3. UNCONSTRAINED MINIMIZATION

Many of the complications encountered in problems of optimization are due to the presence of constraints, but even when there are no constraints a number of important issues arise as to the nature of optimal solutions and the possible ways they might be determined. In treating these issues in the case of a smooth objective function, we will want to take full advantage of the properties incorporated into the standard definition of differentiability for a function of n-variables.

Vector notation: The inner product (or dot product) of two vectors is the value

$$z \cdot w = z_1 w_1 + \cdots + z_n w_n \text{ for } z = (z_1, \ldots, z_n) \text{ and } w = (w_1, \ldots, w_n),$$

as we've already been using. In books with extensive linear algebra this is often expressed instead by $z^T w$ under the convention that vectors are interpreted special matrices—"column vectors"—unless the transpose operation (indicated by a superscript T) turns them into "row vectors." Here we follow the typographically preferable pattern of always writing vectors horizontally but viewing them as "column vectors" in formulas where they get multiplied by matrices.

Angles between vectors: When $z \neq 0$ and $w \neq 0$, one has $z \cdot w = |z||w| \cos\theta$, where θ is the angle between z and w (and $|z|$ and $|w|$ their lengths). Thus, $z \cdot w$ is positive, zero, or negative according to whether the angle is acute, right, or obtuse.

Review of differentiability: The differentiability of a function f on $\mathbb{R}^n$ means more than the existence of its first partial derivatives, which after all would just refer to behavior along various lines parallel to the n coordinate axes. Rather, it's a property expressing the possibility of a kind of approximation of f (namely, "linearization") that is present regardless of any change of coordinates that might be introduced. For our purposes here, we'll avoid subtle distinctions by keeping to the mainstream case where differentiability can be combined with continuity.

Continuous differentiability classes: A function f on $\mathbb{R}^n$ is *continuously differentiable*, or of class $\mathcal{C}^1$, if its first partial derivatives exist everywhere and are continuous everywhere. It's *twice continuously differentiable*, or of class $\mathcal{C}^2$, if this holds for second partial derivatives, and in general of class $\mathcal{C}^k$ its kth partial derivatives exist and are continuous everywhere. Then actually f and all its partial derivatives of orders less than k must be continuous as well.

Localization: Similarly one can speak of f as being a $\mathcal{C}^k$ function relative to some open set, for instance an open neighborhood of some point $\bar{x}$.

Expansions: If f is of class $\mathcal{C}^1$ on a neighborhood of a point $\bar{x}$ it has the first-order expansion

$$f(x) = f(\bar{x}) + \nabla f(\bar{x}) \cdot [x - \bar{x}] + o\big(|x - \bar{x}|\big),$$

where the vector $\nabla f(\bar{x})$ has the partial derivatives $(\partial f / \partial x_j)(\bar{x})$ as its components and is called the *gradient* of f at $\bar{x}$. The classical "o(t)" notation refers to an error term with the property that $o(t)/t \to 0$ as $t \to 0$. This notation is often confusing to students, but really it is just a sort of code for writing down, in a manner deemed more convenient, the assertion that

$$\lim_{\substack{x \to \bar{x} \\ x \neq \bar{x}}} \frac{f(x) - f(\bar{x}) - \nabla f(\bar{x}) \cdot [x - \bar{x}]}{|x - \bar{x}|} = 0,$$

which says that the affine function $l(x) = f(\bar{x}) + \nabla f(\bar{x}) \cdot [x - \bar{x}]$ furnishes a *first-order approximation* $f \approx l$ at $\bar{x}$. Likewise, if f is of class $\mathcal{C}^2$ on a neighborhood of $\bar{x}$ it has the second-order expansion

$$f(x) = f(\bar{x}) + \nabla f(\bar{x}) \cdot [x - \bar{x}] + \tfrac{1}{2} [x - \bar{x}] \cdot \nabla^2 f(\bar{x}) [x - \bar{x}] + o(|x - \bar{x}|^2),$$

where $\nabla^2 f(\bar{x})$ is the symmetric matrix with the partial derivatives $(\partial^2 f / \partial x_i \partial x_j)(\bar{x})$ as its components and is called the *Hessian* of f at $\bar{x}$. This time the "o" notation is code for the assertion that

$$\lim_{\substack{x \to \bar{x} \\ x \neq \bar{x}}} \frac{f(x) - f(\bar{x}) - \nabla f(\bar{x}) \cdot [x - \bar{x}] - \frac{1}{2} [x - \bar{x}] \cdot \nabla^2 f(\bar{x}) [x - \bar{x}]}{|x - \bar{x}|^2} = 0.$$

The quadratic function $q(x) = f(\bar{x}) + \nabla f(\bar{x}) \cdot [x - \bar{x}] + \frac{1}{2} [x - \bar{x}] \cdot \nabla^2 f(\bar{x}) [x - \bar{x}]$ then furnishes a *second-order approximation* $f \approx q$ at $\bar{x}$.

Vector-valued functions: A mapping, or vector-valued function, $F : \mathbb{R}^n \to \mathbb{R}^m$ with $F(x) = \big(f_1(x), \ldots, f_m(x)\big)$ is of class $\mathcal{C}^k$ when its component functions f_i are. As long as F is of class $\mathcal{C}^1$ around $\bar{x}$ it has the first-order expansion

$$F(x) = F(\bar{x}) + \nabla F(\bar{x})[x - \bar{x}] + o\big(|x - \bar{x}|\big),$$

where $\nabla F(\bar{x})$ is the $m \times n$ matrix with the partial derivatives $(\partial f_i / \partial x_j)(\bar{x})$ as its components and is called the *Jacobian* of F at $\bar{x}$.

A connection: In the case of a function f of class $\mathcal{C}^2$ on a neighborhood of $\bar{x}$, the gradient mapping $\nabla f : x \to \nabla f(x)$ has Jacobian $\nabla(\nabla f)(\bar{x}) = \nabla^2 f(\bar{x})$ at $\bar{x}$.

