Mathematical Foundations of Data Sciences

Gabriel Peyré
CNRS & DMA, École Normale Supérieure

CHAPTER 14

Optimization & Machine Learning: Smooth Optimization

September 9, 2026

Training a model often reduces to minimizing a differentiable objective, so the geometry of that objective determines how optimization algorithms progress. Starting from regression and classification, we develop convexity, derivatives, and optimality conditions, then derive gradient descent and practical step-size rules. Quadratic examples guide the convergence analysis, which extends to smooth convex objectives and explains the effects of conditioning, strong convexity, and acceleration.

14.1 Motivation in Machine Learning

14.1.1 Unconstrained Optimization

Throughout most of this chapter, we consider unconstrained convex optimization problems of the form

$$\inf_{x\in\mathbb{R}^p} f(x), \tag{14.1}$$

and seek algorithms that approximate a minimizer, when one exists, with a low cost per iteration. The first-order methods studied here use gradient information. We write

$$\operatorname*{argmin}_x f(x) \overset{\text{def.}}{=} \{x \in \mathbb{R}^p \;;\; f(x) = \inf f\},$$

for the set of minimizers of f, which may contain several points or be empty: $\operatorname{argmin} f = \emptyset$. When a minimizer exists, we write the optimization problem as

$$\min_{x\in\mathbb{R}^p} f(x). \tag{14.2}$$

In learning, $f(x)$ is typically an empirical risk for regression or classification, and p is the number of model parameters. For a linear model, the training sample is $(a_i, y_i)_{i=1}^n$ with feature vectors $a_i \in \mathbb{R}^p$. Let $A \in \mathbb{R}^{n\times p}$ be the matrix with feature vectors a_i as its rows.

14.1.2 Regression

For real-valued responses $y_i \in \mathbb{R}$, the least-squares objective is

$$f(x) = \frac{1}{2}\sum_{i=1}^n (y_i - \langle x, a_i\rangle)^2 = \frac{1}{2}\|Ax - y\|^2, \tag{14.3}$$

as illustrated in Figure 14.1. Here $\langle u, v\rangle = \sum_{i=1}^p u_i v_i$ is the canonical inner product in $\mathbb{R}^p$ and $\|\cdot\|^2 = \langle\cdot,\cdot\rangle$.

14.1.3 Classification

For classification, $y_i \in \{-1, 1\}$, in which case

$$f(x) = \sum_{i=1}^n \ell(-y_i\langle x, a_i\rangle) = L(-\operatorname{diag}(y)Ax) \tag{14.4}$$

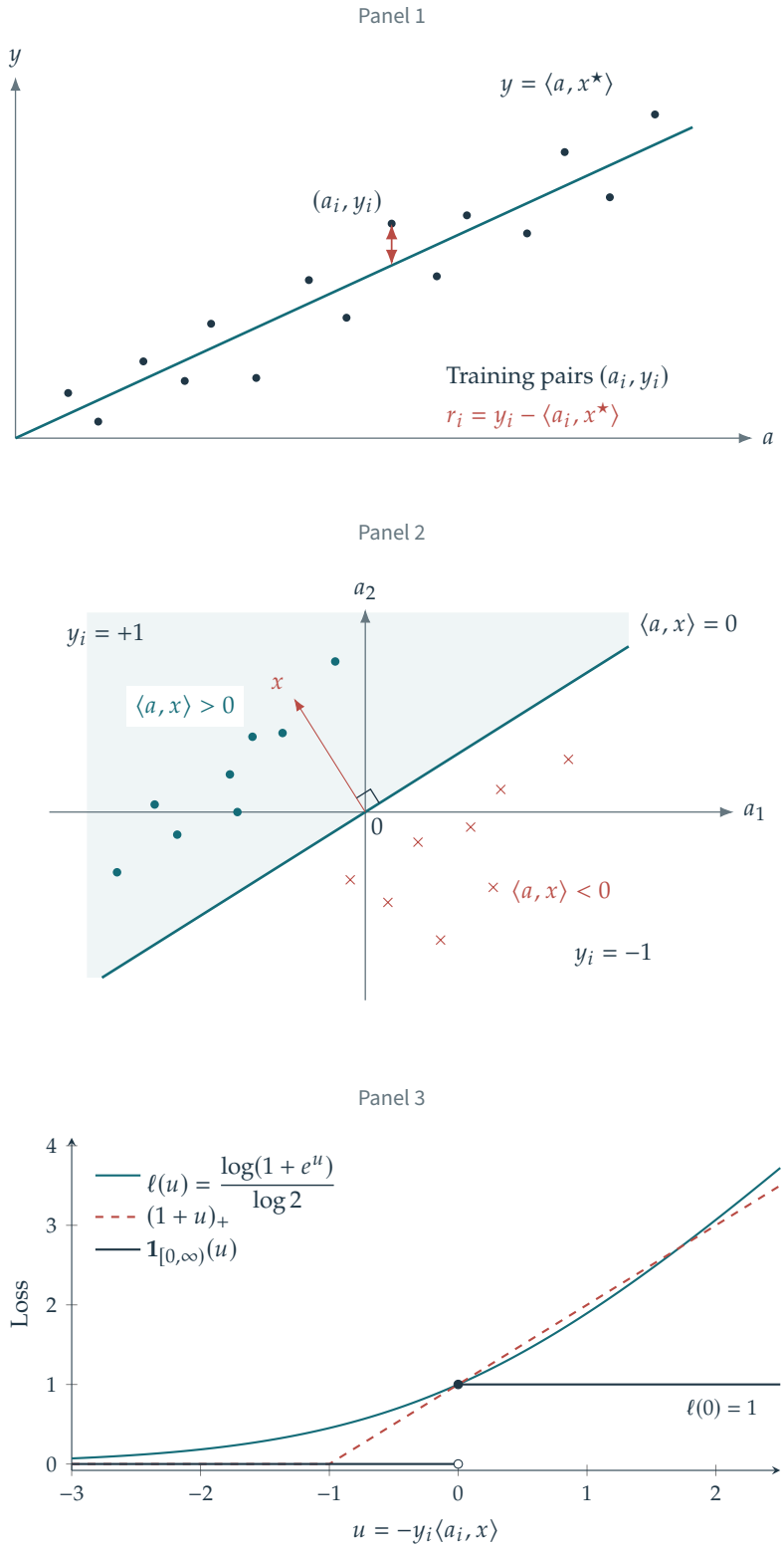

Figure 14.1. Panels 1 and 2: linear regression and a linear classifier. Panel 3: the 0-1 loss and its normalized logistic and hinge upper bounds, both tight at zero margin.

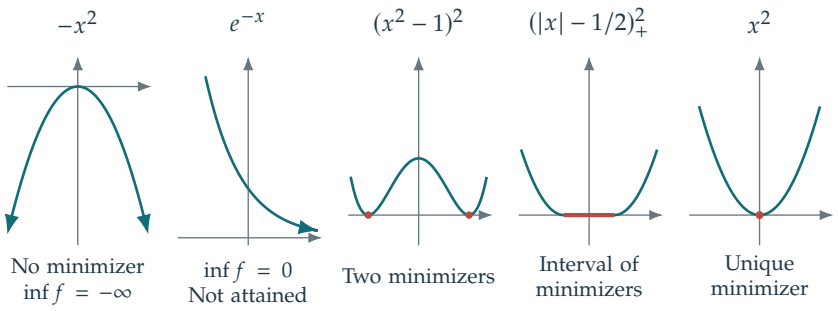

Figure 14.2. Left: no minimizer. Middle: multiple minimizers. Right: a unique minimizer.

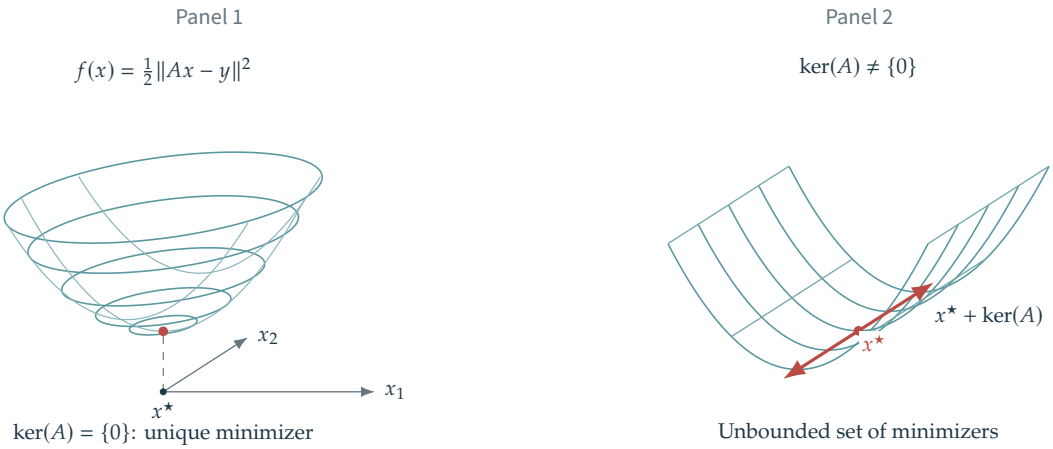

Figure 14.3. Coercivity condition for least squares.

where ℓ is a smooth convex surrogate for the 0-1 loss $\mathbf{1}_{[0,\infty)}$, counting zero margin as an error. For instance, $\ell(u) = \log(1 + \exp(u))/\log 2$ satisfies $\ell(u) \geqslant \mathbf{1}_{[0,\infty)}(u)$, with equality at $u = 0$. The normalization makes this upper bound tight at the decision boundary. The matrix $\mathrm{diag}(y) \in \mathbb{R}^{n\times n}$ is the diagonal matrix with y_i along the diagonal (see Figure 14.1, panel 3). The separable loss function $L : \mathbb{R}^n \to \mathbb{R}$ is, for $z \in \mathbb{R}^n$, $L(z) = \sum_i \ell(z_i)$.

