A Comprehensive View on Optimization: Reasonable Descent

Jan Brinkhuis*

10 June 2005

Econometric Institute Report EI 2005-23

Abstract

Reasonable descent is a novel, transparent approach to a well-established field: the deep methods and applications of the complete analysis of continuous optimization problems. *Standard* reasonable descents give a unified approach to all standard necessary conditions, including *the Lagrange multiplier rule*, *the Karush-Kuhn-Tucker conditions* and *the second order conditions*. *Nonstandard* reasonable descents lead to new necessary conditions. These can be used to give surprising proofs of deep central results outside optimization: *the fundamental theorem of algebra*, *the maximum and the minimum principle of complex function theory*, *the separation theorems for convex sets*, *the orthogonal diagonalization of symmetric matrices* and *the implicit function theorem*. These optimization proofs compare favorably with the usual proofs and are all based on the same strategy. This paper is addressed to all practitioners of optimization methods from many fields who are interested in fully understanding the foundations of these methods and of the central results above.

Keywords: optimization, descent, necessary conditions, Lagrange multiplier, Karush-Kuhn-Tucker, fundamental theorem of algebra, orthogonal diagonalization, implicit function theorem.

AMS subject classification: 90C46, 65K10

OR/MS subject classification: Mathematics.

*Econometric Institute, Erasmus University.

1 Introduction

The idea of this paper is to develop the following geometrical view on the complete analysis of finite-dimensional minimization (and so maximization) problems with continuous variables. The name of the game of this analysis is *to find for as many admissible points P as possible a small continuous curve of admissible points, originating in P and having everywhere outside P a lower value of the objective function than at P*, to be called a *reasonable descent*. These points P cannot be local minima, clearly, and so need not be considered. By a comparison of the—few— remaining admissible points, one finds the global minimum—or minima. To be rigorous, one has to establish the existence of a global minimum in advance, always by means of—a variant of—the theorem of Weierstrass. This theorem states that a continuous function on a non-empty closed and bounded set in $\Re^n$ assumes its global minimum. A useful variant states that a continuous function that is *coercive*—this means that outside a suitable closed and bounded set all values taken are higher than at some point of this set—assumes its minimum. This is for example the case for a continuous function $f : \Re^n \to \Re$ for which $\lim_{|x|\to+\infty} f(x) = +\infty$ (where $|\cdot|$ denotes the *modulus* or *euclidian norm*). In the present paper we will construct standard and non-standard reasonable descents. The former give the standard necessary conditions such as the Fermat theorem, the Lagrange multiplier rule, the Karush-Kuhn-Tucker theorem, and the second order conditions (involving the definiteness of hessians). The latter give *nonstandard* necessary conditions, which can be used to prove the following famous deep results: the fundamental theorem of algebra, the maximum and the minimum principle from complex function theory, the orthogonal diagonalization of symmetric matrices, the implicit function theorem and the separation theorem for convex sets. It might be a surprise that these theorems can be proved at all using optimization methods; in fact optimization leads to proofs that are simpler than the usual ones, and are all based on one and the same strategy: reasonable descent.

Many optimization problems have been solved in the recent and distant past with great effort by some of the greatest scientists. The advantage of the reasonable descent view is its user-friendliness: it allows a wide circle of practitioners to understand fully the methods of optimization and to use these methods to solve optimization problems in a routine way. Thus the high art of great experts is turned into a basic craft.

The reasonable descent approach has a constructive, algorithmic character, and will hopefully be of help in the numerical solution of problems; for example the reasonable descent approach of the theorem on symmetric matrices is closely related to the Jacobi methods; these are efficient algorithms for finding an orthonormal basis of eigenvectors for a symmetric matrix.

The scope of the reasonable descent approach extends to dynamic optimization, that is, to the Calculus of Variations and the Optimal Control. Moreover it would be of interest to discover new necessary conditions for optimality by means of the construction of new reasonable descents.

2 Reasonable descent.

Formally we define the concept reasonable descent as follows. For a set $A \subseteq \Re^n$ and a point $\bar{x} \in A$ a *variation* for a point $\bar{x}$ in the set A is defined to be a curve x_α, $\alpha \geq 0$ in the set A with endpoint $\bar{x} = x_0$ that is continuous at $\alpha = 0$. For a set $A \subseteq \Re^n$ and a function $f : A \to \Re$, we consider the problem to minimize f on A, to be denoted as

$$f(x) \to \min, \ x \in A,$$

and define a *reasonable descent* for an admissible point $\bar{x} \in A$ to be a variation x_α, $\alpha \geq 0$ for $\bar{x}$ inside the set of admissible points A, for which moreover, writing $g(\alpha) = f(x_\alpha)$, the inequality $g(\alpha) < g(0)$ holds for all sufficiently small $\alpha > 0$. Often, but not always, we will show that a concrete variation is a reasonable descent by checking that either g is differentiable at 0 and $g'(0) < 0$, or g is twice differentiable at 0, $g'(0) = 0$ and $g''(0) < 0$. The concept reasonable descent is useful, as reasonable descents are often easy to construct explicitly for almost all points $\bar{x} \in A$, and then these points cannot be local minima, let alone global minima. Therefore we can delete these points from the list of 'suspects' (of the 'crime' of being a global minimum) or 'candidate solutions', which is initially the whole of A, the set of all admissible points. Thus the construction of standard or nonstandard reasonable descents is the most spectacular part of the solution process of an optimization problem. Usually it brings us from a state of complete ignorance about a concrete optimization problem to a situation where we have one candidate solution, which is almost certainly the true solution, and then the Weierstrass theorem leads to complete certainty. For each type of reasonable descent one gets a necessary condition for local optimality: inclusion in the complement of the set of those admissible points for which a reasonable descent has been constructed.

