens when the redundant constraint $x_1 \leq 0$ is adjoined to Example 18 or Example 19. Of course, redundancy depends on whether the constraint is an equality or an inequality (and on the inequality's sense). For example, the single constraint $g_1(x_1, x_2) = x_1^2 + x_2 \leq 0$ is regular, and the system remains regular when the redundant constraint $x_2 \leq 0$ is adjoined. But the constraint $x_2 = 0$ is *not* redundant, and to adjoin it would make the point $(0, 0)$ irregular, with a single point as T but a whole line as L. (The same happens when the constraint $x_2 \geq 0$ is adjoined.) This gives another case in which regularity depends on whether the constraints functions serve as equality or inequality constraints (and on the sense of any inequality).

7 Multiplier rules for maxima with equality constraints only

In terms of the constraint map h with $g = \emptyset$, i.e., with no inequality constraints, the constraint set is $C = \{x : h(x) = 0\}$ and the cone $\mathrm{L}_{\overline{x}}$ is actually a linear space. The FOC for a maximum at a regular point $\overline{x}$, given next, means in geometric terms that $\nabla f(\overline{x})$ is orthogonal to $\mathrm{L}_{\overline{x}}(h, \emptyset)$, i.e., is a linear combination of $(\nabla h_e)_{e=1}^m$. Since $\mathrm{L}_{\overline{x}}(h, \emptyset) = \mathrm{T}_{\overline{x}} C$ by regularity, this means exactly that $\nabla f(\overline{x})$ is orthogonal to C.

Theorem 20 (Necessary FOC in Lagrange Multiplier Rule) *Assume that $h(\overline{x}) = 0$, and that $\overline{x}$ is regular for h (with $g = \emptyset$). If $\overline{x}$ is a local maximum point of f on C, then there exists a $\mu \in \mathbb{R}^m$ with*

$$\nabla f(\overline{x})^{\mathrm{T}} = \mu^{\mathrm{T}} \mathrm{D}h(\overline{x}) \tag{22}$$

or, in expanded form, $\nabla f(\overline{x}) = \sum_{e=1}^m \mu_e \nabla h_e(\overline{x})$. If the vectors $(\nabla h_e(\overline{x}))_{e=1}^m$ are linearly independent, i.e., $\operatorname{rank} \mathrm{D}h(\overline{x}) = m$, then such a μ is unique.[20]

Note that $\nabla_x (f - \mu \cdot h)$, which vanishes at $\overline{x}$, is the gradient of the *Lagrangian* a.k.a. *Lagrange function* L.[21] Its additional arguments, μ, are called *Lagrange multipliers* (though the term is also used more narrowly to mean the particular multiplier values $\overline{\mu}$ in the FOC).

Definition 21 *$L(\mu, x) := f(x) - \mu \cdot h(x)$ for every $\mu \in \mathbb{R}^m$ and $x \in D \subseteq \mathbb{R}^n$.*

[20] This is in [5, Theorem 1.9.1] and [6, Theorem 3.2.2 and p. 166, lines 4–7.].

[21] In variational calculus, "Lagrange function" is a better name because "Lagrangian" means the maximand (or minimand).

Points of C meeting the Necessary FOC, called *stationary points* a.k.a. *critical points*, are examined by using the SOC. This consists in restricted definiteness of the Lagrangian's Hessian, i.e., its definiteness as a quadratic form on the subspace tangent to C. With no inequality constraints, the Lagrangian is equal to the objective function on the *whole* constraint set, i.e., $f(x) = L(\mu, x)$ for every μ and $x \in C$ (in brief, $f = L$ on C). This simplifies both formulation and proof of the SOCs.

Theorem 22 (SOCs in Lagrange Multiplier Rules) *Assume that $\overline{x}$ is a stationary point supported by a multiplier $\overline{\mu} \in \mathbb{R}^m$, i.e., $h(\overline{x}) = 0$ and*

$$0 = \nabla_x L(\overline{\mu}, \overline{x})^{\mathrm{T}} = \nabla f(\overline{x})^{\mathrm{T}} - \overline{\mu}^{\mathrm{T}} \mathrm{D}h(\overline{x})$$

or, in expanded form, $\nabla f(\overline{x}) = \sum_{e=1}^{m} \overline{\mu}_e \nabla h_e(\overline{x})$. Then:[22]

1. *(Necessary SOC) If $\overline{x}$ is a local maximum point of f on C, and $\overline{x}$ is regular for h (with $g = \emptyset$), then*

$$\Delta x^{\mathrm{T}} \mathrm{D}^2_{xx} L(\overline{\mu}, \overline{x}) \, \Delta x \leq 0$$

for every Δx such that $\mathrm{D}h(\overline{x}) \, \Delta x = 0$ (i.e., such that $\nabla h_e(\overline{x}) \cdot \Delta x = 0$ for each e).

2. *(Sufficient SOC) Conversely, if*

$$\Delta x^{\mathrm{T}} \mathrm{D}^2_{xx} L(\overline{\mu}, \overline{x}) \, \Delta x < 0$$

for every nonzero Δx such that $\mathrm{D}h(\overline{x}) \, \Delta x = 0$, then $\overline{x}$ is a strict local maximum point of f on C. What is more, there exist numbers $\delta > 0$ and $\zeta > 0$ such that

$$f(x) < f(\overline{x}) - \zeta \|x - \overline{x}\|^2$$

for every $x \in C$ with $\|x - \overline{x}\| < \delta$ (i.e., for every x in the δ-neighbourhood of $\overline{x}$ relative to C).[23]

To use the SOCs, one needs a computational criterion of definiteness for a quadratic form on a linear subspace of $\mathbb{R}^n$. One such result, given next, extends Sylvester's Determinantal (Principal Minor) Test of unrestricted definiteness. It is stated in terms of the $n \times n$ symmetric matrix H representing the quadratic form and an $m \times n$ matrix B (of full row-rank) representing the subspace as $\ker(B\cdot)$.[24]

[22] This is in [5, Theorem 1.9.2], though without the tangent space spelt out, and in [6, Theorems 3.2.2 and 3.3.1].

[23] One could, of course, equalise δ and ζ downwards to $\min\{\delta, \zeta\}$, but this is meaningless unless both can be measured in the same unit.

[24] An alternative to using the extension is to change the variables, from x to a y, so that the subspace is spanned by a set of $n - m$ coordinate vectors (i.e., so that its locus equations become $\Delta y_{n-m+1} = 0, \ldots, \Delta y_n = 0$ instead of $B\Delta x = 0$): then the original Sylvester's Criterion can be applied to the $(n - m)$-square leading submatrix of the new matrix representing the quadratic.

Definition 23 *The d-th leading submatrix of a square matrix M is the $d \times d$ matrix $[M_{rs}]_{r=1}^{d}{}_{s=1}^{d}$. Its determinant is the d-th leading minor.*[25]

Lemma 24 (Determinantal test of restricted definiteness) *For a symmetric $n \times n$ matrix H and an $m \times n$ matrix B of rank m, the following conditions are equivalent to each other:*

1. *H is negative definite on the right kernel of B (i.e., $v^{\mathrm{T}} H v < 0$ for every nonzero $v \in \mathbb{R}^n$ such that $Bv = 0$).*

2. *For each $d = 2m+1, \ldots, m+n$ the d-th leading minor of the $(m+n)$-square matrix*

$$\begin{bmatrix} 0 & B \\ B^{\mathrm{T}} & H \end{bmatrix} \tag{23}$$

is of the sign $(-1)^{d-m}$ (i.e., the signs alternate and start from $(-1)^{m+1}$ or, equivalently, end with $(-1)^n$ for the determinant of the whole matrix (23)).

Proof. See, e.g., [2]. ■
Comments:

1. In Lemma 24, the ignored leading minors of (23) are those of dimension $d \leq 2m$. Those of dimensions $d \leq 2m-1$ are zero. The one of dimension $2m$ is independent of H; its sign is either $(-1)^m$ or 0.

2. For negative definiteness of H, on $\ker(B\cdot)$, the minor-sign sequence in Condition 2 of Lemma 24 must alternate *and* start from the correct sign $(-1)^{m+1}$ for dimension $d = 2m+1$ or, equivalently, end with $(-1)^n$ for $d = m+n$. A constant sign equal to $(-1)^m$ indicates positive definiteness of H, on $\ker(B\cdot)$. If the sign alternates but starts from $(-1)^m$ and hence ends with $(-1)^{n-1}$, or it is constant but equal to $(-1)^{m+1}$, or it is neither constant nor alternating, then the form is *not* definite.[26]

3. To verify the SOC for a constrained maximum (Theorem 22), the Determinantal Test (Lemma 24) can be applied to $H = \mathrm{D}^2_{xx} L(\overline{\mu}, \overline{x})$ with $B = -\mathrm{D}h(\overline{x})$. With this choice of sign, the compound matrix (23) is the *total* Hessian of L with respect to (μ, x), i.e.,

$$\begin{bmatrix} 0 & B \\ B^{\mathrm{T}} & H \end{bmatrix} = \mathrm{D}^2 L = \begin{bmatrix} 0 & -\mathrm{D}h \\ -(\mathrm{D}h)^{\mathrm{T}} & \mathrm{D}^2_{xx} L \end{bmatrix}. \tag{24}$$

[25] Other matrices obtained by selecting a subset of rows and the corresponding subset of columns (with the same indices) are principal but not leading.

[26] This is easiest to see in the unconstrained case with $n = 2$ (and $m = 0$). When the minor sign sequence is either $(+-)$ or $(--)$, this means that the eigenvalues of H are of opposite signs, i.e., the form H is *not* definite (and so, when $\nabla f(\overline{x}) = 0$ and $H = \mathrm{D}^2_{xx} f(\overline{x})$, the stationary point $\overline{x}$ is a saddle point of f). When the sign sequence is either $(++)$ or $(-+)$, the eigenvalues are of the same sign, i.e., the form is definite (and so $\overline{x}$ is a minimum or a maximum, respectively).

This is also known as the *bordered Hessian*—since the *partial* Hessian of L, w.r.t. x alone, is bordered by the constraint map's Jacobian matrix (and its transpose).

4. The partial Hessian in (24) is that of L, *not* of f; this matters when the constraints are nonlinear. For more explanation, see the Comment after the Proof of Theorem 22.

5. Of course, Lemma 24 can just as well be applied with $B = \mathrm{D}h(\overline{x})$ instead of $-\mathrm{D}h(\overline{x})$: the leading minors of the matrix (23) do not change when B is replaced by $-B$ (since this means changing the signs of m columns and m rows, in the d-th leading submatrix, for $d \geq 2m+1$).

6. The SOC for a constrained maximum (Theorem 22) can also be verified by applying the determinantal test of *positive* definiteness to $-\mathrm{D}^2_{xx}L(\overline{\mu}, \overline{x})$. In other words, a stationary point $\overline{x}$ (supported by $\overline{\mu}$) is a maximum if *all* the leading minors, of dimensions from $2m+1$ to $m+n$, of the $(m+n)$-square matrix

$$-\mathrm{D}^2 L = \begin{bmatrix} 0 & \mathrm{D}h \\ (\mathrm{D}h)^{\mathrm{T}} & -\mathrm{D}^2_{xx}L \end{bmatrix}$$

have the sign $(-1)^m$.