14.2 Basics of Convex Analysis

14.2.1 Existence of Solutions

Problem (14.1) need not have a minimizer. An objective f may be unbounded below: for $f(x) = -x^2$, the infimum is $-\infty$. The objective f can also have a finite unattained infimum: for $f(x) = e^{-x}$, it is $\inf f = 0$.

A continuous function attains its minimum on a nonempty compact set $\Omega \subset \mathbb{R}^p$. In finite dimensions, compactness means that Ω is closed and bounded. Lower semicontinuity suffices for attainment, but we use continuity here. For unconstrained minimization, coercivity provides a compact set to which the search can be restricted: $f(x) \to +\infty$ as $\|x\| \to +\infty$. For any $x_0 \in \mathbb{R}^p$, consider the sublevel set

$$\Omega = \{x \in \mathbb{R}^p \ ; \ f(x) \leqslant f(x_0)\}$$

Coercivity makes this set bounded, and continuity of f makes it closed. For a continuous convex function, coercivity is in fact equivalent to a nonempty bounded minimizer set. Convexity matters: $f(x) = \min(1, x^2)$ has a unique minimizer but is not coercive.

Example 14.1 (Least squares). The least-squares objective $f(x) = \frac{1}{2}\|Ax - y\|^2$ is coercive exactly when $\ker(A) = \{0\}$, meaning the design has full column rank. If $\ker(A) \neq \{0\}$ and $x^\star$ is a minimizer, then $x^\star + u$ is also a minimizer for every $u \in \ker(A)$, so the minimizer set is unbounded. Conversely, if $\ker(A) = \{0\}$, we will show later that the minimizer is unique, see Fig. 14.3. For classification, strict convexity of ℓ and injectivity of A imply strict convexity of the objective, but do not ensure coercivity or existence. Separable logistic regression is a counterexample.

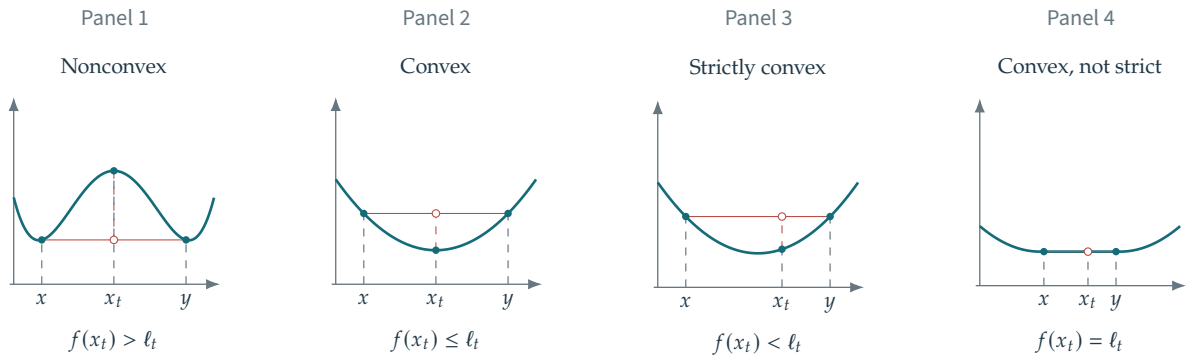

Figure 14.4. Convex and nonconvex functions; strictly convex and convex but not strictly convex functions. Here $x_t = (1-t)x + ty$ and $\ell_t = (1-t)f(x) + tf(y)$, with $0 < t < 1$.

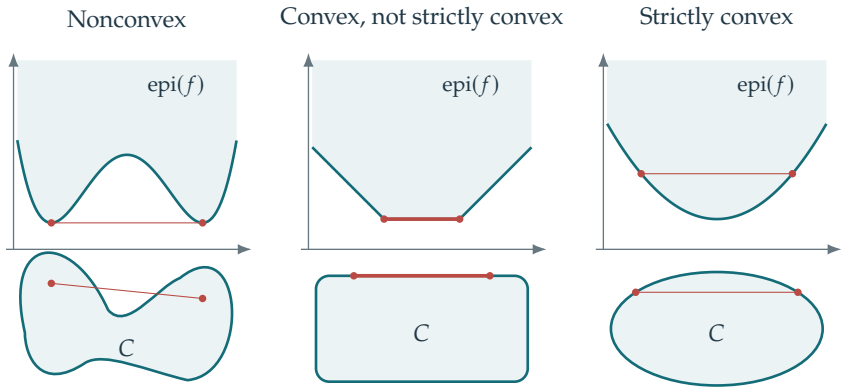

Figure 14.5. Comparison of convex functions $f : \mathbb{R}^p \to \mathbb{R}$ (for $p = 1$) and convex sets $C \subset \mathbb{R}^p$ (for $p = 2$).

14.2.2 Convexity

Convexity is central to optimization because every local minimizer is global, and many convex problems admit efficient algorithms. A function is convex if, for every pair $(x, y) \in (\mathbb{R}^p)^2$,

$$\forall\, t \in [0,1], \quad f((1-t)x + ty) \leqslant (1-t)f(x) + tf(y) \tag{14.5}$$

The graph lies below each secant, and above each tangent when differentiable; see Figure 14.4. If $x^\star$ is a local minimizer of a convex f, then $x^\star$ is a global minimizer, i.e. $x^\star \in \operatorname{argmin} f$.

Convexity is preserved by several useful operations. If f and g are convex and a, b are positive, then $af + bg$ and $\max(f, g)$ are convex. If $g : \mathbb{R}^q \to \mathbb{R}$ is convex and $B \in \mathbb{R}^{q\times p}, b \in \mathbb{R}^q$, then the affine composition $f(x) = g(Bx + b)$ is convex. Thus the squared loss (14.3) is convex, since $\|\cdot\|^2/2$ is a sum of convex squares. Similarly, convexity of ℓ implies convexity of L and of the classification objective (14.4).

Strict convexity. When f is convex, one can strengthen the condition (14.5) and impose that the inequality is strict for distinct x, y and $t \in]0,1[$ (see Figure 14.4, panels 3 and 4), i.e.

$$\forall\, t \in]0,1[, \quad f((1-t)x + ty) < (1-t)f(x) + tf(y). \tag{14.6}$$

A minimizer $x^\star$ is then unique if it exists. Otherwise, two minimizers $x_1^\star \neq x_2^\star$ would satisfy $f(\frac{x_1^\star + x_2^\star}{2}) < f(x_1^\star)$ by strict convexity, contradicting optimality.

Example 14.2 (Least squares). For the quadratic loss function $f(x) = \frac{1}{2}\|Ax - y\|^2$, strict convexity is equivalent to $\ker(A) = \{0\}$. Its Hessian is $\partial^2 f(x) = A^\top A$, which is positive definite exactly when A is injective, since $\langle A^\top Az, z\rangle = \|Az\|^2$. Positive definiteness implies strict convexity. Conversely, a nonzero vector in $\ker(A)$ gives a line along which f is constant, ruling out strict convexity.

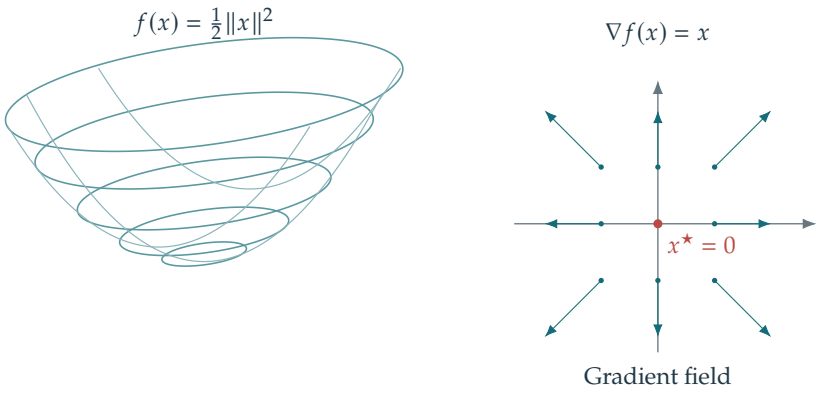

Figure 14.6. The gradient as a vector field.

14.2.3 Convex Sets

A set $\Omega \subset \mathbb{R}^p$ is convex if every pair $(x, y) \in \Omega^2$ satisfies $(1-t)x + ty \in \Omega$ for all $t \in [0,1]$. A function f is convex exactly when its epigraph $\mathrm{epi}(f) \stackrel{\text{def.}}{=} \{(x,t) \in \mathbb{R}^{p+1} \;;\; t \geqslant f(x)\}$ is a convex set.

Remark 14.3 (Convexity of the minimizer set). A minimizer $x^\star$ need not be unique; see Figure 14.3. For convex f, the set $\mathrm{argmin}(f)$ is itself convex. Indeed, if $x_1^\star$ and $x_2^\star$ are minimizers, so that in particular $f(x_1^\star) = f(x_2^\star) = \min(f)$, then $f((1-t)x_1^\star + tx_2^\star) \leqslant (1-t)f(x_1^\star) + tf(x_2^\star) = f(x_1^\star) = \min(f)$, so that $(1-t)x_1^\star + tx_2^\star$ is itself a minimizer. Figure 14.5 shows convex and non-convex sets.

14.3 Derivative and gradient

14.3.1 Gradient

When all coordinate derivatives of f exist, define

$$\nabla f(x) \stackrel{\text{def.}}{=} \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_p} \right)^\top \in \mathbb{R}^p$$

as the gradient vector. The map $\nabla f : \mathbb{R}^p \to \mathbb{R}^p$ is a vector field, and its coordinate derivatives are defined by

$$\frac{\partial f(x)}{\partial x_k} \stackrel{\text{def.}}{=} \lim_{\eta \to 0} \frac{f(x + \eta \delta_k) - f(x)}{\eta}$$

where $\delta_k = (0, \ldots, 0, 1, 0, \ldots, 0)^\top \in \mathbb{R}^p$ is the kth canonical basis vector.