3 Applications of standard reasonable descents without constraints.

In this section we illustrate the concept reasonable descent by means of the following two fundamental results of optimization theory: the first and second order necessary conditions for unconstrained problems. These results are applied in the complete analysis of all concrete unconstrained optimization problems. We emphasize that the core of these results is the construction of the following two types of reasonable descent.

1. If the derivative of a function of n variables at a point $\bar{x}$ is nonzero, then going in a straight line from $\bar{x}$ in the direction opposite to that of the gradient gives a reasonable descent (it is in fact the *steepest descent*). This is the essence of the first order conditions for unconstrained problems.

2. If the derivative of a function of n variables f at a point $\bar{x}$ is zero and the *second derivative* or *hessian matrix* is not positive semidefinite, that is, $h^T f''(\bar{x})h < 0$ for some $h \in \Re^n$, then going in a straight line from $\bar{x}$ in the direction determined by the vector h gives a reasonable descent. This is the essence of the second order conditions for unconstrained problems.

If f is a function defined in a neighborhood of a point $\widehat{x} \in \Re^n$ and differentiable at $\widehat{x}$, we write $f \in D^1(\widehat{x})$. If f is moreover twice differentiable at $\widehat{x}$, we write $f \in D^2(\widehat{x})$.

Theorem. ***First and second order conditions for unconstrained problems.***

1. *Let $\widehat{x} \in \Re^n$ and $f \in D^1(\widehat{x})$. If f is a local minimum at $\widehat{x}$, then $f'(\widehat{x})$ is zero.*
2. *Let $\widehat{x} \in \Re^n$ and $f \in D^2(\widehat{x})$. If f is a local minimum at $\widehat{x}$, then $f'(\widehat{x})$ is zero and $f''(\widehat{x})$ is positive semidefinite.*

The novelty of the proofs below is their presentation in terms of reasonable descent; logically these proofs are equivalent to the usual ones.

Proof.

1. Consider any $\bar{x} \in \Re^n$ and $f \in D^1(\bar{x})$ for which $f'(\bar{x})$ is nonzero; we will construct a reasonable descent for $\bar{x}$. We define the variation $x_\alpha = \bar{x} - \alpha f'(\bar{x})^T,\ \alpha \geq 0$ (where T = transpose). We check that this is a reasonable descent: we write $g(\alpha) = f(x_\alpha)$ and calculate
$$g'(0) = -f'(\bar{x})f'(\bar{x})^T = -|f'(\bar{x})|^2 < 0,$$
where $|\cdot|$ denotes the *modulus* or *euclidian norm* on $\Re^n$.
2. Consider any $\bar{x} \in \Re^n$ and $f \in D^2(\bar{x})$ for which $f'(\bar{x})$ is zero and $f''(\bar{x})$ is not positive semidefinite; we will construct a reasonable descent for $\bar{x}$. Choose a vector $h \in \Re^n$ for which $h^T f''(\bar{x})h < 0$. We define the variation $x_\alpha = \bar{x} + \alpha h,\ \alpha \geq 0$. We check that this is a reasonable descent: we write $g(\alpha) = f(x_\alpha)$ and calculate $g'(0) = f'(\bar{x})h = 0$ and $g''(0) = h^T f''(\bar{x})h < 0$.

One can go on to construct reasonable descents that give higher order conditions. We only give the one variable case, as this has applications. If f is a function defined in a neighborhood of $\widehat{x} \in \Re$ and k times differentiable at $\widehat{x}$, we write $f \in D^k(\widehat{x})$.

Theorem. ***Higher order conditions (one variable case).*** *Let $\widehat{x} \in \Re$ and $f \in D^k(\widehat{x})$. If f is a local minimum at $\widehat{x}$ and k is the smallest positive integer for which $f^{(k)}(\widehat{x})$ is nonzero, then k is even and $f^{(k)}(\widehat{x})$ is positive.*

Proof. Let $\bar{x} \in \Re$ and $f \in D^k(\bar{x})$, and assume that k is the smallest positive integer for which $f^{(k)}(\bar{x})$ is nonzero. To prove the theorem, it suffices to construct reasonable descents in the following two cases.

1. k is odd and $f^{(k)}(\bar{x})$ is positive. We define the variation $x_\alpha = \bar{x} - \alpha,\ \alpha \geq 0$. We check that this is a reasonable descent for $\bar{x}$: we write $g(\alpha) = f(x_\alpha)$ and calculate the k-th order Taylor approximation at $\alpha = 0$

$$g(\alpha) = g(0) + \frac{1}{k!} g^{(k)}(0)\alpha^k + o(\alpha^k)$$

—where o is small Landau-o—and the higher derivative

$$g^{(k)}(0) = (-1)^k f^{(k)}(\bar{x}) < 0.$$

It follows that the variation $x_\alpha,\ \alpha \geq 0$ is a reasonable descent.

