CS/MATH 3414 — Numerical Methods

Lecture 10
Optimization, Part 1

Virginia Tech Department of Computer Science, Fall 2026
Instructor: Sumeet Khatri
Date: October 27, 2026

1 Optimization problems **1**
2 Single-variable optimization **3**
2.1 Using root-finding to locate stationary points . . . 6
3 Multi-variable optimization **6**
3.1 Classifying stationary points . . . 9
3.2 Convexity and global conclusions . . . 12
4 Gradient descent **17**
4.1 Step size, convergence, and conditioning . . . 19
4.2 Quadratic functions, eigenvalues, and slow directions . . . 26
4.3 Exact line search and backtracking . . . 30

Learning Objectives

- Formulate unconstrained minimization and maximization problems in one or several variables.
- State conditions for optimality in terms of stationary points and second derivatives/Hessians.
- Understand gradient descent and explain the effects of step size, curvature, and conditioning.

Primary References

[1] W. Cheney and D. Kincaid, *Numerical Mathematics and Computing*, 7th ed., Section 13.2: Multivariate Case, pp. 573–586 (Brooks/Cole, Cengage Learning, Boston, 2013).

[2] S. Boyd and L. Vandenberghe, *Convex Optimization*, Sections 3.1 and 9.1–9.4: convexity, unconstrained minimization, and descent methods (Cambridge University Press, Cambridge, 2004), https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf (visited on 10/23/2026).

[3] S. Boyd, L. Vandenberghe, and P. Nobel, *Convex Optimization: Unconstrained Minimization*, Revised lecture slides, Chapter 9; especially descent methods, gradient descent, and choice of norm, Stanford University, https://web.stanford.edu/class/ee364a/lectures/unconstrained.pdf (visited on 10/23/2026).

[4] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer Series in Operations Research and Financial Engineering, Chapters 3 and 11: line searches and nonlinear equations (Springer, New York, 2006).

[5] L. Vandenberghe, *Gradient Method*, ECE236C: Optimization Methods for Large-Scale Systems, Lecture 1, Spring 2022, University of California, Los Angeles, (2022) https://www.seas.ucla.edu/~vandenbe/236C/lectures/gradient.pdf (visited on 10/26/2026).

1 Optimization problems

Many numerical problems ask for a *best* choice rather than an exact solution of an equation. We might choose model parameters that minimize disagreement with observations, a configuration that minimizes potential

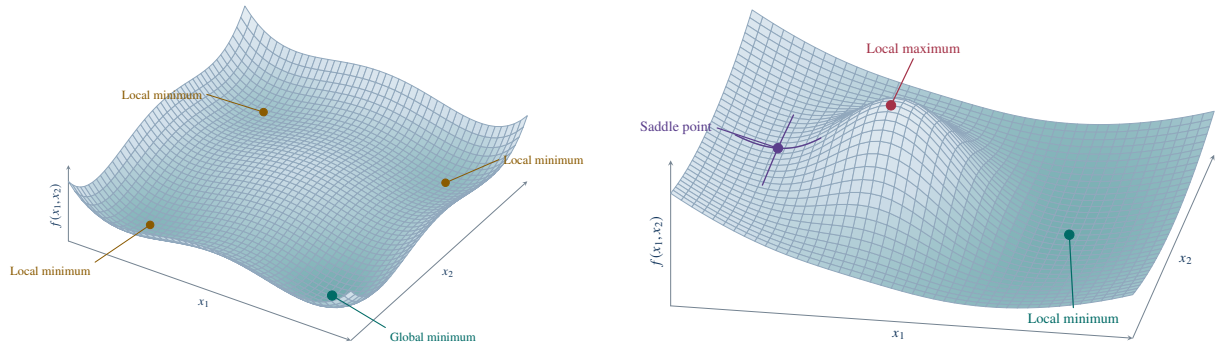


FIGURE 1: A function f can be visualized as a surface (when $n = 2$). In general, the function will have several local maxima and minima. It can also have so-called *saddle points*, points that are neither a maximum nor a minimum.

energy, or a decision that maximizes a numerical measure of performance. The first step is to identify the variables we can change and a scalar quantity that measures their quality. The problem can then often be phrased as an *unconstrained minimization*:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}). \tag{1.1}$$

The function $f : \mathbb{R}^n \to \mathbb{R}$ is the *objective function*, and $\boldsymbol{x}$ contains the decision variables. The case $n = 1$ concerns a function of a single variable. "Unconstrained" means that we impose no additional equations or inequalities restricting the variables. (We will consider constrained optimization problems in the next lecture.) We will principally consider differentiable objectives; whenever second derivatives are used, the required smoothness will be stated.

A maximum can be found by minimizing the negative objective:

$$\arg\max_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}) = \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \big(-f(\boldsymbol{x})\big). \tag{1.2}$$

Here arg min denotes the set of inputs attaining the smallest objective value, whereas min denotes that value. Either set of optimizers can be empty.

Definition 1 (Local and global minimizers)**.** A point $\boldsymbol{x}_*$ is a *global minimizer* if $f(\boldsymbol{x}_*) \leq f(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathbb{R}^n$. It is a *local minimizer* if there is a number $r > 0$ such that the same inequality holds whenever $\|\boldsymbol{x} - \boldsymbol{x}_*\| < r$. A local minimum is *strict* if the inequality is strict for every other point in that neighborhood.

See Figure 1 for a depiction of local maxima and minima. A local minimum is best only among nearby inputs. Consequently, a numerical method that takes small improving steps can settle at a local minimum even when a better value is available elsewhere. Moreover, an unconstrained problem need not possess a minimizer. The function $f(x) = -x^2$ is unbounded below. The function $f(x) = \mathrm{e}^x$ is bounded below by zero, but never attains zero; its infimum is zero and it has no minimum on $\mathbb{R}$.

Optimization and root finding. The equation $f(x) = 0$ asks where a function vanishes; minimizing f asks where it is smallest. These are different questions. Nevertheless, optimization and root-finding are closely connected. To minimize a differentiable function, we will search for zeros of its derivative or gradient.

2 Single-variable optimization

A single-variable optimization problem is of the form

$$\min_{x \in \mathbb{R}} f(x), \tag{2.1}$$

where $f : \mathbb{R} \to \mathbb{R}$ is the objective function and $x \in \mathbb{R}$ is the variable being optimized.

Theorem 2 (Single-variable optimality tests)**.** Let f be twice continuously differentiable near x_*.

1. If x_* is a local minimum, then $f'(x_*) = 0$ and $f''(x_*) \geq 0$.
2. If $f'(x_*) = 0$ and $f''(x_*) > 0$, then x_* is a strict local minimum.
3. If $f'(x_*) = 0$ and $f''(x_*) < 0$, then x_* is a strict local maximum.

A point x_* where $f'(x_*) = 0$ is called a *stationary point*.

The second-derivative tests in items 2. and 3. of Theorem 1 follow from the Taylor expansion:

$$f(x_* + h) = f(x_*) + f'(x_*)h + \frac{1}{2}f''(x_*)h^2 + o(h^2), \tag{2.2}$$

where $h > 0$. At a stationary point, the linear term vanishes. If the second derivative is positive or negative, the quadratic term determines the sign of the change for sufficiently small nonzero h. If $f''(x_*) = 0$, then the second-derivative test is inconclusive. At zero, the functions x^4, $-x^4$, and x^3 all have vanishing first and second derivatives, but exhibit a strict minimum, a strict maximum, and neither, respectively.

Example 3 (A function with two global minima)**.** Consider

$$f(x) = (x^2 - 1)^2, \qquad f'(x) = 4x(x^2 - 1), \qquad f''(x) = 12x^2 - 4. \tag{2.3}$$

The stationary points are -1, 0, and 1. Since $f''(\pm 1) = 8 > 0$, the points ± 1 are strict local minima. Since $f(x) \geq 0$ everywhere and $f(\pm 1) = 0$, they are also global minima. At zero, $f''(0) = -4$, so zero is a strict local maximum. There is no global maximum because $f(x) \to +\infty$ as $|x| \to \infty$.
This example separates three tasks: locate stationary points, classify them locally, and determine whether a local conclusion is global. A derivative-based iteration started exactly at zero can stop there even though zero is not a minimizer. Figure 2 shows the two valleys separated by this stationary maximum.

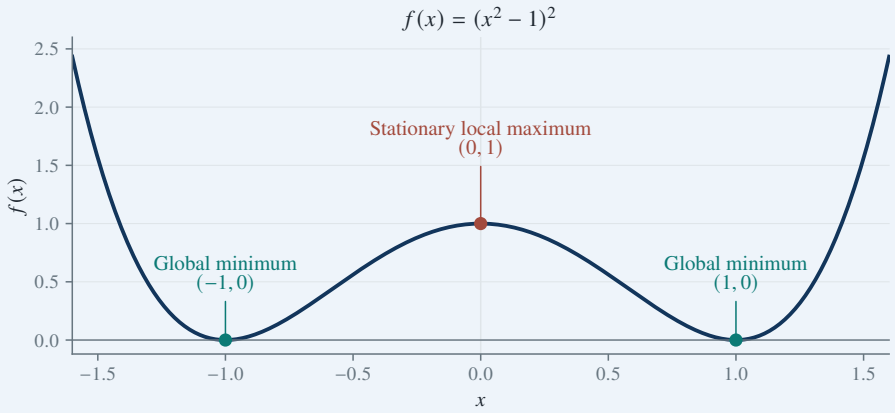


FIGURE 2: The objective $f(x) = (x^2 - 1)^2$ has two global minima and a stationary local maximum. Stationarity alone does not establish optimality.

Example 4 (Stationary points and inflection points)**.** Consider

$$f(x) = x^3, \qquad f'(x) = 3x^2, \qquad f''(x) = 6x, \qquad f'''(x) = 6. \tag{2.4}$$

The only stationary point is $x = 0$. Since $f''(0) = 0$, the second-derivative test is inconclusive. Nevertheless, for every $h > 0$,

$$f(-h) = -h^3 < 0 = f(0) < h^3 = f(h). \tag{2.5}$$

Thus, arbitrarily close to zero, there are points with both smaller and larger function values. Consequently, zero is neither a local minimum nor a local maximum.

An *inflection point* is a point where the concavity changes. Here, $f''(x) < 0$ for $x < 0$ and $f''(x) > 0$ for $x > 0$, so the graph changes from concave down to concave up at zero. Because $f'(0) = 0$, this is a *stationary inflection point*. The nonzero third derivative $f'''(0) = 6$ corresponds to the cubic term, which changes sign across zero.

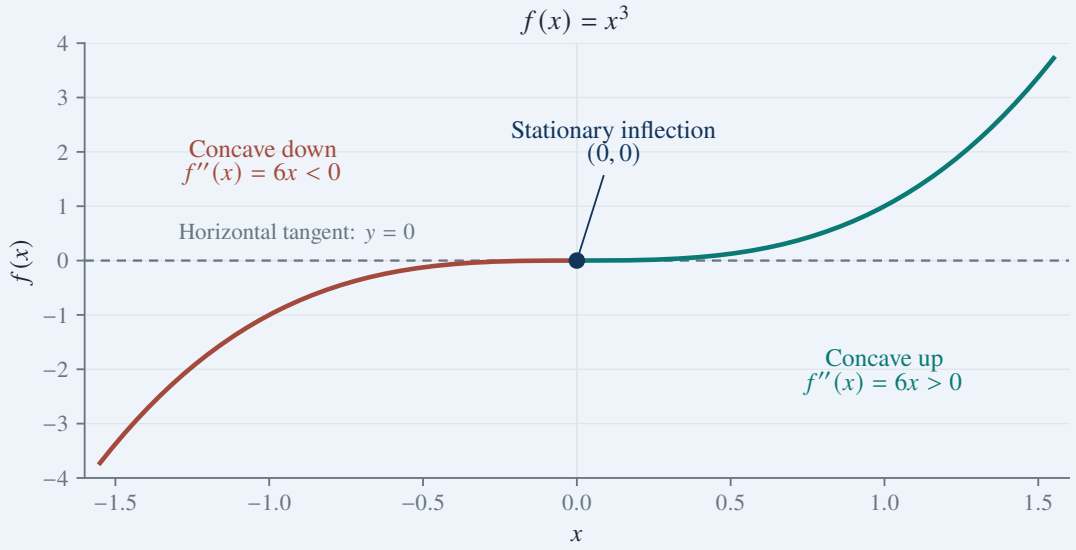


FIGURE 3: The stationary inflection point of $f(x) = x^3$. The tangent at the origin is horizontal, but the function continues increasing through the point as its concavity changes.

A quartic with several types of behavior. Now consider

$$\begin{aligned} g(x) &= x^4 - 4x^3, & g'(x) &= 4x^2(x-3), \\ g''(x) &= 12x(x-2), & g'''(x) &= 24(x-1). \end{aligned} \tag{2.6}$$

The stationary points are $x = 0$ and $x = 3$, but they have different classifications.
At zero, $g'(0) = g''(0) = 0$, while $g'''(0) = -24$. For $0 < h < 1$,

$$g(-h) = h^4 + 4h^3 > 0 = g(0) > h^4 - 4h^3 = g(h). \tag{2.7}$$

Therefore, zero is neither a local minimum nor a local maximum. Moreover, $g''(x) > 0$ for $x < 0$ and $g''(x) < 0$ for $0 < x < 2$, so the concavity changes from up to down at zero. Hence, zero is a stationary inflection point.
There is another inflection point at $x = 2$: the second derivative changes from negative on $(0, 2)$ to positive on $(2, \infty)$. However,

$$g(2) = -16, \qquad g'(2) = -16 \neq 0. \tag{2.8}$$

This inflection point is *not stationary*. An inflection point therefore need not have a horizontal tangent.
At the other stationary point, $x = 3$, we have $g''(3) = 36 > 0$, so the second-derivative test establishes a strict local minimum. In fact, this minimum is uniquely global, because

$$g(x) + 27 = (x-3)^2\big[(x+1)^2 + 2\big] \geq 0. \tag{2.9}$$

The second factor is strictly positive for every real x, so equality holds only at $x = 3$. Thus, $g(x) \geq -27$ everywhere, with equality only at the unique global minimizer.