Local information: For a $\mathcal{C}^2$ function f_0 on $\mathbb{R}^n$, the gradient $\nabla f_0(\bar{x})$ and Hessian $\nabla^2 f_0(\bar{x})$ provide information about f_0 at $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)$ which can well be put to use in numerical methods for minimizing f_0. The main idea is to consider what happens to f_0 along various lines through $\bar{x}$. Any such line can represented parametrically as the set of points of the form $\bar{x} + \tau w$ for $-\infty < \tau < \infty$ for some vector $w \neq 0$, The direction of $w = (w_1, \ldots, w_n)$ gives the direction of the line.

The values of f_0 along such a line can be investigated parametrically through the expression $\varphi(\tau) := f_0(\bar{x} + \tau w)$, where $\varphi(0) = f_0(\bar{x})$. In particular, one can try to minimize $\varphi(\tau)$ in τ, at least part way, in order to come up with a point $\bar{x} + \tau w$ yielding a lower value of f_0 than does $\bar{x}$. A crucial fact in this respect is that

$$\varphi'(0) = \frac{d}{d\tau} f_0(\bar{x} + \tau w)\Big|_{\tau=0} = \nabla f_0(\bar{x}) \cdot w,$$

this being known as the *directional derivative* at f_0 relative to w. By looking at the sign of this quantity we can tell whether the values of f_0 will go up or down as we start to move away from $\bar{x}$ in the direction of w. On the other hand, one has

$$\varphi''(0) = \frac{d^2}{d\tau^2} f_0(\bar{x} + \tau w)\Big|_{\tau=0} = w \cdot \nabla^2 f_0(\bar{x}) w,$$

and this quantity can be crucial in determining second-order effects.

Descent vectors: A vector w is called a *descent vector* for f_0 at $\bar{x}$ if $\nabla f_0(\bar{x}) \cdot w < 0$. This implies that the function $\varphi(\tau) = f_0(\bar{x} + \tau w)$ is a decreasing function on some interval $(-\varepsilon, \varepsilon)$, so that $f_0(\bar{x} + \tau w) < f_0(\bar{x})$ for $0 < \tau < \varepsilon$. When the object is to minimize f_0, we therefore get an improvement in replacing $\bar{x}$ by $\bar{x} + \tau w$ for a value $\tau > 0$ that isn't too large.

Stationary points: A point $\bar{x}$ is *stationary* for f_0 if $\nabla f_0(\bar{x}) = 0$. This is equivalent to the condition that $\nabla f_0(\bar{x}) \cdot w = 0$ for *every* vector w, and it thus means that no descent vector exists at $\bar{x}$.

A trap not to fall into: If $\bar{x}$ minimizes f_0, even just locally, there can't exist a descent vector at $\bar{x}$, and thus $\nabla f_0(\bar{x}) = 0$. But the converse is false: a stationary point doesn't always provide a local minimum. This is obvious as soon as attention is focused on it, since the one-dimensional case already provides numerous examples, but many authors of articles applying mathematics to physical sciences, economics and elsewhere, nonetheless slip up on it unthinkingly and draw wrong conclusions. Conditions that are sufficient for a local or global minimum will be laid out in theorems below.

Optimality considerations: The conditions characterizing a local minimum of an unconstrained function of a single variable in terms of first and second derivatives are known to every student of calculus, but the analogs in higher dimensions are less widely familiar. Since it is not possible beyond the one-dimensional case to speak of a positive second derivative, because the second derivative has been replaced by the Hessian matrix, the following standard notions of linear algebra come into play. Recall that a matrix $A \in \mathbb{R}^{n\times n}$ is

$$\textit{positive definite} \text{ if } w\cdot Aw > 0 \text{ for all } w \neq 0,$$

$$\textit{positive semidefinite} \text{ if } w\cdot Aw \geq 0 \text{ for all } w.$$

THEOREM 3 (local optimality conditions without constraints). *For a function $f_0 : \mathbb{R}^n \to \mathbb{R}$ of class $\mathcal{C}^2$, consider the problem of minimizing f_0 over all $x \in \mathbb{R}^n$.*

(a) (necessary condition). *If $\bar{x}$ is a locally optimal solution, then $\nabla f_0(\bar{x}) = 0$ and $\nabla^2 f_0(\bar{x})$ is positive semidefinite.*

(b) (sufficient condition). *If $\bar{x}$ is such that $\nabla f_0(\bar{x}) = 0$ and $\nabla^2 f_0(\bar{x})$ is positive definite, then $\bar{x}$ is a locally optimal solution. In fact there is a $\delta > 0$ such that*

$$f_0(x) > f_0(\bar{x}) \text{ for all points } x \text{ with } 0 < |x - x_0| < \delta.$$

Proof. (a) If $\bar{x}$ is locally optimal for f_0 on $\mathbb{R}^n$, then in particular it will be true for each choice of $w \in \mathbb{R}^n$ that the function $\varphi(\tau) := f_0(\bar{x} + \tau w)$ has a local minimum at $\tau = 0$. Since $\varphi'(0) = \nabla f_0(\bar{x})\cdot w$ and $\varphi''(0) = w\cdot\nabla^2 f_0(\bar{x})w$, as noted earlier, we conclude that $\nabla f_0(\bar{x})\cdot w = 0$ for every $w \in \mathbb{R}^n$ and $w\cdot\nabla^2 f_0(\bar{x})w \geq 0$ for every $w \in \mathbb{R}^n$. This means that $\nabla f_0(\bar{x}) = 0$ and $\nabla^2 f_0(\bar{x})$ is positive semidefinite.