2. $f^{(k)}(\bar{x})$ is negative. We define the variation $x_\alpha = \bar{x}+\alpha,\ \alpha \geq 0$. We check that this is a reasonable descent for $\bar{x}$: we write $g(\alpha) = f(x_\alpha)$ and calculate the k-th order Taylor approximation at $\alpha = 0$

$$g(\alpha) = g(0) + \frac{1}{k!} g^{(k)}(0)\alpha^k + o(\alpha^k)$$

and the higher derivative $g^{(k)}(0) = f^{(k)}(\bar{x}) < 0$. It follows that the variation $x_\alpha,\ \alpha \geq 0$ is a reasonable descent.

4 Nonstandard reasonable descents and central theorems

Now we will formulate and prove some well-known results. Each one of these is the central result of an entire field. The usual proofs are complicated. None of these results seems to have anything to do with optimization. However we will give straightforward proofs using optimization methods: in each case the heart of the matter is the construction of a suitable type of *non-standard* reasonable descent for a suitable auxiliary optimization problem, giving a *nonstandard* necessary condition for this problem.

Fundamental theorem of algebra. *Each polynomial equation*

$$p(x) = a_0 + \cdots + a_n x^n = 0$$

in one complex variable x—of degree $n \geq 1$ has at least one solution.

This is the central property of the complex numbers. The proofs that are usually given are based on relatively advanced methods such as Galois theory (cf [9]) or complex function theory (cf [10]).

The proof to be given here is based on an idea of d'Alembert. The auxiliary optimization problem is the minimization of the modulus of $p(x)$. We have simplified the recent version of this proof (cf [1]) by replacing the intricate use of trigonometric functions by the easy fact that for each complex number a the equation $x^k - a = 0$ has a solution.

Proof.

1. We consider the problem
$$f(x) = |p(x)| \to \min, \ x \in \mathbf{C}.$$
This two-variable problem ($x = x_1 + ix_2$ with $x_1, x_2 \in \Re$) has a solution $\widehat{x}$, as the function $x \to |p(x)|$ is coercive:
$$|p(x)| = |x|^n |a_n + \cdots + a_0 x^{-n}|,$$
the first factor tends to $+\infty$ and the second one to $|a_n|$ for $|x| \to +\infty$.

2. Consider any complex number $\tilde{x}$ for which $p(\tilde{x}) \neq 0$; we will construct a reasonable descent for $\tilde{x}$. We assume that $\tilde{x} = 0$ as we may without restricting the generality of the argument. Then $a_0 = p(0) \neq 0$. Let k be the smallest positive integer for which $a_k \neq 0$. Choose a solution $u \in \mathbf{C}$ of the equation $a_0 + a_k x^k = 0$: that is, u is one of the k k-th roots of the number $-\frac{a_0}{a_k}$. Define the variation $x_\alpha = \alpha u$, $\alpha \geq 0$. We check that this is a reasonable descent: we write $g(\alpha) = f(x_\alpha)$ and calculate
$$g(\alpha) = |a_0 + a_k(\alpha u)^k + o(\alpha^k)| = |(1 - \alpha^k)a_0 + o(\alpha^k)| < |a_0| = g(0)$$
for $\alpha > 0$ sufficiently small.

3. Deleting from the list of suspects the complex numbers for which we have constructed a reasonable descent, gives that all local solutions of the optimization problem are solutions of the polynomial equation $p(x) = 0$.

4. The theorem is proved, as $p(\widehat{x}) = 0$.

Note that we have constructed a type of reasonable descent and that this gives the nonstandard necessary condition '$p(x) = 0$' for a local solution x of the optimization problem. This reasonable descent can be turned into an algorithm for solving polynomial equations. We note moreover that this reasonable descent is for $k = 1$ (resp. $k = 2$, resp. $k \geq 2$) related to the reasonable descents used in the previous section to establish the first (resp. second, resp. higher) order conditions.

In a similar way one can prove the following more general result; in the proof the role of polynomials is taken over by power series.

Theorem. ***Minimum Principle from complex function theory.*** *The modulus of an analytical function of a complex variable can only assume its minimal value at an interior point if the value of the function at this interior point is zero.*

From this result one can derive the following one by going over from a function $z \mapsto f(z)$ to the function $z \mapsto f(z)^{-1}$.

Theorem. ***Maximum Principle from complex function theory.*** *The modulus of an analytical function of a complex variable on the closure of a bounded open set assumes its maximum value only at boundary points of its domain.*

The minimum and the maximum principle are central results of complex function theory.

For a convex set $C \subseteq \Re^n$ the *relative interior* of C is defined to be the interior of C inside the smallest affine subspace of $\Re^n$ containing C.

Theorem. ***Separation of convex sets.*** *Two convex sets C and D in $\Re^n$ can be separated by a hyperplane if their relative interiors have no point in common.*

This is the central property of *convex analysis* (the subject dealing with convex sets, convex functions and convex minimization problems). The required reasonable descent will be based on the following geometrical observation: if, for two different points P and Q in a two-dimensional plane, we move Q in a direction making an acute angle with the original interval PQ, then the distance to P decreases initially. The auxiliary optimization problem is the shortest distance problem for a closed nonempty convex set C and a point $p \notin C$. The novelty of the proof below is the presentation in terms of reasonable descent; logically this proof is equivalent to the usual one, which is also based on this shortest distance problem.