Existence of the coordinate derivatives, hence of $\nabla f(x)$, does not imply differentiability of f. Differentiability of f at x requires

$$f(x + \varepsilon) = f(x) + \langle \varepsilon, \nabla f(x) \rangle + o(\|\varepsilon\|). \tag{14.7}$$

The remainder notation $R(\varepsilon) = o(\|\varepsilon\|)$ means decay faster than the perturbation ε as it tends to 0: $\frac{R(\varepsilon)}{\|\varepsilon\|} \to 0$ as $\varepsilon \to 0$. Coordinate derivatives concern the behavior of f along the axes, whereas differentiability requires the expansion for every sequence $\varepsilon \to 0$. For example, $f(x) = \frac{2x_1x_2(x_1+x_2)}{x_1^2+x_2^2}$ with $f(0) = 0$ has both coordinate derivatives equal to zero at the origin. Yet $f(t,t) = 2t$, so the linear expansion with zero gradient fails.

The vector $\nabla f(x)$ is uniquely determined by (14.7). To prove differentiability of f and compute $\nabla f(x)$, it therefore suffices to establish an expansion

$$f(x + \varepsilon) = f(x) + \langle \varepsilon, g \rangle + o(\|\varepsilon\|),$$

which identifies $\nabla f(x) = g$.

For differentiable functions, convexity is equivalent to the graph lying above every tangent affine function.

Proposition 14.4. *If f is differentiable, then*

$$f \text{ convex} \quad \Leftrightarrow \quad \forall (x, x'),\ f(x) \geqslant f(x') + \langle \nabla f(x'), x - x' \rangle.$$

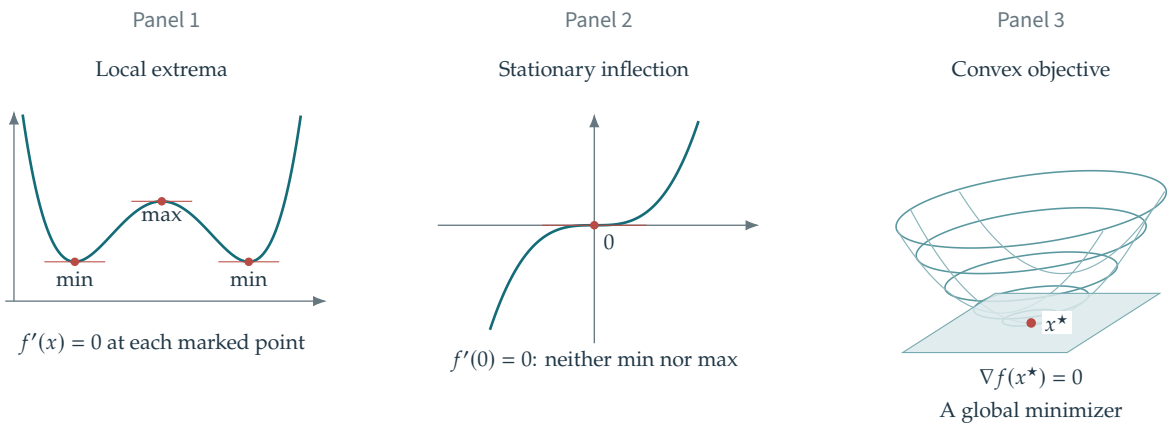

Figure 14.7. Local extrema (panel 1), a stationary inflection point (panel 2), and a global minimizer (panel 3).

Proof. One can write the convexity condition as

$$f((1-t)x + tx') \leqslant (1-t)f(x) + tf(x') \quad \Longrightarrow \quad \frac{f(x + t(x'-x)) - f(x)}{t} \leqslant f(x') - f(x)$$

hence, taking the limit $t \to 0$ one obtains

$$\langle \nabla f(x),\, x' - x \rangle \leqslant f(x') - f(x).$$

Conversely, set $x_t \overset{\text{def.}}{=} (1-t)x + tx'$. Apply the supporting-affine-function inequality at x_t to x and x':

$$f(x) \geqslant f(x_t) + \langle \nabla f(x_t),\, x - x_t \rangle = f(x_t) + t\langle \nabla f(x_t),\, x - x' \rangle$$
$$f(x') \geqslant f(x_t) + \langle \nabla f(x_t),\, x' - x_t \rangle = f(x_t) - (1-t)\langle \nabla f(x_t),\, x - x' \rangle,$$

Multiplying the first inequality by $1-t$ and the second by t, then adding, gives

$$(1-t)f(x) + tf(x') \geqslant f(x_t).$$

□

14.3.2 First Order Conditions

A vanishing gradient is necessary for local optimality, and will also guide the algorithms developed below.

Proposition 14.5. *If f is differentiable and $x^\star$ is a local minimizer, meaning that $f(x^\star) \leqslant f(x)$ throughout some neighborhood of $x^\star$, then*

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

Proof. Fix a direction u and let $\varepsilon \downarrow 0$. Local optimality and differentiability give

$$f(x^\star) \leqslant f(x^\star + \varepsilon u) = f(x^\star) + \varepsilon\langle \nabla f(x^\star),\, u \rangle + o(\varepsilon) \quad \Longrightarrow \quad \langle \nabla f(x^\star),\, u \rangle \geqslant o(1) \quad \Longrightarrow \quad \langle \nabla f(x^\star),\, u \rangle \geqslant 0.$$

So applying this for u and $-u$ in the previous equation shows that $\langle \nabla f(x^\star),\, u \rangle = 0$ for all u, and hence $\nabla f(x^\star) = 0$. □

Note that the converse is not true in general, since one might have $\nabla f(x) = 0$ but x is not a local minimum. For instance $x = 0$ for $f(x) = -x^2$ (here x is a maximizer) or $f(x) = x^3$ (here x is neither a maximizer nor a minimizer; it is a stationary inflection point), see Fig. 14.7. Stationarity alone therefore neither certifies a local minimum nor guarantees convergence of a numerical method to that point. Both depend on the local geometry. Convexity of f makes stationarity sufficient as well as necessary.

Proposition 14.6. *If f is convex and $x^\star$ is a local minimizer, then $x^\star$ is a global minimizer. If f is differentiable and convex,*

$$x^\star \in \operatorname*{argmin}_x f(x) \quad \Longleftrightarrow \quad \nabla f(x^\star) = 0.$$

Proof. Fix x and choose $0 < t < 1$ so that $tx + (1-t)x^\star$ lies in the neighborhood where $x^\star$ is minimal. Local optimality and convexity then give

$$f(x^\star) \leqslant f(tx + (1-t)x^\star) \leqslant tf(x) + (1-t)f(x^\star) \quad \Longrightarrow \quad f(x^\star) \leqslant f(x)$$

proving that $x^\star$ is a global minimizer.

For the second part, we already saw in Proposition 14.5 the $\Rightarrow$ implication. We assume that $\nabla f(x^\star) = 0$. Since the graph of f is above its tangent by convexity (as stated in Proposition 14.4),

$$f(x) \geqslant f(x^\star) + \langle \nabla f(x^\star),\, x - x^\star \rangle = f(x^\star).$$

□

For a differentiable convex objective, minimization is therefore equivalent to solving $\nabla f(x) = 0$, a system of p equations in p unknowns. An explicit solution is often unavailable, but the equation still characterizes the minimizers $x^\star$.

14.3.3 Least Squares

The gradient of the least-squares objective (14.3) follows by expanding the squared norm:

$$\begin{aligned} f(x+\varepsilon) &= \frac{1}{2}\|Ax - y + A\varepsilon\|^2 = \frac{1}{2}\|Ax-y\|^2 + \langle Ax - y,\, A\varepsilon\rangle + \frac{1}{2}\|A\varepsilon\|^2 \\ &= f(x) + \langle \varepsilon,\, A^\top(Ax-y)\rangle + o(\|\varepsilon\|). \end{aligned}$$

We used $\|A\varepsilon\|^2 = o(\|\varepsilon\|)$ and the transpose $A^\top$. The transpose exchanges rows and columns, $A^\top = (A_{j,i})_{i=1,\dots,n}^{j=1,\dots,p}$. Its essential role in gradient calculations is the adjoint identity

$$\forall\, (u,v) \in \mathbb{R}^p \times \mathbb{R}^n, \quad \langle Au,\, v\rangle_{\mathbb{R}^n} = \langle u,\, A^\top v\rangle_{\mathbb{R}^p}.$$

Computing gradients of functions involving linear operators requires such a transposition step. This computation shows that

$$\nabla f(x) = A^\top(Ax - y). \tag{14.8}$$

Hence every minimizer $x^\star$ of $f(x)$ satisfies the normal equations $(A^\top A)x^\star = A^\top y$. If $A^\star A \in \mathbb{R}^{p\times p}$ is invertible, then f has a single minimizer, namely

$$x^\star = (A^\top A)^{-1}A^\top y. \tag{14.9}$$

Thus $x^\star$ depends linearly on y. The operator $(A^\top A)^{-1}A^\star$ is the Moore–Penrose pseudoinverse of A; an ordinary inverse is unavailable when $p \neq n$. The condition that $A^\top A$ is invertible is equivalent to $\ker(A) = \{0\}$, since

$$A^\top Ax = 0 \quad \Longrightarrow \quad \|Ax\|^2 = \langle A^\top Ax,\, x\rangle = 0 \quad \Longrightarrow \quad Ax = 0.$$

When $n < p$, the system is underdetermined and injectivity is impossible. If $n \geqslant p$ and the entries of A have a joint density, then $\ker(A) = \{0\}$ almost surely. In general, injectivity is equivalent to the feature vectors a_i spanning $\mathbb{R}^p$.

14.3.4 Link with PCA

Assume the feature vectors are centered, $\sum_i a_i = 0$. Otherwise, replace each a_i by $a_i - m$, where $m \overset{\text{def.}}{=} n^{-1}\sum_i a_i$ is the empirical mean. Write $C \overset{\text{def.}}{=} A^\top A$. Then C/n is the empirical covariance matrix of the point cloud. With $a_i = (a_{i,1}, \dots, a_{i,p})^\top$ and $A = (a_{i,j})_{i,j}$, its entries are

$$\forall\, (k,\ell) \in \{1,\dots,p\}^2, \quad \frac{C_{k,\ell}}{n} = \frac{1}{n}\sum_{i=1}^n a_{i,k}a_{i,\ell}.$$

In particular, $C_{k,k}/n$ is the variance along the axis k. More generally, for any unit vector $u \in \mathbb{R}^p$, $\langle Cu,\, u\rangle/n \geqslant 0$ is the variance along the axis u.