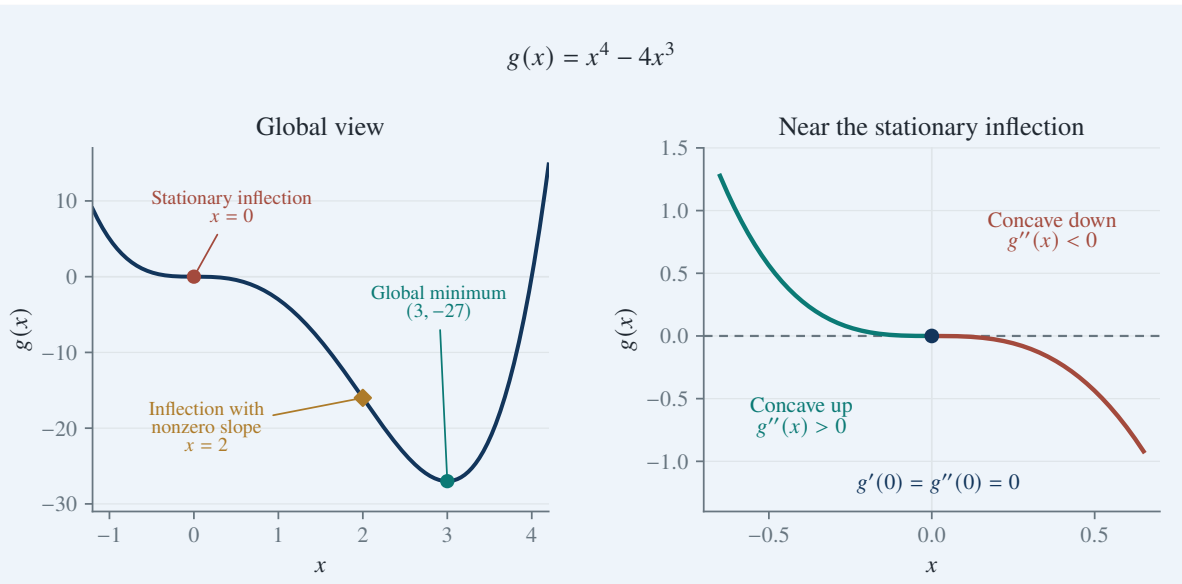


FIGURE 4: For $g(x) = x^4 - 4x^3$, zero is a stationary inflection point, $x = 2$ is an inflection point with nonzero slope, and $x = 3$ is the unique global minimizer. The enlarged view shows the flat crossing at the origin; the dashed line is its horizontal tangent.

2.1 Using root-finding to locate stationary points

Newton's method for a scalar equation $g(x) = 0$ gives $x_{k+1} = x_k - g(x_k)/g'(x_k)$, provided the denominator is nonzero. Choosing $g = f$ searches for a zero of the objective; choosing $g = f'$ searches for a stationary point:

$$\begin{aligned} \text{Solve } f(x) = 0: &\qquad x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \\ \text{Solve } f'(x) = 0: &\qquad x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}. \end{aligned} \tag{2.10}$$

The second update is Newton's method for optimization. It seeks a stationary point and does not, by itself, distinguish a minimum from a maximum. A zero second derivative makes the formula undefined, and negative curvature can produce a step in an undesirable direction for minimization. A sign-changing bracket for f' also locates a stationary point by bisection, but its classification still requires additional information. These observations motivate methods designed explicitly to decrease the objective.

3 Multi-variable optimization

For $f : \mathbb{R}^n \to \mathbb{R}$, the *gradient* vector and *Hessian* matrix collect its first and second derivatives:

$$\nabla f(\boldsymbol{x}) = \begin{pmatrix} \partial f/\partial x_1 \\ \vdots \\ \partial f/\partial x_n \end{pmatrix}, \qquad H(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x}) \right)_{i,j=1}^{n}. \tag{3.1}$$

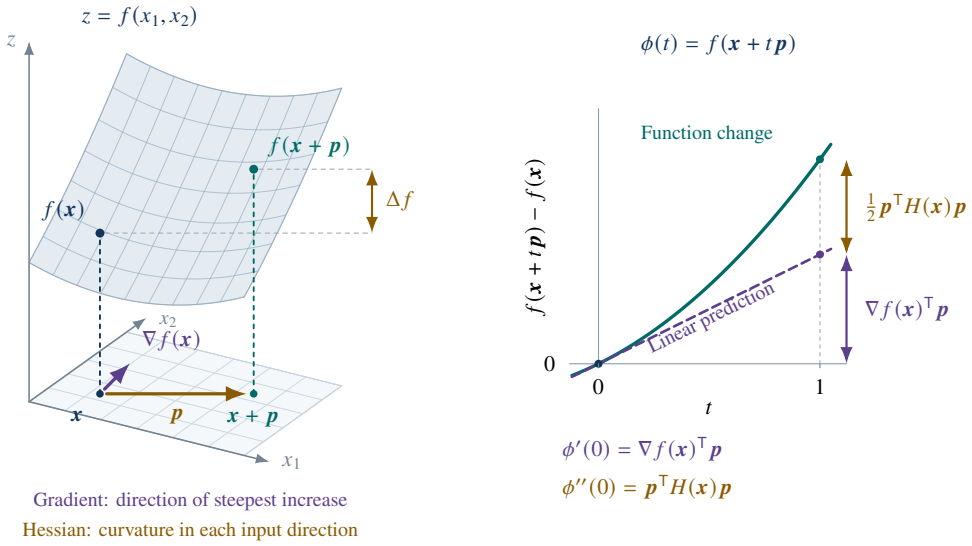


FIGURE 5: The gradient and Hessian generalize the concepts of the first and second derivatives, respectively, to multi-variable functions. The gradient describes the local rate of change, and the Hessian describes the curvature.

For twice continuously differentiable f, the Hessian is symmetric. As with single-variable functions that we saw ezrlier, the Taylor expansion describes how the multi-variable function f behaves near a point $\boldsymbol{x}$:

$$f(\boldsymbol{x} + \boldsymbol{p}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}^\top H(\boldsymbol{x}) \boldsymbol{p} + o(\|\boldsymbol{p}\|^2). \tag{3.2}$$

The gradient determines the linear change. The quadratic form involving the Hessian determines curvature in different directions, generalizing the scalar second derivative [1, Section 13.2, pp. 573–576]; see Figure 5 for a schematic depiction. These local models underlie first- and second-order optimization methods; see [2, Chapter 9] and [3].

As we are particularly interested in optimization, methods like gradient descent rely on starting at a particular value in the domain and "traversing the surface" in order to figure out where the minimum is. While a single-variable function has only one tangent at any given point, a multi-variable function generally as infinitely-many tangent lines at a point, each one pointing in a different direction. When optimizing, we would like to choose a direction that takes us towards the minimum. The gradient formalizes describes this idea, and provides mathematical justification for the gradient descent that we cover later in this lecture.

Lemma 5 (Gradient and directions of steepest increase and decrease)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Its gradient at $\boldsymbol{x} \in \mathbb{R}$, i.e., $\nabla f(\boldsymbol{x})$, points in the direction of steepest ascent at $\boldsymbol{x}$; the negative gradient, $-\nabla f(\boldsymbol{x})$, points in the direction of steepest descent.

Proof. Set $\boldsymbol{x} \in \mathbb{R}^n$. For a fixed unit vector $\boldsymbol{u} \in \mathbb{R}^n$, with $\|\boldsymbol{u}\|_2 = 1$, consider the function value

$$f(\boldsymbol{x} + t\boldsymbol{u}). \tag{3.3}$$

Here, $\boldsymbol{x}$ is the starting point, $\boldsymbol{u}$ specifies a direction, and $t \in \mathbb{R}$ is a *line parameter*. As t varies, $\boldsymbol{x} + t\boldsymbol{u}$ traces the straight line through $\boldsymbol{x}$ in direction $\boldsymbol{u}$; see Figure 6 for a depiction. The value $t = 0$ corresponds to $\boldsymbol{x}$.

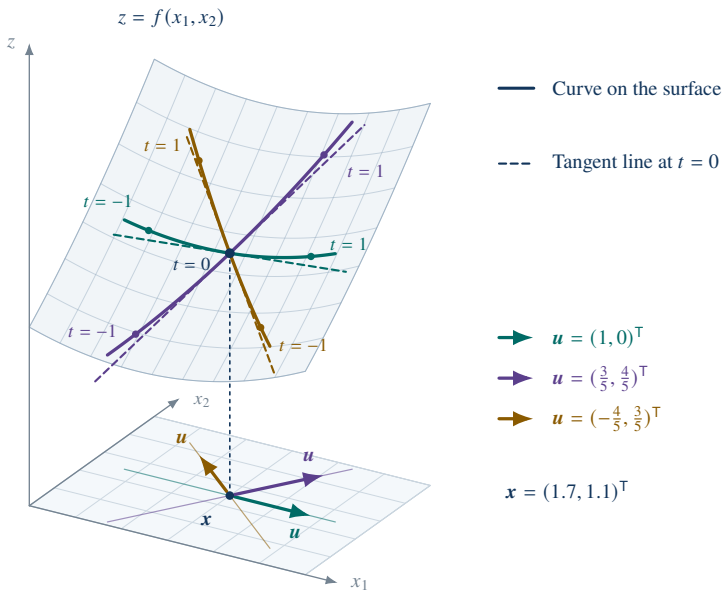


FIGURE 6: For any multi-variable function $f : \mathbb{R}^n \to \mathbb{R}$, at any given point $\boldsymbol{x} \in \mathbb{R}$ there are infinitely many tangents, each one pointing in a different direction. In the setting of optimization, particularly in the gradient descent algorithm, we pick the direction of steepest descent — that direction is given by the gradient of the function at that point.

Positive values of t locate points in direction $\boldsymbol{u}$, while negative values locate points in the opposite direction. Since $\boldsymbol{u}$ is a unit vector, the distance from $\boldsymbol{x}$ to $\boldsymbol{x} + t\boldsymbol{u}$ is

$$\|(\boldsymbol{x} + t\boldsymbol{u}) - \boldsymbol{x}\|_2 = \|t\boldsymbol{u}\|_2 = |t|. \tag{3.4}$$

The derivative with respect to t, evaluated at $t = 0$, measures the initial rate of change of f in direction $\boldsymbol{u}$. It satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x} + t\boldsymbol{u})\bigg|_{t=0} = \nabla f(\boldsymbol{x})^\mathsf{T} \boldsymbol{u}. \tag{3.5}$$

To see this, write $\boldsymbol{x} = (x_1, \ldots, x_n)^\mathsf{T}$ and $\boldsymbol{u} = (u_1, \ldots, u_n)^\mathsf{T}$. Both vectors are fixed while the line parameter t varies. The coordinates of the point $\boldsymbol{x} + t\boldsymbol{u}$ are $x_1 + tu_1, \ldots, x_n + tu_n$, so

$$f(\boldsymbol{x} + t\boldsymbol{u}) = f(x_1 + tu_1, \ldots, x_n + tu_n). \tag{3.6}$$

This is a composition: first, t determines a point in $\mathbb{R}^n$ through the map $t \mapsto \boldsymbol{x} + t\boldsymbol{u}$; then f assigns a function value to that point. The multi-variable chain rule accounts for the dependence of every input coordinate on t. For each coordinate, we multiply the partial derivative of f with respect to that coordinate by the rate at which the coordinate changes, and then sum these contributions. Here, the ith coordinate is $x_i + tu_i$, whose derivative with respect to t is u_i, because x_i and u_i are constants. At $t = 0$, the input point is $\boldsymbol{x}$. Therefore, the

chain rule gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x} + t\boldsymbol{u})\bigg|_{t=0} &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\boldsymbol{x}) \, \frac{\mathrm{d}}{\mathrm{d}t}(x_i + tu_i)\bigg|_{t=0} \\ &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\boldsymbol{x}) u_i \\ &= \nabla f(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{u}.\end{aligned} \tag{3.7}$$

If $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$, the rate in (3.5) is zero for every unit direction.

Now, by the Cauchy–Schwarz inequality and the assumption $\|\boldsymbol{u}\|_2 = 1$,

$$\left|\nabla f(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{u}\right| \leq \|\nabla f(\boldsymbol{x})\|_2 \|\boldsymbol{u}\|_2 = \|\nabla f(\boldsymbol{x})\|_2. \tag{3.8}$$

Consequently,

$$-\|\nabla f(\boldsymbol{x})\|_2 \leq \nabla f(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{u} \leq \|\nabla f(\boldsymbol{x})\|_2. \tag{3.9}$$

When $\nabla f(\boldsymbol{x}) \neq \boldsymbol{0}$, equality in the upper bound holds precisely when $\boldsymbol{u}$ points in the same direction as $\nabla f(\boldsymbol{x})$. The unit-length requirement then gives $\boldsymbol{u} = \nabla f(\boldsymbol{x}) / \|\nabla f(\boldsymbol{x})\|_2$. Equality in the lower bound holds precisely when $\boldsymbol{u}$ points in the opposite direction, giving $\boldsymbol{u} = -\nabla f(\boldsymbol{x}) / \|\nabla f(\boldsymbol{x})\|_2$. These are therefore the unique unit directions attaining the largest and smallest rates, respectively. When $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$, its inner product with every unit vector is zero, so all the directional rates vanish. ■

3.1 Classifying stationary points

For single-variable functions, Theorem 2 gives us necessary conditions on a local minimum of a function in terms of stationary points — points x_* that satisfy $f'(x_*) = 0$ — and points with either a positive or negative second derivative. We can do the analogous thing for multi-variable functions.

A real symmetric matrix H is *positive semidefinite* if $\boldsymbol{p}^{\mathsf{T}} H \boldsymbol{p} \geq 0$ for every $\boldsymbol{p}$, and *positive definite* if the inequality is strict for every nonzero $\boldsymbol{p}$. Equivalently, all its eigenvalues are non-negative or positive, respectively. Negative definiteness reverses the strict inequality; an indefinite matrix has both positive and negative eigenvalues.

Theorem 6 (Multi-variable optimality tests)**.** Suppose f is twice continuously differentiable near $\boldsymbol{x}_* \in \mathbb{R}^n$.

1. If $\boldsymbol{x}_*$ is a local minimum, then $\nabla f(\boldsymbol{x}_*) = \boldsymbol{0}$ and $H(\boldsymbol{x}_*)$ positive semi-definite.
2. If $\nabla f(\boldsymbol{x}_*) = \boldsymbol{0}$ and $H(\boldsymbol{x}_*)$ is positive definite, then $\boldsymbol{x}_*$ is a strict local minimum.
3. If $\nabla f(\boldsymbol{x}_*) = \boldsymbol{0}$ and $H(\boldsymbol{x}_*)$ is negative definite, then $\boldsymbol{x}_*$ is a strict local maximum.
4. If $\nabla f(\boldsymbol{x}_*) = \boldsymbol{0}$ and $H(\boldsymbol{x}_*)$ is indefinite, then $\boldsymbol{x}_*$ is a saddle point: arbitrarily nearby points have both larger and smaller objective values.

As with single-variable functions, a point $\boldsymbol{x}_* \in \mathbb{R}^n$ satisfying $\nabla f(\boldsymbol{x}_*) = \boldsymbol{0}$ is called a *stationary point*.