8 Multiplier rules for maxima with equality and inequality constraints

In terms of the constraint maps h and g, the constraint set is now

$$C = \{x : h(x) = 0, g(x) \leq 0\}$$

and in the FOC there are additional multipliers λ for the inequality constraints. These are always nonnegative, unlike the multipliers μ for the equality constraints, whose sign is a *priori* indefinite.

Theorem 25 (Necessary FOC in Kuhn-Tucker Multiplier Rule) *Assume that $h(\overline{x}) = 0$, $g(\overline{x}) \leq 0$, and that $\overline{x}$ is regular for h and g. If $\overline{x}$ is a local maximum point of f on C, then there exist a $\mu \in \mathbb{R}^m$ and a $\lambda \in \mathbb{R}^l$ with*

$$\lambda \geq 0 \tag{25}$$

$$\lambda \cdot g(\overline{x}) = 0 \tag{26}$$

$$\nabla f(\overline{x})^{\mathrm{T}} = \mu^{\mathrm{T}} \mathrm{D}h(\overline{x}) + \lambda^{\mathrm{T}} \mathrm{D}g(\overline{x}) \tag{27}$$

i.e., in expanded form, $\nabla f(\overline{x}) = \sum_{e=1}^{m} \mu_e \nabla h_e(\overline{x}) + \sum_{i=1}^{l} \lambda_i \nabla g_i(\overline{x})$. *If additionally the vectors* $(\nabla h_e(\overline{x}))_{e=1}^{m}$ *and* $(\nabla g_i(\overline{x}))_{i \in \mathrm{A}(\overline{x})}$ *are linearly independent, i.e.,*

$$\operatorname{rank} \begin{bmatrix} \mathrm{D}h(\overline{x}) \\ \mathrm{D}g_{\mathrm{A}(\overline{x})}(\overline{x}) \end{bmatrix} = m + \operatorname{card} \mathrm{A}(\overline{x})$$

then such a (μ, λ) *is unique.*[27]

With inequalities, the constraints may have a corner point, i.e., a point at which the linearisation cone is pointed (and *a fortiori* the tangent cone is also line-free). For such points, there is a *sufficient* first-order condition.

Theorem 26 (Sufficient FOC in a multiplier rule) *Assume that* $\overline{x}$ *is a stationary point supported by multipliers* $\overline{\mu} \in \mathbb{R}^m$ *and* $\overline{\lambda} \in \mathbb{R}^l$, *i.e.,* $h(\overline{x}) = 0$, $g(\overline{x}) \leq 0$, *and*

$$\begin{aligned} 0 &\leq \overline{\lambda} \\ 0 &= \overline{\lambda} \cdot g(\overline{x}) \\ 0 &= \nabla_x L\left(\overline{\mu}, \overline{\lambda}, \overline{x}\right)^{\mathrm{T}} = \nabla f(\overline{x})^{\mathrm{T}} - \overline{\mu}^{\mathrm{T}} \mathrm{D}h(\overline{x}) - \overline{\lambda}^{\mathrm{T}} \mathrm{D}g(\overline{x}) \end{aligned}$$

or, in expanded form, $\nabla f(\overline{x}) = \sum_{e=1}^{m} \overline{\mu}_e \nabla h_e(\overline{x}) + \sum_{i=1}^{l} \overline{\lambda}_i \nabla g_i(\overline{x})$. *If additionally*[28]

$$\nabla f(\overline{x}) \cdot \Delta x \neq 0 \tag{28}$$

for every nonzero $\Delta x \in \mathrm{L}_{\overline{x}}(h, g)$, *then* $\overline{x}$ *is a strict local maximum point of* f *on* C. *What is more, there exist numbers* $\delta > 0$ *and* $\zeta > 0$ *such that*

$$f(x) \leq f(\overline{x}) - \zeta \|x - \overline{x}\| \tag{29}$$

for every $x \in C$ *with* $\|x - \overline{x}\| < \delta$.

But more usually the Sufficient FOC fails at a stationary point $\overline{x}$ because, although $\nabla f(\overline{x}) \cdot \Delta x \geq 0$ for every $\Delta x \in \mathrm{L}_{\overline{x}}$ by the Necessary FOC, the inequality is not always strict, i.e., $\nabla f(\overline{x}) \cdot \Delta x = 0$ for some nonzero $\Delta x \in \mathrm{L}_{\overline{x}}$. Second-order conditions are then needed.

Definition 27 $L(\mu, \lambda, x) := f(x) - \mu \cdot h(x) - \lambda \cdot g(x)$ *for every* $(\mu, \lambda) \in \mathbb{R}^m \times \mathbb{R}^l$ *and* $x \in D \subseteq \mathbb{R}^n$.

[27] This is in [5, Theorem 1.10.1], [6, Theorem 4.7.1] and [10, 5.2.18].

[28] Since $\mathrm{L}_{\overline{x}}(h, g)$ is a finitely generated convex cone, it suffices to assume (28) for each generator Δx. The assumption means that $\nabla f(\overline{x})$ lies in the interior of $\mathrm{L}_{\overline{x}}(h, g)^{\circ}$, and it obviously implies that $\mathrm{L}_{\overline{x}}(h, g)$ is pointed.

Unlike $\mu \cdot h$, the extra term $\lambda \cdot g$ of the Lagrangian L does not usually vanish on all of C. The condition $\lambda \cdot g(x) = 0$ is called *complementary slackness* (CS). Given a $\lambda \geq 0$, the set of those points of C which meet the CSC is

$$C_{\mathrm{b}}(\lambda) := \{x \in C : \forall i \ (\lambda_i \neq 0 \Rightarrow g_i(x) = 0)\}. \tag{30}$$

Note that the Lagrangian is equal to the objective function only at those points which meet the constraints in a way consistent with the multiplier signs, i.e.,

$$L(\mu, \lambda, x) = f(x) \text{ if } x \in C_{\mathrm{b}}(\lambda).$$

In contrast to the pure equality-constrained case, f does *not* equal L on all of C.

After solving the Necessary FOC system for (μ, λ, x), it is known which inequality constraints are active and which of these are *binding*, i.e., have nonzero multipliers.[29] In the SOCs, binding constraints are treated just like equalities (and any inactive constraints are simply ignored). Active inequalities with *zero* multipliers are called *liminal* constraints; these are active but nonbinding (and hence are also known as "just active" or "degenerate active"). Except when there is just one such constraint, they cannot be ignored like inactive constraints.[30] Nor can liminal constraints be treated as equalities, like binding constraints. This has been persistently overlooked in economics texts: as a result, the so-called "sufficient" SOC of, e.g., [3, Theorems I.2.5 and II.3.4 (pp. 11 and 38)], [12, 19.8] and [13, 1.E.16 (ii)] is in fact *insufficient* (when there is a liminal constraint). See [7] for a counterexample. As for the error made in [3, Proof of Theorem II.3.4 (p. 38)], it is pinpointed at the end of our discussion of that method, in Section 9 before the final Comment.

The set of all the binding constraints is denoted by

$$\mathrm{B}(\lambda) := \{i : \lambda_i \neq 0\}.$$

A binding constraint is active: for $\lambda \geq 0$ and $x \in C$, the CSC means that $\mathrm{B}(\lambda) \subseteq \mathrm{A}(x)$. The set of all the liminal constraints is $\mathrm{A}(x) \setminus \mathrm{B}(\lambda)$. In this notation

$$C_{\mathrm{b}}(\lambda) = \left\{x : h(x) = 0, \ g_{\mathrm{B}(\lambda)}(x) = 0, \ g_{\setminus \mathrm{B}(\lambda)} \leq 0\right\}$$

where $\setminus \mathrm{B}(\lambda) = \{1, \ldots, l\} \setminus \mathrm{B}(\lambda)$. The SOCs are formulated as though the optimisation problem had this larger system of equality constraints, viz., $\left(h, g_{\mathrm{B}(\lambda)}\right) = 0$.

[29] In both mathematics and economics, the terms "active", "effective", "tight" and "binding" are used as synonyms, which is misleading: one feels that "binding" should mean more than merely "active".

[30] A single liminal constraint can be ignored because definiteness of a quadratic form on a half-space is equivalent to its definiteness on the whole space.

Theorem 28 (SOCs in Kuhn-Tucker Multiplier Rules) *Assume that $\overline{x}$ is a stationary point supported by multipliers $\overline{\mu} \in \mathbb{R}^m$ and $\overline{\lambda} \in \mathbb{R}^l$, i.e., $h(\overline{x}) = 0$, $g(\overline{x}) \leq 0$, and*

$$\begin{aligned} 0 &\leq \overline{\lambda} \\ 0 &= \overline{\lambda} \cdot g(\overline{x}) \\ 0 &= \nabla_x L\left(\overline{\mu}, \overline{\lambda}, \overline{x}\right)^{\mathrm{T}} = \nabla f(\overline{x})^{\mathrm{T}} - \overline{\mu}^{\mathrm{T}} \mathrm{D}h(\overline{x}) - \overline{\lambda}^{\mathrm{T}} \mathrm{D}g(\overline{x}) \end{aligned}$$

or, in expanded form, $\nabla f(\overline{x}) = \sum_{e=1}^{m} \overline{\mu}_e \nabla h_e(\overline{x}) + \sum_{i=1}^{l} \overline{\lambda}_i \nabla g_i(\overline{x})$. Then:[31]

1. *(Necessary SOC) If $\overline{x}$ is a local maximum point of f on C, and $\overline{x}$ is regular as a point of $C_{\mathrm{b}}\left(\overline{\lambda}\right)$, i.e., regular for $\left(h, g_{\mathrm{B}\left(\overline{\lambda}\right)}\right)$ and $g_{\setminus \mathrm{B}\left(\overline{\lambda}\right)}$, then*

$$\Delta x^{\mathrm{T}} \mathrm{D}_{xx}^2 L\left(\overline{\mu}, \overline{\lambda}, \overline{x}\right) \Delta x \leq 0 \tag{31}$$

for every Δx such that

$$\mathrm{D}h(\overline{x})\, \Delta x = 0, \text{ i.e., } \nabla h_e(\overline{x}) \cdot \Delta x = 0 \quad \text{for each } e \tag{32}$$

$$\mathrm{D}g_{\mathrm{B}\left(\overline{\lambda}\right)}(\overline{x})\, \Delta x = 0, \text{ i.e., } \nabla g_i(\overline{x}) \cdot \Delta x = 0 \quad \text{for each } i \text{ such that } \overline{\lambda}_i > 0 \tag{33}$$

$$\mathrm{D}g_{\mathrm{A}(\overline{x})\setminus \mathrm{B}\left(\overline{\lambda}\right)}(\overline{x})\, \Delta x \leq 0, \text{ i.e., } \nabla g_i(\overline{x}) \cdot \Delta x \leq 0 \quad \text{for } i \text{ with } \overline{\lambda}_i = 0 \text{ and } g_i(\overline{x}) = 0. \tag{34}$$

2. *(Sufficient SOC) Conversely, if*

$$\Delta x^{\mathrm{T}} \mathrm{D}_{xx}^2 L\left(\overline{\mu}, \overline{\lambda}, \overline{x}\right) \Delta x < 0 \tag{35}$$

for every nonzero Δx meeting (32)–(34), then $\overline{x}$ is a strict local maximum point of f on C. What is more, there exist numbers $\delta > 0$ and $\zeta > 0$ such that

$$f(x) \leq f(\overline{x}) - \zeta \|x - \overline{x}\|^2 \tag{36}$$

for every $x \in C$ with $\|x - \overline{x}\| < \delta$.

