Gradient Descent on a Positive Definite Quadratic

1 Gradient descent on a quadratic function

We study gradient descent applied to the pure quadratic function

$$f(x) = \frac{1}{2} x^\top A x, \qquad x \in \mathbb{R}^n, \tag{1}$$

where $A \in \mathbb{R}^{n \times n}$.

Our aim is to understand:

1. when f has a unique minimiser;
2. why the matrix defining the quadratic may be assumed symmetric;
3. why an orthogonal change of coordinates preserves the optimisation problem;
4. how diagonalisation reduces gradient descent to independent one-dimensional iterations;
5. how the learning rate determines convergence.

The absence of a linear term means that, whenever the quadratic is positive definite, its minimiser is the origin. This makes the essential geometry of gradient descent particularly transparent.

1.1 Why the matrix may be assumed symmetric

The matrix A in (1) need not initially be symmetric. Every real square matrix can be decomposed uniquely into a symmetric part and a skew-symmetric part:

$$A = Q + K, \qquad Q = \frac{A + A^\top}{2}, \qquad K = \frac{A - A^\top}{2}. \tag{2}$$

These matrices satisfy

$$Q^\top = Q \qquad \text{and} \qquad K^\top = -K.$$

The skew-symmetric part does not contribute to a quadratic form. Indeed, for every $x \in \mathbb{R}^n$, the quantity $x^\top K x$ is a scalar and is therefore equal to its transpose. Hence

$$\begin{aligned} x^\top K x &= (x^\top K x)^\top \\ &= x^\top K^\top x \\ &= -x^\top K x. \end{aligned} \tag{3}$$

It follows that

$$x^\top K x = 0 \qquad \text{for every } x \in \mathbb{R}^n. \tag{4}$$

Consequently,

$$\begin{aligned} x^\top Ax &= x^\top (Q+K)x \\ &= x^\top Qx + x^\top Kx \\ &= x^\top Qx. \end{aligned} \tag{5}$$

Thus the original function is exactly equal to

$$f(x) = \frac{1}{2}x^\top Qx, \qquad Q = \frac{A+A^\top}{2}. \tag{6}$$

This is not an approximation and does not alter the function. The matrices A and Q give exactly the same value of $f(x)$ for every x. We may therefore assume, without loss of generality, that the matrix defining the quadratic is symmetric.

There is also a differential interpretation. For a possibly nonsymmetric matrix A,

$$\nabla\left(\frac{1}{2}x^\top Ax\right) = \frac{1}{2}(A+A^\top)x = Qx. \tag{7}$$

Therefore,

$$\nabla f(x) = Qx, \qquad \nabla^2 f(x) = Q. \tag{8}$$

Only the symmetric part of A affects the function, its gradient, and its Hessian.

From now on, we write

$$f(x) = \frac{1}{2}x^\top Qx, \qquad Q^\top = Q. \tag{9}$$

1.2 Positive definiteness and the minimiser

Symmetry alone does not guarantee that f has a minimum. The appropriate additional assumption is that Q is positive definite.

Definition 1. A symmetric matrix $Q \in \mathbb{R}^{n\times n}$ is positive definite, written $Q \succ 0$, if

$$x^\top Qx > 0 \qquad \text{for every } x \neq 0. \tag{10}$$

We assume throughout that

$$Q \succ 0. \tag{11}$$

At the origin,

$$f(0) = 0.$$

For every $x \neq 0$, positive definiteness gives

$$f(x) = \frac{1}{2}x^\top Qx > 0.$$

Therefore,

$$f(0) < f(x) \qquad \text{for every } x \neq 0. \tag{12}$$

It follows immediately that

$$\boxed{x^\star = 0} \tag{13}$$

is the unique global minimiser of f.

The same conclusion follows from the first-order condition. By (8), a stationary point must satisfy

$$Qx = 0.$$

Positive definiteness implies that Q is invertible. To see this, suppose that $Qx = 0$. Then

$$x^\top Qx = 0,$$

and positive definiteness implies $x = 0$. Thus $\ker(Q) = \{0\}$, and so Q is invertible. Consequently,

$$Qx = 0 \quad \Longrightarrow \quad x = 0.$$

Since

$$\nabla^2 f(x) = Q \succ 0,$$

the function is strictly convex, and its unique stationary point is its unique global minimiser.

