

Optimality conditions for unconstrained minimization

Mark S. Gockenbach

1 Introduction

I will now consider the unconstrained minimization problem

$$\min_{x \in \mathbb{R}^n} f(x),$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a smooth function (at least differentiable). I begin with the following fundamental definitions:

Definition 1.1 *Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is given. Then $x^* \in \mathbb{R}^n$ is a local minimizer of f if there exists $\delta > 0$ such that*

$$\|x - x^*\| < \delta \Rightarrow f(x^*) \leq f(x).$$

The point x^ is a strict local minimizer if*

$$0 < \|x - x^*\| < \delta \Rightarrow f(x^*) < f(x).$$

Thus x^* is a local minimizer of f if locally (that is, in a neighborhood of x^*), f does not attain a smaller value.

Definition 1.2 *Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is given. Then $x^* \in \mathbb{R}^n$ is a global minimizer of f if*

$$f(x^*) \leq f(x) \text{ for all } x \in \mathbb{R}^n.$$

A global minimizer yields the absolute minimum of f .

In most cases, given an minimization problem, it is desirable to find the global minimizer. However, this is a very challenging goal, and for many problems, it is beyond the reach of efficient algorithms. For this reason, in numerical optimization, the goal is typically to find a local minimizer. (Indeed, *global optimization* is really a separate research area from numerical optimization.) I will discuss below problems which have *only* global minima, that is, problems in which every local minimizer is in fact a global minimizer.

One point of terminology is potentially confusing. In spite of the fact that the goal of numerical optimization algorithms is to find local optima, the phrase “global convergence” is often used. This phrase does *not* indicate convergence to a global optimum, but rather convergence to a local optimum from any starting point (not necessarily a point close to the solution). As I will explain soon, algorithms that exhibit local convergence (that is, convergence to a solution given a good starting point) must usually be modified if global convergence (that is, convergence from a possibly poor starting point) is to be obtained.

The definition of local minimizer is of little direct use in recognizing a solution. The reader will recall that algorithms for nonlinear problems are usually iterative, and require a starting point $x^{(0)}$ (an initial estimate of the solution). Certainly it seems reasonable that if this starting point is actually the solution, the algorithm should be recognize that fact and not try to improve upon $x^{(0)}$. However, the definition of local minimizer requires that $f(x^{(0)})$ be compared with $f(x)$ for every point x in a neighborhood of $x^{(0)}$. Since a neighborhood contains infinitely many points, this is clearly impossible.¹

¹If f depends on one or even two variables, then the graph of f could be plotted and local minima could be identified, assuming one knew the region of the domain to examine. However, this strategy is of no use for functions of more than two variables.

Because local minima cannot be identified from the definition, *optimality conditions* are absolutely essential in numerical optimization. Optimality conditions come in two varieties. A *necessary condition* must be satisfied by any solution, but it does not guarantee that a point satisfying it is a solution (that is, a necessary condition can be satisfied by a nonsolution). If a point satisfies a *sufficient condition*, on the other hand, it is guaranteed to be a solution.

2 The first-order necessary condition

When discussing local approximations to nonlinear functions, I mentioned in passing the most fundamental optimality condition for unconstrained minimization ($\nabla f(x^*) = 0$). This condition is easily proved.

Theorem 2.1 *Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ has a local minimum at $x = x^*$. If f is differentiable at x^* , then*

$$\nabla f(x^*) = 0.$$

Proof: Since f is differentiable at x^* , it follows that

$$f(x) = f(x^*) + \nabla f(x^*) \cdot (x - x^*) + o(\|x - x^*\|). \quad (1)$$

Suppose $\nabla f(x^*) \neq 0$ and define

$$x_h = x^* - h\nabla f(x^*).$$

Then (1), with $x = x_h$, becomes

$$f(x_h) = f(x^*) - h\nabla f(x^*) \cdot \nabla f(x^*) + o(h).$$

Since $\nabla f(x^*) \neq 0$, it follows that

$$\nabla f(x^*) \cdot \nabla f(x^*) = \|\nabla f(x^*)\|^2 > 0.$$

Since the error term represented by $o(h)$ goes to zero faster than h , it follows that

$$|o(h)| < h\|\nabla f(x^*)\|^2 \text{ for all } h \text{ sufficiently small,}$$

which implies that

$$f(x_h) < f(x^*) \text{ for all } h \text{ sufficiently small.}$$

This contradicts the assumption that x^* is a local minimizer, and shows that $\nabla f(x^*) \neq 0$ is impossible if x^* is a local minimizer. QED

A point x^* satisfying $\nabla f(x^*) = 0$ is called a *stationary point* of f .