Comments (on the SO Multiplier Rules): These can be elucidated by reflecting on C_{b}'s tangent and linearisation cones, and on its relationship to C.

1. Conditions (32)–(34) mean exactly that Δx is in $\mathrm{L}_{\overline{x}}\left(\left(h, g_{\mathrm{B}\left(\overline{\lambda}\right)}\right), g_{\setminus \mathrm{B}\left(\overline{\lambda}\right)}\right)$. This must equal $T_{\overline{x}} C_{\mathrm{b}}\left(\overline{\lambda}\right)$ if the Necessary SOC is indeed to hold at an optimum $\overline{x}$ (supported by $\overline{\mu}$ and $\overline{\lambda}$).

[31] This is in [6, Theorems 4.7.4 and 4.7.5] and, with all the constraints assumed active, also in [5, Theorems 1.10.2 and 1.10.3].

2. In other words, $\overline{x}$ must be regular for the representation of $C_{\mathrm{b}}\left(\overline{\lambda}\right)$ by $\left(h, g_{\mathrm{B}\left(\overline{\lambda}\right)}\right)$ and $g_{\setminus \mathrm{B}\left(\overline{\lambda}\right)}$. This does *not* always follow from regularity of $\overline{x}$ for the representation of C by h and g.[32]

3. But linear independence of constraint gradients (14) implies both of the regularity conditions—since the LICQ does not distinguish between equalities and active inequalities. It then follows that $T_{\overline{x}} C_{\mathrm{b}}\left(\overline{\lambda}\right) = \mathrm{T}_{\overline{x}} C \cap \bigcap_{i \in \mathrm{B}\left(\overline{\lambda}\right)} \ker \nabla g_i\left(\overline{x}\right)$.

4. Assume that the LICQ holds at $\overline{x}$, a stationary point supported by multipliers $\overline{\mu}$ and $\overline{\lambda} \geq 0$. If additionally $\mathrm{A}\left(\overline{x}\right) = \mathrm{B}\left(\overline{\lambda}\right)$ then $C_{\mathrm{b}}\left(\overline{\lambda}\right)$ is, locally around $\overline{x}$, equal to the topological boundary of C relative to the manifold $\{x : h\left(x\right) = 0\}$.

5. Absence of liminal constraints, known also as *strict complementarity*, simplifies the use of the sufficient SOC: if all the active inequalities are binding, then (34) plays no part, and so the range for Δx, defined by the remaining Conditions (32)–(33), is actually a linear subspace (rather than a "proper" cone). So the Determinantal Test (Lemma 24) can be applied to the bordered Hessian of $\left(h, g_{\mathrm{B}\left(\overline{\lambda}\right)}\right)$, with $m + \operatorname{card} \mathrm{B}\left(\overline{\lambda}\right)$ in place of m. The same technique applies when there is just *one* liminal constraint: it can be ignored.[33]

Comments (on strict complementarity):

1. In sensitivity analysis, liminal constraints have to be excluded by assumption if the optimal solution and its multipliers are to have ordinary derivatives with respect to the problem's parameters: see, e.g., [4, Theorems 2.4.4 and 3.2.2] or [8, Theorem 1]. Without the strict complementarity assumption, the solution (and its multipliers) is usually nondifferentiable, but it is still *directionally* differentiable: see, e.g., [4, Theorem 2.4.5] or [8, Theorems 3 and 4].

2. Some topics involve liminal constraints of necessity. An example is the Le Chatelier Principle, discussed in, e.g., [11].

[32] For example, take the constraints $g_1\left(x_1, x_2\right) = -x_1^3 + x_2 \leq 0$ and $g_2\left(x_1, x_2\right) = x_1^3 + x_2 \leq 0$ (with $h = \emptyset$) at $\overline{x} = (0, 0)$; these meet the MFCQ and are therefore regular. But they cease to be regular once one or both inequalities are turned into equality constraints, as the representation of $C_{\mathrm{b}}\left(\overline{\lambda}_1, \overline{\lambda}_2\right)$ requires (if $\overline{\lambda} \neq 0$). For any maximand with $\nabla f\left(0, 0\right) = (0, 1)$, the point $(0, 0)$ is stationary, and every supporting $\overline{\lambda}$ is nonzero.

[33] Matters complicate with two (or more) liminal constraints. In the two-variable case (and no other constraints), the question is whether $\mathrm{D}_{xx}^2 L$ (equal to $\mathrm{D}_{xx}^2 f$, since $\lambda_1 = 0 = \lambda_2$) is negative definite on a planar sector between the tangents to the two borders $\{g_1 = 0\}$ and $\{g_2 = 0\}$ of the sublevel sets. In addition to negative definiteness on the two extreme rays of the sector, one needs to verify that the form does not vanish anywhere in the sector.

9 Square slack reduction to equality constraints and its limitations

Each inequality constraint, $g_i(x) \leq 0$, on the decision variables $x = (x_1, \ldots, x_n)$ can be converted to an equality $g_i(x) + s_i^2/2 = 0$ by introducing a variable whose square takes up any slack. Even though the maximand f does not depend on it, s_i *is* an extra decision variable: the converted problem is to be solved for both x and $s = (s_1, \ldots, s_m)$. This artifice can be used to derive optimality conditions for the *original* inequality-constrained problem from those for the *converted*, purely equality-constrained problem. It is adequate as a way of deriving the Kuhn-Tucker Necessary FOC, although even this necessitates the use of both the FOC and the Necessary SOC (for the converted problem): the multiplier signs (25) do *not* follow from the FOC alone.

By replacing all inequalities with equalities, the method may also seem to somehow by-pass the complication of liminal (active but not binding) inequality constraints, but in fact it offers nothing in this respect. Applied to the equality-constrained converted problem, the second-order Lagrange multiplier rules translate into inferior results for the inequality-constrained original problem: instead of the standard Kuhn-Tucker second-order rules, they yield only a weaker necessary condition and a much stronger sufficient condition. Both consequences are undesirable, and widen the gap between the necessary and the sufficient conditions. The weaker necessary condition is so because, when there are liminal constraints, it asserts the Hessian's negative semidefiniteness on a smaller set of vectors Δx than that defined by (32)–(34) in the standard Necessary SOC. The "much stronger" sufficient condition rules out liminal constraints (i.e., it implies that $g_i(x)$ and the associated multiplier μ_i cannot both be zeros). It is simply the conjunction of strict complementarity and the standard Sufficient SOC (of Theorem 28). It is, pointlessly, even stronger than the so-called Strong Sufficient SOC.[34] The latter, though needlessly stringent as an optimality test,[35] does have a place in the directional-derivative results of solution-sensitivity analysis without strict complementarity: see, e.g., [4, Theorem 2.4.5] or [8, Theorems 3 and 4].

To derive the Kuhn-Tucker Necessary FOC in this way, and to see that it yields inferior *second*-order conditions, take for simplicity a problem with inequality constraints only, i.e., maximisation of $f(x)$ over x subject to $g(x) \leq 0$; even the one-variable, one-constraint case ($n = 1$, $l = 1$) can provide an illustration. After conversion, the problem is to maximise $f(x)$ over $x \in \mathbb{R}^n$ and $s \in \mathbb{R}^l$ subject to $g_i(x) + s_i^2/2 = 0$ for $i = 1, \ldots, l$.

[34] The Strong Sufficient SOC is like the the standard Sufficient SOC *except* for ignoring any liminal constraints, i.e., treating them just like inactive ones. This makes no difference if there is just one liminal constraint, but if there are two or more, to ignore even one of them is to strengthen the SOC: see [7].

[35] It is put as an optimality test in, e.g., [10, 5.3.4].

Its Lagrangian is

$$\begin{aligned}\mathcal{L}(\mu, x, s) &= f(x) - \sum_{i=1}^{l} \mu_i \left(g_i(x) + \frac{s_i^2}{2} \right) = f(x) - \mu \cdot g(x) - \frac{1}{2} s^{\mathrm{T}} \operatorname{diag}(\mu) s \\ &= L(\mu, x) - \frac{1}{2} s^{\mathrm{T}} \operatorname{diag}(\mu) s\end{aligned}$$

where L is the Lagrangian for the *original* problem, and $\operatorname{diag}(\mu)$ is the square matrix with the m-tuple μ as the diagonal and with zero off-diagonal entries. By Theorem 20, the Necessary FOC for the converted problem is

$$0 = -g_i(x) - \frac{s_i^2}{2} \quad \text{for each } i \tag{37}$$

$$0 = \nabla_x \mathcal{L} = \nabla f(x) - \sum_{i=1}^{l} \mu_i \nabla g_i(x) \tag{38}$$

$$0 = \nabla_s \mathcal{L} = -\operatorname{diag}(\mu) s. \tag{39}$$

By (37), $g_i(x) = -s_i^2/2 \leq 0$. And (39) means that $\mu_i s_i = 0$ for each i—i.e., $\mu_i = 0$ or $g_i(x) = 0$ for each i. To show that $\mu \geq 0$ when x is a maximum point, use the Necessary SOC on the Lagrangian's Hessian

$$\mathrm{D}^2_{(x,s)(x,s)} \mathcal{L}(\mu, x, s) = \begin{bmatrix} \mathrm{D}^2_{xx} L(\mu, x) & 0 \\ 0 & -\operatorname{diag}(\mu) \end{bmatrix}. \tag{40}$$

That is, by Part 1 of Theorem 22, $\mathrm{D}^2_{(x,s)(x,s)} \mathcal{L}$ is negative semidefinite on the linear space $\ker \left(\begin{bmatrix} \mathrm{D}g(x) & \operatorname{diag}(s) \end{bmatrix} \cdot \right)$, i.e.,

$$\begin{aligned}0 &\geq \begin{bmatrix} \Delta x^{\mathrm{T}} & \Delta s^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathrm{D}^2_{xx} L(\mu, x) & 0 \\ 0 & -\operatorname{diag}(\mu) \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta s \end{bmatrix} \\ &= \Delta x^{\mathrm{T}} \mathrm{D}^2_{xx} L(\mu, x) \Delta x - \Delta s^{\mathrm{T}} \operatorname{diag}(\mu) \Delta s\end{aligned} \tag{41}$$

for every $(\Delta x, \Delta s)$ such that

$$\nabla g_i(x) \cdot \Delta x + s_i \Delta s_i = 0 \quad \text{for each } i. \tag{42}$$

For $\Delta x = 0$, this means that

$$0 \leq \sum_{i=1}^{l} \mu_i (\Delta s_i)^2 \quad \text{for every } \Delta s \text{ with } \sum_{i=1}^{l} s_i \Delta s_i = 0.$$

For any i with $s_i = 0$, the last condition does *not* restrict Δs_i, and it follows that $\mu_i \geq 0$. So $\mu \geq 0$ (since $\mu_i = 0$ if $s_i \neq 0$). This completes the derivation of the Kuhn-Tucker FOC

(for the original problem): it has been shown that $g(x) \leq 0$, $\mu \geq 0$, $\mu \cdot g(x) = 0$ and $\nabla f(x)^{\mathrm{T}} = \mu^{\mathrm{T}} \mathrm{D} g(x)$. The same argument is in [3, Proof of Theorem II.3.1 (pp. 34–36)].