Remark 1. Positive definiteness is essential for uniqueness. If $Q \succeq 0$ is only positive semidefinite, then

$$f(x) \geq 0,$$

but every $x \in \ker(Q)$ is a minimiser. If Q has a negative eigenvalue, then f is unbounded below and has no global minimiser.

1.3 The spectral theorem

The central advantage of working with a real symmetric matrix is that it can be diagonalised by an orthogonal change of coordinates.

By the spectral theorem, there exist eigenvectors

$$u_1, \ldots, u_n \in \mathbb{R}^n$$

that form an orthonormal basis of $\mathbb{R}^n$. Thus

$$Qu_i = \lambda_i u_i$$

for real eigenvalues λ_i, and

$$u_i^\top u_j = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Let

$$U = \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix} \tag{14}$$

be the matrix whose columns are the eigenvectors, and let

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n). \tag{15}$$

The orthonormality of the eigenvectors gives

$$U^\top U = UU^\top = I. \tag{16}$$

Thus U is orthogonal and

$$U^{-1} = U^\top. \tag{17}$$

The eigenvalue equations can be written simultaneously as

$$QU = U\Lambda.$$

Multiplying on the right by $U^\top$ gives

$$Q = U\Lambda U^\top. \tag{18}$$

Equivalently,

$$U^\top Q U = \Lambda. \tag{19}$$

Positive definiteness of Q implies that all its eigenvalues are positive. Indeed, if $Qu_i = \lambda_i u_i$, then

$$u_i^\top Q u_i = \lambda_i u_i^\top u_i = \lambda_i,$$

where we used $\|u_i\|_2 = 1$. Since $Q \succ 0$, the left-hand side is positive, and therefore

$$\lambda_i > 0 \qquad \text{for every } i. \tag{20}$$

We order the eigenvalues as

$$0 < \mu = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = L. \tag{21}$$

Here

$$\mu = \lambda_{\min}(Q), \qquad L = \lambda_{\max}(Q).$$

1.4 Smoothness and strong convexity constants

The smallest and largest eigenvalues of Q have an important optimisation-theoretic interpretation. In this section, we prove that

$$\mu = \lambda_{\min}(Q) \tag{22}$$

is the largest possible strong-convexity constant of f, while

$$L = \lambda_{\max}(Q) \tag{23}$$

is the smallest possible smoothness constant of f.

Recall that

$$f(x) = \frac{1}{2} x^\top Q x, \qquad \nabla f(x) = Qx, \qquad \nabla^2 f(x) = Q,$$

where Q is symmetric and positive definite.

1.4.1 Eigenvalue bounds and matrix inequalities

Let

$$Q = U \Lambda U^\top$$

be the spectral decomposition of Q, where

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

and

$$0 < \mu = \lambda_1 \leq \lambda_i \leq \lambda_n = L.$$

For any $z \in \mathbb{R}^n$, define

$$w = U^\top z.$$

Since U is orthogonal,

$$\|w\|_2 = \|z\|_2.$$

Using the spectral decomposition, we obtain

$$\begin{aligned} z^\top Q z &= z^\top U \Lambda U^\top z \\ &= w^\top \Lambda w \\ &= \sum_{i=1}^{n} \lambda_i w_i^2. \end{aligned} \tag{24}$$

Because

$$\mu \le \lambda_i \le L$$

for every i, it follows that

$$\mu \sum_{i=1}^{n} w_i^2 \le \sum_{i=1}^{n} \lambda_i w_i^2 \le L \sum_{i=1}^{n} w_i^2. \tag{25}$$

Therefore,

$$\boxed{\mu \|z\|_2^2 \le z^\top Q z \le L \|z\|_2^2} \qquad \text{for every } z \in \mathbb{R}^n. \tag{26}$$

The inequalities in (26) can be written as the matrix inequalities

$$\boxed{\mu I \preceq Q \preceq LI}. \tag{27}$$

Indeed,

$$\mu I \preceq Q$$

means that $Q - \mu I$ is positive semidefinite, which is equivalent to

$$z^\top (Q - \mu I) z \ge 0 \qquad \text{for every } z \in \mathbb{R}^n.$$

This is exactly the lower bound

$$z^\top Q z \ge \mu \|z\|_2^2.$$

Similarly,

$$Q \preceq LI$$

is equivalent to

$$z^\top Q z \le L \|z\|_2^2.$$

Since the Hessian of f is the constant matrix Q, we may also write

$$\mu I \preceq \nabla^2 f(x) \preceq LI \qquad \text{for every } x \in \mathbb{R}^n. \tag{28}$$