For instance, in dimension $p = 2$,

$$\frac{C}{n} = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n a_{i,1}^2 & \sum_{i=1}^n a_{i,1}a_{i,2} \\ \sum_{i=1}^n a_{i,1}a_{i,2} & \sum_{i=1}^n a_{i,2}^2 \end{pmatrix}.$$

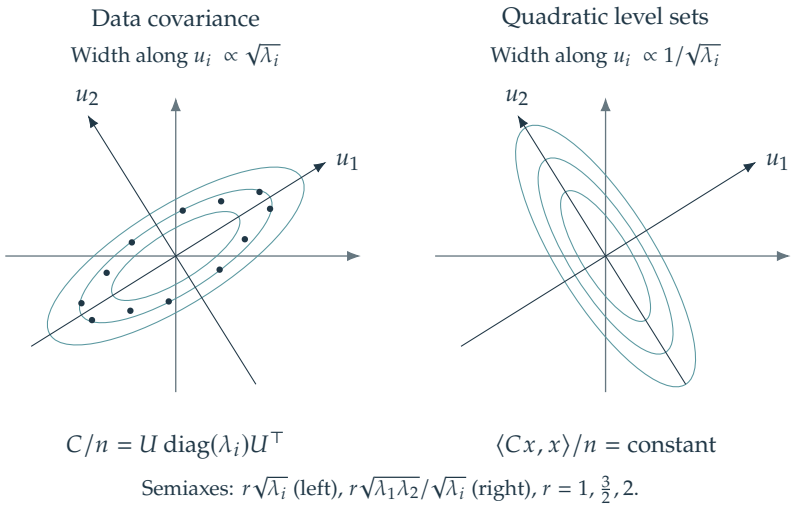

Figure 14.8. Left: point clouds $(a_i)_i$ and their principal directions. Right: the quadratic part of $f(x)$.

Since C is symmetric, choose an orthogonal matrix $U = (u_1, \dots, u_p)$ whose columns are eigenvectors of C/n. If $(C/n)u_k = \lambda_k u_k$, then $C/n = U \operatorname{diag}(\lambda_k) U^\top$ and $U^{-1} = U^\top$. In these coordinates, the quadratic form is a weighted sum of squares:

$$\frac{1}{n}\langle Cx,\, x\rangle = \langle U \operatorname{diag}(\lambda_k) U^\top x,\, x\rangle = \langle \operatorname{diag}(\lambda_k)(U^\top x),\, (U^\top x)\rangle = \sum_{k=1}^{p} \lambda_k \langle x,\, u_k\rangle^2. \tag{14.10}$$

Here $(U^\top x)_k = \langle x,\, u_k\rangle$ is the coordinate k of x in the basis U. Since $\langle Cx,\, x\rangle = \|Ax\|^2$, this shows that all eigenvalues are nonnegative.

Order the eigenvalues as $\lambda_1 \geqslant \lambda_2 \geqslant \dots \geqslant \lambda_p$. Projection of a_i onto the leading m eigenvectors gives principal component analysis (PCA), the optimal linear reconstruction in squared error. It is used for compression, dimensionality reduction, and visualization in two dimensions ($m = 2$) or three dimensions ($m = 3$).

When C is positive definite, its covariance ellipsoid has principal axes $(u_k)_k$ and semiaxes proportional to the standard deviations $\sqrt{\lambda_k}$. This ellipsoid closely describes a Gaussian cloud with density proportional to $\exp(-\frac{n}{2}\langle C^{-1}a,\, a\rangle)$. For the quadratic objective $\frac{1}{2}\langle Cx,\, x\rangle$ in $f(x)$, the level ellipsoid $\left\{x\,;\, \frac{1}{n}\langle Cx,\, x\rangle \leqslant 1\right\}$ has the same principal axes but reciprocal widths $1/\sqrt{\lambda_k}$. Figure 14.8 shows the reciprocal axis widths in dimension $p = 2$. If C is singular, the covariance ellipsoid is degenerate and the quadratic sublevel sets are unbounded along $\ker(C)$.

14.3.5 Classification

For the classification objective (14.4), assume L is differentiable and apply (14.7) at $-\operatorname{diag}(y)Ax$:

$$\begin{aligned} f(x+\varepsilon) &= L(-\operatorname{diag}(y)Ax - \operatorname{diag}(y)A\varepsilon) \\ &= L(-\operatorname{diag}(y)Ax) + \langle \nabla L(-\operatorname{diag}(y)Ax),\, -\operatorname{diag}(y)A\varepsilon\rangle + o(\|\operatorname{diag}(y)A\varepsilon\|). \end{aligned}$$

Using the fact that $o(\|\operatorname{diag}(y)A\varepsilon\|) = o(\|\varepsilon\|)$, one obtains

$$\begin{aligned} f(x+\varepsilon) &= f(x) + \langle \nabla L(-\operatorname{diag}(y)Ax),\, -\operatorname{diag}(y)A\varepsilon\rangle + o(\|\varepsilon\|) \\ &= f(x) + \langle -A^\top \operatorname{diag}(y) \nabla L(-\operatorname{diag}(y)Ax),\, \varepsilon\rangle + o(\|\varepsilon\|), \end{aligned}$$

Using $(AB)^\top = B^\top A^\top$ and $\operatorname{diag}(y)^\top = \operatorname{diag}(y)$ gives

$$\nabla f(x) = -A^\top \operatorname{diag}(y) \nabla L(-\operatorname{diag}(y)Ax).$$

Since $L(z) = \sum_i \ell(z_i)$, its gradient is $\nabla L(z) = (\ell'(z_i))_{i=1}^n$. For the normalized logistic loss, $\ell'(u) = e^u/((1+e^u)\log 2)$. Evaluated at the negative margin $u = -y_i\langle x,\, a_i\rangle$, this scaled sigmoid is large for incorrectly classified observations and weights their contribution to the gradient.

14.3.6 Chain Rule

More generally, for $f(x) = g(Bx)$ with $B \in \mathbb{R}^{q\times p}$ and $g : \mathbb{R}^q \to \mathbb{R}$,

$$f(x+\varepsilon) = g(Bx + B\varepsilon) = g(Bx) + \langle \nabla g(Bx),\, B\varepsilon\rangle + o(\|B\varepsilon\|) = f(x) + \langle \varepsilon,\, B^\top \nabla g(Bx)\rangle + o(\|\varepsilon\|),$$

which shows that

$$\nabla(g \circ B) = B^\top \circ \nabla g \circ B \tag{14.11}$$

where "$\circ$" denotes the composition of functions.

Nonlinear compositions require the differential. For $F : \mathbb{R}^p \to \mathbb{R}^q$, its differential at x is the linear map $\partial F(x) : \mathbb{R}^p \to \mathbb{R}^q$. We represent it by its Jacobian matrix, also denoted $\partial F(x)$, with dimensions $\partial F(x) \in \mathbb{R}^{q\times p}$. Writing $F(x) = (F_1(x), \ldots, F_q(x))$, the Jacobian entries are

$$\forall\, (i,j) \in \{1,\ldots,q\} \times \{1,\ldots,p\}, \quad [\partial F(x)]_{i,j} \overset{\text{def.}}{=} \frac{\partial F_i(x)}{\partial x_j}.$$

The map F is differentiable at x when it admits the expansion

$$F(x+\varepsilon) = F(x) + [\partial F(x)](\varepsilon) + o(\|\varepsilon\|). \tag{14.12}$$

Here $[\partial F(x)](\varepsilon)$ denotes matrix-vector multiplication. The matrix in this expansion is unique, so the expansion can also be used to compute the differential.

For the special case $q = 1$, i.e. if $f : \mathbb{R}^p \to \mathbb{R}$, then the differential $\partial f(x) \in \mathbb{R}^{1\times p}$ and the gradient $\nabla f(x) \in \mathbb{R}^{p\times 1}$ are linked by equating the Taylor expansions (14.12) and (14.7)

$$\forall\, \varepsilon \in \mathbb{R}^p, \quad [\partial f(x)](\varepsilon) = \langle \nabla f(x),\, \varepsilon\rangle \quad \Leftrightarrow \quad \partial f(x) = \nabla f(x)^\top.$$

The differential satisfies the following chain rule

$$\partial(G \circ H)(x) = [\partial G(H(x))] \times [\partial H(x)]$$

where "$\times$" is the matrix product. For instance, if $H : \mathbb{R}^p \to \mathbb{R}^q$ and $G = g : \mathbb{R}^q \mapsto \mathbb{R}$, then $f = g \circ H : \mathbb{R}^p \to \mathbb{R}$ and one can compute its gradient as follows

$$\nabla f(x) = (\partial f(x))^\top = ([\partial g(H(x))] \times [\partial H(x)])^\top = [\partial H(x)]^\top \times [\partial g(H(x))]^\top = [\partial H(x)]^\top \times \nabla g(H(x)).$$

When $H(x) = Bx$ is linear, one recovers formula (14.11).

14.4 Gradient Descent Algorithm

14.4.1 Steepest Descent Direction

The Taylor expansion (14.7) approximates f near x by an affine function:

$$f(z) = T_x(z) + o(\|x - z\|) \quad \text{where} \quad T_x(z) \overset{\text{def.}}{=} f(x) + \langle \nabla f(x),\, z - x\rangle,$$

Figure 14.9 illustrates the approximation. First-order methods use f through its local model T_x.

A nonzero gradient $\nabla f(x)$ points uphill locally, so $-\nabla f(x)$ is a descent direction. At a fixed point x, examine f along the ray

$$\tau \in \mathbb{R}^+ = [0, +\infty[\longmapsto f(x - \tau \nabla f(x)) \in \mathbb{R}.$$

If f is differentiable at x, one has

$$f(x - \tau\nabla f(x)) = f(x) - \tau\langle \nabla f(x),\, \nabla f(x)\rangle + o(\tau) = f(x) - \tau\|\nabla f(x)\|^2 + o(\tau).$$

If $\nabla f(x) = 0$, then x is stationary and is a global minimizer when f is convex. Otherwise, for sufficiently small τ,

$$f(x - \tau\nabla f(x)) < f(x)$$

so the step from x to $x - \tau\nabla f(x)$ decreases the objective.