Proofs of the necessary and sufficient minimum conditions appear in [4, Theorems 2.2–2.4, pp. 14–16]. The maximum condition follows by applying these results to $-f$. The saddle condition follows from Taylor's expansion along eigenvectors with eigenvalues of opposite signs, as shown below.

For necessity, apply the one-variable tests to $t \mapsto f(\boldsymbol{x}_* + t\boldsymbol{p})$ for every direction $\boldsymbol{p}$. For sufficiency, positive definiteness gives a positive lower bound on curvature in every unit direction, so the quadratic term in (3.2) dominates the remainder.

What makes a stationary point a saddle? For a differentiable objective, a *saddle point* is a stationary point that is neither a local minimum nor a local maximum. More precisely, every neighborhood of $\boldsymbol{x}_*$ contains both a point with a smaller objective value and a point with a larger objective value. For every $r > 0$, there are points $\boldsymbol{x}_-$ and $\boldsymbol{x}_+$ such that

$$\|\boldsymbol{x}_- - \boldsymbol{x}_*\| < r, \qquad \|\boldsymbol{x}_+ - \boldsymbol{x}_*\| < r, \qquad f(\boldsymbol{x}_-) < f(\boldsymbol{x}_*) < f(\boldsymbol{x}_+). \tag{3.10}$$

The condition $\nabla f(\boldsymbol{x}_*) = \boldsymbol{0}$ means that all first-order directional slopes vanish. It does not mean that the function is locally constant: second-order or higher-order changes can have different signs in different directions.

The basic model is

$$f(x, y) = x^2 - y^2, \qquad \nabla f(x, y) = \begin{pmatrix} 2x \\ -2y \end{pmatrix}, \qquad H = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}. \tag{3.11}$$

At the origin the gradient vanishes. Along the x-axis, $f(t, 0) = t^2 > 0$ for $t \neq 0$, so the origin is a minimum of that one-dimensional restriction. Along the y-axis, $f(0, t) = -t^2 < 0$, so it is a maximum of the other restriction. Taking t arbitrarily small proves that the origin is a saddle of the full two-variable function. Figure 7 shows the upward and downward cross-sections through the same point. Along $y = x$ and $y = -x$, the value remains zero; a saddle need not increase or decrease along every direction.

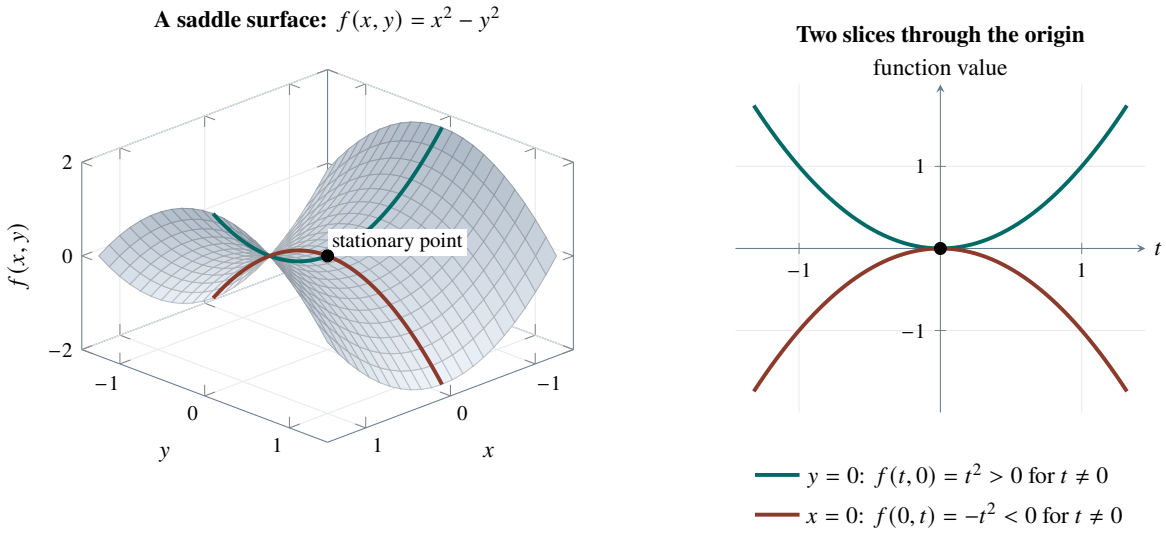


FIGURE 7: A saddle of $f(x, y) = x^2 - y^2$ at the origin. The x-axis cross-section bends upward, while the y-axis cross-section bends downward. Both have zero slope at the origin, but their nearby values lie on opposite sides of $f(0, 0) = 0$.

Curvature can reveal directions that the coordinate axes miss. If the Hessian at a stationary point is indefinite, choose unit eigenvectors $\boldsymbol{v}_+$ and $\boldsymbol{v}_-$ with eigenvalues $\lambda_+ > 0$ and $\lambda_- < 0$. Taylor's expansion along these two lines gives

$$f(\boldsymbol{x}_* + t\boldsymbol{v}_\pm) - f(\boldsymbol{x}_*) = \frac{1}{2}\lambda_\pm t^2 + o(t^2). \tag{3.12}$$

For sufficiently small nonzero t, the quadratic term determines the sign. One eigendirection therefore supplies larger values and the other supplies smaller values, proving the saddle conclusion in the theorem.

It is not enough to test only the original coordinate directions. For instance, $q(x, y) = x^2 + 4xy + y^2$ has positive curvature along both axes: $q(t, 0) = q(0, t) = t^2$. Nevertheless, along the diagonal $y = -x$ we have $q(t, -t) = -2t^2$. The origin is a saddle, and the mixed derivative is essential to detecting it. This is why the multivariable test uses the entire Hessian, rather than just its diagonal entries.

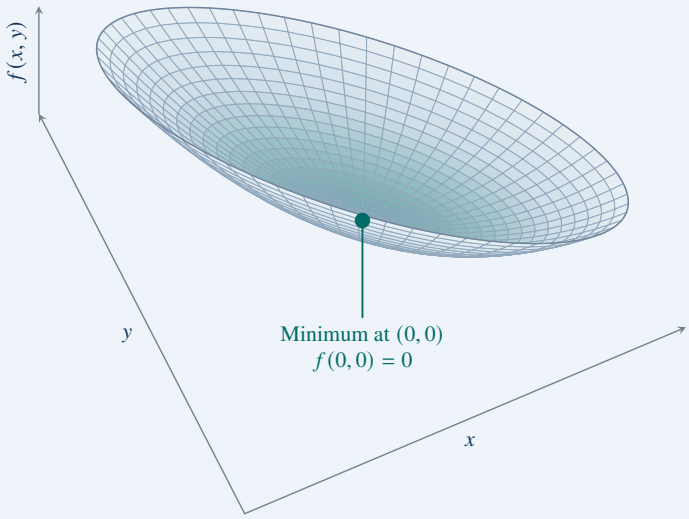
When the second-order test is inconclusive. Positive semidefiniteness allows zero eigenvalues and need not provide a positive lower bound on curvature. Higher-order terms can then decide the behavior. At the origin, $x^4 + y^4$ has a strict minimum, whereas $x^4 - y^4$ has a saddle; both have zero gradient and zero Hessian there. Thus, an indefinite Hessian is sufficient for a saddle but is not necessary. A semidefinite Hessian with a zero eigenvalue therefore requires further analysis of the objective.

Example 7 (Curvature in rotated directions)**.** Consider

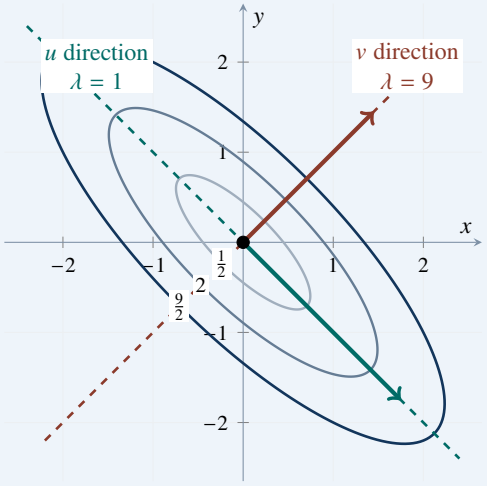
$$f(x, y) = \frac{1}{2}(5x^2 + 8xy + 5y^2), \qquad \nabla f(x, y) = \begin{pmatrix} 5x + 4y \\ 4x + 5y \end{pmatrix}, \qquad H = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}. \tag{3.13}$$

The eigenvectors $(1, -1)^\mathsf{T}$ and $(1, 1)^\mathsf{T}$ have eigenvalues 1 and 9. Introducing the orthonormal coordinates $u = (x - y)/\sqrt{2}$ and $v = (x + y)/\sqrt{2}$ gives $f = \frac{1}{2}(u^2 + 9v^2)$. The origin is the unique global minimizer. The contours are ellipses, with a shallower direction along $y = -x$ and a steeper direction along $y = x$. The curvature ratio is 9, a first indication that one step size must accommodate different rates of change in different directions.

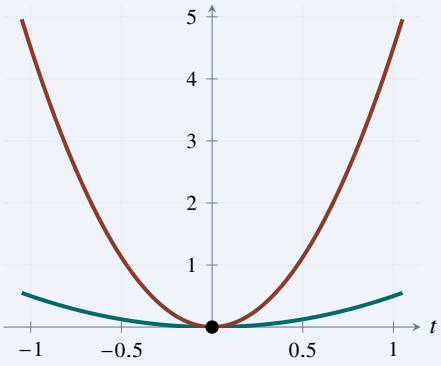
For a positive contour value c, the equation $u^2 + 9v^2 = 2c$ has semiaxis lengths $\sqrt{2c}$ and $\sqrt{2c}/3$. The contour therefore extends three times farther in the shallow direction. In Figure 8, the unit eigenvectors $\boldsymbol{e}_u = (1, -1)^\mathsf{T}/\sqrt{2}$ and $\boldsymbol{e}_v = (1, 1)^\mathsf{T}/\sqrt{2}$ also define the cross-sections $f(t\boldsymbol{e}_u) = t^2/2$ and $f(t\boldsymbol{e}_v) = 9t^2/2$. Both bend upward, unlike the saddle cross-sections, but the same displacement costs nine times as much in the steep direction.



Elliptical level curves



Curvature along unit directions



$f(te_u) = \frac{1}{2}t^2$

$f(te_v) = \frac{9}{2}t^2$

$e_u = (1,-1)^\top/\sqrt{2}, \quad e_v = (1,1)^\top/\sqrt{2}$

FIGURE 8: The rotated quadratic in Example 7. Its elliptical contours align with the Hessian eigenvectors. Cross-sections along the two unit eigenvectors show positive curvatures 1 and 9; the ellipse semiaxis ratio is 3.

3.2 Convexity and global conclusions

There are three separate questions about minimization: does a global minimizer exist, is it unique, and must a local minimizer be global? Convexity, strict convexity, and strong convexity give different answers. Throughout this subsection, the domain is all of $\mathbb{R}^n$, and $f : \mathbb{R}^n \to \mathbb{R}$ is real-valued. This domain assumption will be important for the existence result.

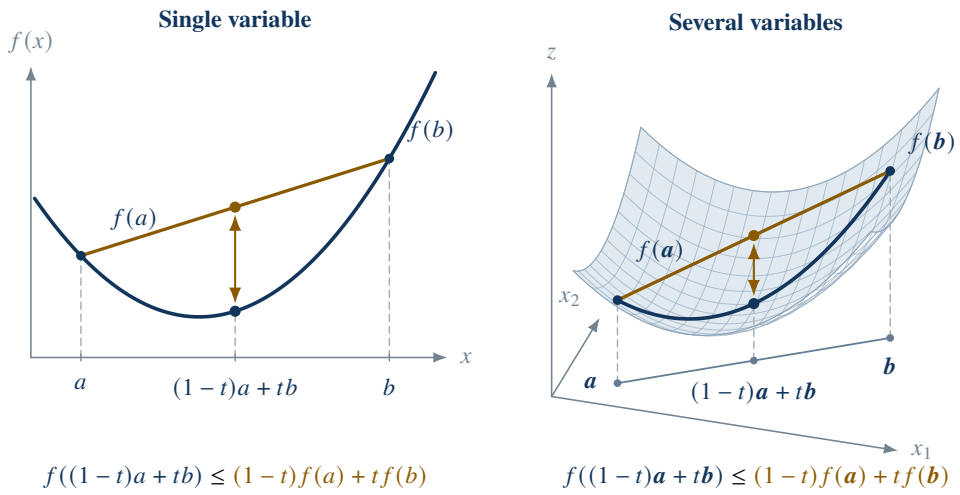


FIGURE 9: Visual depiction of the concept of convexity of a single-variable (left) and multi-variable (right) function.

Convexity and strict convexity. A function f is *convex* if, for every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $0 \le t \le 1$,

$$f((1-t)\boldsymbol{x} + t\boldsymbol{y}) \le (1-t)f(\boldsymbol{x}) + tf(\boldsymbol{y}). \tag{3.14}$$

Geometrically, the graph lies on or below every chord joining two points on the graph. The function is *strictly convex* if this inequality is strict whenever $\boldsymbol{x} \neq \boldsymbol{y}$ and $0 < t < 1$. Thus, strict convexity excludes a straight segment in the graph between distinct input points. See Figure 9 for a depiction of convex functions.

For a differentiable function, convexity is equivalent to the *supporting-plane inequality*

$$f(\boldsymbol{y}) \ge f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}) \qquad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n. \tag{3.15}$$

This says that every tangent plane is a global lower bound for the function; in one variable, the tangent line lies below the graph. A proof of this equivalence appears in [2, Section 3.1.3, pp. 69–70].

Theorem 8 (Optimality under convexity and strict convexity)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex.

1. Every local minimizer of f is a global minimizer.
2. If f is differentiable, then $\boldsymbol{x}_*$ is a global minimizer if and only if $\nabla f(\boldsymbol{x}_*) = \boldsymbol{0}$.
3. If f is strictly convex, then f has at most one global minimizer: if a minimizer exists, it is unique.

Neither convexity nor strict convexity alone guarantees that a global minimizer exists.

For proofs of the local-to-global conclusion and the differentiable optimality condition, see [2, Sections 4.2.2–4.2.3, pp. 138–140]. In particular, setting $\boldsymbol{x} = \boldsymbol{x}_*$ in (3.15) shows directly why a zero gradient gives a global minimum. For uniqueness under strict convexity, two distinct minimizers would have the same minimum value, but the strict inequality in (3.14) would give a smaller value at their midpoint. This is a contradiction; see also the uniqueness statement in [2, Section 4.2.1, p. 137].