The proof of the above theorem shows that, if $\nabla f(x^*) \neq 0$, then the vector $-\nabla f(x^*)$ defines a direction from x^* in which f decreases. To be precise, the vector $p = -\nabla f(x^*)$ satisfies

$$f(x^* + hp) < f(x^*) \text{ for all } h > 0 \text{ sufficiently small.}$$

An examination of the above proof shows that the same is true of any vector p such that $p \cdot \nabla f(x^*) < 0$. Such a vector p is called a *descent direction* for f at x^* . The concept of descent direction is fundamental in defining algorithms for minimization.

It is important to notice that the above first-order condition is only necessary. If x^* is a local minimizer of f , then x^* is a stationary point. However, the fact that x^* is a stationary point does *not* guarantee that x^* is a local minimizer of f . The following simple examples demonstrate this.

Example 2.2 *If $f : \mathbb{R} \rightarrow \mathbb{R}$ is defined by $f(x) = -x^2$, then $f'(0) = 0$ but 0 is a local maximizer of f . If $g : \mathbb{R} \rightarrow \mathbb{R}$ is defined by $g(x) = x^3$, then $g'(0) = 0$ but 0 is neither a local minimizer nor a local maximizer.*

Analogous examples in two variables are $f(x) = -x_1^2 - x_2^2$ and $g(x) = x_1^2 - x_2^2$.

There is no first-order sufficient condition for general nonlinear problems (although I discuss a special case below in which the first-order necessary condition is also sufficient). To obtain a sufficient condition, the curvature of the function must be taken into account, which means examining the second derivative.

3 Second-order conditions

3.1 The second-order necessary condition

If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is twice differentiable at x^* and x^* is a local minimizer of f , then

$$f(x) = f(x^*) + \nabla f(x^*) \cdot (x - x^*) + \frac{1}{2}(x - x^*) \cdot \nabla^2 f(x^*)(x - x^*) + o(\|x - x^*\|^2).$$

Writing $x = x^* + hy$, where h is a scalar and y is a vector, and applying the first-order necessary condition yields

$$f(x^* + hy) = f(x^*) + \frac{h^2}{2}y \cdot \nabla^2 f(x^*)y + o(h^2).$$

Since $o(h^2)$ is negligible compared to

$$\frac{h^2}{2}y \cdot \nabla^2 f(x^*)y$$

when h is sufficiently small,

$$f(x^* + hy) \geq f(x^*) \text{ for all } h \text{ sufficiently small}$$

implies that

$$\frac{h^2}{2}y \cdot \nabla^2 f(x^*)y \geq 0 \text{ for all } h \text{ sufficiently small.}$$

Dividing through by $h^2/2$ yields

$$y \cdot \nabla^2 f(x^*)y \geq 0,$$

and this must hold for all $y \in \mathbb{R}^n$ if x^* is a local minimizer of f . Since $\nabla^2 f(x^*)$ is a symmetric matrix, the following standard definition is relevant.

Definition 3.1 Suppose $A \in \mathbb{R}^{n \times n}$ is symmetric. If

$$y \cdot Ay \geq 0 \text{ for all } y \in \mathbb{R}^n,$$

then A is said to be positive semidefinite. If

$$y \cdot Ay > 0 \text{ for all } y \in \mathbb{R}^n, y \neq 0,$$

then A is said to be positive definite.

The result I proved above is called the second-order necessary condition for a local minimizer.

Theorem 3.2 Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is twice differentiable at x^* and x^* is a local minimizer of f . Then $\nabla^2 f(x^*)$ is positive semidefinite.

As the following example shows, the second-order necessary condition is not sufficient.