(b) The reverse argument is more subtle and can't just be reduced to one dimension, but requires utilizing fully the second-order expansion of f_0 at $\bar{x}$. Our assumptions give for $A := \nabla^2 f_0(\bar{x})$ that $f_0(x) = f_0(\bar{x}) + \frac{1}{2}[x - \bar{x}]\cdot A[x - \bar{x}] + o\big(|x - \bar{x}|^2\big)$. According to the meaning of this, we can find for any $\varepsilon > 0$ a $\delta > 0$ such that

$$\frac{\big|f_0(x) - f_0(\bar{x}) - \frac{1}{2}[x - \bar{x}]\cdot A[x - \bar{x}]\big|}{|x - \bar{x}|^2} < \varepsilon \text{ when } 0 < |x - \bar{x}| < \delta,$$

and in particular,

$$f_0(x) - f_0(\bar{x}) \;>\; \tfrac{1}{2}[x - \bar{x}]\cdot A[x - \bar{x}] - \varepsilon|x - \bar{x}|^2 \text{ when } 0 < |x - \bar{x}| < \delta.$$

Because A is positive definite, the expression $\frac{1}{2}w\cdot Aw$ is positive when $w \neq 0$; it depends continuously on w and therefore achieves its minimum over the closed, bounded set consisting of the vectors w with $|w| = 1$ (the unit sphere in $\mathbb{R}^n$). Denoting this minimum by

λ, we have $\lambda > 0$ and $\frac{1}{2}[\tau w]\cdot A[\tau w] \geq \lambda\tau^2$ for all $\tau \in \mathbb{R}$ when $|w| = 1$. Since any difference vector $x - \bar{x} \neq 0$ can be written as τw for $\tau = |x - \bar{x}|$ and $w = [x - \bar{x}]/|x - \bar{x}|$, we have $\frac{1}{2}[x-\bar{x}]\cdot A[x-\bar{x}] \geq \lambda|x-\bar{x}|^2$ for all x. The estimate from twice differentiability then yields

$$f_0(x) - f_0(\bar{x}) \ > \ (\lambda - \varepsilon)|x - \bar{x}|^2 \ \text{ when } \ 0 < |x - \bar{x}| < \delta.$$

Recalling that ε could have been chosen arbitrarily small, in particular in the interval $(0, \lambda)$, we conclude that there's a $\delta > 0$ such that $f_0(x) > f_0(\bar{x})$ when $0 < |x - \bar{x}| < \delta$. Thus, f has a local minimum at $\bar{x}$. □

Local versus global optimality: Because the results in Theorem 3 relate only to the properties of f_0 in some neighborhood of $\bar{x}$, and give no estimate for the size of that neighborhood (it might be tiny, for all we know), an important question is left unanswered. How can we tell whether a given point $\bar{x}$ furnishes a global minimum to f_0? The best approach to answering this question, and to some extent the only one, is through the concept of convexity.

Convex functions: A function f on $\mathbb{R}^n$ is *convex* if for every choice of points x_0 and x_1 with $x_0 \neq x_1$, and every choice of $\tau \in (0, 1)$, one has

$$f\big((1-\tau)x_0 + \tau x_1\big) \leq (1-\tau)f\big(x_0\big) + \tau f\big(x_1\big) \ \text{ for all } \ \tau \in (0, 1).$$

Interpretation: The expression $x(\tau) := (1-\tau)x_0 + \tau x_1 = x_0 + \tau(x_1 - x_0)$ parameterizes the line through x_0 in the direction of $w = x_1 - x_0$, with $x(0) = x_0$ and $x(1) = x_1$. When $0 < \tau < 1$, $x(\tau)$ is an intermediate point on the line segment joining x_0 with x_1, specifically the point reached in moving the fraction τ of the distance from x_0 to x_1 along this segment. The inequality says that the value of f at this intermediate point doesn't exceed the interpolated value obtained by going the fraction τ of the way from the value $f(x_0)$ to the value $f(x_1)$ (whichever direction that might involve, depending on which of the two values might be larger).

Relativization to lines: Since the condition involves only three collinear points at a time, we have the principle that *f is convex on $\mathbb{R}^n$ if and only if for every line L in $\mathbb{R}^n$, f is convex relative to the L*; in fact, instead of lines it would be enough to speak of line segments. Here a *line* is a set of points in $\mathbb{R}^n$ that can be expressed as $\big\{x + \tau w \,\big|\, -\infty < \tau < \infty\big\}$ for some x and w with $w \neq 0$, whereas a *line segment* is the same thing but with $0 \leq \tau \leq 1$.

Related properties: A function f is

strictly convex: if the inequality always holds with $<$,

concave: if the inequality always holds with $\geq$,

strictly concave: if the inequality always holds with $>$.

Affine functions as an example: It can be shown that f is affine on $\mathbb{R}^n$, as already defined, if and only if f is simultaneously convex and concave.

Jensen's inequality: The definition of convexity implies more generally that for any points x_k and weights $\lambda_k \geq 0$ for $k = 0, 1, \ldots, p$ with $\sum_{k=0}^{p} \lambda_k = 1$, one has

$$f\big(\lambda_0 x_0 + \lambda_1 x_1 + \cdots + \lambda_p x_p\big) \;\leq\; \lambda_0 f\big(x_0\big) + \lambda_1 f\big(x_1\big) + \cdots + \lambda_p f\big(x_p\big).$$

Tests for convexity using derivatives: The following facts help in identifying examples of convex functions that are differentiable. Later there will be other tactics available, which can be used to assertain the convexity of functions that have been put together in certain ways from basic functions whose convexity is already known.

Monotonicity of first derivatives in one dimension: For f differentiable on $\mathbb{R}$,

$$\begin{aligned}
f \text{ is convex} &\iff f' \text{ is nondecreasing,} \\
f \text{ is strictly convex} &\iff f' \text{ is increasing,} \\
f \text{ is concave} &\iff f' \text{ is nonincreasing,} \\
f \text{ is strictly concave} &\iff f' \text{ is decreasing,} \\
f \text{ is affine} &\iff f' \text{ is constant.}
\end{aligned}$$

Incidentally, these generalize also to functions of a single variable that merely have a right derivative and a left derivative at every point. For instance, a piecewise linear cost function is convex if and only if the slope values for consecutive pieces form an increasing sequence. Also as first derivative conditions,

$$\begin{aligned}
f \text{ is convex} &\iff f(y) \geq f(x) + f'(x)(y - x) \text{ for all } x \text{ and } y, \\
f \text{ is strictly convex} &\iff f(y) > f(x) + f'(x)(y - x) \text{ for all } x \text{ and } y,\ x \neq y.
\end{aligned}$$

Signs of second derivatives in one dimension: For f twice differentiable on $\mathbb{R}$,

$$\begin{aligned}
f \text{ is convex} &\iff f''(x) \geq 0 \text{ for all } x, \\
f \text{ is strictly convex} &\impliedby f''(x) > 0 \text{ for all } x.
\end{aligned}$$

Notice that the final condition is *not* an equivalence but only an implication in one direction! An example is $f(x) = x^4$, with $f''(x) = 12x^2$. This function is strictly convex on $\mathbb{R}$ because $f'(x) = 4x^3$ is an increasing function. But $f''(x)$ fails to be positive everywhere: $f''(0) = 0$.