Strong convexity: a uniform quadratic lower bound. A differentiable function f is *μ-strongly convex*, for some fixed $\mu > 0$, if

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \qquad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n. \tag{3.16}$$

The same positive constant μ must work at every point and for every displacement. Ordinary convexity places the graph above its tangent planes; strong convexity places it above an upward-opening quadratic surface based at each point. The extra quadratic term prevents the function from becoming arbitrarily flat. Strong convexity implies strict convexity, and hence convexity; the converse implications do not hold.

Theorem 9 (Hessian criteria for convexity)**.** Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, with Hessian $H(\boldsymbol{x})$.

1. The function is convex if and only if $H(\boldsymbol{x})$ is positive semidefinite at every $\boldsymbol{x}$.
2. If $H(\boldsymbol{x})$ is positive definite at every $\boldsymbol{x}$, then f is strictly convex. This condition is sufficient but not necessary for strict convexity.
3. For a fixed $\mu > 0$, the function is μ-strongly convex if and only if $H(\boldsymbol{x}) - \mu\mathbb{1}$ is positive semidefinite at every $\boldsymbol{x}$. Equivalently, every eigenvalue of every Hessian is at least μ.

The first two criteria are stated in [2, Section 3.1.4, p. 71], where the Hessian characterization is left as a proof exercise. We give the details here. The derivation of the strong-convexity quadratic bound also appears in [2, Section 9.1.2, p. 459].

Proof. Suppose first that f is convex. The point $\boldsymbol{x}$ is the midpoint of $\boldsymbol{x} + t\boldsymbol{p}$ and $\boldsymbol{x} - t\boldsymbol{p}$, so convexity gives $f(\boldsymbol{x} + t\boldsymbol{p}) - 2f(\boldsymbol{x}) + f(\boldsymbol{x} - t\boldsymbol{p}) \geq 0$. Twice continuous differentiability then gives

$$\boldsymbol{p}^{\mathsf{T}} H(\boldsymbol{x})\boldsymbol{p} = \lim_{t \to 0} \frac{f(\boldsymbol{x} + t\boldsymbol{p}) - 2f(\boldsymbol{x}) + f(\boldsymbol{x} - t\boldsymbol{p})}{t^2} \geq 0. \tag{3.17}$$

This holds for every direction $\boldsymbol{p}$, so $H(\boldsymbol{x})$ is positive semidefinite.

Conversely, Taylor's formula with integral remainder along the segment from $\boldsymbol{x}$ to $\boldsymbol{y}$ gives

$$\begin{aligned} &f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}) \\ &\qquad = \int_0^1 (1 - t)(\boldsymbol{y} - \boldsymbol{x})^{\mathsf{T}} H(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x})\, \mathrm{d}t. \end{aligned} \tag{3.18}$$

If every Hessian is positive semidefinite, the integral is nonnegative. Thus, the supporting-plane inequality holds and f is convex. If every Hessian is positive definite and $\boldsymbol{x} \neq \boldsymbol{y}$, the integral is strictly positive. The strict form of the supporting-plane characterization then gives strict convexity; see [2, Section 3.1.3, p. 70].

For the third statement, apply the first statement to $\boldsymbol{x} \mapsto f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$. Its supporting-plane inequality is exactly (3.16), and its Hessian is $H(\boldsymbol{x}) - \mu\mathbb{1}$. This proves both directions of the equivalence. ∎

Positive eigenvalues at each individual point need not have a common positive lower bound over the whole domain; this is the distinction between the second and third statements of Theorem 9.

Theorem 10 (Existence and uniqueness under strong convexity)**.** Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and μ-strongly convex for some $\mu > 0$. Then f attains a unique global minimum at a point $\boldsymbol{x}_*$. Moreover,

$$f(\boldsymbol{y}) - f(\boldsymbol{x}_*) \geq \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}_*\|_2^2 \qquad \text{for every } \boldsymbol{y} \in \mathbb{R}^n. \tag{3.19}$$

Proof. Fix $\boldsymbol{x} = \boldsymbol{0}$ in (3.16) and use the Cauchy–Schwarz inequality to bound the linear term from below:

$$\begin{aligned} f(\boldsymbol{y}) &\geq f(\boldsymbol{0}) + \nabla f(\boldsymbol{0})^{\mathsf{T}}\boldsymbol{y} + \frac{\mu}{2}\|\boldsymbol{y}\|_2^2 \\ &\geq f(\boldsymbol{0}) - \|\nabla f(\boldsymbol{0})\|_2\|\boldsymbol{y}\|_2 + \frac{\mu}{2}\|\boldsymbol{y}\|_2^2. \end{aligned} \tag{3.20}$$

As $\|\boldsymbol{y}\|_2 \to \infty$, the positive quadratic term dominates the linear term, so $f(\boldsymbol{y}) \to +\infty$. This property is called *coercivity*. Consequently, outside a sufficiently large closed ball centered at the origin, every value of f exceeds $f(\boldsymbol{0})$. Inside that ball, continuity and the extreme value theorem guarantee a minimum. Its value is at most $f(\boldsymbol{0})$, so it is also a global minimum.

Strong convexity gives strict convexity because $\boldsymbol{x} \mapsto f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$ is convex and $\frac{\mu}{2}\|\boldsymbol{x}\|_2^2$ is strictly convex. The sum is therefore strictly convex, and Theorem 8 gives uniqueness. Finally, $\nabla f(\boldsymbol{x}_*) = \boldsymbol{0}$ at the minimizer. Substituting $\boldsymbol{x} = \boldsymbol{x}_*$ in (3.16) proves (3.19). ∎

The quadratic-bound argument and its consequences for optimality are developed in [2, Section 9.1.2, pp. 459–460]; the proof above includes the attainment argument explicitly. The assumption that the domain is all of $\mathbb{R}^n$ matters: $f(x) = x^2$ is 2-strongly convex on the open interval $(0, 1)$, but its infimum is not attained there because 0 is excluded.

Example 11 (A convex function with many minimizers)**.** Consider $f(x, y) = x^2$ on $\mathbb{R}^2$. Its Hessian is

$$H(x, y) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.21}$$

The Hessian is positive semidefinite everywhere, so f is convex. Since $f(x, y) \geq 0$ and $f(0, y) = 0$ for every y, the entire line $x = 0$ consists of global minimizers. The function is not strictly convex, because it is constant along that line, and it is not strongly convex, because the Hessian has a zero eigenvalue. Convexity makes these local minima global, but does not make the minimizer unique.

Example 12 (Strict convexity without an attained minimum)**.** Consider $f(x) = \mathrm{e}^x$ on $\mathbb{R}$. Since $f''(x) = \mathrm{e}^x > 0$ everywhere, the function is strictly convex. Nevertheless, $f(x) > 0$ for every finite x, while $f(x) \to 0$ as $x \to -\infty$. Thus, its infimum is 0, but it has no minimizer. It is also not strongly convex on $\mathbb{R}$: the positive second derivative tends to zero as $x \to -\infty$, so no single positive constant μ bounds it from below everywhere. Strict convexity guarantees uniqueness only if a minimizer exists.

Example 13 (Convexity need not guarantee a minimum)**.** Consider the function

$$f(x) = \sqrt{1 + x^2} - x, \qquad x \in \mathbb{R}. \tag{3.22}$$

Its second derivative is

$$f''(x) = \frac{1}{(1+x^2)^{3/2}} > 0, \tag{3.23}$$

so f is strictly convex. Nevertheless, it has no global minimum. To see why, multiply the numerator and denominator by $\sqrt{1+x^2}+x$ to obtain

$$f(x) = \frac{1}{\sqrt{1+x^2}+x} > 0. \tag{3.24}$$

As $x \to +\infty$, the denominator grows without bound, so $f(x)$ approaches zero. Thus, the function takes values arbitrarily close to zero, but it never equals zero at any finite point. Every candidate for a minimum can therefore be improved by taking x sufficiently large.
Now add a quadratic term:

$$g(x) = f(x) + x^2 = \sqrt{1+x^2} - x + x^2. \tag{3.25}$$

The second derivative satisfies

$$g''(x) = \frac{1}{(1+x^2)^{3/2}} + 2 \geq 2 \qquad \text{for every } x \in \mathbb{R}. \tag{3.26}$$

Hence, g is 2-strongly convex on all of $\mathbb{R}$ and therefore attains a unique global minimum. We know that this minimizer exists and is unique without first solving $g'(x) = 0$. Its location is approximately $x_* = 0.3393$. Figure 10 compares the two functions.

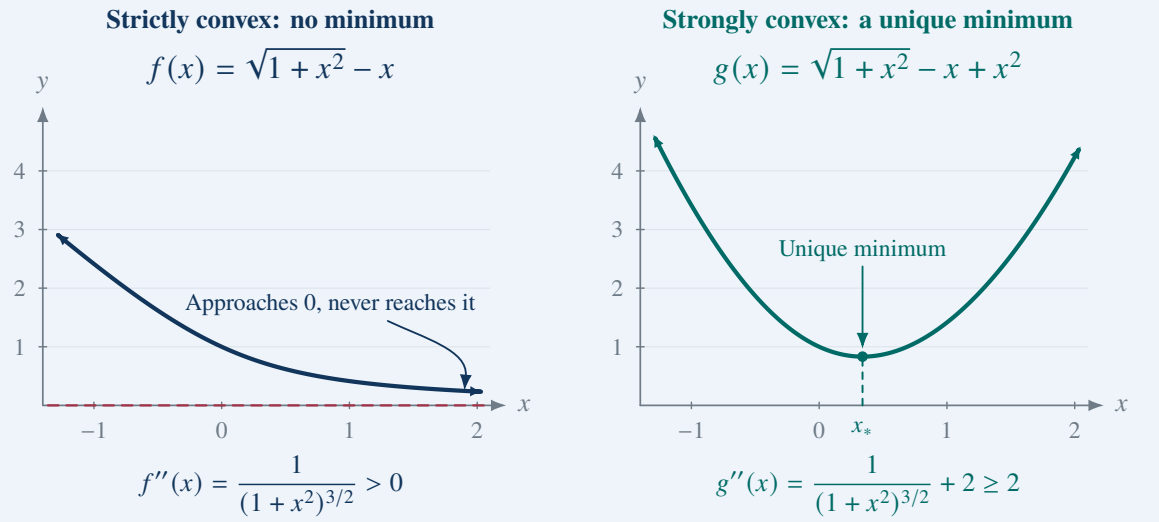


FIGURE 10: The strictly convex function $f(x) = \sqrt{1+x^2} - x$ approaches zero without ever reaching it, so it has no global minimum. Adding x^2 gives the strongly convex function g, whose unique global minimum is attained at $x_* \approx 0.3393$. Both functions are defined on all of $\mathbb{R}$, and the two panels use identical scales.

The rotated quadratic in Example 7 is a multivariable strongly convex example: its constant Hessian has eigenvalues 1 and 9, so it is 1-strongly convex. Later, lower and upper bounds on curvature will help explain how quickly gradient descent converges.

Key Idea

Convexity guarantees that every local minimizer is global. Strict convexity guarantees at most one minimizer. Strong convexity guarantees an attained, unique global minimum for a differentiable real-valued function on all of $\mathbb{R}^n$, together with a quadratic lower bound on the increase away from that minimum.

4 Gradient descent

The optimality conditions describe what a solution must look like, but they do not yet tell us how to find it. In a nonlinear problem, solving $\nabla f(\boldsymbol{x}) = \mathbf{0}$ symbolically can be just as difficult as the original minimization. We therefore construct a sequence of approximations. At each iteration we ask two separate questions: in which direction should we move, and how far should we move? The gradient descent algorithm moves in the direction of steepest descent at every point — which we now know is given by the gradient at that point.

Definition 14 (Gradient descent)**.** Given an initial point $\boldsymbol{x}_0$ and positive step sizes α_k, gradient descent generates

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k), \qquad k = 0, 1, 2, \ldots. \tag{4.1}$$

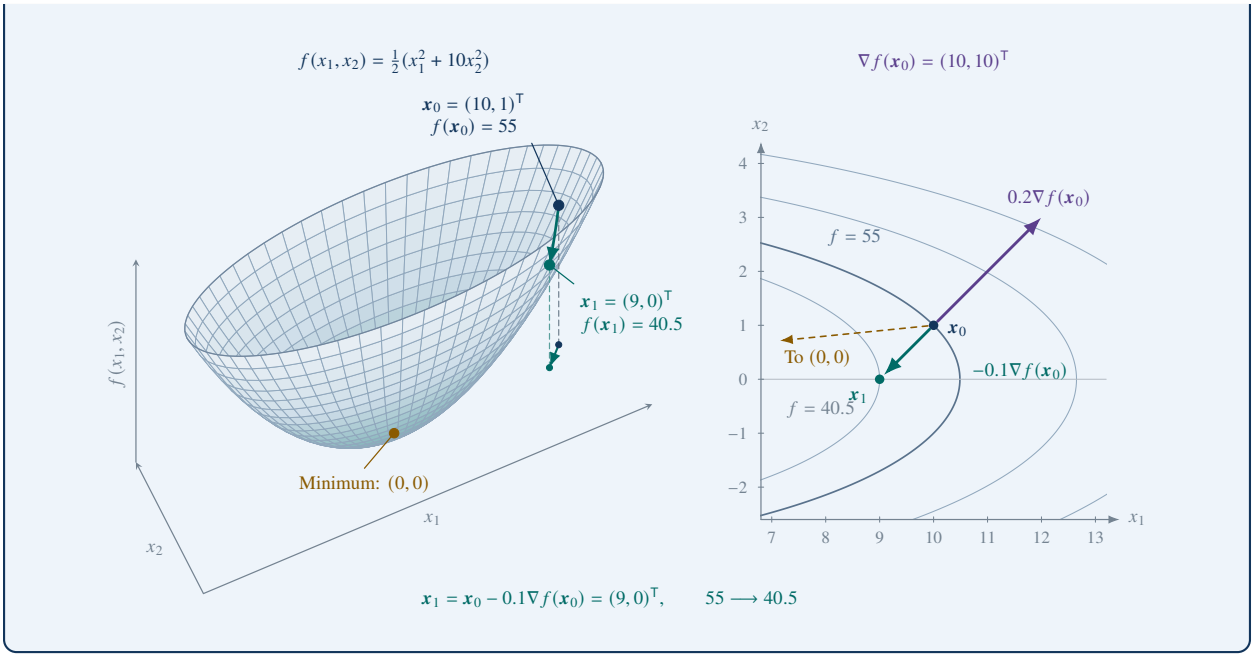
For one variable, this becomes $x_{k+1} = x_k - \alpha_k f'(x_k)$. For maximization, gradient ascent uses $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \nabla f(\boldsymbol{x}_k)$, equivalently gradient descent applied to $-f$.

The gradient contains all the partial derivatives at the *same current point*. Thus a multivariable update changes all coordinates simultaneously. In a program, compute the gradient before overwriting any component of the current vector. Updating one component and then recomputing another partial derivative at the partially updated vector defines a different iteration.