Proof. We only give the proof in the following special case: C is closed, D consists of one point p only and p is not contained in C. The general case can be derived from this special case.

1. We consider the problem

   $$f(x) = |x - p|^2 \to \min, \ x \in C,$$

   that is, the *shortest distance problem* for p and C. It has a global solution $\widehat{x}$: this follows from the Weierstrass theorem after adding the constraint $f(x) \leq f(x_0)$ for some $x_0 \in C$ (this does not change the solution set and makes the admissible set bounded).

2. Consider any $\bar{x} \in C$ that has the following property: the hyperplane

   $$\mathcal{H}_{\bar{x}} = \{x : \langle p - \bar{x}, x - \bar{x} \rangle = 0\}$$

—where $\langle\, , \rangle$ denotes the standard inner product on $\Re^n$—through the point $\bar{x}$ and orthogonal to the line through $\bar{x}$ and p, does *not* separate C from p; we will construct a reasonable descent for $\bar{x}$. We choose a point $x_0 \in C$ that lies strictly on the same side of the hyperplane $\mathcal{H}_{\bar{x}}$ as p. That is,

$$\langle p - \bar{x}, x_0 - \bar{x} \rangle > 0. \qquad (*)$$

Define the curve $x_\alpha = (1-\alpha)\bar{x} + \alpha x_0$, $\alpha \geq 0$, that is, the straight line connecting $\bar{x}$ and x_0. We check that this is a reasonable descent for $\bar{x}$: we write $g(\alpha) = f(x_\alpha)$ and calculate $g'(0) = 2\langle \bar{x} - p, x_0 - \bar{x} \rangle < 0$ using $(*)$.

3. Deleting from the list of suspects all points in C for which we have constructed a reasonable descent, gives that all local solutions of the problem separate C from p.

4. The theorem is proved: the hyperplane $\mathcal{H}_{\widehat{x}}$ separates p from C.

Note that we have constructed a reasonable descent and that it gives the nonstandard necessary condition '*$\mathcal{H}_x$ separates p from C*' for a local solution x.

Theorem. ***Orthogonal diagonalization of symmetric matrices.*** *For each symmetric $n \times n$ matrix A there exists an orthogonal $n \times n$-matrix P such that $P^T AP$ is a diagonal matrix.*

This is the central property of symmetric matrices (an alternative formulation of this result is that each symmetric matrix has an orthonormal basis of eigenvectors). We will give an optimization formulation of this theorem and then we will construct a type of reasonable descent inside the set of orthogonal $n \times n$-matrices. The descent curve will be constructed by multiplication of a given orthogonal $n \times n$-matrix with suitable orthogonal matrices representing rotations in $\Re^n$ over small angles. This reasonable descent is closely related to the *Jacobi methods*; this is a class of efficient algorithms to calculate an orthogonal diagonalization of a given symmetric matrix (cf. [4]). The usual proof makes use of complex numbers(cf. [4]). There is an alternative proof that makes use of optimization methods (cf [1]): for the most convenient presentation of it one has to view a symmetric matrix as a linear operator on an abstract finite-dimensional vector space; it proceeds by induction with respect to the dimension of the vector space, each step involving an optimization problem. Below we present a *one-blow proof*, requiring only one optimization problem to be solved, and not making use of complex numbers. The auxiliary optimization problem is the problem to minimize the sum of the squares of the off-diagonal elements of the matrix that is the result of conjugating the given symmetric matrix with a variable orthogonal matrix.

Proof.

1. We consider the problem

$$f(P) = s(P^T AP) \to \min, \ P \text{ orthogonal},$$

   where we write $s(B)$ for the sum of the squares of the off-diagonal elements of a symmetric $n \times n$ matrix B. Note that $s(B) = 0$ precisely if B is a diagonal matrix. This problem has a global solution $\widehat{P}$: this follows from the Weierstrass theorem, as the orthogonal $n \times n$-matrices form a nonempty closed and bounded set in the space of $n \times n$-matrices (boundedness follows from the fact that all columns of an orthogonal matrix have unit length).

2. Consider any orthogonal $n \times n$-matrix $\bar{P}$ for which $C = \bar{P}^T A\bar{P}$ is not a diagonal matrix; we will construct a reasonable descent for $\bar{P}$. We choose an off-diagonal element c_{ij} of C that is nonzero. For each $\alpha \in \Re$ we let $Q(\alpha)$ denote the $n \times n$-matrix for which

$$q_{ii}(\alpha) = q_{jj}(\alpha) = \cos\alpha, \ q_{ji}(\alpha) = -q_{ij}(\alpha) = \sin\alpha, \ p_{kk} = 1 \ \forall k \neq i, j,$$

   and all other entries are zero; note that $Q(\alpha)$ is an orthogonal matrix: a counterclockwise rotation over angle α in the two-dimensional i, j-plane. Define the curve $P_\alpha = \bar{P}Q(\alpha), \ \alpha \in \Re$ of $\bar{P}$ inside the set of orthogonal $n \times n$-matrices. We write $g(\alpha) = f(\bar{P}P(\alpha))$ and calculate

$$g'(0) = 2c_{ij}(c_{ii} - c_{jj})$$

   and

$$g''(0) = (c_{ii} - c_{jj})^2 - 4c_{ij}^2.$$

   For later use we note the following consequence:

$$g'(0) = 0 \Rightarrow g''(0) < 0. \qquad (**)$$

   It follows that the variation $P_\alpha, \ \alpha \geq 0$ (resp. $P_{-\alpha}, \ \alpha \geq 0$) is a reasonable descent if $g'(0) \leq 0$—taking into account (**)—(resp. if $g'(0) > 0$).