The inferior second-order results obtainable in this way are derived next. Given a stationary point (x, s) and a supporting multiplier μ, which meet (37)–(39), recall the notation (6) for the set of active constraints (for the original problem), i.e.,

$$\mathrm{A}(x) = \{i : g_i(x) = 0\} = \{i : s_i = 0\}.$$

The Necessary SOC for the converted problem—i.e., the negative semidefiniteness (41) under the restrictions (42) on $(\Delta x, \Delta s)$—can be restated as

$$\Delta x^{\mathrm{T}} \mathrm{D}^2_{xx} L(\mu, x) \Delta x \leq \sum_{i \in \mathrm{A}(x)} \mu_i (\Delta s_i)^2 \quad \text{if } \nabla g_i(x) \cdot \Delta x + s_i \Delta s_i = 0 \text{ for each } i \tag{43}$$

since $\mu_i = 0$ for any $i \notin \mathrm{A}(x)$. Furthermore, for any $i \notin \mathrm{A}(x)$ its restriction on $(\Delta x, \Delta s)$ can be simply deleted from (43). This is because it can be solved for $\Delta s_i = -\nabla g_i(x) \cdot \Delta x / s_i$ given any Δx, and because the summation in (43) excludes any such i (i.e., the sum does not depend on Δs_i). So (43) is equivalent to

$$\Delta x^{\mathrm{T}} \mathrm{D}^2_{xx} L(\mu, x) \Delta x \leq \sum_{i \in \mathrm{A}(x)} \mu_i (\Delta s_i)^2 \quad \text{if } \nabla g_i(x) \cdot \Delta x = 0 \text{ for each } i \in \mathrm{A}(x).$$

Finally, since each Δs_i in the above sum is unrestricted and $\mu \geq 0$, the Necessary SOC for the converted problem is equivalent to

$$\Delta x^{\mathrm{T}} \mathrm{D}^2_{xx} L(\mu, x) \Delta x \leq 0 \quad \text{if } \nabla g_i(x) \cdot \Delta x = 0 \text{ for each } i \in \mathrm{A}(x). \tag{44}$$

The same argument is in [3, Proof of Theorem II.3.3 (pp. 37–38)]. Note that (44) is weaker than the standard Kuhn-Tucker Necessary SOC for the original problem because it treats liminal constraints just like equality constraints: when i is liminal (i.e., $g_i(x) = 0$ and $\mu_i = 0$), the equality restriction on Δx in (44) is more stringent than the corresponding inequality in (34). For example, when there is just one liminal constraint, the standard formulation shows negative semidefiniteness on a space of one dimension higher than does (44). In the very simplest case of just one variable x and one constraint with $\mathrm{d}g/\mathrm{d}x \neq 0$, if $g(x) = 0$ and $\mu = 0$ then (44) becomes the tautology $0 \leq 0$ (since Δx can only be 0), whereas the standard Kuhn-Tucker Necessary SOC means then that $\mathrm{d}^2 f/\mathrm{d}x^2 \leq 0$.

The Sufficient SOC for the converted problem is that (40) be negative definite under the restrictions (42), i.e., that

$$\Delta x^{\mathrm{T}} \mathrm{D}^2_{xx} L(\mu, x) \Delta x \leq \sum_{i=1}^{l} \mu_i (\Delta s_i)^2 \quad \text{if } \begin{cases} \nabla g_i(x) \cdot \Delta x + s_i \Delta s_i = 0 \text{ for each } i \\ (\Delta x, \Delta s) \neq (0, 0) \end{cases}. \tag{45}$$

This can be similarly restated (by recalling that $\mu_i = 0$ for any $i \notin \mathrm{A}(x)$, and by solving again for $\Delta s_i = -\nabla g_i(x) \cdot \Delta x / s_i$, which is 0 if $\Delta x = 0$) as

$$\Delta x^{\mathrm{T}} \mathrm{D}^2_{xx} L(\mu, x)\, \Delta x < \sum_{i \in \mathrm{A}(x)} \mu_i (\Delta s_i)^2 \quad \text{if} \begin{cases} \nabla g_i(x) \cdot \Delta x = 0 \text{ for each } i \in \mathrm{A}(x) \\ \left(\Delta x, \Delta s_{\mathrm{A}(x)}\right) \neq (0,0) \end{cases} . \tag{46}$$

Next, by considering three cases: (i) $\Delta x \neq 0$ but $\Delta s_{\mathrm{A}} = 0$, (ii) $\Delta x = 0$ but $\Delta s_{\mathrm{A}} \neq 0$, and (iii) both $\Delta x \neq 0$ and $\Delta s_{\mathrm{A}} \neq 0$, (46) is shown to be equivalent to the conjuction of[36]

$$\Delta x^{\mathrm{T}} \mathrm{D}^2_{xx} L(\mu, x)\, \Delta x < 0 \quad \text{if } \Delta x \neq 0 \text{ and } \nabla g_i(x) \cdot \Delta x = 0 \text{ for each } i \in \mathrm{A}(x) \tag{47}$$

$$\mu_i > 0 \quad \text{for each } i \in \mathrm{A}(x). \tag{48}$$

Without (48), Condition (47) is generally weaker than the standard Kuhn-Tucker Sufficient SOC (for the original problem) because, like (44), it treats liminal constraints just like equality constraints: when i is a liminal constraint (i.e., $g_i(x) = 0$ and $\mu_i = 0$), the equality restriction on Δx in (47) is more stringent than the corresponding inequality in (34). In other words, it is weaker because it requires the Hessian to be negative definite on a smaller set of nonzero vectors Δx than in (32)–(34). Under (48), which is strict complementarity, (47) becomes of course equivalent to the Kuhn-Tucker Sufficient SOC. Thus the method reproduces the standard sufficiency result—but *only* under the strong and extraneous assumption of strict complementarity.

Being weaker than the standard Sufficient SOC, Condition (47) is by itself insufficient for a local maximum—contrary to the assertion of [3, Theorems I.2.5 and II.3.4 (pp. 11 and 38)], which is reproduced in, e.g., [12, 19.8] and [13, 1.E.16 (ii)]. Although the argument in [3, Proof of Theorem II.3.4 (p. 38)] is largely the same as the above one, it goes wrong at the end [3, p. 38, line 3 f.b.]: in our notation, from $\left(\Delta x, \Delta s_{\mathrm{A}(x)}\right) \neq (0,0)$ it obviously does *not* follow that $\Delta x \neq 0$ (which *would* have given $\Delta x^{\mathrm{T}} \mathrm{D}^2_{xx} L \Delta x < 0$). When $\Delta x = 0$, this argument must rely on positivity of $\sum_{i \in \mathrm{A}(x)} \mu_i (\Delta s_i)^2$ for $\Delta s_{\mathrm{A}(x)} \neq 0$—and it necessitates the missing assumption of strict complementarity. This oversight produces a *false* "sufficient" SOC, which is in fact *insufficient* when there is a liminal constraint. When all inequality constraints are liminal,[37] its use can lead to a strict minimum point being misidentified as a strict maximum: see [7] for an example.

Comment: That the square-slack approach to inequalities cannot deal properly with liminal constraints is easiest to see in the one-variable, one-constraint case: the Sufficient SOC for the converted problem is that the 2×2 Hessian matrix $\mathrm{D}^2_{(x,s)(x,s)} \mathcal{L}$ be negative

[36] Case (iii) does not add another condition because the inequality in (46) then follows from either of (47) and (48).

[37] When there *is* a binding inequality, maximum and minimum points are distinguished from each other already by their multiplier signs in the FOCs: for a point to be stationary for both maximisation and minimisation, all the active inequality constraints must be liminal.

definite on $\ker\left(\left(\mathrm{d}g/\mathrm{d}x, s\right)\cdot\right)$, i.e., on nonzero scalar multiples of the vector $(-s, \mathrm{d}g/\mathrm{d}x)$. This means that

$$0 > s^2\left(\frac{\mathrm{d}^2 f}{\mathrm{d}x^2} - \mu\frac{\mathrm{d}^2 g}{\mathrm{d}x^2}\right) - \mu(\frac{\mathrm{d}g}{\mathrm{d}x})^2 \tag{49}$$

which implies that $(\mu, s) \neq (0,0)$, i.e., it implies strict complementarity. This can also be viewed as the simplest application of the Determinantal Test (Lemma 24): the expression in (49) equals

$$\det\begin{bmatrix} 0 & \mathrm{d}g/\mathrm{d}x & s \\ \mathrm{d}g/\mathrm{d}x & -\mathrm{d}^2 f/\mathrm{d}x^2 + \mu\mathrm{d}^2 g/\mathrm{d}x^2 & 0 \\ s & 0 & \mu \end{bmatrix}.$$

10 Maxima with abstract constraints and derivation of multiplier rules

The multiplier rules derive from the FOC and SOCs for a maximum on an arbitrary set C, which are given next.

Theorem 29 (Abstract Necessary FOC) *If $\overline{x}$ is a local maximum point of f on C, then $\nabla f(\overline{x})\cdot \Delta x \leq 0$ for every $\Delta x \in \mathrm{T}_{\overline{x}}C$ (i.e., $\nabla f(\overline{x}) \in \mathrm{N}_{\overline{x}}C$). So $\nabla f(\overline{x})\cdot\Delta x = 0$ for every $\Delta x \in \mathrm{T}_{\overline{x}}C \cap (-\mathrm{T}_{\overline{x}}C)$.*[38]

Proof. One can assume that $\|\Delta x\| = 1$. Take a sequence $(x(k))_{k=1}^{\infty}$ in $C \setminus \{\overline{x}\}$ converging to $\overline{x}$ from the direction Δx, i.e., $(x(k) - \overline{x})/\|x(k) - \overline{x}\| \to \Delta x$ as $k \to \infty$. Then $f(x(k)) \leq f(\overline{x})$ for every sufficiently large k, and so

$$0 \geq \lim_{k\to\infty} \frac{f(x(k)) - f(\overline{x})}{\|x(k) - \overline{x}\|} = \nabla f(\overline{x})\cdot\Delta x$$

by (3). ■

Proof of Theorem 20 (NFOC in Lagrange Multiplier Rule, equalities only). Here

$$C = \{x : h(x) = 0\}.$$

By Theorem 29, $\nabla f(\overline{x})\cdot\Delta x \leq 0$ for $\Delta x \in \mathrm{T}_{\overline{x}}C$. By the regularity assumption, $\mathrm{T}_{\overline{x}}C$ equals $\mathrm{L}_{\overline{x}}(h)$, which is a linear space (by its definition (7), since there are no inequality constraints). So actually

$$\nabla f(\overline{x})\cdot\Delta x = 0 \quad \text{for every } \Delta x \in \mathrm{T}_{\overline{x}}C = \mathrm{L}_{\overline{x}}(h) := \ker\left(\mathrm{D}h(\overline{x})\cdot\right).$$

To complete the proof, apply Lemma 33 (the Factorisation Lemma) to $p = \nabla f(\overline{x})$ and $B(e) = \nabla h_e(\overline{x})$. ■

[38] This is in [5, Theorem 1.8.1] and in [6, Theorem 4.6.1].