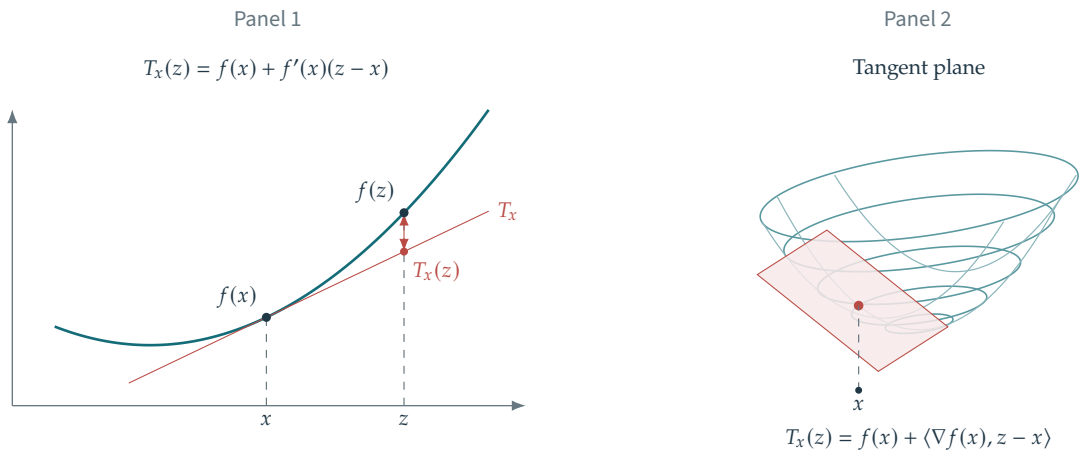

Figure 14.9. Panels 1 and 2: first-order Taylor approximations in one and two dimensions. Panels 3 and 4: the gradient is orthogonal to the level set; the diagram illustrates the proof.

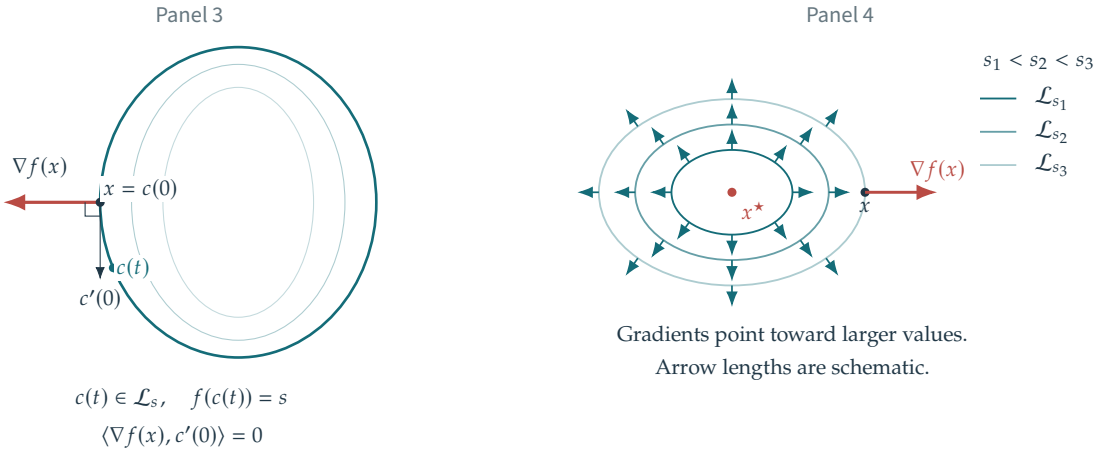

Figure 14.9 (continued). Panels 1 and 2: first-order Taylor approximations in one and two dimensions. Panels 3 and 4: the gradient is orthogonal to the level set; the diagram illustrates the proof.

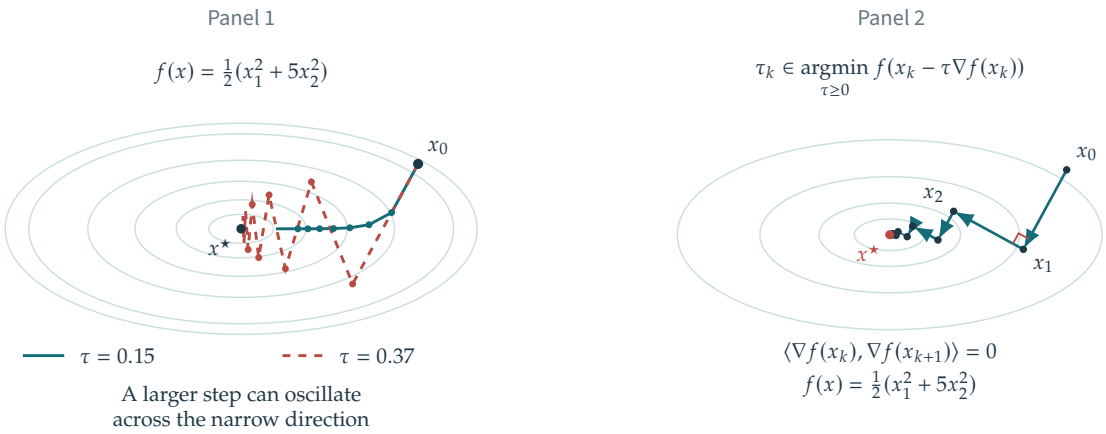

Figure 14.10. Effect of the step size τ on gradient descent (panel 1) and exact line search (panel 2).

Remark 14.7 (Orthogonality to level sets). A level set of f contains points with the same value of f. For $s \in \mathbb{R}$, define

$$\mathcal{L}_s \overset{\text{def.}}{=} \{x \; ; \; f(x) = s\}.$$

Assume f is continuously differentiable near x and $\nabla f(x) \neq 0$. Then x is a regular point of the level set $\mathcal{L}_{f(x)}$. The gradient is normal to its tangent space and points toward increasing values of f; see Figure 14.9, panels 3 and 4. Take a smooth curve through x in $\mathcal{L}_s$, parameterized by $t \in \mathbb{R} \mapsto c(t)$ with $c(0) = x$. The function $h(t) \overset{\text{def.}}{=} f(c(t))$ has constant value $h(t) = s$ because $c(t)$ lies in the level set. Thus $h'(t) = 0$. Expanding at $t = 0$ also gives

$$h(t) = f(c(0) + tc'(0) + o(t)) = h(0) + t\langle c'(0), \nabla f(c(0))\rangle + o(t)$$

i.e. $h'(0) = \langle c'(0), \nabla f(x)\rangle = 0$, so that $\nabla f(x)$ is orthogonal to the tangent $c'(0)$ of the curve c, which lies in the tangent plane of $\mathcal{L}_s$ (as shown in Figure 14.9, panel 3). Since the curve c is arbitrary, the whole tangent plane is thus orthogonal to $\nabla f(x)$.

Remark 14.8 (Local optimal descent direction). Assume f is continuously differentiable near x and $\nabla f(x) \neq 0$. Among displacements of length r, every minimizing direction approaches the normalized negative gradient as $r \downarrow 0$:

$$\frac{1}{r} \operatorname*{argmin}_{\|u\|=r} f(x+u) \xrightarrow{r \to 0} -\frac{\nabla f(x)}{\|\nabla f(x)\|}.$$

To verify the sign, let u_r minimize $f(x+u)$ on the radius-r sphere. Compare it with $-r\nabla f(x)/\|\nabla f(x)\|$. The uniform first-order remainder gives

$$\langle \nabla f(x), \, u_r/r \rangle \leqslant -\|\nabla f(x)\| + o(1).$$

Since $\|u_r/r\| = 1$, equality in the limiting Cauchy–Schwarz bound forces the stated direction.

14.4.2 Gradient Descent

Starting from $x_0 \in \mathbb{R}^p$, gradient descent iterates

$$x_{k+1} \overset{\text{def.}}{=} x_k - \tau_k \nabla f(x_k) \tag{14.13}$$

The parameter $\tau_k > 0$ is the step size, or learning rate. A sufficiently small τ_k decreases f at a nonstationary iterate. Choosing τ_k balances stable descent against progress per step. One may fix $\tau_k = \tau$ or adapt τ_k at each iteration; see Figure 14.10.

Remark 14.9 (Greedy choice). Exact line search chooses the best step along the current direction, although computing it can be expensive:

$$\tau_k \in \operatorname*{argmin}_{\tau \geqslant 0} h(\tau) \overset{\text{def.}}{=} f(x_k - \tau \nabla f(x_k)).$$

The line-search objective $h(\tau)$ is scalar. Its derivative follows from a Taylor expansion of h:

$$h(\tau + \delta) = f(x_k - \tau\nabla f(x_k) - \delta \nabla f(x_k)) = f(x_k - \tau \nabla f(x_k)) - \delta \langle \nabla f(x_k - \tau \nabla f(x_k)), \, \nabla f(x_k)\rangle + o(\delta).$$

Note that at $\tau = \tau_k$, $\nabla f(x_k - \tau \nabla f(x_k)) = \nabla f(x_{k+1})$ by definition of x_{k+1} in (14.13). If the current gradient is nonzero and the minimum is attained, an optimal step satisfies $\tau_k > 0$, since small positive steps decrease f. The first-order condition is therefore

$$h'(\tau_k) = -\langle \nabla f(x_k),\, \nabla f(x_{k+1}) \rangle = 0.$$

Thus successive gradients, and hence successive descent directions, are orthogonal; see Figure 14.10.