3. Deleting from the list of suspects all orthogonal matrices $\bar{P}$ for which we have constructed a reasonable descent, gives that the matrix $P^T AP$ is diagonal for all local solutions P of the problem.

4. The theorem is proved: $\widehat{P}^T A\widehat{P}$ is a diagonal matrix.

Note that we have constructed reasonable descents and that these give the nonstandard necessary condition '*$P^T AP$ is a diagonal matrix*' for a local solution P of the optimization problem.

We need some definitions in order to formulate the next result. For a neighborhood $W \subseteq \Re^n$ of a point $w \in \Re^n$ and a mapping $F = (f_1, \ldots, f_m)^T : W \to \Re^m$, one says that F is *differentiable* at w with *derivative* or *Jacobi-matrix* $F'(w) \in M_{m\times n}$—$F \in D^1(w, \Re^m)$—if

$$\lim_{h\to 0} |h|^{-1}|F(w+h) - F(w) - F'(w)h| = 0.$$

A mapping $F \in D^1(0_n, \Re^m)$ is called *continuously differentiable* at 0_n—$F \in C^1(0_n, \Re^m)$—if F is differentiable in some neighborhood of 0_n and if moreover the derivative mapping F' is continuous at 0_n. Usually it is easy to check this property for concrete mappings F: it holds if all *partial derivatives* of the functions f_i, $1 \le i \le m$, exist in a neighborhood of 0_n and are continuous at 0_n; moreover, then $F'(0_n) = (\frac{\partial f_i(0_n)}{\partial x_j})_{i,j}$. We define for a mapping $F \in C^1(0_n, \Re^m)$ the remainder functions $r(x_1, x_2)$ and $s(x)$ implicitly by

$$F(x_1) - F(x_2) = F'(0_n)(x_1 - x_2) + r(x_1, x_2)$$

and

$$F'(x) = F'(0_n) + s(x).$$

The following properties hold true:

$$\lim_{x_1,\, x_2 \to 0_n} \frac{|r(x_1, x_2)|}{|x_1 - x_2|} = 0, \tag{1}$$

$$\lim_{x\to 0_n} \|s(x)\| = 0. \tag{2}$$

Property (2) expresses the continuity of F' at 0_n; property (1) is called the *strict differentiability* of F at 0_n. Property (1) follows readily from (2) and the following chain of inequalities, which hold for any $\varepsilon > 0$ and any convex neighborhood W of 0_n for which $\|s(x)\| \le \varepsilon \ \forall x \in W$:

$$|r(x_1, x_2)| = |\int_0^1 s(tx_1 + (1-t)x_2)(x_1 - x_2)dt|$$

$$\le |x_1 - x_2| \int_0^1 \|s(tx_1 + (1-t)x_2)\|dt \le \varepsilon |x_1 - x_2| \ \ \forall x_1, x_2 \in W. \tag{3}$$

Implicit function theorem. *Let $F \in C^1(0_n, \Re^m)$ be given and assume that $F(0_n) = 0_m$ and that the last m columns of $F'(0_n)$ are linearly independent. Then the vector equation $F(x) = 0_m$ determines in a suitable neighborhood of 0_n the last m variables $x_{n-m+1}, \ldots, x_n$ as functions of the remaining variables $x_1, \ldots, x_{n-m}$, say $z = \varphi(y)$, where $y = (x_1, \ldots, x_{n-m})$ and $z = (x_{n-m+1}, \ldots, x_n)$. The mapping φ is differentiable at 0_{n-m}, and*

$$\varphi'(0_{n-m}) = -F'_z(0_n)^{-1} F'_y(0_n).$$

This theorem is the central result of *differential calculus*. We will prove it by means of an analysis of the following optimization problem: minimize in a neighborhood of 0_n the modulus of $F(x_1,\ldots,x_n)$ for fixed $y=(x_1,\ldots,x_{n-m})$ and variable $z=(x_{n-m+1},\ldots,x_n)$. The usual proof of the implicit function makes use of the *contraction principle*. The optimization proof given below is essentially different.

Proof. To begin with, we give a reformulation of the continuity of F' at 0_n that is convenient for the present purpose. Let U be a neighborhood of 0_n on which F is defined and differentiable and on which the modulus of F' is bounded. For each $r>0$ for which U contains the 'block'

$$B_r=\{(y,z):|y|\le r,\ |z|\le(|F'_z(0_n)^{-1}F'_y(0_n)|+1)r\}$$

we let $\varepsilon(r)>0$ be the supremum of $|F'(x)-F'(0_n)|$, where x runs over B_r. Then $\lim_{r\downarrow 0}\varepsilon(r)=0$, as $F\in C^1(0_n,\Re^n)$.