Comment: So, without inequality constraints, conversion of the Abstract Necessary FOC into a Lagrange multiplier rule requires only the Factorisation Lemma, which is a purely algebraic result: it does not rely on separation of convex sets. The case of inequality constraints, dealt with next, does require a separation argument (viz., Farkas' Lemma).

Proof of Theorem 25 (NFOC in Kuhn-Tucker Multiplier Rule, with inequalities). Here

$$C = \{x : h(x) = 0, g(x) \leq 0\}.$$

By Theorem 29 and the regularity assumption,

$$\nabla f(\overline{x}) \cdot \Delta x \leq 0 \quad \text{for every } \Delta x \in \mathrm{T}_{\overline{x}} C = \mathrm{L}_{\overline{x}}(h, g).$$

To complete the proof, apply Lemma 34 (Farkas' Lemma) to $p = \nabla f(\overline{x})$ with $B = \mathrm{D}h_e(\overline{x})$ and $A = \mathrm{D}g_i(\overline{x})$. ■

Theorem 30 (Abstract Sufficient FOC) *If $\nabla f(\overline{x}) \cdot \Delta x < 0$ for every nonzero $\Delta x \in \mathrm{T}_{\overline{x}} C$ (which implies that $\mathrm{T}_{\overline{x}} C$ is line-free), then $\overline{x}$ is a is a strict local maximum point of f on C.*[39] *What is more, there exist numbers $\delta > 0$ and $\zeta > 0$ such that*

$$f(x) \leq f(\overline{x}) - \zeta \|x - \overline{x}\|$$

for every $x \in C$ with $\|x - \overline{x}\| < \delta$.[40]

Proof. Suppose contrarily that there is a sequence $(x(k))_{k=1}^{\infty}$ in C such that

$$x(k) \to \overline{x} \text{ as } k \to \infty \quad \text{and} \quad f(x(k)) > f(\overline{x}) - \frac{1}{k} \|x(k) - \overline{x}\| \text{ for each } k.$$

Then $x(k) \neq \overline{x}$, and since the unit sphere is compact, one can assume (by passing to a subsequence) that $(x(k) - \overline{x}) / \|x(k) - \overline{x}\|$ converges to some Δx (a unit vector). So, by (3),

$$\nabla f(x) \cdot \Delta x = \lim_{k \to \infty} \frac{f(x(k)) - f(\overline{x})}{\|x(k) - \overline{x}\|} \geq -\lim_{k \to \infty} \frac{1}{k} = 0$$

which contradicts the assumption. ■

Proof of Theorem 26 (SFOC, for corner points). For every $\Delta x \in \mathrm{L}_{\overline{x}}(h, g)$

$$\nabla f(\overline{x}) \cdot \Delta x = \overline{\mu}^{\mathrm{T}} \mathrm{D}h(\overline{x}) \Delta x + \overline{\lambda}^{\mathrm{T}} \mathrm{D}g(\overline{x}) \Delta x = 0 + \overline{\lambda}^{\mathrm{T}} \mathrm{D}g(\overline{x}) \Delta x \leq 0.$$

Since equality is excluded by the assumption (28), it follows that actually $\nabla f(\overline{x}) \cdot \Delta x < 0$ every nonzero $\Delta x \in \mathrm{L}_{\overline{x}}(h, g)$, and *a fortiori* for every nonzero $\Delta x \in \mathrm{T}_{\overline{x}} C$. So, by Theorem 30, $\overline{x}$ is a strict local maximum point of f on $C = \{x : h(x) = 0, g(x) \leq 0\}$. ■

The next result will be applied to the Lagrangian L as a function of the decision variables x (with the multipliers $\overline{\mu}$ and $\overline{\lambda}$ fixed).

[39] If, as in the case of regular constraints, $\mathrm{T}_{\overline{x}} C$ is a finitely generated convex cone, then it obviously suffices that $\nabla f(\overline{x}) \cdot \Delta x < 0$ for each generator Δx.

[40] This is in [6, Theorem 4.6.3].

Theorem 31 (Abstract SOCs) *Assume that $\nabla L(\overline{x}) = 0$, for a real-valued function L that is twice differentiable at a point $\overline{x} \in C$. Then:*[41]

1. *(Necessary SOC) If $\overline{x}$ is a local maximum point of L on C, then $\Delta x^{\mathrm{T}} \mathrm{D}^2 L(\overline{x}) \Delta x \leq 0$ for every $\Delta x \in \mathrm{T}_{\overline{x}} C$.*

2. *(Sufficient SOC) Conversely, if $\Delta x^{\mathrm{T}} \mathrm{D}^2 L(\overline{x}) \Delta x < 0$ for every nonzero $\Delta x \in \mathrm{T}_{\overline{x}} C$, then $\overline{x}$ is a strict local maximum point of L on C. What is more, there exist numbers $\delta > 0$ and $\zeta > 0$ such that*

$$L(x) \leq L(\overline{x}) - \zeta \|x - \overline{x}\|^2$$

for every $x \in C$ with $\|x - \overline{x}\| < \delta$.

Proof. For Part 1, after scaling Δx to unit length, take a sequence $(x(k))_{k=1}^{\infty}$ in $C \setminus \{\overline{x}\}$ converging to $\overline{x}$ from the direction Δx. Then $L(x(k)) \leq L(\overline{x})$ for every sufficiently large k, and so

$$\begin{aligned} 0 &\geq \lim_{k\to\infty} \frac{L(x(k)) - L(\overline{x})}{\|x(k) - \overline{x}\|^2} = \lim_{k\to\infty} \frac{L(x(k)) - L(\overline{x}) - \nabla L(\overline{x}) \cdot (x(k) - \overline{x})}{\|x(k) - \overline{x}\|^2} \\ &= \frac{1}{2} \Delta x^{\mathrm{T}} \mathrm{D}^2 L(\overline{x}) \Delta x \end{aligned} \tag{50}$$

by (4).

For Part 2, suppose contrarily that there is a sequence $(x(k))_{k=1}^{\infty}$ in C such that

$$x(k) \to \overline{x} \text{ as } k \to \infty \quad \text{and} \quad L(x(k)) > L(\overline{x}) - \frac{1}{k} \|x(k) - \overline{x}\|^2 \text{ for each } k.$$

Then $x(k) \neq \overline{x}$, and since the unit sphere is compact, one can assume (by passing to a subsequence) that $(x(k) - \overline{x}) / \|x(k) - \overline{x}\|$ converges to some Δx (a unit vector). So, again by (4),

$$\begin{aligned} \frac{1}{2} \Delta x^{\mathrm{T}} \mathrm{D}^2 L(\overline{x}) \Delta x &= \lim_{k\to\infty} \frac{L(x(k)) - L(\overline{x}) - \nabla L(\overline{x}) \cdot (x(k) - \overline{x})}{\|x(k) - \overline{x}\|^2} \\ &= \lim_{k\to\infty} \frac{L(x(k)) - L(\overline{x})}{\|x(k) - \overline{x}\|^2} \geq -\lim_{k\to\infty} \frac{1}{k} = 0 \end{aligned} \tag{51}$$

which contradicts the assumption. ■

In the case of equality constraints only, to maximise f on C means exactly the same as to maximise $L(\overline{\mu}, \cdot)$ on C (since the two functions are simply equal on all of C).[42]

[41] This is in [5, Theorem 1.8.2]; Part 1 is also in [6, Theorem 4.6.5].

[42] In [5], this gives [5, Theorem 1.9.2] immediately from [5, Theorem 1.8.2].

So both of the above Abstract SOCs can be applied (to L) to prove the corresponding second-order multiplier rules without further ado.

Proof of Theorem 22 (SOCs as Multiplier Rules, equality constraints only). With $\overline{\mu}$ fixed, abbreviate $L(\overline{\mu}, \cdot)$ to L; then $\nabla L(\overline{x}) = 0$ by assumption. So Theorem 31 applies, and the result transcribes exactly into Theorem 22, since $L = f$ on $C = \{x : h(x) = 0\}$, and since $\ker(\mathrm{D}h(\overline{x})\cdot) = \mathrm{L}_{\overline{x}}(h) = \mathrm{T}_{\overline{x}}C$ by regularity (which is needed only for Part 1, since $\mathrm{L}_{\overline{x}}$ always contains $T_{\overline{x}}$ by Lemma 9). ■

Comment: The preceding proofs are next examined to explain why the SOCs must use the Hessian of the Lagrangian $L(\overline{\mu}, \cdot)$ and *not* that of the maximand f (unless all the constraints h are linear functions, in which case the two Hessians are equal).

1. Theorem 31 *cannot* be applied to f instead of $L(\overline{\mu}, \cdot)$ because $\nabla f(\overline{x})$ does not vanish (unlike $\nabla_x L(\overline{\mu}, \overline{x})$, which vanishes by the FOC).

2. And Theorem 31 does depend on the assumption that $\nabla L(\overline{x}) = 0$: this *cannot* be weakened to
$$\nabla L(\overline{x}) \cdot \Delta x = 0 \quad \text{for } \Delta x \in \mathrm{T}_{\overline{x}}C. \tag{52}$$
When $\mathrm{T}_{\overline{x}}C$ is a linear space, the weaker condition (52) is exactly the FOC for $\overline{x}$ to be a maximum point of L (and hence it would hold in Part 1 of Theorem 31 even if $\nabla L(\overline{x})$ were nonzero). But (52) would not suffice (for the Proof of Theorem 31 if $\nabla L(\overline{x})$ were nonzero) because the term $\nabla L(\overline{x}) \cdot (x(k) - \overline{x})$ in (50) and (51) would not vanish (unless the constraints are linear, in which case C and $\overline{x} + \mathrm{T}_{\overline{x}}C$ are locally identical, so $x(k) - \overline{x} \in \mathrm{T}_{\overline{x}}C$ and hence $\nabla L(\overline{x}) \cdot (x(k) - \overline{x}) = 0$ purely by (52)). Nor would this term generally vanish in the limit after the division, i.e., it would not be of higher order than $\|x(k) - \overline{x}\|^2$ (although it would still be of higher order than $\|x(k) - \overline{x}\|$, purely by (3) and (52)).

Even with inequality constraints, the Necessary Abstract SOC is still applicable enough to prove the corresponding multiplier rule. It is applied, to $L(\overline{\mu}, \overline{\lambda}, \cdot)$, on the smaller set $C_{\mathrm{b}}(\overline{\lambda})$—on which f and L are still equal. Elsewhere on the constraint set, $f \leq L$. So a maximum point $\overline{x}$ for f, which always lies in C_{b}, need *not* be a maximum point for L on C (since $L(\overline{x}) = f(\overline{x}) \geq f(x) \leq L(x)$ for $x \in C$). But it is, of course, a maximum point for L on C_{b}.