THEOREM 4 (derivative tests for convexity in higher dimensions). *For a twice differentiable function f on $\mathbb{R}^n$,*

$$\begin{aligned}
f \text{ is convex} &\iff f(y) \geq f(x) + \nabla f(x)\cdot[y - x] \text{ for all } x \text{ and } y,\\
f \text{ is strictly convex} &\iff f(y) > f(x) + \nabla f(x)\cdot[y - x] \text{ for all } x \text{ and } y,\ x \neq y,\\
f \text{ is convex} &\iff \nabla^2 f(x) \text{ is positive semidefinite for all } x,\\
f \text{ is strictly convex} &\impliedby \nabla^2 f(x) \text{ is positive definite for all } x.
\end{aligned}$$

Proof. The trick in every case is to reduce to the corresponding one-dimensional criterion through the principle that f has the property in question if and only if it has it relative to every line segment. The first of the conditions will suffice in illustrating this technique. To say that f is convex is to say that for every choice of points x_0 and x_1 with $x_0 \neq x_1$ the function $\varphi(\tau) := f\big((1-\tau)x_0 + \tau x_1\big) = f\big(x_0 + \tau(x_1 - x_0)\big)$ is convex on the interval $(0, 1)$. From the chain rule one calculates that

$$\varphi''(\tau) = w\cdot\nabla^2 f(x)w \qquad \text{for } x = (1-\tau)x_0 + \tau x_1, \quad w = x_1 - x_0.$$

The convexity of f is thus equivalent to having $w\cdot\nabla^2 f(x)w \geq 0$ for every possible choice of x and $w \neq 0$ such that x is an intermediate point of some line segment in the direction of w. This holds if and only if $\nabla^2 f(x)$ is positive semidefinite for every x. The arguments for the other three conditions are very similar in character. □

Local strict convexity: A function f is strictly convex *locally* around $\bar{x}$ if the strict convexity inequality holds over the line segment joining x_0 and x_1 whenever these points lie within a certain neighborhood of $\bar{x}$. The proof of Theorem 3 show that for this to be true in the case of a function f of class $\mathcal{C}^2$ it suffices to have the Hessian $\nabla^2 f(x)$ be positive definite at all points x in some neighborhood of $\bar{x}$. In fact it suffices to have $\nabla^2 f(\bar{x})$ itself be positive definite, because any matrix having entries close enough to those of a positive definite matrix must likewise be positive definite, and here the entries of $\nabla^2 f(x)$ depend continuously on x. (The stability of positive definiteness under perturbations follows from identifying the positive definiteness of a matrix A with the positivity of the function $q(w) = w\cdot Aw$ on the compact set consisting of the vectors w with $|w| = 1$.)

Tests of positive definiteness: For a symmetric matrix $A \in \mathbb{R}^{n\times n}$ there are many tests of whether A is positive definite or positive semidefinite, as may be found in texts on linear algebra, but they aren't always easy to apply. Computer tests are available as well. Perhaps the main thing to remember is that any symmetric matrix A is similar to a diagonal matrix having as its diagonal entries the n eigenvalues of A (with multiplicities). *Positive definiteness holds if and only if all the eigenvalues are positive, whereas positive semidefiniteness holds if and only of all the eigenvalues are nonnegative.*

Two-dimensional criterion: Positive definiteness implies that both the determinant of A and the trace of A (the sum of the diagonal entries of A) are positive. When $n = 2$ the converse holds as well, although not when $n > 2$.

Consequences of convexity in unconstrained minimization:

Global optimality of stationary points: For a $\mathcal{C}^1$ function f_0 on $\mathbb{R}^n$ that's convex, the condition $\nabla f_0(\bar{x}) = 0$ implies that $\bar{x}$ gives the *global minimum* of f_0 on $\mathbb{R}^n$.

Argument: Convexity implies by Theorem 4 that $f_0(x) \geq f_0(\bar{x}) + \nabla f_0(\bar{x})\cdot[x - \bar{x}]$ for all x. When $\nabla f_0(\bar{x}) = 0$, this reduces to having $f_0(x) \geq f_0(\bar{x})$ for all x.

Uniqueness from strict convexity: If a strictly convex function f_0 has its minimum at $\bar{x}$, then $\bar{x}$ is the *only* point where f_0 has its minimum. In fact, for this conclusion it's enough that f_0 be a convex function that's strictly convex locally around $\bar{x}$.

Argument: If there were another point $\hat{x}$ where f_0 had its minimum value, say α, the intermediate points $x_\tau = (1-\tau)\bar{x} + \tau\hat{x}$ for $\tau \in (0,1)$ would have $f_0(x_\tau) \leq (1-\tau)f_0(\bar{x}) + \tau f_0(\hat{x}) = (1-\tau)\alpha + \tau\alpha = \alpha$, hence $f_0(x_\tau) = \alpha$, since nothing lower than α is possible. Then f_0 would be constant on the line segment joining $\bar{x}$ and $\hat{x}$, so it couldn't be strictly convex any portion of it.

Convexity of quadratic functions: If $f(x) = \frac{1}{2}x\cdot Ax + b\cdot x + c$ for a symmetric matrix $A \in \mathbb{R}^{n\times n}$, a vector $b \in \mathbb{R}^n$, and a constant $c \in \mathbb{R}$, we have $\nabla f(x) = Ax + b$ and $\nabla^2 f(x) = Ax$ for all x. Therefore, such a function is convex if and only if A is positive semidefinite. It is strictly convex if and only if A is positive definite. This second assertion doesn't fully follow from the second-order condition in Theorem 4, which only gives the implication in one direction, but it can be deduced from the *first-order* condition for strict convexity.