Example 15. Let $f(x_1, x_2) = \frac{1}{2}(x_1^2 + 10x_2^2)$ and start at $\boldsymbol{x}_0 = (10, 1)^\mathsf{T}$. The minimizer is the origin, while

$$\nabla f(\boldsymbol{x}_0) = (10, 10)^\mathsf{T}, \qquad -\boldsymbol{x}_0 = (-10, -1)^\mathsf{T}. \tag{4.2}$$

The negative gradient and the displacement toward the minimizer are not parallel. With $\alpha_0 = 0.1$, the new point is $(9, 0)^\mathsf{T}$ and the objective drops from 55 to 40.5. The method removes the second component quickly, but has barely reduced the first. Different curvatures in the two directions explain this behavior.



For nonconvex functions, the initial point can determine which minimum is found, whether an iteration approaches a saddle, or whether it remains at a stationary point. For $f(x) = (x^2 - 1)^2$, starting at $x_0 = 0$ leaves every gradient-descent iterate at zero even though zero is a local maximum. Starting at several points can reveal different outcomes, but a collection of successful local runs is not a proof of global optimality.

Algorithm: Basic gradient descent

1. Choose $\boldsymbol{x}_0$, a maximum iteration count, and a tolerance for the gradient norm.
2. At iteration k, evaluate $f(\boldsymbol{x}_k)$ and $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k)$.
3. If $\|\boldsymbol{g}_k\|_2$ is below the tolerance, report approximate stationarity and stop.
4. Choose a positive step α_k, for example by the backtracking rule in Section 4.1.
5. Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k$ and repeat. Report separately if the iteration budget is exhausted or no acceptable step can be found.

What the stopping test actually certifies. The gradient norm measures first-order optimality. A small value says that no unit direction has a large first-order slope; it does not distinguish a minimum from a maximum or a saddle. When f is differentiable and μ-strongly convex on $\mathbb{R}^n$, with minimizer $\boldsymbol{x}_\star$, the gradient provides quantitative information:

$$f(\boldsymbol{x}) - f(\boldsymbol{x}_\star) \le \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{2\mu}, \qquad \|\boldsymbol{x} - \boldsymbol{x}_\star\|_2 \le \frac{\|\nabla f(\boldsymbol{x})\|_2}{\mu}. \tag{4.3}$$

The first inequality follows by minimizing the strong-convexity lower quadratic model over its argument; the second follows from strong monotonicity of the gradient and Cauchy–Schwarz. These are meaningful error certificates when a valid positive μ is known. Without such assumptions, an equally small gradient can have a very different interpretation.

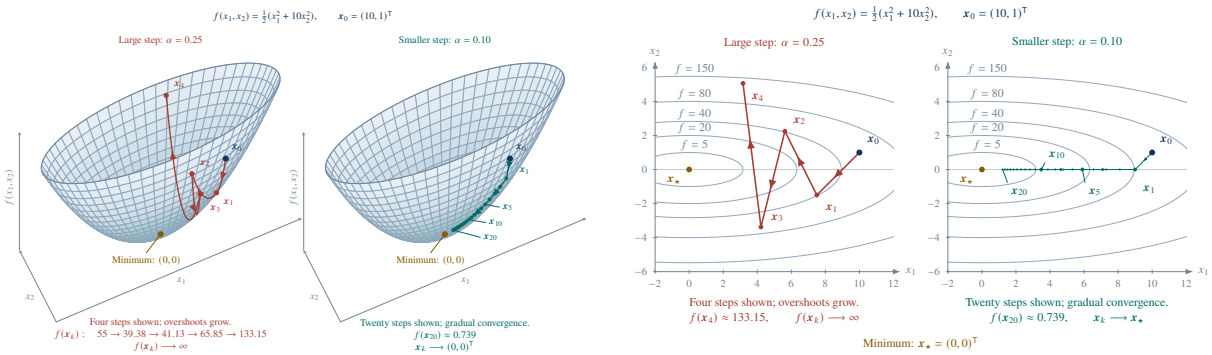


FIGURE 11: Steps of gradient descent for the function of Example 15. If the step size is too large, then the gradient descent algorithm can actually overshoot the minimum and *increase* the value of the objective function.

4.1 Step size, convergence, and conditioning

Choosing the negative gradient settles the direction of motion, but it does not settle how far to move. The gradient describes the slope at the current point; after a long move, the slope may have changed substantially. A step can therefore point initially downhill and still end at a point with a larger function value. Conversely, an unnecessarily small step may produce very little progress. Consider the function in Example 15. Starting at the same point, two different step sizes can have drastically different behavior; see Figure 11.

To obtain a concrete understanding of how to pick the step size, let us examine a quadratic, where every iterate can be calculated exactly, and then identify assumptions that give useful guarantees for more general functions.

Consider $f(x) = \frac{a}{2}(x - b)^2$, where $a > 0$. The unique minimizer is b, and the constant second derivative $f''(x) = a$ measures the curvature. With a fixed positive step size α, gradient descent becomes $x_{k+1} = x_k - \alpha a(x_k - b)$. The quantity we want to reduce is the error $x_k - b$, rather than the value of x_k itself.

Theorem 16 (Constant steps for a scalar quadratic)**.** For $f(x) = \frac{a}{2}(x - b)^2$ with $a > 0$ and a constant step size $\alpha > 0$, the gradient-descent errors satisfy

$$\begin{aligned} x_{k+1} - b &= (1 - \alpha a)(x_k - b), \\ x_k - b &= (1 - \alpha a)^k (x_0 - b), \qquad k \geq 1. \end{aligned} \tag{4.4}$$

The iterates converge to b from every initial point if and only if

$$0 < \alpha < \frac{2}{a}. \tag{4.5}$$

Proof. Subtract b from the update and collect the terms containing $x_k - b$:

$$x_{k+1} - b = x_k - b - \alpha a(x_k - b) = (1 - \alpha a)(x_k - b). \tag{4.6}$$

Each update therefore multiplies the previous error by the same number. Repeated substitution, or induction on k, gives the second equality in (4.4). If $|1 - \alpha a| < 1$, its kth power tends to zero and hence $x_k \to b$ for

every x_0. If $|1 - \alpha a| \geq 1$, any nonzero initial error fails to tend to zero. Finally,

$$|1 - \alpha a| < 1 \quad \Longleftrightarrow \quad -1 < 1 - \alpha a < 1 \quad \Longleftrightarrow \quad 0 < \alpha a < 2. \tag{4.7}$$

Dividing by $a > 0$ gives (4.5). ■

The error multiplier also explains the different paths that the iterates can take.

- If $0 < \alpha a < 1$, the error keeps its sign and decreases in magnitude: the iterates approach the minimizer from one side.
- If $\alpha a = 1$, the next error is zero, so the method reaches the minimizer in one step.
- If $1 < \alpha a < 2$, the error changes sign at every step but decreases in magnitude: the iterates overshoot the minimizer in alternating directions and still converge.
- If $\alpha a = 2$, a nonzero error changes sign without shrinking. The iterates alternate between two points with the same objective value.
- If $\alpha a > 2$, a nonzero error changes sign and grows in magnitude, so the iteration diverges.

If $x_0 = b$, the gradient is already zero and the point stays fixed for every step size. This exceptional initialization is why the theorem specifies convergence from *every* initial point.

Since $f(b) = 0$, the objective error is $f(x_k) - f(b) = \frac{a}{2}(x_k - b)^2$. Thus, changing the sign of the point error does not change the objective value. Figure 12 compares the regimes for the same objective and initial point. In that example, $\alpha = 0.1$ and $\alpha = 0.9$ have error multipliers 0.8 and -0.8, so their objective histories coincide even though one path stays on one side of the minimizer and the other oscillates.

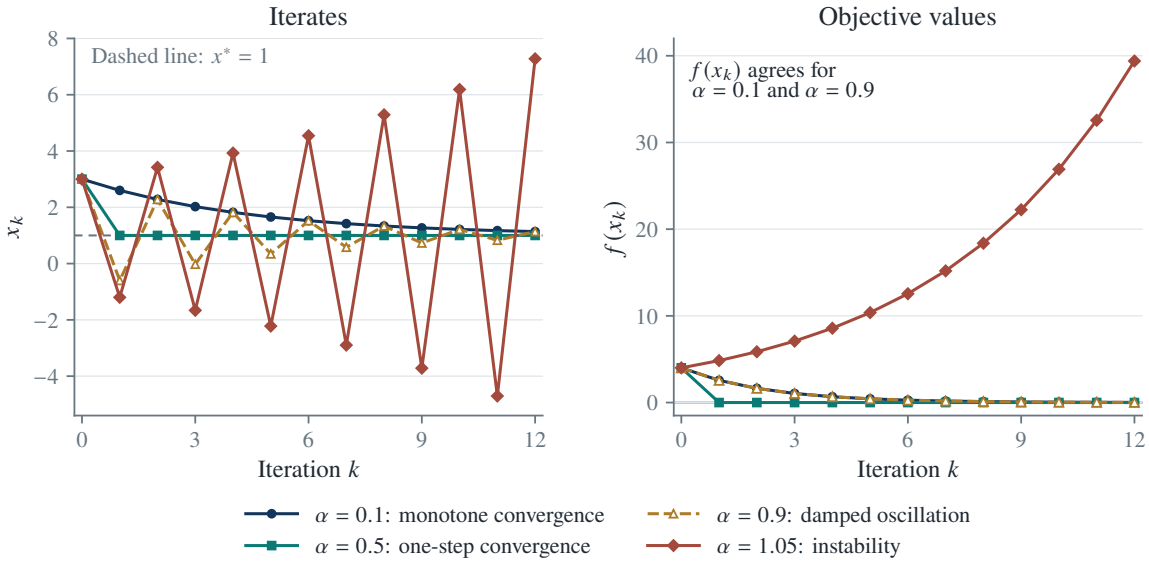


FIGURE 12: Constant-step gradient descent for $f(x) = (x - 1)^2$, starting at $x_0 = 3$. Here $a = 2$, so the stability range is $0 < \alpha < 1$. Small steps converge slowly, $\alpha = 0.5$ reaches the minimizer in one step, $\alpha = 0.9$ gives decaying oscillations, and $\alpha = 1.05$ is unstable. Oscillation of the iterates does not by itself mean that the objective is increasing.

Connection with fixed-point iteration. For a general one-variable function, gradient descent is the fixed-point iteration associated with $T(x) = x - \alpha f'(x)$. A stationary point $x_\star$ satisfies $T(x_\star) = x_\star$, and, if f is twice continuously differentiable near that point,

$$T'(x_\star) = 1 - \alpha f''(x_\star). \tag{4.8}$$

This recovers the quadratic error multiplier as a local description of a nonlinear iteration. To justify the local conclusion, suppose $|T'(x_\star)| < 1$. By continuity of T', there is an interval centered at $x_\star$ and a number q with $0 < q < 1$ such that $|T'(x)| \le q$ throughout that interval. The mean value theorem gives, for every x in the interval,

$$|T(x) - x_\star| = |T(x) - T(x_\star)| \le q|x - x_\star|. \tag{4.9}$$

The next point is closer to $x_\star$, so it remains in the same interval. Applying this bound repeatedly proves $|x_k - x_\star| \le q^k |x_0 - x_\star|$ for sufficiently close initial points. This is a local guarantee; a distant starting point need not enter that interval.

At a stationary point with $f''(x_\star) < 0$, a positive step instead gives $T'(x_\star) > 1$. Continuity supplies a smaller interval and a constant $r > 1$ for which $T'(x) \ge r$ throughout the interval. The mean value theorem now gives $|T(x) - x_\star| \ge r|x - x_\star|$. A nonzero error grows while the iterates remain nearby, so the stationary point repels nearby iterates. An iterate placed exactly at the stationary point still remains there because its gradient is zero.

Smoothness controls the error in a linear prediction. For the quadratic, the curvature a determines how large a step is safe. A general function needs a bound on how quickly its gradient can change. We say that the gradient is *L-Lipschitz continuous* if there is a constant $L > 0$ such that

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \le L\|\boldsymbol{x} - \boldsymbol{y}\|_2 \qquad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n. \tag{4.10}$$

A function with this property is called *L-smooth*. The condition limits the change in slope over a given displacement. It does not require convexity: a function can have bounded positive and negative curvature and still satisfy (4.10).

For a twice continuously differentiable function, a sufficient condition is $\|H(\boldsymbol{x})\|_2 \le L$ everywhere, where the matrix norm is defined by $\|H\|_2 = \max_{\|\boldsymbol{p}\|_2 = 1} \|H\boldsymbol{p}\|_2$. To verify this, apply the fundamental theorem of calculus to each component of the gradient along the segment from $\boldsymbol{x}$ to $\boldsymbol{y}$:

$$\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) = \int_0^1 H(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x})\, \mathrm{d}t. \tag{4.11}$$

Taking norms, using the triangle inequality for the integral, and then using the matrix norm bound gives

$$\begin{aligned} \|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|_2 &\le \int_0^1 \|H(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))\|_2 \|\boldsymbol{y} - \boldsymbol{x}\|_2 \, \mathrm{d}t \\ &\le \int_0^1 L\|\boldsymbol{y} - \boldsymbol{x}\|_2 \, \mathrm{d}t = L\|\boldsymbol{y} - \boldsymbol{x}\|_2. \end{aligned} \tag{4.12}$$

Thus, a bound on the Hessian provides a usable value of L. In the scalar quadratic example, we can take $L = a$.

Theorem 17 (Descent lemma and a gradient step)**.** Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and has an L-Lipschitz gradient. Then, for every point $\boldsymbol{x}$ and displacement $\boldsymbol{p}$,

$$f(\boldsymbol{x} + \boldsymbol{p}) \le f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{p} + \frac{L}{2} \|\boldsymbol{p}\|_2^2. \tag{4.13}$$

In particular, for every $\alpha > 0$,

$$f(\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})) \le f(\boldsymbol{x}) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla f(\boldsymbol{x})\|_2^2. \tag{4.14}$$

Consequently, $0 < \alpha < 2/L$ guarantees a strict decrease whenever $\nabla f(\boldsymbol{x}) \neq \boldsymbol{0}$.