Remark 14.10 (Armijo rule). Instead of finding the optimal τ, Armijo backtracking seeks an admissible τ that produces sufficient decrease. Given $0 < \alpha < 1$ (values below $1/2$ are customary, especially for Newton-type methods), one considers a τ to be valid for a descent direction d_k (for instance $d_k = -\nabla f(x_k)$) if it satisfies

$$f(x_k + \tau d_k) \leqslant f(x_k) + \alpha \tau \langle d_k,\, \nabla f(x_k) \rangle \tag{14.14}$$

For small τ, one has $f(x_k + \tau d_k) = f(x_k) + \tau \langle d_k,\, \nabla f(x_k) \rangle + o(\tau)$, so that, assuming d_k is a valid descent direction (i.e $\langle d_k,\, \nabla f(x_k) \rangle < 0$), condition (14.14) will always be satisfied for τ small enough (if f is convex, the set of allowable τ is of the form $[0, \tau_{\max}]$). Start from a positive trial step and repeatedly reduce it by $\tau \leftarrow \beta \tau$, with $0 < \beta < 1$, until (14.14) holds. This procedure is called backtracking line search.

14.5 Convergence Analysis

14.5.1 Quadratic Case

Convergence analysis for the quadratic case. We first analyze gradient descent for the quadratic objective

$$f(x) = \frac{1}{2}\|Ax - y\|^2 = \frac{1}{2}\langle Cx,\, x \rangle - \langle x,\, b \rangle + \text{cst} \quad \text{where} \quad \begin{cases} C \overset{\text{def.}}{=} A^\top A \in \mathbb{R}^{p \times p}, \\ b \overset{\text{def.}}{=} A^\top y \in \mathbb{R}^p. \end{cases}$$

We already saw that in (14.9) if $\ker(A) = \{0\}$, which is equivalent to C being invertible, then there exists a single global minimizer $x^\star = (A^\top A)^{-1} A^\top y = C^{-1} b$.

The quadratic $\frac{1}{2}\langle Cx,\, x \rangle - \langle x,\, b \rangle$ is convex exactly when the symmetric matrix C is positive semidefinite, as follows from the spectral decomposition (14.10).

Proposition 14.11. *Let $f(x) = \frac{1}{2}\langle Cx,\, x \rangle - \langle b,\, x \rangle$, where C is symmetric and its eigenvalues lie in $[\mu, L]$ with $0 < \mu \leqslant L$. If the step sizes satisfy*

$$0 < \tau_{\min} \leqslant \tau_\ell \leqslant \tau_{\max} < \frac{2}{L}$$

then there exists $0 \leqslant \tilde{\rho} < 1$ such that

$$\|x_k - x^\star\| \leqslant \tilde{\rho}^k \|x_0 - x^\star\|. \tag{14.15}$$

The smallest uniform one-step contraction bound based only on μ and L is obtained with the constant step

$$\tau_\ell = \frac{2}{L + \mu} \quad \Longrightarrow \quad \tilde{\rho} \overset{\text{def.}}{=} \frac{L - \mu}{L + \mu} = 1 - \frac{2\varepsilon}{1 + \varepsilon} \quad \text{where} \quad \varepsilon \overset{\text{def.}}{=} \mu / L. \tag{14.16}$$

Proof. One iterate of gradient descent reads

$$x_{k+1} = x_k - \tau_k (C x_k - b).$$

The minimizer $x^\star$ satisfies $C x^\star = b$, so

$$x_{k+1} - x^\star = x_k - x^\star - \tau_k C (x_k - x^\star) = (\mathrm{Id}_p - \tau_k C)(x_k - x^\star).$$

If $S \in \mathbb{R}^{p \times p}$ is a symmetric matrix, one has

$$\|Sz\| \leqslant \|S\|_{\text{op}} \|z\| \quad \text{where} \quad \|S\|_{\text{op}} \overset{\text{def.}}{=} \max_k |\lambda_k(S)|,$$

where $\lambda_k(S)$ are the eigenvalues of S and $\sigma_k(S) \overset{\text{def.}}{=} |\lambda_k(S)|$ are its singular values. Indeed, S can be diagonalized in an orthogonal basis U, so that $S = U \operatorname{diag}(\lambda_k(S)) U^\top$, and $S^\top S = S^2 = U \operatorname{diag}(\lambda_k(S)^2) U^\top$ so that

$$\begin{aligned} \|Sz\|^2 &= \langle S^\top S z,\, z \rangle = \langle U \operatorname{diag}(\lambda_k^2) U^\top z,\, z \rangle = \langle \operatorname{diag}(\lambda_k^2) U^\top z,\, U^\top z \rangle \\ &= \sum_k \lambda_k^2 (U^\top z)_k^2 \leqslant \max_k(\lambda_k^2) \|U^\top z\|^2 = \max_k(\lambda_k^2) \|z\|^2. \end{aligned}$$

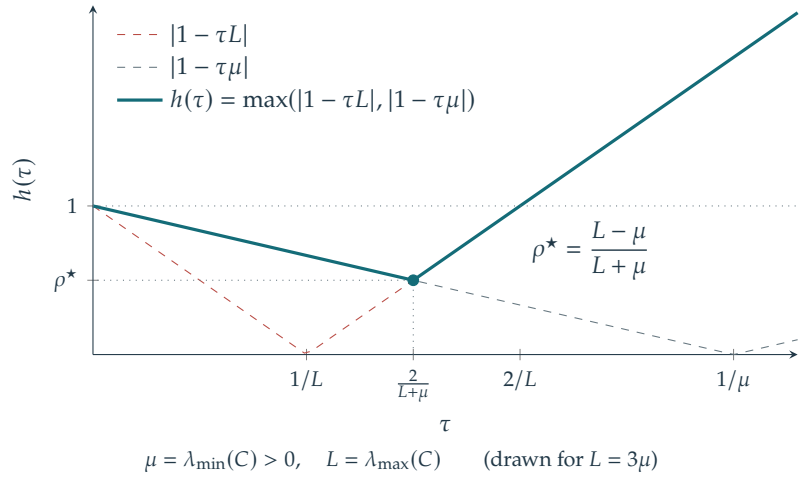

Figure 14.11. Contraction constant $h(\tau)$ for a quadratic function.

Applying this to $S = \mathrm{Id}_p - \tau C$ gives the contraction factor as a function of the step τ:

$$h(\tau) \overset{\text{def.}}{=} \|\mathrm{Id}_p - \tau C\|_{\mathrm{op}} = \max_j |1 - \tau\lambda_j(C)| \leqslant H(\tau) \overset{\text{def.}}{=} \max\big(|1-\tau L|, |1-\tau\mu|\big).$$

If μ and L are the extreme eigenvalues, equality holds and Figure 14.11 plots $h = H$. In general, $H(\tau) < 1$ for $0 < \tau < 2/L$. Continuity on $[\tau_{\min}, \tau_{\max}]$ gives a common bound $\tilde{\rho} < 1$, proving (14.15). The bound H is minimized at $\tau^\star = \frac{2}{L+\mu}$, giving

$$H(\tau^\star) = \left|1 - \frac{2L}{L+\mu}\right| = \frac{L-\mu}{L+\mu}.$$

□

Note that when the inverse condition number $\xi \overset{\text{def.}}{=} \mu/L \ll 1$ is small (which is the typical setup for ill-posed problems), then the contraction constant appearing in (14.16) scales like

$$\tilde{\rho} \sim 1 - 2\xi. \tag{14.17}$$

We also use ε below for this inverse condition number. The condition number is the ratio of the largest to the smallest singular value, and its minimum value is 1, attained by orthogonal matrices.

The geometric decay $O(\rho^k)$ in (14.15) is called linear convergence in optimization. The bound is global because it holds for every k, rather than only sufficiently large k.

If $\ker(A) \neq \{0\}$, then C has zero eigenvalues and the minimizer set is infinite. A linear rate still follows because the increments $x_{k+1} - x_k$ are orthogonal to $\ker(A)$, while the kernel component remains equal to that of x_0. The preceding proof applies on the orthogonal complement, with μ replaced by the smallest positive eigenvalue of C. This eigenvalue may be very small, giving a contraction factor close to one. The spectral argument also relies on the quadratic structure. We therefore give a different analysis that provides a sublinear bound on the objective value.

Proposition 14.12. *Let $f(x) = \frac{1}{2}\langle Cx, x\rangle - \langle b, x\rangle$ with $C \succeq 0$, $b \in \mathrm{Im}(C)$, and eigenvalues at most $L > 0$. For a constant step $0 < \tau_k = \tau < 1/L$ and $k \geqslant 1$,*

$$f(x_k) - f(x^\star) \leqslant \frac{\mathrm{dist}(x_0, \operatorname{argmin} f)^2}{\tau 8k}.$$

where

$$\mathrm{dist}(x_0, \operatorname{argmin} f) \overset{\text{def.}}{=} \min_{x^\star \in \operatorname{argmin} f} \|x_0 - x^\star\|.$$

Proof. We have $Cx^\star = b$ for any minimizer $x^\star$ and $x_{k+1} = x_k - \tau(Cx_k - b)$ so that as before

$$x_k - x^\star = (\mathrm{Id}_p - \tau C)^k (x_0 - x^\star).$$

Now one has

$$\frac{1}{2}\langle C(x_k - x^\star),\, x_k - x^\star\rangle = \frac{1}{2}\langle Cx_k,\, x_k\rangle - \langle Cx_k,\, x^\star\rangle + \frac{1}{2}\langle Cx^\star,\, x^\star\rangle$$

and we have $\langle Cx_k,\, x^\star\rangle = \langle x_k,\, Cx^\star\rangle = \langle x_k,\, b\rangle$ and also $\langle Cx^\star,\, x^\star\rangle = \langle x^\star,\, b\rangle$ so that

$$\frac{1}{2}\langle C(x_k - x^\star),\, x_k - x^\star\rangle = \frac{1}{2}\langle Cx_k,\, x_k\rangle - \langle x_k,\, b\rangle + \frac{1}{2}\langle x^\star,\, b\rangle = f(x_k) + \frac{1}{2}\langle x^\star,\, b\rangle.$$

Note also that

$$f(x^\star) = \frac{1}{2}\langle Cx^\star,\, x^\star\rangle - \langle x^\star,\, b\rangle = \frac{1}{2}\langle x^\star,\, b\rangle - \langle x^\star,\, b\rangle = -\frac{1}{2}\langle x^\star,\, b\rangle.$$