Example 3.3 Define $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ by $f(x) = x_1^4 - x_2^4$. Then

$$\begin{aligned} \nabla f(x) &= \begin{bmatrix} 4x_1^3 \\ -4x_2^3 \end{bmatrix}, \\ \nabla^2 f(x) &= \begin{bmatrix} 12x_1^2 & 0 \\ 0 & -12x_2^2 \end{bmatrix}, \end{aligned}$$

so

$$\begin{aligned} \nabla f(0) &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \\ \nabla^2 f(0) &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

Since the zero matrix is certainly positive semidefinite, both the first- and second-order necessary conditions hold. However, f does not have a local minimum at $x = 0$, since $f(x) < f(0) = 0$ for every x of the form $x = (0, x_2)$.

However, a slightly stronger second-order condition is sufficient.

3.2 The second-order sufficient condition

Theorem 3.4 Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is twice differentiable at x^* , $\nabla f(x^*) = 0$, and $\nabla^2 f(x^*)$ is positive definite. Then x^* is a strict local minimizer of f .

Proof: The function

$$y \mapsto y \cdot \nabla^2 f(x^*)y$$

is continuous and positive over the compact set $\{y \in \mathbb{R}^n : \|y\| = 1\}$. Therefore, there exists $\alpha > 0$ such that

$$\|y\| = 1 \Rightarrow y \cdot \nabla^2 f(x^*)y \geq \alpha.$$

Since f is twice differentiable at x^* and $\nabla f(x^*) = 0$ by assumption,

$$f(x^* + hy) = f(x^*) + \frac{h^2}{2}y \cdot \nabla^2 f(x^*)y + o(h^2). \quad (2)$$

By definition,

$$|o(h^2)| < \frac{h^2}{2}\alpha \text{ for all } h \text{ sufficiently small.}$$

Hence there exists $h_0 > 0$ such that

$$0 < h < h_0, \|y\| = 1 \Rightarrow \frac{h^2}{2}y \cdot \nabla^2 f(x^*)y + o(h^2) > 0.$$

Together with (2), this shows that $f(x^* + hy) > f(x^*)$ for all h, y with $0 < h < h_0, \|y\| = 1$. Since

$$\{x^* + hy : 0 < h < h_0, \|y\| = 1\} = \{x : 0 < \|x - x^*\| < h_0\},$$

it follows that

$$0 < \|x - x^*\| < h_0 \Rightarrow f(x^*) < f(x)$$

and hence that x^* is a strict local minimizer of f . QED

The above results can be summarized as follows:

1. If x^* is a local minimizer of f , then

$$\nabla f(x^*) = 0 \text{ and } \nabla^2 f(x^*) \text{ is positive semidefinite.}$$

2. If

$$\nabla f(x^*) = 0 \text{ and } \nabla^2 f(x^*) \text{ is positive definite,}$$

then x^* is a strict local minimizer of f .

It is important to realize that the sufficient condition is not necessary; that is, x^* can be a local minimizer of f (even a strict local minimizer of f) and yet $\nabla^2 f(x^*)$ not be positive definite.

Example 3.5 Define $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ by $f(x) = x_1^2 + x_2^4$. Clearly $x^* = 0$ is a strict local minimizer of f . A simple calculation yields

$$\begin{aligned} \nabla f(x) &= \begin{bmatrix} 2x_1 \\ 4x_2^3 \end{bmatrix}, \\ \nabla^2 f(x) &= \begin{bmatrix} 2 & 0 \\ 0 & 12x_2^2 \end{bmatrix}, \end{aligned}$$

so

$$\begin{aligned} \nabla f(0) &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \\ \nabla^2 f(0) &= \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

For any $y \in \mathbb{R}^2$,

$$y \cdot \nabla^2 f(0)y = 2y_1^2 \geq 0,$$

which shows that the first- and second-order necessary conditions are indeed satisfied. However,

$$y = (0, 1) \neq 0 \text{ and } y \cdot \nabla^2 f(0)y = 0.$$

This shows that $\nabla^2 f(0)$ is not positive definite, and so the second-order sufficient condition fails.

4 Convex functions

I will now discuss a class of functions f for which the first-order necessary condition $\nabla f(x^*) = 0$ is both necessary and sufficient, namely, the class of convex functions.