By contrast, the above Sufficient Abstract SOC is not readily applicable if the binding inequality constraints are indeed to be treated just like equality constraints.[43] It is to achieve this important objective that the Hessian's negative definiteness is assumed,

[43] There are at least two advantages from being able to handle the binding constraints like equality constraints. First, the sufficient SOC matches the necessary SOC (i.e., is obtained just by making the definiteness strict). Second, when there is no liminal constraint, the negative definiteness condition is to be verified on a linear subspace (rather than on a cone).

in Part 2 of Theorem 28, only for those $\Delta x \in \mathrm{T}_{\overline{x}}C$ with $\nabla g_i(\overline{x}) \cdot \Delta x = 0$ for each binding constraint i. Since Part 2 of Theorem 31 requires negative definiteness for *every* $\Delta x \in \mathrm{T}_{\overline{x}}C$, it does not apply as it stands. What is needed, then, is a variant (of Part 2 of Theorem 31) that, from the weaker premise, draws the weaker but adequate conclusion that $\overline{x}$ is a (local) maximum point for f, though not necessarily for L (on C).[44] Such a variant is given next.

Theorem 32 (Modified Abstract Sufficient SOC) *In addition to $\nabla L(\overline{x}) = 0$, for a real-valued L that is twice differentiable at $\overline{x} \in C$, assume that $L \geq f$ on C and $L(\overline{x}) = f(\overline{x})$. If, furthermore,*

$$\Delta x^{\mathrm{T}} \mathrm{D}^2 L(\overline{x}) \Delta x < 0$$

for every nonzero $\Delta x \in \mathrm{T}_{\overline{x}}C$ such that $\nabla(L-f)(\overline{x})\Delta x = 0$ (i.e., $\nabla f(\overline{x}) \cdot \Delta x = 0$), then $\overline{x}$ is a (strict) local maximum point of f on C. What is more, there exist numbers $\delta > 0$ and $\zeta > 0$ such that

$$f(x) \leq f(\overline{x}) - \zeta \|x - \overline{x}\|^2$$

for every $x \in C$ with $\|x - \overline{x}\| < \delta$.[45]

Proof. Suppose contrarily that there is a sequence $(x(k))_{k=1}^{\infty}$ in C such that

$$x(k) \to \overline{x} \text{ as } k \to \infty \text{ and } f(x(k)) > f(\overline{x}) - \frac{1}{k}\|x(k) - \overline{x}\|^2 \text{ for every } k.$$

Then $x(k) \neq \overline{x}$; and one can assume that $(x(k) - \overline{x}) / \|x(k) - \overline{x}\|$ converges to a unit vector Δx. So, by (4),

$$\begin{aligned} \frac{1}{2}\Delta x^{\mathrm{T}} \mathrm{D}^2 L(\overline{x}) \Delta x &= \lim_{k\to\infty} \frac{L(x(k)) - L(\overline{x})}{\|x(k) - \overline{x}\|^2} \geq \limsup_{k\to\infty} \frac{f(x(k)) - f(\overline{x})}{\|x(k) - \overline{x}\|^2} \\ &\geq \liminf_{k\to\infty} \frac{f(x(k)) - f(\overline{x})}{\|x(k) - \overline{x}\|^2} \geq -\lim_{k\to\infty} \frac{1}{k} = 0. \end{aligned}$$

This also shows that the sequence $(L-f)(x(k)) / \|x(k) - \overline{x}\|^2$ is bounded. Since $L(\overline{x}) = f(\overline{x})$, it follows that

$$0 = \lim_{k\to\infty} \frac{(L-f)(x(k))}{\|x(k) - \overline{x}\|} = \lim_{k\to\infty} \frac{(L-f)(x(k)) - (L-f)(\overline{x})}{\|x(k) - \overline{x}\|} = \nabla(L-f)(\overline{x}) \cdot \Delta x$$

by (3). This, together with $\Delta x^{\mathrm{T}} \mathrm{D}^2 L(\overline{x}) \Delta x \geq 0$, contradicts the assumption. ■

[44] Optimality of $\overline{x}$ for L on C would imply its optimality for f on C, from $f(\overline{x}) = L(\overline{x}) \geq L(x) \geq f(x)$.

[45] This is in [6, Theorem 4.6.4], and implicitly in [5, Proof of Theorem 1.10.3].

Proof of Theorem 28 (SOCs as Multiplier Rules, with inequality constraints). Here

$$C = \{x : h(x) = 0, g(x) \leq 0\}.$$

With $\overline{\mu}$ and $\overline{\lambda}$ fixed, abbreviate $L\left(\overline{\mu}, \overline{\lambda}, \cdot\right)$ to L; then $L = f$ on the set $C_{\mathrm{b}}\left(\overline{\lambda}\right)$ given by (30). Also, $\nabla L(\overline{x}) = 0$ by assumption.

For Part 1, since $\overline{x}$ is a (local) maximum point of f on C, it is *a fortiori* a maximum point of L on $C_{\mathrm{b}}\left(\overline{\lambda}\right)$. So Theorem 31 applies, with C_{b} in place of C. This gives the semi-definiteness (31) of $\mathrm{D}^2_{xx}L$ for every $\Delta x \in \mathrm{T}_{\overline{x}}C_{\mathrm{b}}\left(\overline{\lambda}\right)$, which equals $\mathrm{L}_{\overline{x}}\left(\left(h, g_{\mathrm{B}\left(\overline{\lambda}\right)}\right), g_{\setminus \mathrm{B}\left(\overline{\lambda}\right)}\right)$ by the regularity assumption.

For Part 2, note first that $L \geq f$ on C and $L(\overline{x}) = f(\overline{x})$. So, to apply Theorem 32, one needs only to verify the definiteness (35) of $\mathrm{D}^2_{xx}L$ for every nonzero $\Delta x \in \mathrm{T}_{\overline{x}}C$ such that $\nabla(L - f)(\overline{x}) \cdot \Delta x = 0$. Since $\overline{\lambda}_i = 0$ for $i \notin \mathrm{B}\left(\overline{\lambda}\right)$, and since $\mathrm{D}h(\overline{x}) \cdot \Delta x = 0$,

$$\nabla(L - f)(\overline{x}) \cdot \Delta x = \sum_{i \in \mathrm{B}\left(\overline{\lambda}\right)} \overline{\lambda}_i \nabla g_i(\overline{x}) \cdot \Delta x.$$

In other words, it suffices to verify (35) for every nonzero Δx in the cone

$$\mathrm{T}_{\overline{x}}C \cap \bigcap_{i \in \mathrm{B}\left(\overline{\lambda}\right)} \ker \nabla g_i(\overline{x}) = \mathrm{T}_{\overline{x}}C \cap \ker\left(\mathrm{D}g_{\mathrm{B}\left(\overline{\lambda}\right)}(\overline{x}) \cdot\right).$$

Since, by Lemma 9, this is contained in

$$\mathrm{L}_{\overline{x}}(h, g) \cap \ker\left(\mathrm{D}g_{\mathrm{B}\left(\overline{\lambda}\right)}(\overline{x}) \cdot\right) = \mathrm{L}_{\overline{x}}\left(\left(h, g_{\mathrm{B}\left(\overline{\lambda}\right)}\right), g_{\setminus \mathrm{B}\left(\overline{\lambda}\right)}\right) \tag{53}$$

the requirement is met by the assumption for Part 2 (of Theorem 28)—which is exactly that (35) holds for every nonzero Δx in the last cone in (53).[46] So Theorem 32 applies; this yields the inequality (36), which shows that $\overline{x}$ is a strict maximum point. ■

11 Remarks on image linearisation as another approach

The above derivation of the FOCs consists in linearising the objective f and the inequality constraints g as functions of the decision variables, x (assuming for simplicity that there are no equality constraints). For a *small* increment Δx to the point in question, $\overline{x}$, this guarantees a close approximation to the functions' values and thus, for regular

[46] Under the regularity assumption of Part 1, $\mathrm{L}_{\overline{x}}\left((h, g_{\mathrm{B}}), g_{\setminus \mathrm{B}}\right) = \mathrm{T}_{\overline{x}}C_{\mathrm{b}}$. This is not needed for Part 2, though.

constraints, an adequate approximation of the original problem by optimisation of the linear functional $p\cdot = \nabla f(\overline{x}) \cdot$ over the cone $\mathrm{L}_{\overline{x}}(g)$, which is equal to the tangent at $\overline{x}$ to the constraint set $C = \{g \leq 0\}$. When $\overline{x}$ is a corner point of C (i.e., $\mathrm{L}_{\overline{x}}$ is a pointed cone), this actually gives a *sufficient* FOC: if $\overline{x}$ is a *unique* optimum for the linearised problem, then it is also a local optimum for the original problem (Theorems 26 and 30). Of more importance is the corresponding *necessary* FOC: if $\overline{x}$ is any local optimum for the original problem, then it is also an optimum for the linearised problem (Theorem 29). This is converted into a multiplier rule (Theorem 25) by applying Farkas' Lemma, which can be proved by separating a point from a closed convex cone with a hyperplane. The separation argument has two versions dual to each other. In the First Proof of Lemma 34, the point $p = \nabla f(\overline{x})$ is separated from the relevant cone by means of a $v \in \mathbb{R}^n$, i.e., a vector of decision variables with the interpretation of an increment to $\overline{x}$.[47] In the Second Proof of Lemma 34, an orthant with its vertex at $(1, 0) \in \mathbb{R} \times \mathbb{R}^l$, where l is the number of constraints, is separated from the range of the *linearised* objective and constraint functions, i.e., from the image of $\mathbb{R}^n$ under the map $v \mapsto (p \cdot v, Av)$ with $p = \nabla f(\overline{x})$ and $A = \mathrm{D}g(\overline{x})$. This time, the sets being separated lie in the space of values of the objective and constraint functions, and they are separated by means of a vector $\left(-1, \overline{\lambda}\right)$ that gives the multipliers $\overline{\lambda} \in \mathbb{R}^l$.

The latter separation argument can also be set up by linearising only *after* mapping the variables space (here, $\mathbb{R}^n$ or a subset D) to the values space (here, $\mathbb{R}^{1+l}$) by means of the nonlinear objective and constraint functions (f, g). The image set $\{(f(x), g(x)) : x \in D\}$ is then linearised, around a point $(f(\overline{x}), g(\overline{x}))$, by taking either the whole tangent cone, if it is convex, or a suitably chosen convex subcone of the tangent: see [5, Chapter 4] or [6, Chapter 6].[48] By dealing primarily with the functions' values rather than their variables, this method allows any variation of $\overline{x}$, say to $\widetilde{x}(\epsilon)$, that results in small changes of the relevant values (f and g). This is particularly useful when the decision variables form an infinite-dimensional space X, since it allows also the use of a nondirectional variation $\overline{x} - x(\epsilon)$ that may converge to zero in some topology on X, as $\epsilon \searrow 0$, but not necessarily from any particular direction in the specified space X. When X is a space of absolutely continuous functions, one well-known example is Weierstrass' needle-like variation. Used originally to derive his necessary condition for a strong extremum, the method was later adapted for a derivation of Pontriagin's Maximum Principle. Thus the image approach provides a unifying framework for the classical calculus of variations and the more recent optimal control theory: see [5, Chapters 4 ff.].

Another approach to multiplier rules consists in convexifying the Lagrangian by augmenting it with quadratic penalty terms. This leads to the principle of constraint removal and to an algorithm using both penalties and multipliers: see [6, Chapter 5].

[47] The separated cone and point (p) lie in what is the dual parameter space when the linearised problem is regarded as the primal, in the duality framework of linear or convex programming.

[48] Even the case of a convex image set captures some nonconvex programmes: see [6, Example 6.6.1].