1. For each y with

$$\varepsilon(|y|)<\frac{|F'_z(0_n)^{-1}|^{-1}}{2+|F'_z(0_n)^{-1}F'_y(0_n)|},\qquad (***)$$

   we consider the problem

$$f_y(z)=|F(y,z)|^2\to\min,\ |z|\le(|F'_z(0_n)^{-1}F'_y(0_n)|+1)|y|.$$

   This problem has a global solution $\widehat{z}=z(y)$ by the Weierstrass theorem.

2. Consider an interior point $\bar{z}$ of the admissible set for which $F(y,\bar{z})$ is nonzero; we will construct a reasonable descent for $\bar{z}$. Define the variation $z_\alpha=\bar{z}-\alpha f'(\bar{z})^T,\ \alpha\ge 0$ inside the admissible set. We check that this is a reasonable descent: we write $g(\alpha)=f(z_\alpha)$ and calculate

$$g'(0)=f'(\bar{z})(-f'(\bar{z})^T)=-|f'(\bar{z})|^2,$$

   so it suffices to verify that $f'(\bar{z})$ is nonzero. One has

$$f'(\bar{z})=2F(y,\bar{z})^TF'_z(y,\bar{z})$$

   and this shows that $f'(\bar{z})$ is nonzero, as $F(y,\bar{z})$ is nonzero by assumption and $F'_z(y,\bar{z})$ is invertible: this follows from the inequality

$$\|F'_z(y,\bar{z})F'_z(0_n)^{-1}-I_n\|\le\|F'_z((y,\bar{z}))-F'_z(0_n)\|\|F'_z(0)^{-1}\|<1.$$

3. Deleting from the list of suspects all points for which a reasonable descent has been constructed, it follows that $F(y,z)=0_m$ for each *interior* local solution z of the problem.

   Now we establish two additional properties of minima.

(a) For all global minima $\bar{x}$ the following inequality holds (this will give the differentiability property of the implicit function at $y = 0_{n-m}$):

$$|\bar{z} - (-F_z'(0_n)^{-1}F_y'(0_n)y)| \leq \frac{\varepsilon(|y|)(1 + \|F_z'(0_n)^{-1}F_y'(0_n)\|}{|F_z'(0_n)^{-1}|^{-1} - \varepsilon(|y|)}|y|.$$

We argue by contradiction: assume that this inequality does not hold.
Applying the inequality (3) twice, once with $x_1 = (y, \bar{z}),\ x_2 = (y, -F_z'(0_n)^{-1}F_y'(0_n)y)$ and once with $x_1 = (y, -F_z'(0_n)^{-1}F_y'(0_n)y),\ x_2 = 0_n$ leads to the inequality

$$f(\bar{z}) > f(-F_z'(0_n)^{-1}F_y'(0_n)y);$$

this shows that $(y, \bar{z})$ is not a global minimum.

(b) The problem has at most one local minimum. If $F(y, z_1) = F(y, z_2) = 0_m$, then by (3) we get $|F'(0_n)(x_1 - x_2)| \leq \varepsilon(|y|)|x_1 - x_2|$ and by $(***)$ this implies $z_1 = z_2$.

4. The theorem is proved and $\varphi(y)$ can be characterized as the unique global minimum $\widehat{z}(y)$ of the problem above.

Note that we have constructed a reasonable descent and that these gives the nonstandard necessary condition '*z is a solution of the vector equation $F(y, z) = 0_m$*' for a local interior solution z of the optimization problem. It is of interest that here the existence of a unique solution of a system of non-linear equations is proved without using strict convexity (which does in fact not necessarily hold here).

Now we formulate a result that is equivalent to the implicit function theorem, but is geometrically more intuitive; in the next section we will use it to prove the Lagrange multiplier rule. We need some definitions. If S is a subset of $\Re^n$ containing the origin and L is a linear subspace of $\Re^n$, then we say that L is the *tangent space of S at the origin* if there exists a neighborhood U of the origin such that the orthogonal projection π on L sends the intersection $S \cap U$ bijectively onto $L \cap U$, and $v = \pi(v) + o(|\pi(v)|)$ where v runs over $S \cap U$. The value of this definition is that a tangent space gives a parametrization and approximation of a complicated set S by means of a simpler set L near the origin: each point $v \in S$ near the origin is approximated by the point $\pi(v)$ near the origin, which is moreover very close to v, as $v = \pi(v) + o(|\pi(v)|)$. Warning: there are other—weaker—concepts of tangent spaces.

Tangent space theorem. *Let $F \in C^1(0_n, \Re^m)$ with $F(0_n) = 0_m$, and nondegenerate at 0_n in the sense that the Jacobi-matrix $F'(0_n)$ has full rank m. Then the linear subspace* $\ker F'(0_n) = \{x : F'(0_n)x = 0_m\}$ *is the tangent space at the origin to the set $\{x : F(x) = 0_m\}$.*

The tangent space theorem gives a parametrization and approximation of the solution set of a system of nonlinear equations $F(x) = 0_m$ in a neighborhood of a given nondegenerate solution 0_n in terms of the solution set of the system of linear equations $F'(0_n)x = 0_m$, which arises by linearizing at 0_n the given nonlinear system. Note that this result is essentially the most general result of this type: it implies the result with 0_n replaced by an arbitrary point $\widehat{x} \in \Re^n$, and 0_m by an arbitrary point $\widehat{y} \in \Re^m$ (by means of translations in $\Re^n$ and $\Re^m$).