Minimizing a quadratic function: A quadratic function can't attain its minimum anywhere if it isn't a convex function. It attains its minimum at a unique point if and only if it's strictly convex—with positive definite Hessian.

Argument: If a quadratic function q attains its minimum at a point $\bar{x}$, its Hessian at $\bar{x}$ must be positive semidefinite by Theorem 3. But, because it's quadratic, q has this same Hessian at every point. Then by Theorem 4, q is convex. If the Hessian matrix is A, the fact that the gradient of q at $\bar{x}$ is 0 means we have the expansion $q(x) = q(\bar{x}) + \frac{1}{2}[x - \bar{x}]\cdot A[x - \bar{x}]$. Under the assumption that the minimum is attained uniquely at $\bar{x}$ there can't be a vector $x - \bar{x} \neq 0$ such that $A[x - \bar{x}] = 0$. Then A is nonsingular. But from linear algebra, a positive semidefinite matrix is nonsingular if and only if it's positive definite. Then q is strictly convex by Theorem 4.

Conversely, if q has Hessian A, it has the expression $q(x) = \frac{1}{2}x\cdot Ax + b\cdot x + c$ for $b = \nabla q(0)$ and $c = q(0)$. If A is positive definite there is a $\lambda > 0$ such that $\frac{1}{2}x\cdot Ax \geq \lambda|x|^2$ for all x, by reasoning given in the proof of part of Theorem 3(b). Then $|q(x)| \geq \lambda|x|^2 - |b||x| - |c|$, so that for any $\rho > 0$ the norms $|x|$ of the vectors in the level set $\{x \,|\, q(x) \leq \rho\}$ all lie in the interval $\{t \,|\, \lambda t^2 - |b|t - [c + \rho] \leq 0\}$, which is bounded because of λ being positive. These level sets are therefore all bounded, so the problem of minimizing q is well posed and by Theorem 1 has a solution.

Applications to numerical optimization: Most numerical methods for the unconstrained minimization of a twice continuously differentiable function rely at least to some degree on the facts we've been developing. Here, for purposes of illustration, we'll look at some of the most popular approaches based on utilization of local information.

Descent methods: A large class of methods for minimizing a smooth function f_0 on $\mathbb{R}^n$ fits the following description. A sequence of points x^ν such that $f_0(x^0) > f_0(x^1) > \cdots > f_0(x^\nu) > f_0(x^{\nu+1}) > \cdots$ is generated from a chosen starting point x^0 by selecting, through some special scheme in each iteration, a descent vector w^ν and a corresponding value $\tau^\nu > 0$, called a *step size*, such that $f_0(x^\nu + \tau^\nu w^\nu) < f_0(x^\nu)$. The improved point $x^\nu + \tau^\nu w^\nu$ is taken to be $x^{\nu+1}$.

Of course, if in a given iteration no descent vector exists at all, this means that $\nabla f_0(x^\nu) = 0$. In that case the method terminates with $\bar{x} = x^\nu$ as a stationary point. Usually, however, an infinite sequence $\{x^\nu\}_{\nu=1}^\infty$ is generated, and the question for analysis is whether this is an optimal sequence, or more soberly in the nonconvex case, at least a sequence for which every cluster point $\bar{x}$ is a stationary point.

Line search: One way to choose a step size τ^ν yielding $f_0(x^\nu + \tau^\nu w^\nu) < f(x^\nu)$ is to execute some kind of *line search* in the direction of w^ν, which refers to an exploration of the values of f_0 along the half-line emanating from x^ν in the direction of w^ν. In the notation $\varphi^\nu(\tau) := f(x^\nu + \tau w^\nu)$ we have $(\varphi^\nu)'(0) < 0$, and the task is to select a value $\tau^\nu > 0$ such that $\varphi^\nu(\tau^\nu) < \varphi^\nu(0)$, yet not one so small that progress might stagnate.

Exact line search: An approach with natural appeal is to choose τ^ν to be a value of τ that minimizes φ^ν on the interval $[0, \infty)$. Techniques are available for carrying out the one-dimensional minimization of φ^ν to whatever accuracy is desired, at least when f_0 is convex. Of course, in numerical work hardly anything is really "exact."

Backtracking line search: Professional opinion now favors a different approach, which depends on a choice of parameters β and γ with $0 < \beta < \gamma < 1$. Calculating the first integral power $\kappa \geq 0$ such that $\big[\varphi^\nu(\gamma^\kappa) - \varphi^\nu(0)\big]/\gamma^\kappa < -\beta(\varphi^\nu)'(0)$, one sets $\tau^\nu = \gamma^\kappa \in (0, 1]$. (Such a κ exists, because $\gamma^\kappa \to 0$ as $\kappa \to \infty$, while $\big[\varphi^\nu(\tau) - \varphi^\nu(0)\big]/\tau$ tends to $(\varphi^\nu)'(0) < 0$ as τ decreases to 0.)

Example: Cauchy's method ("steepest descent")

A descent method can be obtained by choosing $w^\nu = -\nabla f_0(x^\nu)$ in every iteration (as long as this vector is nonzero), since $\nabla f_0(x^\nu) \cdot w^\nu = -\big|\nabla f_0(x^\nu)|^2 < 0$ unless $\nabla f_0(x^\nu) = 0$. One speaks then of following the direction of *steepest descent*, because of all the vectors w with $|w| = 1$, the one giving the lowest (i.e., most negative) value to the directional derivative $\nabla f_0(x^\nu) \cdot w$ is $w = -\nabla f_0(x^\nu)/|\nabla f_0(x^\nu)|$. (In backtracking line search, one takes the latter vector as w^ν instead of $-\nabla f_0(x^\nu)$.)