A Theorems of the Alternative

The linear algebra tool for converting the abstract FOC into a multiplier rule is the Factorisation Lemma with a "signed" extension known as Farkas' Lemma. For multi-objective optimisation,[49] there are "multi-vector" extensions of Farkas' Lemma, known as Tucker's and Motzkin's Lemmas. The latter is used here for a different purpose, viz., to show the equivalence of two forms of the Mangasarian-Fromovitz Constraint Qualification (Lemma 14).

The Factorisation Lemma can be stated as a criterion for a nonhomogeneous system of linear equations to have a solution—i.e., for a given vector p to be a linear combination of, say, the rows of a given matrix B (when the system's variables μ are arranged in a row to the left of B). For the system $p^{\mathrm{T}} = \mu^{\mathrm{T}} B$ to be soluble for μ, it is obviously necessary that the system $Bv = 0$ and $p^{\mathrm{T}} v \neq 0$ have no solution for v. The point is that this condition is also sufficient. Thus the lemma is a *theorem of the alternative*—stating that always either one system or the other has a solution (but obviously never both).

Lemma 33 (Factorisation) *Given a vector $p \in \mathbb{R}^n$ and an $m \times n$ (real) matrix B, exactly one of the following two systems has a solution: either*

$$p^{\mathrm{T}} = \mu^{\mathrm{T}} B \quad \text{for some } \mu \tag{54}$$

or

$$\begin{cases} p^{\mathrm{T}} v \neq 0 \\ Bv = 0 \end{cases} \quad \text{for some } v \tag{55}$$

(but not both (54) and (55)).

In other words, with $B_{e\bullet}$ denoting the e-th row of B, there exists a $\mu \in \mathbb{R}^m$ such that

$$p^{\mathrm{T}} = \sum_{e=1}^{m} \mu_e B_{e\bullet} \tag{56}$$

if (and only if)

$$\ker (B\cdot) \subseteq \ker \left(p^{\mathrm{T}}\cdot\right). \tag{57}$$

Comments:

1. In (57), the data are viewed as linear operations, $v \mapsto Bv$ and $v \mapsto p^{\mathrm{T}} v$. The inclusion between their kernels is not only necessary but also sufficient for (56) to be met by some μ. (In the Proof of Theorem 20), this is applied to the maximand's gradient ∇f as p and the constraints' Jacobian $\mathrm{D}h$ as B.)

[49] See, e.g., [10, (7.5.14)].

2. In the language of linear operations, whose composition corresponds to matrix multiplication, Lemma 33 means that, given a linear functional p: $\mathbb{R}^n \to \mathbb{R}$ and a linear map B: $\mathbb{R}^n \to \mathbb{R}^m$, the inclusion $\ker(B) \subseteq \ker(p)$ is sufficient (and obviously necessary) for p to factorise into the composition, $\mu \circ B$, of B and some linear functional μ: $\mathbb{R}^m \to \mathbb{R}$ (hence the lemma's name). The lemma extends to linear operations between any linear spaces; in particular, p may be vector-valued.[50]

Farkas' Lemma gives a similar criterion for a (nonhomogeneous) system of linear equations to have a *nonnegative* solution (for λ)—i.e., for a given vector p to be a linear combination with nonnegative coefficients of a given set of vectors, say the rows of a given matrix A. For the system $p^{\mathrm{T}} = \lambda^{\mathrm{T}} A$ to be soluble for λ, it is obviously necessary that the system of homogeneous linear inequalities $Av \leq 0$ and $p^{\mathrm{T}} v > 0$ have no solution for v. The point is that this condition is also sufficient. Lemma 33 can be deduced from this by rewriting the equality $Bv = 0$ as a pair of opposite inequalities ($Bv \leq 0$ and $Bv \geq 0$). Indeed, as stated next, Farkas' Lemma contains the Factorisation Lemma.

Lemma 34 (Farkas' Alternative) *Given a vector $p \in \mathbb{R}^n$, an $m \times n$ matrix B and an $l \times n$ matrix A (both real), exactly one of the following two systems has a solution: either*

$$p^{\mathrm{T}} = \mu^{\mathrm{T}} B + \lambda^{\mathrm{T}} A \quad \text{for some } \mu \text{ and } \lambda \geq 0 \tag{58}$$

or

$$\begin{cases} p^{\mathrm{T}} v > 0 \\ Bv = 0 \qquad \text{for some } v \\ Av \leq 0 \end{cases} \tag{59}$$

(but not both (58) and (59)).[51]

In other words, , with $A_{i\bullet}$ denoting the i-th row of A, there exists a $\mu \in \mathbb{R}^m$ and a nonnegative $\lambda \in \mathbb{R}^l_+$ such that

$$p^{\mathrm{T}} = \sum_{e=1}^{m} \mu_e B_{e\bullet} + \sum_{i=1}^{l} \lambda_i A_{i\bullet} \tag{60}$$

if (and only if) for every $v \in \mathbb{R}^n$

$$(Bv = 0 \text{ and } Av \leq 0) \Rightarrow p \cdot v \leq 0. \tag{61}$$

First Proof. To start with, one can assume that $m = 0$, i.e., one can omit B (replacing $\mu^{\mathrm{T}} B$ and Bv by zeros in (58) and (59)). Since both systems cannot be

[50] Also, the Factorisation Lemma has its counterparts for other algebraic structures—groups, rings, etc. It is also known as the Homomorphism Theorem or Sard's Quotient Theorem.

[51] This is in [5, Lemma 1.5.3], [6, Theorem 4.3.4], [9, pp. 11, 68–69, 115, 141] and [10, 5.2.3].

soluble (simultaneously), it suffices to choose either system and show that its insolubility implies solubility of the other. With either choice, this can be done by separation with a hyperplane. If (59) is chosen to be insoluble, suitable subsets of $\mathbb{R}^{1+l}$ are separated. If (58) is chosen to be insoluble then p can be separated, by a $v \in \mathbb{R}^n$, from the range of $\cdot A$ on $\mathbb{R}^l_+$. The latter argument, detailed next, is slightly simpler.

By assumption, p does not lie in $\mathbb{R}^l_+ A$, the image of $\mathbb{R}^l_+$ under the linear map $\lambda \mapsto \lambda^{\mathrm{T}} A$. Since this is a finitely generated convex cone,[52] it is a closed set: : see, e.g., [5, Lemma 1.5.5] or [6, Theorem 4.3.2].[53] Therefore, p can be separated strongly from $\mathbb{R}^l_+ A$ by a hyperplane, i.e., there exists a $v \in \mathbb{R}^n$ and a scalar z such that

$$p^{\mathrm{T}} v > z \geq \lambda^{\mathrm{T}} A v \tag{62}$$

for every $\lambda \geq 0$. It follows that $z \geq 0$ (by setting $\lambda = 0$). Furthermore, z can be chosen to be 0 (since if $\lambda^{\mathrm{T}} A v$ were positive, it could be made arbitrarily large by scaling λ up, and so it could not be bounded from above). So $\lambda^{\mathrm{T}} A v \leq 0$ for every $\lambda \geq 0$, i.e., $Av \leq 0$. And $p^{\mathrm{T}} v > 0$, as required.

The case of $m > 0$ is reduced to the case of $m = 0$ (with $2m + l$ instead of l) by rewriting $Bv = 0$ as $Bv \leq 0$ and $-Bv \leq 0$. This is because the existence of $\lambda' \geq 0$, $\lambda'' \geq 0$ and $\lambda \geq 0$ such that $p = (\lambda' - \lambda'')^{\mathrm{T}} B + \lambda^{\mathrm{T}} A$ is equivalent to (58), by setting $\mu = \lambda' - \lambda''$. ■

Comments:

1. Farkas' Lemma contains the Factorisation Lemma because if a linear functional has a semidefinite sign on a linear space then it actually vanishes on it. Here, this means that the condition $p \cdot v \leq 0$ for every $v \in \ker(B\cdot)$ is actually equivalent to the apparently stronger condition $p \cdot v = 0$ for $v \in \ker(B\cdot)$.

   Second Proof of Lemma 34. This time, it is (59) that is assumed to be insoluble, and solubility of (58) is to be deduced. As before, one can set $m = 0$.

   By assumption, the image of $\mathbb{R}^n$ under the linear map $v \mapsto \left(p^{\mathrm{T}} v, Av\right) \in \mathbb{R}^{1+l}$ is disjoint from the set

   $$\mathbb{R}_{++} \times \mathbb{R}^l_- := \{(w, -u) : w > 0, u \geq 0\}.$$

   This means that the image space $\left(p^{\mathrm{T}}, A\right)\mathbb{R}^n$ is disjoint from the closed orthant $(1, 0) + \left(\mathbb{R}_+ \times \mathbb{R}^l_-\right)$. Equivalently

   $$(1; 0, \ldots, 0) \notin \left(p^{\mathrm{T}}, A\right)\mathbb{R}^n + \left(\mathbb{R}_- \times \mathbb{R}^l_+\right).$$

[52] It is the convex hull of the cone generated by $\{\mathrm{I}_{i\bullet} A : i = 1, \ldots, l\}$, the images under $\cdot A$ of the coordinate unit vectors $\begin{bmatrix} 0 & \ldots & 1 & 0 & \ldots \end{bmatrix}$.

[53] What is more, a finitely generated convex cone is the same as a polyhedral cone (intersection of a finite number of half-spaces): see, e.g., [9, 4.6.1 and 4.6.2].

Being the (algebraic) sum of a linear space and an orthant, this set is a finitely generated convex cone, so it is closed: see, e.g., [5, Lemma 1.5.5] or [6, Theorem 4.3.2]. Therefore it can be separated strongly from $(1, 0)$ by a hyperplane perpendicular to a vector $(-\nu, \lambda)$, i.e., there exists a scalar ν, a $\lambda \in \mathbb{R}^l$ and a scalar z such that

$$-\nu < z \leq \left(-\nu p^{\mathrm{T}} + \lambda^{\mathrm{T}} A\right) v + (\nu w + \lambda \cdot u) \tag{63}$$

for every v, $u \geq 0$ and (scalar) $w \geq 0$. It follows that $z \leq 0$ (by setting $v = 0$, $w = 0$ and $u = 0$). Furthermore, z can be chosen to be 0 (since if the r.h.s. of the second inequality in (63) were negative, it could be made arbitrarily large in absolute value by scaling up v, w and u). It follows that $\lambda \geq 0$ (since $\lambda \cdot u \geq 0$ for every $u \geq 0$, by setting $v = 0$ and $w = 0$), and similarly $\nu \geq 0$ (by setting $v = 0$ and $u = 0$). Actually $\nu > 0$ (by the first inequality in (63)) and so, by scaling (ν, λ), one can set $\nu = 1$. Then, finally,

$$p^{\mathrm{T}} = \lambda^{\mathrm{T}} A$$

(because if not, then the term $\left(-p^{\mathrm{T}} + \lambda^{\mathrm{T}} A\right) v$ in (63) could be made negative and arbitrarily large in absolute value by a choice of v). ■

Lemma 35 (Motzkin's Alternative) *Given a $q \times n$ matrix C, an $m \times n$ matrix B and an $l \times n$ matrix A (all real), exactly one of the following two systems has a solution: either*

$$0 = \nu^{\mathrm{T}} C + \mu^{\mathrm{T}} B + \lambda^{\mathrm{T}} A \quad \text{for some } \nu > 0 \text{ and } \mu \text{ and } \lambda \geq 0 \tag{64}$$

or[54]

$$\begin{cases} Cv \ll 0 \\ Bv = 0 \qquad \text{for some } v \\ Av \leq 0 \end{cases} \tag{65}$$

(but not both (64) and (65)).[55]

First Proof. This can be proved like Farkas' Lemma (which it contains): as in the Proof of Lemma 34, it suffices to choose either system and show that its insolubility implies solubility of the other; and one can assume that $m = 0$ (replacing $\mu^{\mathrm{T}} B$ and Bv by zeros in (64) and (65)).