5 Applications of standard reasonable descents with constraints.

In this section we present further illustrations of the construction of reasonable descents in order to establish first order necessary conditions for optimization problems with constraints. To begin with we consider the *Lagrange multiplier rule*—the first order conditions for equality constrained problems. The novelty of the proof is the presentation in terms of reasonable descent as well as the use of the tangent space theorem. This use leads to a more intuitive proof than the usual one, which proceeds by a mechanical verification, using the implicit function theorem. The core of the result is the construction of the following type of reasonable descent. Let continuously differentiable functions $f_i,\ 0 \le i \le m$, of n variables, and a solution $\bar{x}$ of the system of equations $f_i(x) = 0,\ 1 \le i \le m$, be given. Consider the problem

$$f_0(x) \to \min,\ f_i(x) = 0,\ 1 \le i \le m.$$

If there is no nontrivial linear relation between the derivatives of the functions $f_i,\ 0 \le i \le m$, at $\bar{x}$, then the following variation for $\bar{x}$ inside the admissible set is a reasonable descent : the curve $x_\alpha,\ \alpha \ge 0$, where x_α is the admissible point close to $\bar{x}$ that is mapped to the same point as $\bar{x} - \alpha f_0'(\bar{x})^T$ under the orthogonal projection onto the tangent space at $\bar{x}$ to the admissible set. This fact is the essence of the Lagrange multiplier rule in terms of reasonable descent.

Theorem. ***Lagrange multiplier rule.*** *Consider the problem with equality constraints*

$$f_0(x) \to \min,\ f_i(x) = 0,\ 1 \le i \le m,$$

where $f_0 \in D^1(\widehat{x}),\ f_i \in C^1(\widehat{x}, \Re),\ 1 \le i \le m$, *for some* $\widehat{x} \in \Re^n$. *Then for each local solution* $\widehat{x}$, *there is a nonzero selection of multipliers* $(\lambda_0, \ldots, \lambda_m)$ *for which* $\mathcal{L}'(\widehat{x}) = 0_n$ *is zero, where the Lagrange function* $\mathcal{L}$ *is defined by* $\mathcal{L}(x) = \sum_{i=0}^m \lambda_i f_i(x)$.

Proof. Consider any admissible point $\bar{x}$ for which the derivatives $f_i'(\bar{x}),\ 0 \le i \le m$, are linearly independent; we will construct a reasonable descent for $\bar{x}$. We assume that $\bar{x} = 0$, as we may without restricting the generality of the argument. We take the orthogonal projection v of $f_0'(\bar{x})$ on the linear subspace of vectors that are orthogonal to the vectors $f_i'(\bar{x}),\ 1 \leq i \leq m$. This projection v is a nonzero vector: by the linear independence assumption $f_0'(\bar{x})$ is not contained in the linear span of the vectors $f_i'(\widehat{x}),\ 1 \leq i \leq m$, and so the orthogonal projection v of $f_0'(\widehat{x})$ on the orthogonal complement of this linear span is nonzero. Now we apply the tangent space theorem to the mapping $(f_1, \ldots, f_m)^T \in C^1(0_n, \Re^m)$. This implies that there is a variation $x_\alpha,\ \alpha \geq 0$ with endpoint $x_0 = 0_n$ in the admissible set of the given optimization problem for which the orthogonal projection of x_α on the tangent space at $\bar{x} = 0_n$ to this admissible set is $-\alpha v$ and such that moreover $x_\alpha = -\alpha v + o(\alpha)$. We check that this is a reasonable descent: we write $g(\alpha) = f_0(x_\alpha)$ and calculate

$$g(\alpha) = f_0(x_\alpha) = f_0(0_n) - f_0'(0_n)v\alpha + o(\alpha) = g(0) - |v|^2\alpha + o(\alpha) < g(0),$$

for $\alpha > 0$ sufficiently small.

In addition we give for three types of mixed smooth-convex problems the necessary conditions. These results can also be established by the use of suitable reasonable descents, but we will only display the reasonable descent proof of the Karush-Kuhn-Tucker theorem. It is based on the separation theorem. The proofs of the other two results require a combination of the techniques used to prove the central results of differentiable and convex calculus: the implicit function theorem and the separation theorem.

For each closed, pointed, solid, convex cone C in $\Re^n$ we define an ordering $\succeq$ on $\Re^n$ by $x \succeq y$ precisely if $x - y \in C$; such an ordering will be called a *conic ordering*.

Theorem.