Definition 4.1 Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$. If

$$x^{(1)}, x^{(2)} \in \mathbb{R}^n, \alpha_1, \alpha_2 \geq 0, \alpha_1 + \alpha_2 = 1 \Rightarrow f(\alpha_1 x^{(1)} + \alpha_2 x^{(2)}) \leq \alpha_1 f(x^{(1)}) + \alpha_2 f(x^{(2)}),$$

then f is said to be convex. If

$$x^{(1)}, x^{(2)} \in \mathbb{R}^n, \alpha_1, \alpha_2 > 0, \alpha_1 + \alpha_2 = 1 \Rightarrow f(\alpha_1 x^{(1)} + \alpha_2 x^{(2)}) < \alpha_1 f(x^{(1)}) + \alpha_2 f(x^{(2)}),$$

then f is said to be strictly convex.

A little geometry makes the above definition easy to understand. Given two points $x^{(1)}, x^{(2)}$ in \mathbb{R}^n ,

$$\left\{ \alpha_1 x^{(1)} + \alpha_2 x^{(2)} : \alpha_1, \alpha_2 \geq 0, \alpha_1 + \alpha_2 = 1 \right\}$$

is the line segment with endpoints $x^{(1)}, x^{(2)}$, and points of the form

$$\left(\alpha_1 x^{(1)} + \alpha_2 x^{(2)}, f(\alpha_1 x^{(1)} + \alpha_2 x^{(2)}) \right)$$

comprise the *secant line* through the points $(x^{(1)}, f(x^{(1)})), (x^{(2)}, f(x^{(2)}))$ on the graph of f . A convex function “curves up” in such a way that the secant line lies (on or) above the graph of the function. Figure 1 shows an example.

Here is a fundamental property of smooth convex functions.

Theorem 4.2 Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex. If f is differentiable at $a \in \mathbb{R}^n$, then the following inequality holds:

$$f(x) \geq f(a) + \nabla f(a) \cdot (x - a) \text{ for all } x \in \mathbb{R}^n.$$

If f is strictly convex, then the inequality is strict for $x \neq a$. Thus the graph of a convex function f lies on or above the plane tangent to the graph of f at $(a, f(a))$.

Proof: I begin with the case that f is convex I suppose, by way of contradiction, that there exists $b \in \mathbb{R}^n$ such that

$$f(b) < f(a) + \nabla f(a) \cdot (b - a). \quad (3)$$

The convexity of f implies the following inequality:

$$f((1 - \alpha)a + \alpha b) \leq (1 - \alpha)f(a) + \alpha f(b) \Rightarrow f(a + \alpha(b - a)) \leq f(a) + \alpha(f(b) - f(a)).$$

I define $\phi : \mathbb{R} \rightarrow \mathbb{R}$ by

$$\phi(\alpha) = f(a + \alpha(b - a)).$$

Then, by the chain rule,

$$\phi'(0) = \nabla f(a) \cdot (b - a).$$

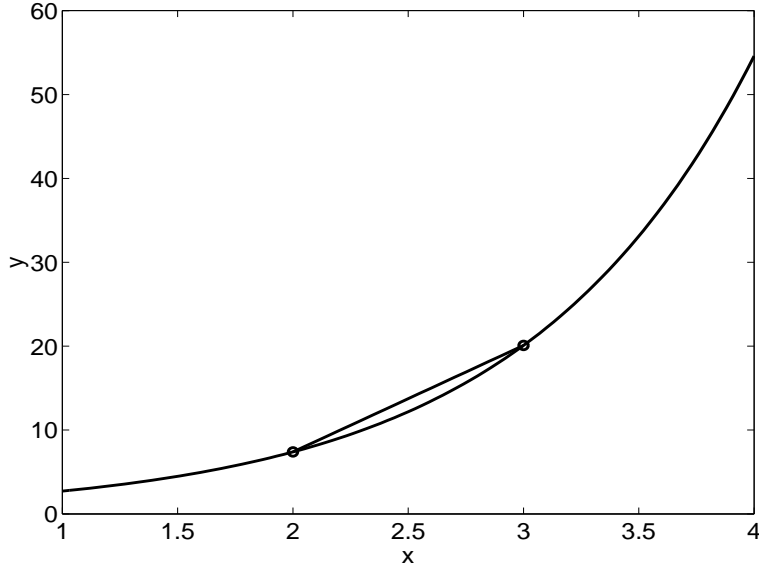


Figure 1: The convex function $f(x) = e^x$ and a secant line.