1.4.2 Smoothness

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called L-smooth with respect to the Euclidean norm if its gradient is L-Lipschitz:

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L \|x - y\|_2 \qquad \text{for every } x, y \in \mathbb{R}^n. \tag{29}$$

For the quadratic function under consideration,

$$\begin{aligned} \nabla f(x) - \nabla f(y) &= Qx - Qy \\ &= Q(x - y). \end{aligned} \tag{30}$$

Let

$$z = x - y.$$

Using the spectral decomposition $Q = U \Lambda U^\top$ and defining $w = U^\top z$, we have

$$\begin{aligned} \|Qz\|_2^2 &= \|U \Lambda U^\top z\|_2^2 \\ &= \|\Lambda w\|_2^2 \\ &= \sum_{i=1}^{n} \lambda_i^2 w_i^2. \end{aligned} \tag{31}$$

Because $\lambda_i \le L$, we obtain

$$\begin{aligned}\|Qz\|_2^2 &\le L^2 \sum_{i=1}^{n} w_i^2 \\ &= L^2 \|w\|_2^2 \\ &= L^2 \|z\|_2^2. \end{aligned} \tag{32}$$

Taking square roots gives

$$\|Qz\|_2 \le L\|z\|_2. \tag{33}$$

Substituting $z = x - y$, we conclude that

$$\boxed{\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2}. \tag{34}$$

Thus f is L-smooth with

$$L = \lambda_{\max}(Q).$$

Equivalently, the induced Euclidean operator norm of Q is

$$\|Q\|_2 = \max_{z \ne 0} \frac{\|Qz\|_2}{\|z\|_2} = \lambda_{\max}(Q) = L. \tag{35}$$

The equality holds because Q is symmetric and positive definite.

We can also derive the usual smoothness inequality for function values. For any $x, y \in \mathbb{R}^n$, let

$$d = y - x.$$

Then $y = x + d$, and direct expansion gives

$$\begin{aligned} f(y) &= \frac{1}{2}(x + d)^\top Q(x + d) \\ &= \frac{1}{2}x^\top Qx + x^\top Qd + \frac{1}{2}d^\top Qd \\ &= f(x) + \nabla f(x)^\top d + \frac{1}{2}d^\top Qd. \end{aligned} \tag{36}$$

Using

$$d^\top Qd \le L\|d\|_2^2,$$

we obtain

$$\boxed{f(y) \le f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2}. \tag{37}$$

This is the standard quadratic upper bound associated with L-smoothness.

1.4.3 Strong convexity

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called μ-strongly convex with respect to the Euclidean norm if

$$f(y) \ge f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2 \tag{38}$$

for every $x, y \in \mathbb{R}^n$.

Using the exact expansion (36), we have

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top Q(y - x). \tag{39}$$

The lower eigenvalue bound in (26) gives

$$(y-x)^\top Q(y-x) \geq \mu \|y-x\|_2^2.$$

Substituting this inequality into (39), we obtain

$$\boxed{f(y) \geq f(x) + \nabla f(x)^\top (y-x) + \frac{\mu}{2}\|y-x\|_2^2}. \tag{40}$$

Therefore, f is μ-strongly convex with

$$\mu = \lambda_{\min}(Q).$$

Strong convexity can also be seen directly from the Hessian. A twice differentiable function is μ-strongly convex if

$$\nabla^2 f(x) \succeq \mu I \qquad \text{for every } x.$$

In the present case,

$$\nabla^2 f(x) = Q \succeq \mu I,$$

so the Hessian criterion gives the same conclusion.

The gradient also satisfies the strong-monotonicity inequality

$$\big(\nabla f(x) - \nabla f(y)\big)^\top (x-y) \geq \mu \|x-y\|_2^2. \tag{41}$$

Indeed,

$$\begin{aligned} \big(\nabla f(x) - \nabla f(y)\big)^\top (x-y) &= \big(Q(x-y)\big)^\top (x-y) \\ &= (x-y)^\top Q(x-y) \\ &\geq \mu \|x-y\|_2^2. \end{aligned} \tag{42}$$

1.4.4 Tightness of the constants

We have shown that f is L-smooth for

$$L = \lambda_{\max}(Q)$$

and μ-strongly convex for

$$\mu = \lambda_{\min}(Q).$$

We now show that these constants are tight.