1. ***Theorem of Karush-Kuhn-Tucker.*** *Consider the convex problem with inequality constraints*

$$f_0(x) \to \min,\ f_i(x) \leq 0,\ 1 \leq i \leq m,\ x \in A,$$

where $A \subseteq \Re^n$ is a convex set, $f_i : A \to \Re,\ 1 \leq i \leq m$ are convex functions. Then each local solution $\widehat{x}$ is a global solution, and there is a nonzero selection of multipliers $(\lambda_0, \ldots, \lambda_m)$ such that the following conditions hold true:

- $\mathcal{L}(\widehat{x}, \lambda) = \min_{x \in A} \mathcal{L}(x, \lambda)$, *where the Lagrange function $\mathcal{L}$ is defined by $\mathcal{L}(x) = \sum_{i=0}^m \lambda_i f_i(x)$.*
- $\lambda_i \geq 0,\ 0 \leq i \leq m,$
- $\lambda_i f_i(\widehat{x}) = 0,\ 1 \leq i \leq m.$

2. ***Lagrange multiplier rule for problems with inequality constraints.*** *Consider the problem with inequality constraints*

$$f_0(x) \to \min,\ f_i(x) \leq 0,\ 1 \leq i \leq m,$$

where $f_0 \in D^1(\widehat{x}),\ f_i \in C^1(\widehat{x}, \Re),\ 1 \le i \le m$, *for some* $\widehat{x} \in \Re^n$. *Then for each local solution* $\widehat{x}$, *there is a nonzero selection of multipliers* $(\lambda_0, \ldots, \lambda_m)$ *such that the following conditions hold true:*

- $\mathcal{L}'(\widehat{x}) = 0_n$, *where the Lagrange function* $\mathcal{L}$ *is defined by* $\mathcal{L}(x) = \sum_{i=0}^m \lambda_i f_i(x)$.
- $\lambda_i \ge 0,\ 0 \le i \le m$,
- $\lambda_i f_i(\widehat{x}) = 0,\ 1 \le i \le m$.

3. ***Necessary conditions for conic problems.*** *Consider the conic problem*

$$f_0(x) \to \min,\ F(x) = 0_m,\ x \succeq 0_n,$$

where $f_0 \in D^1(\widehat{x}),\ F = (f_1, \ldots, f_m)^T \in C^1(\widehat{x}, \Re^m)$ *and* $\succeq$ *is a conic ordering on* $\Re^n$. *Then for each local solution* $\widehat{x}$, *there is a nonzero selection of multipliers* $(\lambda_0, \ldots, \lambda_m)$ *with* $\lambda_i \ge 0,\ 0 \le i \le m$ *such that the following condition holds true:*

$$\min_{h \succeq -\widehat{x}} \mathcal{L}'(\widehat{x})h = 0,$$

where the Lagrange function $\mathcal{L}$ *is defined by* $\mathcal{L}(x) = \sum_{i=0}^m \lambda_i f_i(x)$.

Proof of the Karush-Kuhn-Tucker theorem. Consider any admissible point $\bar{x}$ for which the KKT-conditions do not hold; we will construct a reasonable descent for $\bar{x}$. We assume that all constraints are tight at $\bar{x}$ as we may without restricting the generality of the argument (by omitting the constraints that are not tight at $\bar{x}$). Moreover we assume that $\bar{x} = 0_n$, also without restricting the generality of the argument. That the KKT-conditions do not hold is equivalent to the fact that the point 0_{m+1} and the set $C = \{r \in \Re^{m+1} : f_i(x) \le r_i,\ 0 \le i \le m\}$ cannot be separated by a hyperplane; by the separation theorem this implies that 0_{m+1} is an interior point of C. In particular C contains a point $(\beta, 0, \ldots, 0)$ for some $\beta < 0$. That is, there exists a point $\tilde{x} \in \Re^n$ for which $f_0(\tilde{x}) < 0$ and $f_i(\tilde{x}) \le 0,\ 1 \le i \le m$. Define the variation $x_\alpha = (1 - \alpha)\bar{x} + \alpha\tilde{x},\ \alpha \ge 0$. This variation is a reasonable descent, clearly.

All other first, second—and higher—order conditions for optimization problems with equality and inequality constraints can be established in a similar way by means of suitable reasonable descents.

References

[1] V.M. Alekseev, V.M. Tikhomirov, and S.V. Fomin, *Optimal Control,* Consultants Bureau, New York (1987).

[2] D.P. Bertsekas, *Constrained Optimization and Lagrange multipliers,* Athena Scientific, Belmont, Massachusetts (1996).

[3] J. Brinkhuis, V.M. Tikhomirov, *Optimization: Insights and Applications,* Princeton University, to appear in 2005.

[4] R.A. Horn and C.R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press (1991).

[5] A.D. Ioffe and V.M. Tikhomirov, *Theory of Extremal Problems*, North-Holland, Amsterdam-New York-Oxford (1979).

[6] F. John, *Extremum problems with inequalities as subsidiary conditions*, in: Studies and Essays. Courant Anniversary Volume, Interscience, New York (1948), pp 187-204.

[7] W.E. Karush, *Minima of Functions of Several Variables with Inequalities as Side Conditions*, Univ. of Chicago Press, Chicago (1939).

[8] H.W. Kuhn and A.W. Tucker, *Nonlinear programming*, in: Proc. of Second Berkeley Symp., Univ. of California Press, Berkeley (1951), pp. 481-492.

[9] S. Lang, *Algebra*, Addison-Wesley Publishing Company, Reading (1971).

[10] R. Remmert, *Theory of Complex Functions*, Springer Verlag, New York (1989).

[11] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, New Yersey (1970)

Jan Brinkhuis, Erasmus Universiteit Rotterdam, Econometrisch Instituut,
Faculteit der Economische Wetenschappen, H 11-16, Postbus 1738,
3000 DR Rotterdam, The Netherlands,
e-mail: brinkhuis@few.eur.nl