This shows that

$$\frac{1}{2}\langle C(x_k - x^\star),\, x_k - x^\star\rangle = f(x_k) - f(x^\star).$$

This thus implies

$$f(x_k) - f(x^\star) = \frac{1}{2}\langle (\mathrm{Id}_p - \tau C)^k C(\mathrm{Id}_p - \tau C)^k (x_0 - x^\star),\, x_0 - x^\star\rangle \leqslant \frac{\sigma_{\max}(M_k)}{2}\|x_0 - x^\star\|^2$$

where we have denoted

$$M_k \overset{\text{def.}}{=} (\mathrm{Id}_p - \tau C)^k C(\mathrm{Id}_p - \tau C)^k.$$

Since $x^\star$ can be chosen arbitrarily, one can replace $\|x_0 - x^\star\|$ by $\mathrm{dist}(x_0, \operatorname{argmin} f)$. One has, for any ℓ, the following bound

$$\sigma_\ell(M_k) \leqslant \max_{0 \leqslant s \leqslant L} s(1 - \tau s)^{2k} \leqslant \frac{1}{\tau 4k}$$

because the eigenvalues of M_k have the form $s(1 - \tau s)^{2k}$ for eigenvalues s of C. Setting $t = \tau s \in [0, 1]$, use

$$\forall\, t \in [0, 1], \quad (1 - t)^{2k} t \leqslant \frac{1}{4k}.$$

Indeed, one has

$$(1 - t)^{2k} t \leqslant (e^{-t})^{2k} t = \frac{1}{2k}(2kt)e^{-2kt} \leqslant \frac{1}{2k}\sup_{u \geqslant 0} ue^{-u} = \frac{1}{2ek} \leqslant \frac{1}{4k}.$$

□

14.5.2 General Case

We now extend the convergence analysis to general smooth convex functions. The Hessian replaces the constant matrix C of a quadratic objective.

Hessian. For a twice continuously differentiable function, the Hessian matrix is

$$(\partial^2 f)(x) = \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right)_{1 \leqslant i,j \leqslant p} \in \mathbb{R}^{p \times p}.$$

Each entry differentiates the ith component of the gradient with respect to the jth coordinate. Equality of mixed partial derivatives makes $\partial^2 f(x)$ symmetric.

Twice differentiability of f at x implies the Taylor expansion

$$f(x + \varepsilon) = f(x) + \langle \nabla f(x),\, \varepsilon\rangle + \frac{1}{2}\langle \partial^2 f(x)\varepsilon,\, \varepsilon\rangle + o(\|\varepsilon\|^2). \tag{14.18}$$

Thus f has a local quadratic approximation near x. The Hessian is the unique symmetric matrix in this expansion. If an expansion has the form below with a symmetric matrix H,

$$f(x + \varepsilon) = f(x) + \langle \nabla f(x),\, \varepsilon\rangle + \frac{1}{2}\langle H\varepsilon,\, \varepsilon\rangle + o(\|\varepsilon\|^2).$$

comparison with (14.18) identifies $\partial^2 f(x) = H$, avoiding separate computation of all p^2 partial derivatives. Equivalently, differentiate the gradient:

$$\nabla f(x + \varepsilon) = \nabla f(x) + [\partial^2 f(x)](\varepsilon) + o(\|\varepsilon\|)$$

where $[\partial^2 f(x)](\varepsilon) \in \mathbb{R}^p$ is the product of the Hessian $\partial^2 f(x)$ with ε.

A twice differentiable function f on $\mathbb{R}^p$ is convex exactly when, at every x, the Hessian $\partial^2 f(x)$ is positive semidefinite. Positive definiteness everywhere implies strict convexity of f, but is not necessary: x^4 is strictly convex on $\mathbb{R}$ although its second derivative vanishes at $x = 0$.

For a quadratic objective $f(x) = \frac{1}{2}\langle Cx,\, x\rangle - \langle x,\, u\rangle$, the gradient $\nabla f(x) = Cx - u$ gives the constant Hessian $\partial^2 f(x) = C$. For the classification function, one has

$$\nabla f(x) = -A^\top \operatorname{diag}(y)\nabla L(-\operatorname{diag}(y)Ax).$$

and thus

$$\begin{aligned}\nabla f(x+\varepsilon) &= -A^\top \operatorname{diag}(y)\nabla L(-\operatorname{diag}(y)Ax - \operatorname{diag}(y)A\varepsilon)\\ &= \nabla f(x) - A^\top \operatorname{diag}(y)[\partial^2 L(-\operatorname{diag}(y)Ax)](-\operatorname{diag}(y)A\varepsilon) + o(\|\varepsilon\|)\end{aligned}$$

Since $\nabla L(u) = (\ell'(u_i))$ one has $\partial^2 L(u) = \operatorname{diag}(\ell''(u_i))$. This means that

$$\partial^2 f(x) = A^\top \operatorname{diag}(y) \times \operatorname{diag}(\ell''(-\operatorname{diag}(y)Ax)) \times \operatorname{diag}(y)A.$$

This matrix is symmetric positive semidefinite when ℓ is convex, because ℓ'' is nonnegative.

Remark 14.13 (Second order optimality condition). The Hessian can classify a stationary point $x^\star$ with $\nabla f(x^\star) = 0$. If $\partial^2 f(x^\star)$ is positive definite, then $x^\star$ is a strict local minimizer. Positive semidefiniteness of $\partial^2 f(x^\star)$ alone does not classify the point; for example, x^3 on $\mathbb{R}$ has a vanishing Hessian at its stationary inflection point. Conversely, if $x^\star$ is a local minimum, then $\partial^2 f(x^\star)$ is positive semidefinite.

Remark 14.14 (Second order algorithms). The Hessian also yields second-order methods such as Newton's method. These can converge faster locally than gradient descent, at a greater cost per iteration. A variable-metric gradient step has the form

$$x_{k+1} = x_k - H_k \nabla f(x_k)$$

where $H_k \in \mathbb{R}^{p\times p}$ is symmetric positive definite. Gradient descent uses $H_k = \tau_k \mathrm{Id}_p$; Newton's method uses $H_k = [\partial^2 f(x_k)]^{-1}$ when the Hessian is positive definite. Note that

$$f(x_{k+1}) = f(x_k) - \langle H_k \nabla f(x_k),\, \nabla f(x_k)\rangle + o(\|H_k \nabla f(x_k)\|).$$

Positive definiteness of H_k ensures that, if x_k is nonstationary with $\nabla f(x_k) \neq 0$, then $\langle H_k \nabla f(x_k),\, \nabla f(x_k)\rangle > 0$. Scaling H_k sufficiently small therefore guarantees descent, $f(x_{k+1}) < f(x_k)$. We do not develop second-order algorithms further here.

For convergence analysis, uniform Hessian bounds play the role of the spectral bounds on C in the quadratic case. We require these bounds on $\partial^2 f(x)$ at every x. Integrating these local bounds yields global inequalities that support an analysis similar to the quadratic one.

Smoothness and strong convexity. For $L > 0$, quantify the smoothness of f by requiring an L-Lipschitz gradient:

$$\forall\, (x, x') \in (\mathbb{R}^p)^2, \quad \|\nabla f(x) - \nabla f(x')\| \leqslant L\|x - x'\|. \tag{$\mathcal{R}_L$}$$

A linear convergence guarantee for the iterates follows from a uniform lower curvature bound. We impose this by assuming that f is μ-strongly convex

$$\forall\, (x, x') \in (\mathbb{R}^p)^2, \quad \langle \nabla f(x) - \nabla f(x'),\, x - x'\rangle \geqslant \mu\|x - x'\|^2. \tag{$\mathcal{S}_\mu$}$$

For C^2 functions, these conditions are equivalent to uniform Hessian bounds.

Proposition 14.15. *Conditions* ($\mathcal{R}_L$) *and* ($\mathcal{S}_\mu$) *imply*

$$\forall\, (x, x'), \quad f(x') + \langle \nabla f(x'),\, x - x'\rangle + \frac{\mu}{2}\|x - x'\|^2 \leqslant f(x) \leqslant f(x') + \langle \nabla f(x'),\, x - x'\rangle + \frac{L}{2}\|x - x'\|^2. \tag{14.19}$$

If f is of class C^2, conditions ($\mathcal{R}_L$) *and* ($\mathcal{S}_\mu$) *are equivalent to*

$$\forall\, x, \quad \mu \mathrm{Id}_p \preceq \partial^2 f(x) \preceq L\mathrm{Id}_p \tag{14.20}$$

where $\partial^2 f(x) \in \mathbb{R}^{p\times p}$ is the Hessian and $\preceq$ denotes the semidefinite order:

$$A \preceq B \quad \Longleftrightarrow \quad \forall\, u \in \mathbb{R}^p, \quad \langle Au,\, u\rangle \leqslant \langle Bu,\, u\rangle.$$

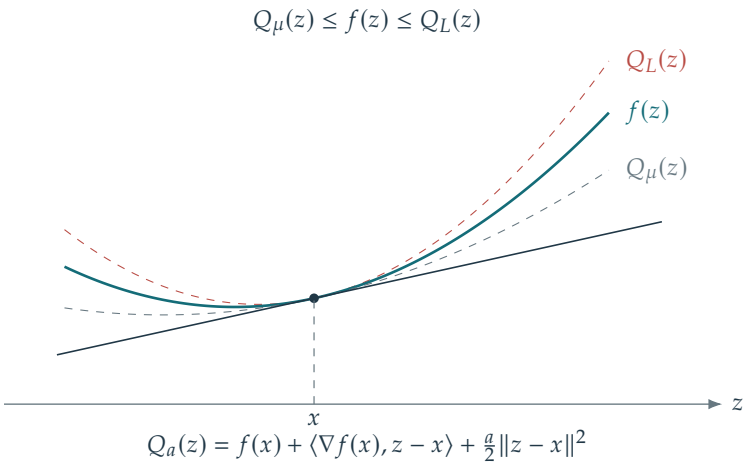

Figure 14.12. Quadratic upper and lower bounds for a smooth strongly convex function.