Example: Newton's method in optimization

From the definition of twice differentiability we know that the quadratic function

$$q^\nu(x) := f_0(x^\nu) + \nabla f_0(x^\nu) \cdot [x - x^\nu] + \tfrac{1}{2}[x - x^\nu] \cdot \nabla^2 f_0(x^\nu)[x - x^\nu]$$

(whose Hessian everywhere is $A = \nabla^2 f_0(x^\nu)$) furnishes a second-order local approximation of f_0 around x^ν. This suggests that by investigating the minimum of $q^\nu(x)$ we can learn something about where to look in trying to minimize f_0. Specifically, *assume that $q^\nu(x)$ attains its minimum at a unique point different from x^ν*, this point being denoted by $\hat{x}^\nu$; then $\hat{x}^\nu \neq x^\nu$, and $q^\nu(\hat{x}^\nu) < q^\nu(x^\nu) = f(x^\nu)$. (From the above, this is equivalent to assuming that the matrix $\nabla^2 f(x^\nu)$ is positive definite, hence in particular nonsingular, while the vector $\nabla f(x^\nu)$ isn't 0.) The vector $w^\nu = \hat{x}^\nu - x^\nu \neq 0$

is then called the *Newton vector* for f_0 at $\bar{x}$. It satisfies

$$w^\nu = -\nabla^2 f_0(x^\nu)^{-1} \nabla f_0(x^\nu).$$

It is a descent vector, and descent methods based on using it are called versions of *Newton's method* in optimization.

Argument: Because $\hat{x}^\nu$ minimizes $\nabla q^\nu(\hat{x}^\nu)$, it must be a stationary point of q^ν:

$$0 = \nabla q^\nu(\hat{x}^\nu) = \nabla f_0(x^\nu) + \nabla^2 f_0(x^\nu)[\hat{x}^\nu - x^\nu] = \nabla f_0(x^\nu) + \nabla^2 f_0(x^\nu) w^\nu.$$

In solving this equation for w^ν, utilizing our assumption, which implies that the inverse matrix $\nabla^2 f_0(x^\nu)^{-1}$ exists, we get the formula claimed. To verify that w^ν is then a descent vector, observe that because $q^\nu(\hat{x}^\nu) < q^\nu(x^\nu)$ we have $\nabla f_0(x^\nu) \cdot w^\nu + \frac{1}{2} w^\nu \cdot \nabla^2 f_0(x^\nu) w^\nu < 0$. We wish to conclude that $\nabla f_0(x^\nu) \cdot w^\nu < 0$. If this weren't true, we'd have to have from the preceding inequality that $w^\nu \cdot \nabla^2 f_0(x^\nu) w^\nu < 0$. But this would contradict the positive definiteness of $\nabla^2 f_0(x^\nu)$, which was observed to follow from our assumption about q^ν attaining its minimum at a unique point.

Relation to Newton's method for equation solving: Newton's method in classical form refers not to minimizing a function but solving an equation $F(x) = 0$ for a smooth mapping $F : \mathbb{R}^n \to \mathbb{R}^n$. In principle, a sequence $\{x^\nu\}_{\nu \in \boldsymbol{N}}$ is generated from an initial point x^0 as follows. In iteration ν, the given equation is replaced by its first-order approximation $F(x^\nu) + \nabla F(x^\nu)(x - x^\nu) = 0$. The unique solution to this approximate equation is $\hat{x}^\nu = -\nabla F(x^\nu)^{-1} F(x^\nu)$, as long as the inverse matrix $\nabla F(x^\nu)^{-1}$ exists, and one takes $x^{\nu+1} = \hat{x}^\nu$.

Newton's method in optimization corresponds closely to the case of this where $F(x) = \nabla f_0(x)$. It resembles applying the classical form of Newton's method to solving the equation $\nabla f_0(x) = 0$. But it differs in not just automatically taking $x^{\nu+1} = \hat{x}^\nu = x^\nu + w^\nu$ but $x^{\nu+1} = x^\nu + \tau^\nu w^\nu$ for some step size τ^ν determined through a form of line search.

Effectiveness and validation of descent methods: Describing an approach to minimizing a function, or for that matter solving a vector equation, is a far cry from establishing the circumstances in which it can be counted upon to work effectively, or providing an analysis that helps comparison with other approaches. For such purposes it is essential at the very least to make use of the conditions characterizing a point at which a minimum occurs as well as, in some situations aspects of convexity.

Convergence questions: The theory of numerical methods of optimization and why (or whether) they work is full of ingenious ideas and pleasing geometry, as well as rigorous, technical developments. For a small taste of what it involves, let's consider more closely the question of whether a descent method (with a particular scheme for choosing decent vectors and executing line searches) for the unconstrained minimization of a function f_0 generates a sequence of points x^ν that in some way "solves" the problem. Any such method does generate a decreasing sequence of function values $f_0(x^0) > f_0(x^1) > f_0(x_2) \ldots$, and any decreasing sequence of real numbers does have a limit $\alpha \in \overline{\mathbb{R}}$, but unfortunately α could fall short of furnishing the optimal value in the problem unless f_0 has certain rather special properties. Nonetheless we can search for guidance on when a method can sensibly be implemented and what it might accomplish even if it doesn't determine an optimal or locally optimal solution.

Well posed problems: On the basis of the observation after Theorem 1, an unconstrained problem of minimizing f_0 over $\mathbb{R}^n$ is well posed as long as f_0 is continuous and all its level sets $\{x \mid f_0(x) \leq \alpha\}$ are bounded. Certainly f_0 is continuous when it's differentiable, as in the descent methods we've been investigating. In unconstrained minimization there's no distinction between feasible and asymptotically feasible sequences (every sequences is such), nor any between optimal and asymptotically optimal sequences. As long as the problem is well posed, we know then from Theorem 2 that every optimal sequence is bounded, and all its cluster points are optimal solutions. However, this doesn't necessarily make it easier to *generate* an optimal sequence.