Let $u_{\max}$ be a unit eigenvector corresponding to the largest eigenvalue L:

$$Qu_{\max} = Lu_{\max}, \qquad \|u_{\max}\|_2 = 1.$$

Choose

$$x = u_{\max}, \qquad y = 0.$$

Then

$$\begin{aligned} \|\nabla f(x) - \nabla f(y)\|_2 &= \|Qu_{\max}\|_2 \\ &= \|Lu_{\max}\|_2 \\ &= L, \end{aligned} \tag{43}$$

while

$$\|x-y\|_2 = 1.$$

Therefore,

$$\frac{\|\nabla f(x) - \nabla f(y)\|_2}{\|x - y\|_2} = L. \tag{44}$$

No constant smaller than L can satisfy the smoothness inequality for all x and y. Hence

$$\boxed{L = \lambda_{\max}(Q) \text{ is the smallest valid smoothness constant}}. \tag{45}$$

Now let $u_{\min}$ be a unit eigenvector corresponding to the smallest eigenvalue μ:

$$Qu_{\min} = \mu u_{\min}, \qquad \|u_{\min}\|_2 = 1.$$

Taking $x = 0$ and $y = u_{\min}$, we obtain

$$f(0) = 0, \qquad \nabla f(0) = 0,$$

and

$$\begin{aligned} f(u_{\min}) &= \frac{1}{2} u_{\min}^\top Q u_{\min} \\ &= \frac{1}{2} u_{\min}^\top (\mu u_{\min}) \\ &= \frac{\mu}{2}. \end{aligned} \tag{46}$$

If f were strongly convex with some constant $\widetilde{\mu} > \mu$, then strong convexity with $x = 0$ and $y = u_{\min}$ would require

$$f(u_{\min}) \geq f(0) + \nabla f(0)^\top u_{\min} + \frac{\widetilde{\mu}}{2} \|u_{\min}\|_2^2.$$

This would imply

$$\frac{\mu}{2} \geq \frac{\widetilde{\mu}}{2},$$

which contradicts $\widetilde{\mu} > \mu$. Consequently,

$$\boxed{\mu = \lambda_{\min}(Q) \text{ is the largest valid strong-convexity constant}}. \tag{47}$$

The terminology "smallest valid smoothness constant" and "largest valid strong-convexity constant" is important. If f is L-smooth, then it is also L'-smooth for every

$$L' \geq L.$$

Similarly, if f is μ-strongly convex, then it is also μ'-strongly convex for every

$$0 < \mu' \leq \mu.$$

The eigenvalues $\lambda_{\max}(Q)$ and $\lambda_{\min}(Q)$ give the tight constants.

1.4.5 Combined quadratic bounds

Combining smoothness and strong convexity gives

$$\boxed{\frac{\mu}{2}\|y - x\|_2^2 \leq f(y) - f(x) - \nabla f(x)^\top (y - x) \leq \frac{L}{2}\|y - x\|_2^2}. \tag{48}$$

For the quadratic function, the expression in the middle is exactly

$$f(y) - f(x) - \nabla f(x)^\top (y - x) = \frac{1}{2}(y - x)^\top Q (y - x).$$

Thus the combined bounds are precisely the eigenvalue bounds on the quadratic form.

Taking $x = x^\star = 0$, where

$$f(x^\star) = 0 \qquad \text{and} \qquad \nabla f(x^\star) = 0,$$

gives

$$\boxed{\frac{\mu}{2}\|x\|_2^2 \le f(x) \le \frac{L}{2}\|x\|_2^2}. \tag{49}$$

These inequalities quantify how the objective value grows as one moves away from the minimiser.

Similarly, the gradient satisfies

$$\mu\|x\|_2 \le \|\nabla f(x)\|_2 \le L\|x\|_2. \tag{50}$$

To verify the lower bound, write $x = Uw$. Then

$$\begin{aligned} \|Qx\|_2^2 &= \sum_{i=1}^n \lambda_i^2 w_i^2 \\ &\ge \mu^2 \sum_{i=1}^n w_i^2 \\ &= \mu^2 \|x\|_2^2. \end{aligned} \tag{51}$$

The upper bound follows in the same way using $\lambda_i \le L$.