Proof. Fix $\boldsymbol{x}$ and $\boldsymbol{p}$. As the scalar parameter t runs from 0 to 1, the point $\boldsymbol{x} + t\boldsymbol{p}$ runs along the segment from $\boldsymbol{x}$ to $\boldsymbol{x} + \boldsymbol{p}$. Apply the chain rule to the composition $t \mapsto f(\boldsymbol{x} + t\boldsymbol{p})$. Each input coordinate $x_i + tp_i$ changes at rate p_i, so

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x} + t\boldsymbol{p}) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\boldsymbol{x} + t\boldsymbol{p}) p_i = \nabla f(\boldsymbol{x} + t\boldsymbol{p})^{\mathsf{T}} \boldsymbol{p}. \tag{4.15}$$

The fundamental theorem of calculus therefore gives

$$f(\boldsymbol{x} + \boldsymbol{p}) - f(\boldsymbol{x}) = \int_0^1 \nabla f(\boldsymbol{x} + t\boldsymbol{p})^{\mathsf{T}} \boldsymbol{p} \, \mathrm{d}t. \tag{4.16}$$

The linear prediction $\nabla f(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{p}$ is constant with respect to t. Subtracting it from both sides gives

$$\begin{aligned} &f(\boldsymbol{x} + \boldsymbol{p}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{p} \\ &\qquad = \int_0^1 \big(\nabla f(\boldsymbol{x} + t\boldsymbol{p}) - \nabla f(\boldsymbol{x})\big)^{\mathsf{T}} \boldsymbol{p} \, \mathrm{d}t. \end{aligned} \tag{4.17}$$

For each $t \in [0, 1]$, the Cauchy–Schwarz inequality and then (4.10) imply

$$\begin{aligned} \big(\nabla f(\boldsymbol{x} + t\boldsymbol{p}) - \nabla f(\boldsymbol{x})\big)^{\mathsf{T}} \boldsymbol{p} &\le \|\nabla f(\boldsymbol{x} + t\boldsymbol{p}) - \nabla f(\boldsymbol{x})\|_2 \|\boldsymbol{p}\|_2 \\ &\le L\|t\boldsymbol{p}\|_2 \|\boldsymbol{p}\|_2 = Lt\|\boldsymbol{p}\|_2^2. \end{aligned} \tag{4.18}$$

Integrating this upper bound yields $\int_0^1 Lt\|\boldsymbol{p}\|_2^2 \, \mathrm{d}t = (L/2)\|\boldsymbol{p}\|_2^2$, which proves (4.13).

For a gradient step, substitute $\boldsymbol{p} = -\alpha \nabla f(\boldsymbol{x})$. The linear and quadratic terms become

$$\begin{aligned} \nabla f(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{p} + \frac{L}{2} \|\boldsymbol{p}\|_2^2 &= -\alpha \|\nabla f(\boldsymbol{x})\|_2^2 + \frac{L\alpha^2}{2} \|\nabla f(\boldsymbol{x})\|_2^2 \\ &= -\alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla f(\boldsymbol{x})\|_2^2. \end{aligned} \tag{4.19}$$

This gives (4.14). If $0 < \alpha < 2/L$, the coefficient $\alpha(1 - L\alpha/2)$ is positive. A nonzero gradient then makes the right-hand side strictly smaller than $f(\boldsymbol{x})$. ∎

The bound separates a predicted benefit from a possible curvature penalty. The negative term $-\alpha\|\nabla f(\boldsymbol{x})\|_2^2$ comes from the slope at the starting point; the positive term $(L\alpha^2/2)\|\nabla f(\boldsymbol{x})\|_2^2$ accounts for how much that slope can change during the move. The penalty grows quadratically with the step size, whereas the predicted benefit grows linearly. This explains why making the step arbitrarily large is unsafe. For additional discussion of the quadratic upper bound, see [5, pp. 1.12–1.13].

What repeated decrease guarantees. A decreasing sequence of objective values does not automatically establish convergence to a minimizer. We must distinguish the function values $f(\boldsymbol{x}_k)$, the gradients $\nabla f(\boldsymbol{x}_k)$, and the points $\boldsymbol{x}_k$. The following results show what additional assumptions allow us to conclude about each of them. We use the fixed step $\alpha = 1/L$, for which (4.14) becomes the particularly useful estimate

$$f(\boldsymbol{x}_{k+1}) \le f(\boldsymbol{x}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2, \qquad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k). \tag{4.20}$$

Thus a large gradient forces a substantial decrease. If the objective has a finite lower bound, such decreases cannot continue indefinitely without the gradients becoming small.

Theorem 18 (A stationarity guarantee)**.** Suppose f has an L-Lipschitz gradient on $\mathbb{R}^n$ and $f(\boldsymbol{x}) \ge f_{\text{low}}$ for all $\boldsymbol{x}$, where f_{low} is finite. For gradient descent with step size $1/L$ and every integer $N \ge 1$,

$$\begin{aligned} \sum_{k=0}^{N-1} \|\nabla f(\boldsymbol{x}_k)\|_2^2 &\le 2L\big(f(\boldsymbol{x}_0) - f_{\text{low}}\big), \\ \min_{0\le k<N} \|\nabla f(\boldsymbol{x}_k)\|_2^2 &\le \frac{2L\big(f(\boldsymbol{x}_0) - f_{\text{low}}\big)}{N}. \end{aligned} \tag{4.21}$$

Moreover, $\|\nabla f(\boldsymbol{x}_k)\|_2 \to 0$, and the objective values converge to a finite limit. Convexity is not required for these conclusions.

Proof. Rearranging (4.20) gives $\|\nabla f(\boldsymbol{x}_k)\|_2^2 \le 2L(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}))$. Sum this inequality over the first N iterations. The intermediate objective values cancel:

$$\begin{aligned} \sum_{k=0}^{N-1} \|\nabla f(\boldsymbol{x}_k)\|_2^2 &\le 2L \sum_{k=0}^{N-1} \big(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1})\big) \\ &= 2L\big(f(\boldsymbol{x}_0) - f(\boldsymbol{x}_N)\big) \\ &\le 2L\big(f(\boldsymbol{x}_0) - f_{\text{low}}\big). \end{aligned} \tag{4.22}$$

The smallest of N nonnegative quantities is no greater than their average. Dividing this sum bound by N proves the second inequality in (4.21).

The partial sums of the nonnegative series $\sum_{k=0}^{\infty}\|\nabla f(\boldsymbol{x}_k)\|_2^2$ are bounded by a constant independent of N. Therefore the series converges and its terms tend to zero. Explicitly, infinitely many terms larger than a fixed positive number would make the partial sums unbounded. This proves $\|\nabla f(\boldsymbol{x}_k)\|_2 \to 0$. Finally, the objective values are nonincreasing by (4.20) and bounded below by f_{low}, so they converge to a finite limit. ∎

The second bound in (4.21) guarantees that *at least one* of the first N iterates has a small gradient. To guarantee a gradient norm at most $\varepsilon > 0$, it suffices to take $N \ge 2L(f(\boldsymbol{x}_0) - f_{\text{low}})/\varepsilon^2$. This is a bound on the best gradient norm seen so far, not a claim that every gradient norm decreases at each iteration. A stationary point can also be a maximum or a saddle. If a subsequence of the iterates has a finite limit, continuity of the gradient makes that limit stationary, but these assumptions alone do not classify it as a minimum.

Example 19 (Small gradients without a finite limiting point)**.** The function $f(x) = \sqrt{1 + x^2} - x$ from the preceding subsection is bounded below by zero and satisfies $0 < f''(x) \leq 1$, so it has a 1-Lipschitz derivative. Gradient descent with step size 1 gives

$$x_{k+1} = x_k + 1 - \frac{x_k}{\sqrt{1 + x_k^2}} > x_k. \tag{4.23}$$

The inequality holds because $x/\sqrt{1 + x^2} < 1$ for every finite x. If this increasing sequence were bounded above, it would converge to some finite number ℓ. Its successive differences would then tend to zero. However, the update formula would give

$$x_{k+1} - x_k \longrightarrow 1 - \frac{\ell}{\sqrt{1 + \ell^2}} > 0, \tag{4.24}$$

a contradiction. Hence $x_k \to +\infty$, even though $f'(x_k) \to 0$ and $f(x_k) \to 0$. Small gradients and convergent objective values do not, by themselves, guarantee convergence to a finite minimizer.

Convexity turns decrease into a bound on the objective error. For a convex function, the supporting-plane inequality relates the gradient to the difference between the current objective value and the minimum value. To use this information, we assume that a minimizer actually exists. This excludes the preceding example. We measure the objective error by $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star)$, where $\boldsymbol{x}_\star$ is any global minimizer.

Theorem 20 (Convergence of objective values for a convex function)**.** Suppose f is convex, has an L-Lipschitz gradient on $\mathbb{R}^n$, and attains a global minimum at $\boldsymbol{x}_\star$. Gradient descent with step size $1/L$ satisfies

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star) \leq \frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2}{2k}, \qquad k \geq 1. \tag{4.25}$$

Proof. Convexity gives $f(\boldsymbol{x}_\star) \geq f(\boldsymbol{x}_j) + \nabla f(\boldsymbol{x}_j)^\mathsf{T}(\boldsymbol{x}_\star - \boldsymbol{x}_j)$. Rearranging,

$$f(\boldsymbol{x}_j) - f(\boldsymbol{x}_\star) \leq \nabla f(\boldsymbol{x}_j)^\mathsf{T}(\boldsymbol{x}_j - \boldsymbol{x}_\star). \tag{4.26}$$

Subtract $f(\boldsymbol{x}_\star)$ from (4.20) and use this upper bound on the current objective error:

$$\begin{aligned} f(\boldsymbol{x}_{j+1}) - f(\boldsymbol{x}_\star) &\leq f(\boldsymbol{x}_j) - f(\boldsymbol{x}_\star) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_j)\|_2^2 \\ &\leq \nabla f(\boldsymbol{x}_j)^\mathsf{T}(\boldsymbol{x}_j - \boldsymbol{x}_\star) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_j)\|_2^2. \end{aligned} \tag{4.27}$$

The expression on the last line is a difference of squared distances. Indeed, expanding the gradient update gives

$$\begin{aligned} \|\boldsymbol{x}_{j+1} - \boldsymbol{x}_\star\|_2^2 &= \|\boldsymbol{x}_j - \boldsymbol{x}_\star - \tfrac{1}{L}\nabla f(\boldsymbol{x}_j)\|_2^2 \\ &= \|\boldsymbol{x}_j - \boldsymbol{x}_\star\|_2^2 - \frac{2}{L}\nabla f(\boldsymbol{x}_j)^\mathsf{T}(\boldsymbol{x}_j - \boldsymbol{x}_\star) + \frac{1}{L^2}\|\nabla f(\boldsymbol{x}_j)\|_2^2. \end{aligned} \tag{4.28}$$

Rearranging this equality and substituting it into the preceding bound yields

$$f(\boldsymbol{x}_{j+1}) - f(\boldsymbol{x}_\star) \leq \frac{L}{2}\left(\|\boldsymbol{x}_j - \boldsymbol{x}_\star\|_2^2 - \|\boldsymbol{x}_{j+1} - \boldsymbol{x}_\star\|_2^2\right). \tag{4.29}$$

Now sum over $j = 0, \dots, k-1$. The intermediate squared distances cancel, leaving

$$\begin{aligned}\sum_{j=0}^{k-1} \big(f(\boldsymbol{x}_{j+1}) - f(\boldsymbol{x}_\star)\big) &\le \frac{L}{2}\left(\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2 - \|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2^2\right) \\ &\le \frac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2.\end{aligned} \tag{4.30}$$

Because the objective values are nonincreasing, each of the k terms on the left is at least $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star)$. Consequently,

$$k\big(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star)\big) \le \frac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2. \tag{4.31}$$

Dividing by k proves (4.25). ■

The bound is of order $1/k$: doubling the iteration count halves the stated upper bound on the objective error. This describes a guarantee, not an exact formula for every run; a particular objective may converge faster. The proof uses both convexity and an attained minimum, whereas the stationarity theorem required neither convexity nor attainment. Further discussion of this estimate appears in [5, pp. 1.24–1.26].

Strong convexity gives a geometric decrease guarantee. Strong convexity provides a stronger connection between the gradient and the objective error. If the function value is still substantially above its minimum, the gradient cannot already be arbitrarily small. Combining that fact with the decrease obtained at each step produces a fixed fractional reduction of the remaining error.

Theorem 21 (Convergence under strong convexity)**.** Suppose f is μ-strongly convex and has an L-Lipschitz gradient on $\mathbb{R}^n$, where $\mu > 0$. Then $\mu \le L$, and f has a unique global minimizer $\boldsymbol{x}_\star$. Gradient descent with step size $1/L$ satisfies

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star) \le \left(1 - \frac{\mu}{L}\right)^k \big(f(\boldsymbol{x}_0) - f(\boldsymbol{x}_\star)\big), \qquad k \ge 1. \tag{4.32}$$

The points $\boldsymbol{x}_k$ also converge to $\boldsymbol{x}_\star$.

Proof. Existence and uniqueness follow from Theorem 10. To compare μ and L, apply the strong-convexity lower bound and the descent lemma to the same point and any nonzero displacement:

$$\frac{\mu}{2}\|\boldsymbol{p}\|_2^2 \le f(\boldsymbol{x} + \boldsymbol{p}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{p} \le \frac{L}{2}\|\boldsymbol{p}\|_2^2. \tag{4.33}$$

Dividing by $\|\boldsymbol{p}\|_2^2/2$ gives $\mu \le L$.

Next, apply strong convexity with comparison point $\boldsymbol{x}_\star$ and complete the square:

$$\begin{aligned} f(\boldsymbol{x}_\star) &\ge f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{x}_\star - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x}_\star - \boldsymbol{x}\|_2^2 \\ &= f(\boldsymbol{x}) - \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{2\mu} + \frac{\mu}{2}\left\|\boldsymbol{x}_\star - \boldsymbol{x} + \frac{\nabla f(\boldsymbol{x})}{\mu}\right\|_2^2 \\ &\ge f(\boldsymbol{x}) - \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{2\mu}. \end{aligned} \tag{4.34}$$

The last step drops a nonnegative squared norm. Rearranging proves the first gradient certificate in (4.3), in the equivalent form

$$\|\nabla f(\boldsymbol{x})\|_2^2 \geq 2\mu\big(f(\boldsymbol{x}) - f(\boldsymbol{x}_\star)\big). \tag{4.35}$$

Insert this lower bound on the squared gradient norm into (4.20). Since the squared norm is subtracted, replacing it by its lower bound gives

$$\begin{aligned} f(\boldsymbol{x}_{j+1}) - f(\boldsymbol{x}_\star) &\leq f(\boldsymbol{x}_j) - f(\boldsymbol{x}_\star) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_j)\|_2^2 \\ &\leq \left(1 - \frac{\mu}{L}\right)\big(f(\boldsymbol{x}_j) - f(\boldsymbol{x}_\star)\big). \end{aligned} \tag{4.36}$$

Repeated application gives (4.32). If $\mu < L$, the factor lies strictly between zero and one, so the objective error tends to zero geometrically. If $\mu = L$, the one-step bound forces the objective error to be zero after the first update, and uniqueness identifies that point as $\boldsymbol{x}_\star$.