THEOREM 5 (convergence of descent methods; exact line search). *Consider a well posed problem of minimizing a function f_0 over $\mathbb{R}^n$, with f_0 not just continuous but differentiable, and let S be the set of all stationary points of f_0 (the points $\bar{x}$ where $\nabla f_0(\bar{x}) = 0$). Consider a descent method that starts from a point x^0 and generates subsequent points by exact line search relative to vectors w^ν determined by a formula $w^\nu = D(x^\nu)$ having the property that, for each $x \notin S$ with $f_0(x) \leq f_0(x^0)$, $D(x)$ is a uniquely determined descent vector for f_0 at x, and $D(x)$ depends continuously on x. (The method terminates if $x^\nu \in S$.)*

(a) *If the method generates an infinite sequence $\{x^\nu\}_{\nu=0}^\infty$ (by not attaining a point of S in finitely many iterations), this sequence must be bounded, and all of its cluster points must belong to S.*

(b) *If actually there is only one point $\bar{x} \in S$ with $f_0(\bar{x}) \leq f_0(x^0)$, the sequence is indeed optimal and converges to $\bar{x}$, this being the unique optimal solution.*

Proof. In each iteration with $x^\nu \notin S$, the vector w^ν is well defined according to our hypothesis, and it is not the zero vector (because it is a descent vector). We minimize $\varphi^\nu(\tau) := f_0(x^\nu + \tau w^\nu)$ over $\tau \in [0, \infty)$ to get τ^ν and then set $x^{\nu+1} = x^\nu + \tau^\nu w^\nu$. This line search subproblem is itself a well posed problem of optimization because the sets $\left\{\tau \geq 0 \,\middle|\, \varphi^\nu(\tau) \leq \alpha\right\}$ are all bounded by virtue of the level sets of f_0 all being bounded. Thus it does have an optimal solution τ^ν (perhaps not unique) by Theorem 1.

From the definition of w^ν being a descent vector, we know moreover that $f_0(x^{\nu+1}) < f_0(x^\nu)$ always. Thus the sequence $\{f_0(x^\nu)\}_{\nu=1}^\infty$ is decreasing and therefore converges to some value α. Also, the sequence $\{x^\nu\}_{\nu=1}^\infty$ is contained in the set $\left\{x \,\middle|\, f_0(x) \leq f_0(x^0)\right\}$, which by hypothesis is bounded. Consider any cluster point $\bar{x}$ of this sequence; there is a subsequence $\{x^{\nu_\kappa}\}_{\kappa=1}^\infty$ such that $x^{\nu_\kappa} \to \bar{x}$ as $\kappa \to \infty$. In particular we have $f_0(\bar{x}) = \alpha$, because f_0 is continuous. We wish to show that $\bar{x} \in S$ in order to establish (a).

Suppose $\bar{x} \notin S$. Then the vector $\bar{w} := D(\bar{x})$ is a descent vector for f_0 at $\bar{x}$, and the vectors $w^{\nu_\kappa} := D(x^{\nu_\kappa})$ are such that $w^{\nu_\kappa} \to \bar{w}$ (by our assumption in (a) that the mapping D specifying the method is well defined everywhere outside of S and continuous there). Because $\bar{w}$ is a descent vector, we know there is a value $\bar{\tau} > 0$ such that

$$f_0(\bar{x} + \bar{\tau}\bar{w}) < f_0(\bar{x}).$$

On the other hand, for each κ we know that $f_0(x^{\nu_\kappa+1}) = f_0(x^{\nu_\kappa} + \tau^{\nu_\kappa} w^{\nu_\kappa}) \leq f_0(x^{\nu_\kappa} + \bar{\tau} w^{\nu_\kappa})$ because $\bar{\tau}$ is one of the candidates considered in the minimization subproblem solved by τ^{ν_κ}. Taking the limit in the outer expressions in this inequality, we get

$$\alpha \leq f_0(\bar{x} + \bar{\tau}\bar{w})$$

because $f_0(x^\nu) \to \alpha$, $x^{\nu_\kappa} \to \bar{x}$ and $w^{\nu_\kappa} \to \bar{w}$ (again utilizing the continuity of f_0). This result is incompatible with the fact that $f_0(\bar{x} + \bar{\tau}\bar{w}) < f_0(\bar{x}) = \alpha$. The contradiction establishes (a).

The extra assumption in (b) gives the existence of a unique optimal solution to the unconstrained minimization problem, because (1) an optimal solution exists by Theorem 1, (2) any optimal solution must in particular belong to S by Theorem 3, and of course any optimal solution must belong to the set $\left\{x \,\middle|\, f_0(x) \leq f_0(x^0)\right\}$. From (a), this optimal solution $\bar{x}$ is the only candidate for a cluster point of $\{x^\nu\}_{\nu=1}^\infty$. As noted earlier, a bounded sequence with no more than one cluster point must be convergent. Thus, $x^\nu \to \bar{x}$ and, by the continuity of f_0, also $f_0(x^\nu) \to f_0(\bar{x})$. Since $f_0(\bar{x})$ is the optimal value in the problem, we conclude in this case that the sequence is optimal. □

Specializations: Particular applications of the convergence result in Theorem 5 are obtained by considering various choices of the mapping D.

Cauchy's method with exact line search: Under the assumption that f_0 is a $\mathcal{C}^1$ function, so that $f_0(x)$ and $\nabla f_0(x)$ depend continuously on x, let $D(x) = -\nabla f_0(x)$. This is a descent vector as long as x is not a stationary point (cf. Example 1). The assumptions of Theorem 5 are satisfied, and we can conclude that if all the level sets $\left\{x \,\middle|\, f_0(x) \leq \alpha\right\}$ are bounded the method will generate a bounded sequence $\{x^\nu\}_{\nu=1}^\infty$, all of whose cluster points are stationary points of f_0. If in addition f_0 is convex, these stationary points give the global minimum of f_0. In that case the method has generated an optimal sequence $\{x^\nu\}_{\nu=1}^\infty$.

Newton's method with exact line search: Under the assumption that f_0 is a $\mathcal{C}^2$ function, let $D(x)$ denote the Newton vector *under the condition that it is well defined* for every $x \notin S$ with $f_0(x) \leq f_0(x^0)$; we've seen this is tantamount to $\nabla f_0(x)^2$ being positive definite for all such x. Then $D(x) = -\nabla^2 f_0(x)^{-1} \nabla f_0(x)$, so $D(x)$ depends continuously on x (because if a nonsingular matrix varies continuously, its inverse varies continuously, a fact derivable from determinant formulas for the inverse). As long as the level sets $\left\{x \,\middle|\, f_0(x) \leq \alpha\right\}$ of f_0 are bounded, so that the problem is well posed, Theorem 5 is applicable and tells us that the method will generate a bounded sequence $\{x^\nu\}_{\nu=1}^\infty$, all of whose cluster points are stationary points of f_0. In fact, because of the positive definiteness of the Hessians, any cluster point must be a *locally optimal* solution to the problem of minimizing f_0, due to Theorem 3(b). Around any such cluster point, f_0 is strictly convex, so if f_0 is convex as a whole there can only be one cluster point, $\bar{x}$, this being the only point where f_0 attains its minimum. Then the sequence $\{x^\nu\}_{\nu=1}^\infty$ is optimal and converges to $\bar{x}$.