I can also estimate $\phi'(0)$ as follows:

$$\begin{aligned}
 \phi'(0) &= \lim_{\alpha \rightarrow 0^+} \frac{\phi(\alpha) - \phi(0)}{\alpha} \\
 &= \lim_{\alpha \rightarrow 0^+} \frac{f(a + \alpha(b - a)) - f(a)}{\alpha} \\
 &\leq \lim_{\alpha \rightarrow 0^+} \frac{f(a) + \alpha(f(b) - f(a)) - f(a)}{\alpha} \\
 &= \lim_{\alpha \rightarrow 0^+} (f(b) - f(a)) \\
 &= f(b) - f(a).
 \end{aligned}$$

Since $\phi'(0) = \nabla f(a) \cdot (b - a)$, I obtain

$$\nabla f(a) \cdot (b - a) \leq f(b) - f(a).$$

But (3) implies that

$$f(b) - f(a) < \nabla f(a) \cdot (b - a).$$

This contradiction completes the proof of the first result.

Now I suppose that f is strictly convex. The convexity of f implies that

$$\alpha \in [0, 1] \Rightarrow f(a) + \alpha \nabla f(a) \cdot (b - a) \leq f(a + \alpha(b - a)) \leq f(a) + \alpha(f(b) - f(a)).$$

However, $\alpha \mapsto f(a) + \alpha \nabla f(a) \cdot (b - a)$ and $\alpha \mapsto f(a) + \alpha(f(b) - f(a))$ are two affine functions of α agreeing at $\alpha = 0$. If $f(b) = f(a) + \nabla f(a) \cdot (b - a)$, then the two affine functions also agree at $\alpha = 1$ and hence must be equal for all α . But then

$$\alpha \in [0, 1] \Rightarrow f(a) + \alpha \nabla f(a) \cdot (b - a) = f(a + \alpha(b - a)) = f(a) + \alpha(f(b) - f(a)),$$

which contradicts the strict convexity of f . QED

The converse of Theorem 4.2 is also true. I will prove it in the next section.

I can now prove the main result of this section.

Theorem 4.3 *Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex. If f is differentiable at $x^* \in \mathbb{R}^n$ and $\nabla f(x^*) = 0$, then f has a global minimizer at x^* . If f is strictly convex, then x^* is the unique global minimizer of f .*

Proof: By the previous theorem,

$$f(x) \geq f(x^*) + \nabla f(x^*) \cdot (x - x^*) \text{ for all } x \in \mathbb{R}^n.$$

If $\nabla f(x^*) = 0$, then this reduces to

$$f(x) \geq f(x^*) \text{ for all } x \in \mathbb{R}^n,$$

which shows that x^* is a global minimizer of f . If f is strictly convex, then the inequalities above are strict, and x^* is the only global minimizer of f . QED

4.1 Sufficient conditions for convexity

The first sufficient condition I present is the converse of Theorem 4.2.

Theorem 4.4 *Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable everywhere and*

$$f(x) \geq f(a) + \nabla f(a) \cdot (x - a) \text{ for all } a, x \in \mathbb{R}^n. \quad (4)$$

Then f is convex. If the inequality in (4) is strict for $x \neq a$, then f is strictly convex.

Proof: I suppose, by way of contradiction, that (4) holds but f is not convex. Then there exist $a, b \in \mathbb{R}^n$ and $\theta \in (0, 1)$ such that

$$f((1 - \theta)a + \theta b) > (1 - \theta)f(a) + \theta f(b).$$

By assumption, with $c = (1 - \theta)a + \theta b$,

$$\begin{aligned} f(a) &\geq f(c) + \nabla f(c) \cdot (a - c), \\ f(b) &\geq f(c) + \nabla f(c) \cdot (b - c) \end{aligned}$$

It follows that

$$\begin{aligned} f(c) &> (1 - \theta)f(a) + \theta f(b) \\ &\geq (1 - \theta)(f(c) + \nabla f(c) \cdot (a - c)) + \theta(f(c) + \nabla f(c) \cdot (b - c)) \\ &= (1 - \theta + \theta)f(c) + \nabla f(c) \cdot ((1 - \theta)a + \theta b - (1 - \theta + \theta)c) \\ &= f(c) + \nabla f(c) \cdot (c - c) \\ &= f(c). \end{aligned}$$