Finally, the ratio

$$\kappa = \frac{L}{\mu} = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)} \tag{52}$$

is the condition number of the quadratic. It measures the disparity between the largest and smallest curvatures of f. As the later convergence analysis will show, this disparity governs the worst-case rate of gradient descent.

1.5 The change of coordinates

Every $x \in \mathbb{R}^n$ can be expressed uniquely in the orthonormal eigenvector basis:

$$x = y_1 u_1 + \cdots + y_n u_n.$$

In matrix notation, this is

$$x = Uy. \tag{53}$$

Multiplying by $U^\top$ and using $U^\top U = I$, we obtain

$$y = U^\top x. \tag{54}$$

Thus $x \mapsto y = U^\top x$ is an invertible change of coordinates, with inverse $y \mapsto x = Uy$. The entries of y are

$$y_i = u_i^\top x,$$

so y_i is the component of x in the eigendirection u_i.

Because U is orthogonal, this transformation preserves the Euclidean inner product:

$$(Uy)^\top (Uz) = y^\top U^\top U z = y^\top z.$$

In particular, it preserves Euclidean lengths:

$$\|Uy\|_2 = \|y\|_2. \tag{55}$$

Geometrically, an orthogonal transformation is a rotation, a reflection, or a combination of the two. It changes the coordinates used to describe a point, but it does not change the underlying point set or the distances between points.

1.6 The transformed objective

Define the transformed objective by

$$g(y) = f(Uy). \tag{56}$$

Using the spectral decomposition of Q, we find

$$\begin{aligned} g(y) &= \frac{1}{2}(Uy)^\top Q(Uy) \\ &= \frac{1}{2}y^\top U^\top QUy \\ &= \frac{1}{2}y^\top \Lambda y \\ &= \frac{1}{2}\sum_{i=1}^{n} \lambda_i y_i^2. \end{aligned} \tag{57}$$

The transformed problem is therefore

$$\min_{y\in\mathbb{R}^n} \frac{1}{2}\sum_{i=1}^{n} \lambda_i y_i^2. \tag{58}$$

The original quadratic may contain interactions between different coordinates. In the eigenvector basis, these interactions disappear. The objective becomes a sum of independent one-dimensional quadratics.

1.7 Why convexity is preserved

A change of input coordinates cannot turn this convex quadratic into a concave quadratic.

More generally, suppose f is convex and define

$$g(y) = f(Uy),$$

where U is linear. Let $y, z \in \mathbb{R}^n$ and $t \in [0,1]$. By linearity,

$$U\big(ty + (1-t)z\big) = tUy + (1-t)Uz.$$

Convexity of f therefore gives

$$\begin{aligned} g\big(ty + (1-t)z\big) &= f\big(U(ty + (1-t)z)\big) \\ &= f\big(tUy + (1-t)Uz\big) \\ &\le tf(Uy) + (1-t)f(Uz) \\ &= tg(y) + (1-t)g(z). \end{aligned} \tag{59}$$

Hence g is convex.

For the quadratic, the same result can be checked from the Hessian:

$$\nabla^2 g(y) = U^\top QU = \Lambda. \tag{60}$$

For every $y \neq 0$,

$$\begin{aligned} y^\top \Lambda y &= y^\top U^\top QUy \\ &= (Uy)^\top Q(Uy). \end{aligned} \tag{61}$$

Since U is invertible,

$$y \neq 0 \quad \Longrightarrow \quad Uy \neq 0.$$

Positive definiteness of Q therefore implies

$$(Uy)^\top Q(Uy) > 0.$$

Thus

$$\Lambda = U^\top QU \succ 0. \tag{62}$$

This argument applies more generally to any invertible linear coordinate transformation $x = Ty$. Under this transformation, the matrix of the quadratic becomes

$$Q \longmapsto T^\top QT,$$

and

$$y^\top T^\top QTy = (Ty)^\top Q(Ty) > 0$$

for every $y \neq 0$. Therefore, positive definiteness is preserved by invertible changes of input coordinates.

A change of input coordinates should not be confused with replacing f by $-f$. The latter operation negates the objective values and turns minimisation into maximisation. It is a change of the optimisation problem, not merely a new coordinate representation of the same problem.