Proof. We prove (14.19), using Taylor expansion with integral remainder

$$f(x') - f(x) = \int_0^1 \langle \nabla f(x_t),\, x' - x \rangle \mathrm{d}t = \langle \nabla f(x),\, x' - x \rangle + \int_0^1 \langle \nabla f(x_t) - \nabla f(x),\, x' - x \rangle \mathrm{d}t$$

where $x_t \overset{\text{def.}}{=} x + t(x' - x)$. Using Cauchy–Schwarz, and then the smoothness hypothesis ($\mathcal{R}_L$)

$$f(x') - f(x) \leqslant \langle \nabla f(x),\, x' - x \rangle + \int_0^1 L\|x_t - x\|\|x' - x\| \mathrm{d}t \leqslant \langle \nabla f(x),\, x' - x \rangle + L\|x' - x\|^2 \int_0^1 t \mathrm{d}t$$

which is the desired upper-bound. Using directly ($\mathcal{S}_\mu$) gives

$$f(x') - f(x) = \langle \nabla f(x),\, x' - x \rangle + \int_0^1 \langle \nabla f(x_t) - \nabla f(x),\, \frac{x_t - x}{t} \rangle \mathrm{d}t \geqslant \langle \nabla f(x),\, x' - x \rangle + \mu \int_0^1 \frac{1}{t}\|x_t - x\|^2 \mathrm{d}t$$

which gives the desired result since $\|x_t - x\|^2/t = t\|x' - x\|^2$. □

Equation (14.19) places the objective between a lower quadratic tangent model from strong convexity and an upper quadratic tangent model from smoothness.

Condition (14.20) places every eigenvalue of $\partial^2 f(x)$ in $[\mu, L]$. The upper bound is also equivalent to $\|\partial^2 f(x)\|_{\text{op}} \leqslant L$ where $\|\cdot\|_{\text{op}}$ is the operator norm, i.e. the largest singular value. In the special case of a quadratic function of the form $\frac{1}{2}\langle Cx,\, x \rangle - \langle b,\, x \rangle$ (with C symmetric positive semidefinite), $\partial^2 f(x) = C$ is constant, so that $[\mu, L]$ can be chosen to be the range of the eigenvalues of C.

Convergence analysis. For a general smooth convex function, gradient descent admits a sublinear objective bound. Strong convexity gives a linear rate for the iterates. Other structural assumptions can also yield faster rates, but are outside this analysis. Without strong convexity, the minimizer need not be unique.

Theorem 14.16. *If f is convex, satisfies ($\mathcal{R}_L$), and has a minimizer, assume there are constants $\tau_{\min}, \tau_{\max}$ such that*

$$0 < \tau_{\min} \leqslant \tau_\ell \leqslant \tau_{\max} < \frac{2}{L},$$

then x_k converges to a solution $x^\star$ of (14.1) and there exists $C > 0$ such that

$$f(x_k) - f(x^\star) \leqslant \frac{C}{k+1}. \tag{14.21}$$

If furthermore f is μ-strongly convex, then there exists $0 \leqslant \rho < 1$ such that $\|x_k - x^\star\| \leqslant \rho^k \|x_0 - x^\star\|$.

Proof. We first derive the objective bound and the strong-convexity estimate for the step $1/L$. We then explain convergence for the full step-size range in the theorem. Take $\tau_\ell = 1/L$. The L-smoothness property implies (14.19), which reads

$$f(x_{k+1}) \leqslant f(x_k) + \langle \nabla f(x_k),\, x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2.$$

Using the fact that $x_{k+1} - x_k = -\frac{1}{L}\nabla f(x_k)$, one obtains

$$f(x_{k+1}) \leqslant f(x_k) - \frac{1}{L}\|\nabla f(x_k)\|^2 + \frac{1}{2L}\|\nabla f(x_k)\|^2 \leqslant f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \tag{14.22}$$

Thus the objective sequence $(f(x_k))_k$ is nonincreasing. By convexity

$$f(x_k) + \langle \nabla f(x_k),\, x^\star - x_k \rangle \leqslant f(x^\star)$$

and plugging this in (14.22) shows

$$f(x_{k+1}) \leqslant f(x^\star) - \langle \nabla f(x_k),\, x^\star - x_k \rangle - \frac{1}{2L}\|\nabla f(x_k)\|^2 \tag{14.23}$$

$$= f(x^\star) + \frac{L}{2}\left(\|x_k - x^\star\|^2 - \|x_k - x^\star - \frac{1}{L}\nabla f(x_k)\|^2\right) \tag{14.24}$$

$$= f(x^\star) + \frac{L}{2}\left(\|x_k - x^\star\|^2 - \|x^\star - x_{k+1}\|^2\right). \tag{14.25}$$

Summing these inequalities for $\ell = 0, \ldots, k$, one obtains

$$\sum_{\ell=0}^{k} f(x_{\ell+1}) - (k+1) f(x^\star) \leqslant \frac{L}{2}\left(\|x_0 - x^\star\|^2 - \|x^{(k+1)} - x^\star\|^2\right)$$

Since $f(x_{k+1})$ is nonincreasing, $\sum_{\ell=0}^{k} f(x_{\ell+1}) \geqslant (k+1) f(x^{(k+1)})$. Hence

$$f(x^{(k+1)}) - f(x^\star) \leqslant \frac{L\|x_0 - x^\star\|^2}{2(k+1)}$$

which gives (14.21) for $C \overset{\text{def.}}{=} L\|x_0 - x^\star\|^2$.

If we now assume f is μ-strongly convex, then, using $\nabla f(x^\star) = 0$, one has $\frac{\mu}{2}\|x^\star - x\|^2 \leqslant f(x) - f(x^\star)$ for all x. Rearranging (14.25) gives

$$\frac{\mu}{2}\|x_{k+1} - x^\star\|^2 \leqslant f(x_{k+1}) - f(x^\star) \leqslant \frac{L}{2}\left(\|x_k - x^\star\|^2 - \|x^\star - x_{k+1}\|^2\right),$$

and hence

$$\|x_{k+1} - x^\star\| \leqslant \sqrt{\frac{L}{L+\mu}}\|x_k - x^\star\|, \tag{14.26}$$

which is the desired result.

For the general step-size range, use the cocoercivity inequality for a convex function with an L-Lipschitz gradient:

$$\|\nabla f(x) - \nabla f(y)\|^2 \leqslant L\langle \nabla f(x) - \nabla f(y),\, x - y \rangle.$$

It follows by applying the descent bound to $f - \langle \nabla f(y), \cdot \rangle$, whose minimum is at y, and then adding the inequality with x and y exchanged. Set $g_k = \nabla f(x_k)$ and $a = \tau_{\min}(1 - L\tau_{\max}/2) > 0$. The descent bound and cocoercivity give

$$f(x_k) - f(x_{k+1}) \geqslant a\|g_k\|^2,$$

$$\|x_{k+1} - x^\star\|^2 \leqslant \|x_k - x^\star\|^2 - \frac{2a}{L}\|g_k\|^2.$$

The second inequality makes the iterates bounded and their distance to every minimizer nonincreasing. Summation gives $g_k \to 0$. Every accumulation point is therefore a minimizer; the nonincreasing distance to such a point proves convergence of the entire sequence.

For the objective rate, write $d_k = f(x_k) - f(x^\star)$ and $r_k = \|x_k - x^\star\|$. Convexity and the update imply

$$d_k \leqslant \frac{r_k^2 - r_{k+1}^2}{2\tau_{\min}} + \frac{\tau_{\max}}{2a}(d_k - d_{k+1}).$$

Summing from 0 to k and using monotonicity of d_k proves $(k+1)d_k \leqslant r_0^2/(2\tau_{\min}) + \tau_{\max} d_0/(2a)$. Under strong convexity, the same distance calculation gives

$$r_{k+1}^2 \leqslant \big(1 - \mu\tau_{\min}(2 - L\tau_{\max})\big) r_k^2,$$

whose factor belongs to $[0, 1)$. This establishes all claims for the stated step sizes. □

For poor conditioning, $\varepsilon \ll 1$, the rate (14.26) has the same qualitative dependence as the quadratic rate (14.17):

$$\sqrt{\frac{L}{L+\mu}} = (1+\varepsilon)^{-\frac{1}{2}} \sim 1 - \frac{1}{2}\varepsilon.$$

14.5.3 Acceleration

For a convex function with an L-Lipschitz gradient, ordinary gradient descent has a worst-case $O(1/k)$ objective bound. Accelerated methods combine a gradient step with extrapolation. Starting from $y_0 = x_0$ and $0 < s \leqslant 1/L$, one example is

$$x_{k+1} = y_k - s\nabla f(y_k), \qquad y_{k+1} = x_{k+1} + \beta_k(x_{k+1} - x_k), \qquad \beta_k = \frac{k}{k+3}.$$

This is a Nesterov-type scheme. It differs from the heavy-ball method, which evaluates the gradient at x_k and adds momentum to that step. The accelerated scheme satisfies

$$f(x_k) - f(x^\star) = O\left(\frac{\|x_0 - x^\star\|^2}{sk^2}\right).$$

This is a worst-case guarantee; it does not ensure that every individual trajectory is faster. If strong convexity is known, appropriately tuned acceleration or restart can give a linear rate.

A formal continuous-time limit explains the momentum schedule. Set $\tau = \sqrt{s}$ and $t = k\tau$. Since $\beta_{k-1} = (k-1)/(k+2)$,

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{\tau^2} + \frac{3}{(k+2)\tau}\frac{x_k - x_{k-1}}{\tau} + \nabla f(y_k) = 0.$$

If the interpolated iterates converge smoothly as $\tau \to 0$, the limit solves

$$x''(t) + \frac{3}{t}x'(t) + \nabla f(x(t)) = 0, \qquad x(0) = x_0, \quad x'(0) = 0.$$

The term $3x'(t)/t$ is a time-dependent friction. Its decay permits inertial motion while retaining an accelerated energy bound. The coefficient 3 is not the only admissible choice: related dynamics with friction α/t for $\alpha \geqslant 3$ also admit $O(t^{-2})$ objective estimates under the usual convexity assumptions.