Finally, strong convexity based at $\boldsymbol{x}_\star$, where $\nabla f(\boldsymbol{x}_\star) = \boldsymbol{0}$, gives

$$\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2^2 \leq \frac{2}{\mu}\big(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star)\big). \tag{4.37}$$

The right-hand side tends to zero, so the points themselves converge to the unique minimizer. ■

A geometric error bound is called a *linear convergence guarantee*. Here "linear" refers to reducing the remaining error by a fixed factor, not to a straight line on ordinary error-versus-iteration axes. A geometric bound appears as a straight line on a semilogarithmic plot. The factor $1 - \mu/L$ also explains why curvature matters: L limits the safe step, while μ controls the minimum progress that the gradient can force. If L is much larger than μ, this factor is close to one and the guarantee can be slow. The next subsection makes these competing effects explicit in the different eigendirections of a quadratic; see also [2, Sections 9.1–9.3].

4.2 Quadratic functions, eigenvalues, and slow directions

The scalar quadratic shows that a step size must be chosen in relation to curvature. In several variables, the same function can have very different curvatures in different directions, while the gradient-descent update uses only one step size. Quadratic functions let us examine this conflict exactly: their Hessians are constant, and a suitable rotation of coordinates separates the iteration into independent scalar recurrences.

Let A be a symmetric positive definite (SPD) $n \times n$ matrix and let $\boldsymbol{b} \in \mathbb{R}^n$. Consider

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\mathsf{T} A\boldsymbol{x} - \boldsymbol{b}^\mathsf{T}\boldsymbol{x}. \tag{4.38}$$

The matrix A determines the curvature, while $\boldsymbol{b}$ determines the location of the minimum. Minimizing this function is also a way of solving a linear system, as the following calculation shows.

Theorem 22 (The minimum of an SPD quadratic)**.** The function in (4.38) has gradient $\nabla f(\boldsymbol{x}) = A\boldsymbol{x} - \boldsymbol{b}$ and Hessian A. It has a unique global minimizer $\boldsymbol{x}_\star$, determined by $A\boldsymbol{x}_\star = \boldsymbol{b}$. Moreover,

$$f(\boldsymbol{x}) - f(\boldsymbol{x}_\star) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_\star)^\mathsf{T} A(\boldsymbol{x} - \boldsymbol{x}_\star). \tag{4.39}$$

Proof. Writing the quadratic term as $\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_i x_j$, differentiation with respect to x_ℓ gives

$$\frac{\partial f}{\partial x_\ell}(\boldsymbol{x}) = \frac{1}{2}\sum_{j=1}^{n} A_{\ell j}x_j + \frac{1}{2}\sum_{i=1}^{n} A_{i\ell}x_i - b_\ell = \sum_{j=1}^{n} A_{\ell j}x_j - b_\ell. \tag{4.40}$$

The last equality uses symmetry: $A_{i\ell} = A_{\ell i}$. These partial derivatives form the vector $A\boldsymbol{x} - \boldsymbol{b}$, and differentiating once more gives the Hessian A.

Positive definiteness implies that A is invertible. Indeed, if $A\boldsymbol{v} = \boldsymbol{0}$ for a nonzero vector $\boldsymbol{v}$, then $\boldsymbol{v}^\mathsf{T}A\boldsymbol{v} = 0$, contradicting positive definiteness. Thus, $A\boldsymbol{x}_\star = \boldsymbol{b}$ has exactly one solution. Substitute $\boldsymbol{x} = \boldsymbol{x}_\star + (\boldsymbol{x} - \boldsymbol{x}_\star)$ into the objective and expand:

$$\begin{aligned} f(\boldsymbol{x}) - f(\boldsymbol{x}_\star) &= (A\boldsymbol{x}_\star - \boldsymbol{b})^\mathsf{T}(\boldsymbol{x} - \boldsymbol{x}_\star) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_\star)^\mathsf{T}A(\boldsymbol{x} - \boldsymbol{x}_\star) \\ &= \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_\star)^\mathsf{T}A(\boldsymbol{x} - \boldsymbol{x}_\star). \end{aligned} \tag{4.41}$$

This difference is positive whenever $\boldsymbol{x} \neq \boldsymbol{x}_\star$, again by positive definiteness. Hence, $\boldsymbol{x}_\star$ is the unique global minimizer. ■

This equivalence connects unconstrained optimization to the numerical solution of SPD linear systems [1, Section 13.2, p. 580]. We can now study how an error in the current approximation changes under gradient descent.

Each eigenvector direction has its own scalar recurrence. With a constant step size $\alpha > 0$, the update is $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha(A\boldsymbol{x}_k - \boldsymbol{b})$. Subtract $\boldsymbol{x}_\star$ and use $\boldsymbol{b} = A\boldsymbol{x}_\star$:

$$\begin{aligned} \boldsymbol{x}_{k+1} - \boldsymbol{x}_\star &= \boldsymbol{x}_k - \boldsymbol{x}_\star - \alpha A(\boldsymbol{x}_k - \boldsymbol{x}_\star) \\ &= (\mathbb{1} - \alpha A)(\boldsymbol{x}_k - \boldsymbol{x}_\star). \end{aligned} \tag{4.42}$$

Thus, the error evolves by repeated multiplication by the same matrix. The purpose of introducing eigenvectors is to find directions in which this matrix multiplication is simply multiplication by a number.

Because A is symmetric, it has orthonormal eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ with positive eigenvalues $0 < \lambda_1 \leq \cdots \leq \lambda_n$. In particular, $A\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$. Expand the initial error in this basis:

$$\boldsymbol{x}_0 - \boldsymbol{x}_\star = \sum_{i=1}^{n} c_i \boldsymbol{v}_i, \qquad c_i = \boldsymbol{v}_i^\mathsf{T}(\boldsymbol{x}_0 - \boldsymbol{x}_\star). \tag{4.43}$$

Here, c_i is the initial error component in direction $\boldsymbol{v}_i$. Since $(\mathbb{1} - \alpha A)\boldsymbol{v}_i = (1 - \alpha\lambda_i)\boldsymbol{v}_i$, applying (4.42) repeatedly gives

$$\boldsymbol{x}_k - \boldsymbol{x}_\star = \sum_{i=1}^{n} (1 - \alpha\lambda_i)^k c_i \boldsymbol{v}_i. \tag{4.44}$$

This identity follows by induction: it holds at $k = 0$, and each additional update multiplies the ith coefficient by $1 - \alpha\lambda_i$. Orthogonality removes all cross terms when taking the squared Euclidean norm, so

$$\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2^2 = \sum_{i=1}^{n} (1 - \alpha\lambda_i)^{2k} c_i^2. \tag{4.45}$$

Theorem 23 (The exact stability range for an SPD quadratic)**.** For the function in (4.38), constant-step gradient descent converges to $\boldsymbol{x}_\star$ from every initial point if and only if

$$\max_{1\le i\le n} |1-\alpha\lambda_i| < 1, \qquad \text{equivalently} \qquad 0 < \alpha < \frac{2}{\lambda_{\max}(A)}. \tag{4.46}$$

Proof. If every factor satisfies $|1-\alpha\lambda_i| < 1$, then every term in the finite sum (4.45) tends to zero. Hence, the entire error tends to zero for any initial coefficients c_i.

Conversely, suppose $|1-\alpha\lambda_j| \ge 1$ for some j. Choose the initial point $\boldsymbol{x}_0 = \boldsymbol{x}_\star + \boldsymbol{v}_j$, so that the only nonzero coefficient in (4.43) is $c_j = 1$. Then the error norm is $|1-\alpha\lambda_j|^k$, which does not tend to zero. Therefore convergence from every initial point requires all the factors to have magnitude less than one. Finally, because each $\lambda_i > 0$,

$$|1-\alpha\lambda_i| < 1 \quad\Longleftrightarrow\quad -1 < 1-\alpha\lambda_i < 1 \quad\Longleftrightarrow\quad 0 < \alpha < \frac{2}{\lambda_i}. \tag{4.47}$$

All these inequalities hold simultaneously precisely when the inequality corresponding to the largest eigenvalue holds, proving (4.46). ∎

The eigenvalues describe curvature along the eigenvectors in a literal one-variable sense. Because each $\boldsymbol{v}_i$ has unit length, (4.39) gives

$$f(\boldsymbol{x}_\star + t\boldsymbol{v}_i) - f(\boldsymbol{x}_\star) = \frac{\lambda_i}{2}t^2. \tag{4.48}$$

Thus, the largest eigenvalue is the largest curvature and restricts the step size. The smallest eigenvalue is the smallest curvature; when the permitted step is small, its error multiplier $1-\alpha\lambda_{\min}(A)$ can be close to one. Thus, the step needed to avoid overshooting in a steep direction may make very little progress in a shallow direction. An initial error that has no component in a particular eigendirection never acquires one, so an individual starting point may behave better than the worst-case bound.

Choosing the best constant step. For a fixed α, define the worst-case error factor

$$q(\alpha) = \max_{1\le i\le n} |1-\alpha\lambda_i|. \tag{4.49}$$

The word "worst-case" refers to the initial error direction. Equation (4.45) gives $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_\star\|_2 \le q(\alpha)\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2$, and equality is attained by an error along an eigenvector whose multiplier has magnitude $q(\alpha)$. Therefore minimizing $q(\alpha)$ is exactly the problem of choosing a constant step with the best uniform guarantee over all error directions.

Theorem 24 (Best constant step and the condition number)**.** For a symmetric positive definite matrix A, let $\kappa_2(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$. The positive constant step minimizing (4.49) is

$$\alpha_\star = \frac{2}{\lambda_{\max}(A) + \lambda_{\min}(A)}, \qquad q_\star = q(\alpha_\star) = \frac{\kappa_2(A)-1}{\kappa_2(A)+1}. \tag{4.50}$$

With this step, the distance to the minimizer and the objective gap satisfy

$$\begin{aligned} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_\star\|_2 &\le q_\star \|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2, \\ f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_\star) &\le q_\star^2 \big(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star)\big). \end{aligned} \tag{4.51}$$

Proof. For fixed α, the function $\lambda \mapsto |1 - \alpha\lambda|$ is convex, so its value anywhere in the interval $[\lambda_{\min}(A), \lambda_{\max}(A)]$ is at most the larger endpoint value. Both endpoints are eigenvalues. Consequently,

$$q(\alpha) = \max\big\{|1 - \alpha\lambda_{\min}(A)|, |1 - \alpha\lambda_{\max}(A)|\big\}. \tag{4.52}$$

To identify which endpoint controls the maximum, compare their squares:

$$\begin{aligned} &(1 - \alpha\lambda_{\min}(A))^2 - (1 - \alpha\lambda_{\max}(A))^2 \\ &\quad = \alpha\big(\lambda_{\max}(A) - \lambda_{\min}(A)\big)\big(2 - \alpha(\lambda_{\max}(A) + \lambda_{\min}(A))\big). \end{aligned} \tag{4.53}$$

Suppose first that $\lambda_{\min}(A) < \lambda_{\max}(A)$. For $0 < \alpha \le \alpha_\star$, the last expression is nonnegative, so the smallest-eigenvalue endpoint controls the maximum. Also $\alpha_\star\lambda_{\min}(A) < 1$, and therefore $q(\alpha) = 1 - \alpha\lambda_{\min}(A)$ on this interval; this expression decreases as α increases. For $\alpha \ge \alpha_\star$, the largest-eigenvalue endpoint controls the maximum. Since $\alpha_\star\lambda_{\max}(A) > 1$, we have $q(\alpha) = \alpha\lambda_{\max}(A) - 1$ on this interval; this expression increases as α increases. The minimum occurs where the two endpoint magnitudes agree, at $\alpha = \alpha_\star$. Substitution gives

$$q(\alpha_\star) = \frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)} = \frac{\kappa_2(A) - 1}{\kappa_2(A) + 1}. \tag{4.54}$$

If the smallest and largest eigenvalues are equal, all eigenvalues equal some $\lambda > 0$. Then $q(\alpha) = |1 - \alpha\lambda|$ is minimized at $\alpha = 1/\lambda$, with value zero. Formula (4.50) gives exactly this step and factor.

The Euclidean error bound follows from (4.45), or directly from the definition of q. For the objective gap, use (4.39) and expand in the same orthonormal eigenbasis:

$$\begin{aligned} f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star) &= \frac{1}{2}\sum_{i=1}^n \lambda_i (1 - \alpha\lambda_i)^{2k} c_i^2, \\ f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_\star) &= \frac{1}{2}\sum_{i=1}^n \lambda_i (1 - \alpha\lambda_i)^{2(k+1)} c_i^2 \\ &\le q(\alpha)^2 \frac{1}{2}\sum_{i=1}^n \lambda_i (1 - \alpha\lambda_i)^{2k} c_i^2 \\ &= q(\alpha)^2 \big(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star)\big). \end{aligned} \tag{4.55}$$

Set $\alpha = \alpha_\star$ to obtain the second inequality in (4.51). ∎

The condition number measures how difficult it is for one step size to serve all directions. When $\kappa_2(A) = 1$, all curvatures are equal and the best step reaches the minimizer in one update. When $\kappa_2(A)$ is large, the factor $q_\star = 1 - 2/(\kappa_2(A) + 1)$ is close to one: even the best constant step can remove only a small fraction of the worst-case error in each iteration. This statement concerns a uniform guarantee, not the best step for a particular current point. The squared factor for objective values arises because the objective gap is quadratic in the error.

For the earlier objective $f(x_1, x_2) = \frac{1}{2}(x_1^2 + 10x_2^2)$, the Hessian eigenvalues are 1 and 10. With $\alpha = 0.1$, the error in the second coordinate disappears in one step, while the error in the first coordinate is multiplied by 0.9 at every step. The best constant step is $2/11$; its two error multipliers are $9/11$ and $-9/11$. It accepts an alternating error in the steep direction in exchange for a better uniform contraction factor. This explains why oscillation can accompany faster convergence rather than signal failure. The role of curvature bounds and conditioning in gradient descent is developed further in [2, Sections 9.1–9.3].

4.3 Exact line search and backtracking

A fixed step size must be chosen before seeing how much the function changes along the current search direction. The preceding quadratic analysis shows the difficulty: a step that is suitable in a direction of small curvature may be too large in a direction of large curvature. A *line search* chooses the step using function values along the direction that has already been selected. At iteration k, we hold the point $\boldsymbol{x}_k$ and the vector $-\nabla f(\boldsymbol{x}_k)$ fixed while varying only the scalar α. Thus, a multivariable problem produces a one-variable problem at each iteration.