1.8 Why the minimisation problems are equivalent

The original and transformed problems are

$$\min_{x\in\mathbb{R}^n} f(x) \qquad \text{and} \qquad \min_{y\in\mathbb{R}^n} g(y), \quad g(y) = f(Uy). \tag{63}$$

These are equivalent because the map

$$y \longmapsto x = Uy$$

is a bijection. Every y determines exactly one $x = Uy$, and every x can be represented uniquely as

$$x = U(U^\top x).$$

Furthermore, corresponding points have exactly the same objective value:

$$g(y) = f(Uy). \tag{64}$$

Suppose $y^\star$ minimises g. Then

$$g(y^\star) \leq g(y) \qquad \text{for every } y \in \mathbb{R}^n.$$

Define

$$x^\star = Uy^\star.$$

Given any $x \in \mathbb{R}^n$, choose $y = U^\top x$, so that $x = Uy$. Then

$$\begin{aligned} f(x^\star) &= f(Uy^\star) \\ &= g(y^\star) \\ &\leq g(y) \\ &= f(Uy) \\ &= f(x). \end{aligned} \tag{65}$$

Since this holds for every x, the point $x^\star = Uy^\star$ minimises f.

Conversely, if $x^\star$ minimises f, then

$$y^\star = U^\top x^\star$$

minimises g. Consequently,

$$\arg\min_{x\in\mathbb{R}^n} f(x) = U\left(\arg\min_{y\in\mathbb{R}^n} g(y)\right). \tag{66}$$

For the diagonalised quadratic,

$$g(y) = \frac{1}{2}\sum_{i=1}^{n} \lambda_i y_i^2.$$

Because every $\lambda_i > 0$, each term is nonnegative, and

$$g(y) = 0$$

if and only if

$$y_1 = \cdots = y_n = 0.$$

Hence

$$y^\star = 0.$$

Transforming back gives

$$x^\star = Uy^\star = U0 = 0,$$

in agreement with the direct argument based on positive definiteness.

1.9 Gradient descent

Gradient descent with a constant learning rate $\eta > 0$ is defined by

$$x_{k+1} = x_k - \eta\nabla f(x_k). \tag{67}$$

Since

$$\nabla f(x) = Qx,$$

the update becomes

$$x_{k+1} = x_k - \eta Qx_k = (I - \eta Q)x_k. \tag{68}$$

Repeated application gives the exact expression

$$x_k = (I - \eta Q)^k x_0. \tag{69}$$

The matrix expression is correct, but the behaviour of the iteration becomes much clearer after diagonalisation.

1.10 Gradient descent in eigenvector coordinates

Define

$$y_k = U^\top x_k, \qquad x_k = Uy_k. \tag{70}$$

Multiplying (68) by $U^\top$, we obtain

$$\begin{aligned} y_{k+1} &= U^\top x_{k+1} \\ &= U^\top (I - \eta Q)x_k \\ &= U^\top (I - \eta Q)Uy_k \\ &= \big(I - \eta U^\top QU\big)y_k \\ &= (I - \eta\Lambda)y_k. \end{aligned} \tag{71}$$

Because Λ is diagonal, the coordinates evolve independently:

$$y_{i,k+1} = (1 - \eta\lambda_i)y_{i,k}, \qquad i = 1, \dots, n. \tag{72}$$

Solving this scalar recurrence gives

$$y_{i,k} = (1 - \eta\lambda_i)^k y_{i,0}. \tag{73}$$

Thus gradient descent on an n-dimensional positive definite quadratic is equivalent to applying gradient descent simultaneously to the one-dimensional functions

$$g_i(y_i) = \frac{1}{2}\lambda_i y_i^2.$$

The eigenvalue λ_i measures the curvature in the direction u_i.

1.11 The range of convergent learning rates

The i-th coordinate converges to zero for every initial value $y_{i,0}$ if and only if

$$|1 - \eta\lambda_i| < 1.$$

Because $\lambda_i > 0$, this is equivalent to

$$\begin{aligned} |1 - \eta\lambda_i| < 1 &\iff -1 < 1 - \eta\lambda_i < 1 \\ &\iff -2 < -\eta\lambda_i < 0 \\ &\iff 0 < \eta\lambda_i < 2 \\ &\iff 0 < \eta < \frac{2}{\lambda_i}. \end{aligned} \tag{74}$$

For all coordinates to converge, this inequality must hold for every i. The most restrictive condition comes from the largest eigenvalue L. Therefore,

$$\boxed{0 < \eta < \frac{2}{L}} \tag{75}$$

is necessary and sufficient for convergence from every initial point.