Since $f(c) > f(c)$ is impossible, it must be the case that f is convex. The proof of the second part of the theorem is left to the reader. QED

If f is twice-differentiable, then the second derivative of f reflects the curvature of the graph of f . I want to show that f is convex if $\nabla^2 f$ is everywhere positive semidefinite and f is strictly convex if $\nabla^2 f$ is everywhere positive definite. To do this, I need Taylor's theorem. The reader will recall that, if $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is twice continuously differentiable, then, given any $\alpha, \beta \in \mathbb{R}$, there exists γ between α and β such that

$$\phi(\beta) = \phi(\alpha) + \phi'(\alpha)(\beta - \alpha) + \frac{\phi''(\gamma)}{2}(\beta - \alpha)^2.$$

This is a consequence of Taylor's theorem for functions of one variable; it give an explicit form of the error in the local linear approximation (although the number γ is generally unknown).

I can use Taylor's theorem for functions of one variable to derive the analogous result for a function of n variables. Given a twice continuously differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and any $a, x \in \mathbb{R}^n$, I define $\phi : \mathbb{R} \rightarrow \mathbb{R}$ by

$$\phi(\alpha) = f(a + \alpha(x - a)).$$

The derivatives of ϕ are

$$\begin{aligned}\phi'(\alpha) &= \nabla f(a + \alpha(x - a)) \cdot (x - a), \\ \phi''(\alpha) &= (x - a) \cdot \nabla^2 f(a + \alpha(x - a))(x - a).\end{aligned}$$

Therefore, applying Taylor's theorem to ϕ yields, for some γ between α and β ,

$$\phi(\beta) = \phi(\alpha) + (\beta - \alpha)\phi'(\alpha) + \frac{\phi''(\gamma)}{2}(\beta - \alpha)^2$$

or

$$f(a + \beta(x - a)) = f(a + \alpha(x - a)) + (\beta - \alpha)\nabla f(a + \alpha(x - a)) \cdot (x - a) + \frac{(\beta - \alpha)^2}{2}(x - a) \cdot \nabla^2 f(a + \gamma(x - a))(x - a).$$

Taking $\alpha = 0$, $\beta = 1$ yields

$$f(x) = f(a) + \nabla f(a) \cdot (x - a) + \frac{1}{2}(x - a) \cdot \nabla^2 f(a + \gamma(x - a))(x - a),$$

where γ is between 0 and 1. This proves the following theorem.

Theorem 4.5 *If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is twice continuously differentiable, then, for any $a, x \in \mathbb{R}^n$, there exists $\gamma \in [0, 1]$ such that*

$$f(x) = f(a) + \nabla f(a) \cdot (x - a) + \frac{1}{2}(x - a) \cdot \nabla^2 f(a + \gamma(x - a))(x - a).$$

I can now prove the following results.

Theorem 4.6 *If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is twice continuously differentiable and $\nabla^2 f$ is everywhere positive semidefinite, then f is convex. If $\nabla^2 f$ is everywhere positive definite, then f is strictly convex.*

Proof: If $\nabla^2 f$ is everywhere positive semidefinite, then, for any $a, x \in \mathbb{R}^n$, there exists $\gamma \in [0, 1]$ such that

$$f(x) = f(a) + \nabla f(a) \cdot (x - a) + \frac{1}{2}(x - a) \cdot \nabla^2 f(a + \gamma(x - a))(x - a).$$

Since $\nabla^2 f(a + \gamma(x - a))$ is positive semidefinite,

$$(x - a) \cdot \nabla^2 f(a + \gamma(x - a))(x - a) \geq 0,$$

which yields

$$f(x) \geq f(a) + \nabla f(a) \cdot (x - a).$$

This holds for all $a, x \in \mathbb{R}^n$, so by Theorem 4.4, it follows that f is convex. If $\nabla^2 f$ is everywhere positive definite, then the two previous inequalities are strict, which implies (again by Theorem 4.4) that f is strictly convex. QED