Exact line search: choose the best point on the current ray. Define the one-variable function

$$\begin{aligned} h_k(\alpha) &= f\big(\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k)\big), \qquad \alpha \geq 0, \\ \alpha_k &\in \operatorname*{arg\,min}_{\alpha \geq 0} h_k(\alpha). \end{aligned} \tag{4.56}$$

The second line defines *exact line search*: among the points on the ray starting at $\boldsymbol{x}_k$ in direction $-\nabla f(\boldsymbol{x}_k)$, choose one with the lowest objective value. It does not search over every direction. If the gradient is zero, the ray consists of a single point and there is no step to choose. If the gradient is nonzero, sufficiently small positive steps decrease the objective, but an exact minimizing step need not exist. For example, for $f(x) = -x$ at $x_0 = 0$, the negative gradient is 1 and $h_0(\alpha) = -\alpha$ decreases indefinitely. Even when a minimum along the ray exists, solving this additional optimization problem can be costly. One-dimensional techniques from Section 2 can be used after a suitable interval has been bracketed. This reduction to one variable is discussed in [1, Section 13.2, p. 577].

Theorem 25 (Exact line search for a positive definite quadratic)**.** For the quadratic function in (4.38), suppose $\nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0}$. The exact line-search step is positive and unique, and is given by

$$\alpha_k = \frac{\|\nabla f(\boldsymbol{x}_k)\|_2^2}{\nabla f(\boldsymbol{x}_k)^\top A \nabla f(\boldsymbol{x}_k)}. \tag{4.57}$$

Proof. Substitute $\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k)$ into $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top A\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}$. The two cross terms from the quadratic expression combine because A is symmetric. Grouping the terms by their powers of α gives

$$\begin{aligned} h_k(\alpha) &= f(\boldsymbol{x}_k) - \alpha \nabla f(\boldsymbol{x}_k)^\top (A\boldsymbol{x}_k - \boldsymbol{b}) \\ &\quad + \frac{\alpha^2}{2} \nabla f(\boldsymbol{x}_k)^\top A \nabla f(\boldsymbol{x}_k) \\ &= f(\boldsymbol{x}_k) - \alpha \|\nabla f(\boldsymbol{x}_k)\|_2^2 + \frac{\alpha^2}{2} \nabla f(\boldsymbol{x}_k)^\top A \nabla f(\boldsymbol{x}_k), \end{aligned} \tag{4.58}$$

where the last equality uses $A\boldsymbol{x}_k - \boldsymbol{b} = \nabla f(\boldsymbol{x}_k)$. The gradient at $\boldsymbol{x}_k$ is fixed during the search, so differentiating with respect to α gives

$$\begin{aligned} h_k'(\alpha) &= -\|\nabla f(\boldsymbol{x}_k)\|_2^2 + \alpha \nabla f(\boldsymbol{x}_k)^{\mathsf{T}} A \nabla f(\boldsymbol{x}_k), \\ h_k''(\alpha) &= \nabla f(\boldsymbol{x}_k)^{\mathsf{T}} A \nabla f(\boldsymbol{x}_k) > 0. \end{aligned} \tag{4.59}$$

The strict inequality follows from positive definiteness of A and the nonzero-gradient assumption. Thus, h_k is a strictly convex quadratic in α. Its unique minimizer solves $h_k'(\alpha) = 0$, which gives (4.57). Both its numerator and denominator are positive, so this minimizer lies in the allowed range $\alpha > 0$. ■

The step formula adapts to the curvature in the current gradient direction. Its reciprocal is the quotient $\nabla f(\boldsymbol{x}_k)^{\mathsf{T}} A \nabla f(\boldsymbol{x}_k) / \|\nabla f(\boldsymbol{x}_k)\|_2^2$. In an eigenbasis of A, this quotient is a weighted average of the eigenvalues, with weights given by the squared components of the gradient divided by its squared norm. A gradient concentrated in a direction of large curvature therefore produces a smaller step, whereas a gradient concentrated in a direction of small curvature permits a larger one.

Why exact line search produces perpendicular successive directions. This fact follows from the one-variable optimality condition and does not require a quadratic objective. Suppose f is differentiable and the exact line minimum occurs at an interior point $\alpha_k > 0$. To differentiate h_k, first map α to the point $\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k)$, then apply f to that point. The derivative of this first map is the fixed vector $-\nabla f(\boldsymbol{x}_k)$. The chain rule therefore gives

$$h_k'(\alpha) = -\nabla f\big(\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k)\big)^{\mathsf{T}} \nabla f(\boldsymbol{x}_k). \tag{4.60}$$

At the minimizing step, the one-variable necessary condition is $h_k'(\alpha_k) = 0$. Since $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)$, we obtain

$$\nabla f(\boldsymbol{x}_{k+1})^{\mathsf{T}} \nabla f(\boldsymbol{x}_k) = 0. \tag{4.61}$$

When both gradients are nonzero, the successive search directions are consequently perpendicular. This does not mean that the method reaches the solution in two steps. After a step across a narrow valley, the next gradient can point back across the valley, producing many turns while making only modest progress along it. Exact line search selects the best step along the current direction; the geometry of the objective still determines that direction.

Figure 13 illustrates this behavior for $f(x_1, x_2) = \frac{1}{2}(x_1^2 + 10x_2^2)$ and $\boldsymbol{x}_0 = (10, 1)^{\mathsf{T}}$. Here $A = \operatorname{diag}(1, 10)$ and $\nabla f(\boldsymbol{x}_0) = (10, 10)^{\mathsf{T}}$, so the first exact step is $\alpha_0 = 200/1100 = 2/11$. In fact, the iterates have the explicit form

$$\boldsymbol{x}_k = \begin{pmatrix} 10(9/11)^k \\ (-1)^k (9/11)^k \end{pmatrix}. \tag{4.62}$$

To check this for every iteration, it holds at $k = 0$. If it holds at iteration k, the two components of the gradient have equal magnitude $10(9/11)^k$, and (4.57) again gives $\alpha_k = 2/11$. The next update multiplies the first coordinate by $1 - 2/11 = 9/11$ and the second by $1 - 20/11 = -9/11$, proving the formula at $k + 1$. Thus, the second coordinate changes sign at every step while the first remains positive, explaining the zigzag. For this particular initial point, exact line search coincides with the best constant step in (4.50). Other initial points need not preserve the equal gradient-component magnitudes, so the coincidence is not a general identity between the two methods.

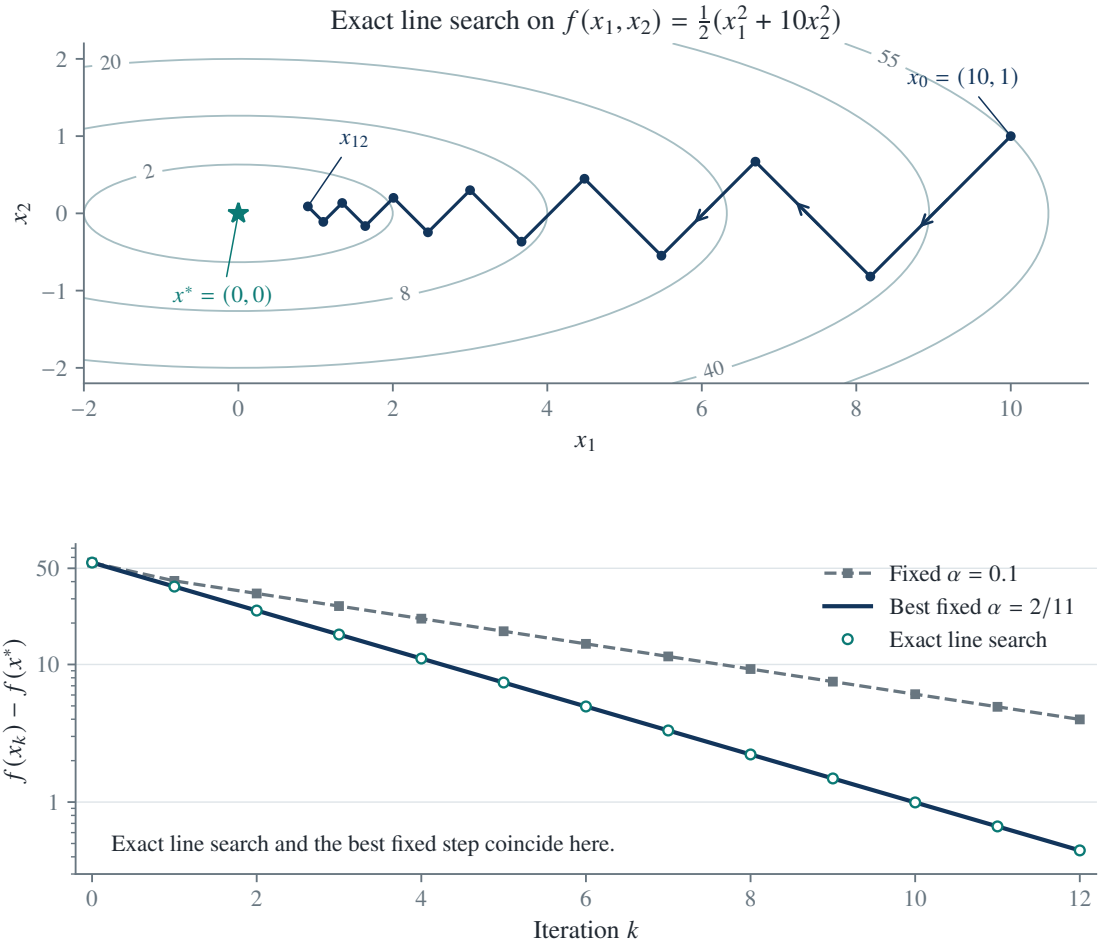


FIGURE 13: Gradient descent on $f(x_1, x_2) = \frac{1}{2}(x_1^2 + 10x_2^2)$, starting at $(10, 1)^\top$. The contour plot shows the first twelve exact-line-search steps. The convergence plot compares exact line search with two constant steps from the same initial point. The long valley forces repeated corrections across it; a locally optimal step length does not remove the effect of unequal curvature.

Backtracking: require sufficient progress instead of an exact minimum. Exact line search can spend substantial work finding the best point on a ray that will change at the next iteration. Backtracking asks a simpler question: does the proposed step reduce the objective by a reasonable amount? The linear model at $\boldsymbol{x}_k$ predicts

$$f\big(\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k)\big) \approx f(\boldsymbol{x}_k) - \alpha \|\nabla f(\boldsymbol{x}_k)\|_2^2. \tag{4.63}$$

Thus, $\alpha\|\nabla f(\boldsymbol{x}_k)\|_2^2$ is the predicted decrease. The approximation becomes more reliable for small steps, but curvature can make it inaccurate for large steps. Choose a number $c \in (0, 1)$ and require the actual decrease to be at least the fraction c of this prediction. This gives the *Armijo condition*

$$f\big(\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k)\big) \le f(\boldsymbol{x}_k) - c\alpha \|\nabla f(\boldsymbol{x}_k)\|_2^2. \tag{4.64}$$

For a nonzero gradient and a positive step, the right-hand side is strictly smaller than $f(\boldsymbol{x}_k)$. The condition therefore requires a definite improvement rather than merely accepting a step that does not increase the objective.

Choose an initial trial step $\bar{\alpha} > 0$ and a reduction factor $\beta \in (0, 1)$. Test $\alpha = \bar{\alpha}$. If (4.64) fails, replace α by $\beta\alpha$ and test again, keeping $\boldsymbol{x}_k$ and $\nabla f(\boldsymbol{x}_k)$ fixed throughout the search. The possible trial values are $\bar{\alpha}, \beta\bar{\alpha}, \beta^2\bar{\alpha}, \ldots$. The first accepted value becomes α_k in the gradient-descent update. A smaller c asks for a smaller fraction of the predicted decrease, whereas a smaller β reduces rejected trial steps more aggressively. A common choice is $\beta = 1/2$ with a small positive c, such as 10^{-4}. This construction also appears in [3].

Theorem 26 (Finite termination of backtracking)**.** Suppose f has an L-Lipschitz gradient on $\mathbb{R}^n$ and $\nabla f(\boldsymbol{x}_k) \neq \mathbf{0}$. For $c, \beta \in (0, 1)$ and $\bar{\alpha} > 0$, backtracking accepts a positive step after finitely many trials in exact arithmetic. The accepted step strictly decreases f.

Proof. The descent bound (4.14) gives

$$f\big(\boldsymbol{x}_k - \alpha\nabla f(\boldsymbol{x}_k)\big) \leq f(\boldsymbol{x}_k) - \alpha\left(1 - \frac{L\alpha}{2}\right)\|\nabla f(\boldsymbol{x}_k)\|_2^2. \tag{4.65}$$

To obtain the Armijo condition, it is enough that the factor $1 - L\alpha/2$ multiplying the predicted decrease is at least c. Solving this inequality for α gives

$$1 - \frac{L\alpha}{2} \geq c \quad \Longleftrightarrow \quad \alpha \leq \frac{2(1-c)}{L}. \tag{4.66}$$

The number on the right is positive because $c < 1$. Since $0 < \beta < 1$, the geometric sequence $\beta^j\bar{\alpha}$ tends to zero as the number j of reductions increases. Therefore, some finite trial satisfies this sufficient bound and must be accepted; an earlier trial may already have passed. Every trial is positive. Once the Armijo condition holds, its decrease term $c\alpha\|\nabla f(\boldsymbol{x}_k)\|_2^2$ is strictly positive, which proves strict decrease. ■

The argument does not assume convexity. It also does not require the algorithm to know L: this constant is used to prove that sufficiently small trials succeed, whereas the algorithm tests the actual function values. For an objective with a restricted domain, a trial point must lie in that domain before its objective can be evaluated. In floating-point arithmetic, an implementation needs a trial limit and a separate failure report, because rounding or numerical evaluation errors can interfere with a strict decrease test.

Example 27 (Backtracking rejects a step that makes no progress)**.** For $f(x) = (x-1)^2$ at $x = 3$, we have $f(3) = 4$ and $f'(3) = 4$. Choose $\bar{\alpha} = 1$, $\beta = 1/2$, and $c = 10^{-4}$. The first trial point is

$$3 - 1 \cdot f'(3) = -1, \qquad f(-1) = 4. \tag{4.67}$$

The Armijo condition would require

$$4 \leq f(3) - 10^{-4} \cdot 1 \cdot |f'(3)|^2 = 4 - 0.0016 = 3.9984, \tag{4.68}$$

which is false. This step crosses the minimizer and returns to the same objective value on the opposite side. Repeatedly taking this step would alternate between 3 and -1, making no progress. Backtracking

therefore rejects it and tries $\alpha = 1/2$. The new trial point is $3 - (1/2) \cdot 4 = 1$, and the test becomes

$$f(1) = 0 \leq 4 - 10^{-4} \cdot \frac{1}{2} \cdot 16 = 3.9992. \tag{4.69}$$

The smaller step is accepted and reaches the minimizer. This calculation shows why the acceptance rule asks for an actual amount of decrease, rather than only checking whether the objective has failed to increase.