[54] The symbols $<$ and $\ll$ denote semistrict and strict vector inequalities, i.e., $<$ means "$\leq$ but $\neq$", whilst $\ll$ means strict inequality for each pair of entries. Also, it is usually assumed that the matrix C is nonempty, i.e., that $q \geq 1$ (in addition to $n \geq 1$). Formally, this is unnecessary because when $q = 0$, there is no $\nu > 0$ (so (64) is insoluble, whilst (65) has $v = 0$ as a solution).

[55] This is in [9, p. 135] and [10, 7.5.3]. In [10], Proof 3 needs a correction: instead of being scaled to have 1 as one of its entries, the y_1^* in [10, (7.5.6)] should be scaled to lie in a compact base for $\mathbb{R}^q_+$—such as the unit simplex in (66) here. This will make the set H in [10, (7.5.6)] convex and compact. As it stands, H has neither property, but both are needed for the strong separation argument.

Suppose that (64) is insoluble, i.e., the ranges of $\cdot C$ on $-\left(\mathbb{R}^q_+ \setminus \{0\}\right)$ and of $\cdot A$ on $\mathbb{R}^l_+$ are disjoint. Equivalently, $\mathbb{R}^l_+ A$ is disjoint from $-S^q_1 C$, where

$$S^q_1 := \left\{ \nu \in \mathbb{R}^q : \nu \geq 0, \sum_{j=1}^{n} \nu_j = 1 \right\} \tag{66}$$

is the unit simplex in $\mathbb{R}^q_+$ (i.e., the image of $\mathbb{R}^l_+$ under the linear map $\lambda \mapsto \lambda^{\mathrm{T}} A$ is disjoint from the image of $-S^q_1$ under the linear map $\nu \mapsto \nu^{\mathrm{T}} C$). Since S^q_1 is compact and convex, so is its linear image; and therefore $-S^q_1 C$ can be strongly separated, by a hyperplane, from the closed convex cone $\mathbb{R}^l_+ A$. In other words, there exists a $v \in \mathbb{R}^n$ and a scalar z such that

$$-\nu^{\mathrm{T}} C v > z \geq \lambda^{\mathrm{T}} A v \tag{67}$$

for every $\nu \in S^q_1$ and $\lambda \geq 0$. It follows that $z \geq 0$ (by setting $\lambda = 0$). Furthermore, z can be chosen to be 0 (since if $\lambda^{\mathrm{T}} A v$ were positive, it could be made arbitrarily large by scaling λ up). So $\lambda^{\mathrm{T}} A v \leq 0$ for every $\lambda \geq 0$, i.e., $Av \leq 0$.

Similarly $Cv \ll 0$ (i.e., Cv is strictly negative), since $\nu^{\mathrm{T}} C v < 0$ for every $\nu \in S^q_1$ (or, equivalently, for every $\nu > 0$, i.e., for semipositive v). ■

Comments:

1. Like Farkas' Lemma, Motzkin's Lemma can also be proved by separation in the other space, viz., $\mathbb{R}^{q+l}$.

   Second Proof of Lemma 35. This time, it is (65) that is assumed to be insoluble, and solubility of (64) is to be deduced. As before, one can set $m = 0$.

   By assumption, the image of $\mathbb{R}^n$ under the linear map $v \mapsto (Cv, Av) \in \mathbb{R}^{q+l}$ is disjoint from the set

   $$\mathbb{R}^q_{--} \times \mathbb{R}^l_{-} := -\left\{ (w, u) : w \gg 0, u \geq 0 \right\}.$$

   This means that the image space $(C, A)\,\mathbb{R}^n$ is disjoint from the closed orthant $(-1; 0) + \left(\mathbb{R}^q_+ \times \mathbb{R}^l_-\right)$. Equivalently

   $$(-1, \ldots, -1; 0, \ldots, 0) \notin (C, A)\,\mathbb{R}^n + \left(\mathbb{R}_+ \times \mathbb{R}^l_+\right).$$

   Therefore this closed convex cone can be separated strongly from $(-1; 0)$ by a hyperplane (in $\mathbb{R} \times \mathbb{R}^l$), i.e., there exists a $\nu \in \mathbb{R}^q$, a $\lambda \in \mathbb{R}^l$ and a scalar z such that

   $$-\sum_{k=1}^{q} \nu_k < z \leq \left(\nu^{\mathrm{T}} C + \lambda^{\mathrm{T}} A\right) v + (\nu \cdot w + \lambda \cdot u) \tag{68}$$

   for every v, $u \geq 0$ and $w \geq 0$. It follows that $z \leq 0$ (by setting $v = 0$, $w = 0$ and $u = 0$). Furthermore, z can be chosen to be 0 (since if the r.h.s. of the second

inequality in (68) were negative, it could be made arbitrarily large in absolute value by scaling up v, w and u). So $\lambda \geq 0$ (by setting $v = 0$ and $w = 0$), and similarly $\nu \geq 0$ (by setting $v = 0$ and $u = 0$). And $\nu \neq 0$ (by the first inequality in (68)); so $\nu > 0$. Finally,

$$0 = \nu^{\mathrm{T}} C + \lambda^{\mathrm{T}} A$$

(because if not, then the term $\left(\nu^{\mathrm{T}} C + \lambda^{\mathrm{T}} A\right) v$ in (68) could be made negative and arbitrarily large in absolute value by a choice of v). ■

2. The proof by separation in $\mathbb{R}^{q+l} = \mathbb{R}^q \times \mathbb{R}^l$ (with $m = 0$) can also be split into two stages, as in [10, p. 170, Proof 2]: first, a separation argument in $\mathbb{R}^q$ produces a ν, which is then be used to "scalarise" one dimension of C, i.e., to reduce Motzkin's Lemma to Farkas' Lemma with $p^{\mathrm{T}} = -\nu^{\mathrm{T}} C$.

3. Another theorem of the alternative, Tucker's, is obtained "by swapping the strictness and semi-strictness" in Motzkin's, i.e., by having $\nu \gg 0$ in (64) and $Cv < 0$ in (65)—which together imply that $\nu^{\mathrm{T}} C v < 0$, as before. See, e.g., [9] or [10, 7.5.7]. The two, Motzkin's and Tucker's, are combined in what is known as Slater's Alternative (which contains all those given here) : see, e.g., [9] or [10, 7.5.11].

A special cases of Motzkin's Lemma gives a criterion for a homogeneous system of linear equations to have a semipositive solution.

Lemma 36 (Gordan's Alternative) *Given a $q \times n$ (real) matrix C, exactly one of the following two systems has a solution: either*

$$0 = \nu^{\mathrm{T}} C \quad \text{for some } \nu > 0 \tag{69}$$

or

$$Cv \ll 0 \quad \text{for some } v \tag{70}$$

(but not both).[56]

Proof. Set $l = 0$ and $m = 0$ in Lemma 35 (i.e., take both A and B to be empty). ■

Definition 37 *A set of vectors is positively independent if none of its semipositive linear combinations equals zero (i.e., if a nonnegative combination vanishes, then all its coefficients are zeros).*

By Gordan's Lemma, a finite set of vectors $p(1)$, $p(2), \ldots, p(q)$ is positively independent if and only if there exists a v with $p(i) \cdot v < 0$ for each i.[57] Geometrically, a

[56] This is in [9, pp. 71 and 137] and [10, pp. 171–172].

[57] The rows of a matrix C are positively independent if and only if system (69) is insoluble for ν, i.e., if and only if system (70) has a solution v.

finite set of nonzero vectors is positively independent if and only if the convex cone they generate is pointed (i.e., line-free): see, e.g., [6, Exercise 4.3.13]. And a linear functional has a negative scalar product with each of a finite number of nonzero vectors if and only if it lies in the interior of the polar to the convex cone they generate. So, in geometric terms, Gordan's Lemma means that a finitely generated convex cone is pointed if and only if its polar is solid (i.e., has a nonempty interior).

Another special case of Motzkin's Lemma gives a similar criterion for a homogeneous system of linear equations to have a strictly positive solution.

Lemma 38 (Stiemke's Alternative) *Given a $m \times n$ (real) matrix B, exactly one of the following two systems has a solution: either*

$$0 < \mu^{\mathrm{T}} B \quad \text{for some } \mu \tag{71}$$

or

$$\begin{cases} v \gg 0 \\ Bv = 0 \end{cases} \quad \text{for some } v \tag{72}$$

(but not both).[58]

Proof. Set $q = n$ and $C = -\mathrm{I}\,(n)$ and an empty A (i.e., $l = 0$) in Lemma 35. ■

References

[1] Bigelow, J. H., and N. E. Shapiro (1974): "Implicit function theorems for mathematical programming and for systems of inequalities", *Mathematical Programming*, 6, 141–156.

[2] Debreu, G. (1952): "Definite and semidefinite quadratic forms", *Econometrica*, 20, 295–300.

[3] El-Hodiri, M. A. (1971): *Constrained extrema: introduction to the differentiable case with economic applications*. (Lecture Notes in Operations Research and Mathematical Systems, vol. 56). Berlin-Heidelberg-New York: Springer.

[4] Fiacco, A. V. (1983): *Introduction to sensitivity and stability analysis in nonlinear programming*. New York: Academic Press.

[5] Hestenes, M. R. (1966): *Calculus of variations and optimal control theory*. New York-London-Sydney: Wiley.

[58] This is in [6, Theorem 4.3.5] and, as a case of Tucker's Lemma, in [9, pp. 138] and [10, p. 172].

[6] Hestenes, M. R. (1975): *Optimization theory: the finite-dimensional case*. New York-London-Sydney: Wiley.

[7] Horsley, A., and A. J. Wrobel (2003): "Liminal inequality constraints and second-order optimality conditions", CDAM Research Report LSE-CDAM-2003-16.

[8] Jittorntrum, A. V. (1984): "Solution point differentiability without strict complementarity in nonlinear programming", *Mathematical Programming Study*, 21, 127–138.

[9] Panik, M. J. (1993): *Fundamentals of convex analysis*. Dordrecht-Boston-London: Kluwer.

[10] Ponstein, J. (1980): *Approaches to the theory of optimisation*. Cambridge-New York-Melbourne: Cambridge University Press.

[11] Silberberg, E. (1971): "The Le Chatelier Principle as a corollary to a generalized envelope theorem", *Journal of Economic Theory*, 3, 146–155.

[12] Simon, C., and L. Blume (1994): *Mathematics for economists*. New York-London: Norton.

[13] Takayama, A. (1985): *Mathematical economics*. Cambridge-London-New York: Cambridge University Press.