Comparison: Cauchy's method works quite generally, but Newton's method requires positive definiteness of the Hessians and therefore local strict convexity. But Newton's method has a much better *rate* of convergence than Cauchy's method: typically Newton's method converges *quadratically*, whereas Cauchy's method only converges *linearly*. We won't go into the theory of that here, however.

Compromise: Because Cauchy's method and Newton's method have complementary strengths and weaknesses, they are often combined in a single descent method in which, roughly speaking, the Cauchy descent vector is used early on, but eventually a switch is made to the Newton descent vector. In some versions, this approach would likewise fit into Theorem 5 for a certain formula for $D(x)$.

Extensions: Various features of Theorem 5 can readily be generalized. For instance, other forms of line search rather than exact line search can be handled. While we won't be treating this or rates of convergence here, two other directions of extension deserve to be indicated.

Example: Quasi-Newton methods

These popular methods try to span between the properties of Cauchy's method and Newton's method of optimization in an especially interesting way. They select the direction vector by $w^\nu = -A^\nu \nabla f_0(x^\nu)$, where the matrix A^ν, generated in each iteration by some further rule, is symmetric and positive semidefinite. The case of $A^\nu = I$ gives Cauchy's method, while the case of $A^\nu = \nabla^2 f_0(x^\nu)^{-1}$ (when this matrix is positive definite) gives Newton's method. For reasons already suggested, a simple choice like the one for Cauchy's method is favored as long as the current point is considered likely to be far from the solution, in order to take advantage of the global convergence properties of that method without making too many demands on the function f_0, but a choice approximating the one for Newton's method is favored near a locally optimal solution $\bar{x}$ at which $\nabla^2 f_0(\bar{x})$ is positive definite (cf. the sufficient second-order optimality condition in Theorem 3(b)). A central question is how to select and update A^ν by gleaning information about second-derivative properties of f_0 that may be present in the computations carried out up to a certain stage. This is a large topic in itself with many clever schemes that have been developed through years of research.

Example: Trust region methods

Newton's method obtains the descent vector w^ν from the fact that $x^\nu + w^\nu$ is the point that minimizes the quadratic function q^ν giving the local second-order approximation to f_0 at x^ν (cf. Example 2). This is a potential source of trouble, because, as we have seen, q^ν doesn't achieve a minimum at a unique point unless the Hessian $\nabla f_0(x^\nu)$ is positive definite. Instead of minimizing q^ν over all of $\mathbb{R}^n$, one can minimize it over a certain bounded neighborhood X^ν of x^ν, which is called a *trust region*. In denoting a minimizing point by $\hat{x}^\nu$ and defining $w^\nu = \hat{x}^\nu - x^\nu$, one gets a descent vector w^ν. The trust region can in particular be specified by linear constraints, like upper and lower bounds on the variables to keep their values near the component values in the current vector x^ν, and the subproblem for producing w^ν is then one of *quadratic programming*. This idea can be hybridized with the one behind quasi-Newton methods.

Optimization versus equation solving: The equation $\nabla f_0(x) = 0$ is a first step toward identifying points that minimize a smooth function f_0. This leads to the notion that solving an unconstrained optimization problem might be reduced to solving a system of equations, which has some degree of merit, but tends to misguide beginners, to whom equation-solving is a more familiar idea. The best approaches to equation-solving are through optimization, rather than the other way around.

Linear equations: In numerical analysis, the solving of $Ax = b$ when A is *symmetric* and *positive semidefinite* (and possibly quite large) plays a big role. Such an equation gives the condition that is both necessary and sufficient for the minimization of the quadratic convex function $f_0(x) = \frac{1}{2}x{\cdot}Ax - b{\cdot}x$ over $\mathbb{R}^n$. Thus, this branch of numerical analysis is in truth a branch of numerical optimization.

Linear least squares: The problem of solving a general system $Ax = b$ when A is not necessarily symmetric or positive semidefinite can be approached as the problem of minimizing the function $f(x) = \frac{1}{2}|Ax - b|^2$. This has the advantage of making sense even when the system is overdetermined (more equations than unknowns), as in applications to parameter identification from accumulated data; if a true solution can't be obtained, the problem focuses instead on finding a vector x which minimizes an error expression for the difference between the two sides of the desired equation. Here f is a convex function with

$$\nabla f_0(x) = A^*[Ax - b], \qquad \nabla^2 f_0(x) \equiv A^*A \qquad (A^* = \text{transpose of } A).$$

Thus, solving $\nabla f_0(x) = 0$ means solving $A^*Ax = A^*b$, where the matrix A^*A is symmetric and positive semidefinite. This fits the pattern mentioned first.

Nonlinear least squares: An approach often taken to solving $F(x) = 0$ in the case of a general smooth mapping $F : \mathbb{R}^n \to \mathbb{R}^n$, with $F(x) = \big(f_1(x), \ldots, f_n(x)\big)$, is to translate it to optimization by a trick that's always available:

$$\text{minimize } g(x) := \tfrac{1}{2}|F(x)|^2 = \tfrac{1}{2}f_1(x)^2 + \cdots + \tfrac{1}{2}f_n(x)^2 \text{ over all } x \in \mathbb{R}^n.$$

If the optimal value in this problem is 0, the optimal solutions are precisely the solutions to $F(x) = 0$. On the other hand, if the optimal value is positive, there are no solutions to $F(x) = 0$.

Criticism: While this device is often useful, it's likely to be off-track when the equation $F(x) = 0$ *already* corresponds to an optimization problem in the sense that $F = \nabla f_0$ for some function f_0, a circumstance that all too often goes unrecognized along with its potential for a more direct approach.