At the boundary $\eta = 2/L$, a component associated with the eigenvalue L satisfies

$$y_{i,k+1} = -y_{i,k}.$$

It oscillates without decreasing in magnitude. The endpoint $2/L$ must therefore be excluded.

If $\eta > 2/L$, then

$$|1 - \eta L| > 1,$$

so a component in an eigendirection corresponding to L grows in magnitude.

1.12 Behaviour in a single eigendirection

The factor

$$1 - \eta\lambda_i$$

determines the behaviour in the i-th eigendirection.

1. If

   $$0 < \eta < \frac{1}{\lambda_i},$$

   then

   $$0 < 1 - \eta\lambda_i < 1.$$

   The coordinate approaches zero without changing sign.

2. If
$$\eta = \frac{1}{\lambda_i},$$
then
$$1 - \eta\lambda_i = 0.$$
That coordinate reaches its minimiser in one iteration.

3. If
$$\frac{1}{\lambda_i} < \eta < \frac{2}{\lambda_i},$$
then
$$-1 < 1 - \eta\lambda_i < 0.$$
The coordinate alternates in sign but decreases in magnitude.

4. If
$$\eta = \frac{2}{\lambda_i},$$
then
$$1 - \eta\lambda_i = -1.$$
The coordinate oscillates with constant magnitude.

5. If
$$\eta > \frac{2}{\lambda_i},$$
then
$$|1 - \eta\lambda_i| > 1,$$
and the coordinate diverges unless its initial value is zero.

1.13 The convergence factor

Define
$$\rho(\eta) = \max_{1 \le i \le n} |1 - \eta\lambda_i|. \tag{76}$$
For $0 < \eta < 2/L$, we have $\rho(\eta) < 1$.

Using (73),
$$\begin{aligned} \|y_k\|_2^2 &= \sum_{i=1}^{n} (1 - \eta\lambda_i)^{2k} y_{i,0}^2 \\ &\le \rho(\eta)^{2k} \sum_{i=1}^{n} y_{i,0}^2. \end{aligned} \tag{77}$$

Taking square roots gives
$$\|y_k\|_2 \le \rho(\eta)^k \|y_0\|_2. \tag{78}$$

Because U preserves the Euclidean norm,
$$\|x_k\|_2 = \|y_k\|_2,$$
and therefore
$$\boxed{\|x_k\|_2 \le \rho(\eta)^k \|x_0\|_2}. \tag{79}$$

The objective value also converges geometrically. Since

$$f(x_k) = g(y_k) = \frac{1}{2}\sum_{i=1}^{n} \lambda_i y_{i,k}^2,$$

we have

$$\begin{aligned} f(x_k) &= \frac{1}{2}\sum_{i=1}^{n} \lambda_i (1-\eta\lambda_i)^{2k} y_{i,0}^2 \\ &\leq \rho(\eta)^{2k} \frac{1}{2}\sum_{i=1}^{n} \lambda_i y_{i,0}^2 \\ &= \rho(\eta)^{2k} f(x_0). \end{aligned} \tag{80}$$

Thus

$$\boxed{f(x_k) \leq \rho(\eta)^{2k} f(x_0)}. \tag{81}$$

1.14 The learning rate $\eta = 1/L$

A common conservative choice is

$$\eta = \frac{1}{L}. \tag{82}$$

For every eigenvalue $\lambda_i \in [\mu, L]$,

$$0 \leq 1 - \frac{\lambda_i}{L} \leq 1 - \frac{\mu}{L}.$$

Hence

$$\rho\left(\frac{1}{L}\right) = 1 - \frac{\mu}{L}. \tag{83}$$

Define the spectral condition number of Q by

$$\kappa = \frac{L}{\mu}. \tag{84}$$

Then

$$\rho\left(\frac{1}{L}\right) = 1 - \frac{1}{\kappa}. \tag{85}$$

Therefore,

$$\|x_k\|_2 \leq \left(1 - \frac{1}{\kappa}\right)^k \|x_0\|_2.$$

If κ is large, the contraction factor is close to one and convergence can be slow.

1.15 The optimal constant learning rate

We now determine the constant learning rate that minimises the worst-case contraction factor. Recall that, in the eigendirection associated with λ_i, one iteration of gradient descent multiplies the corresponding component by

$$1 - \eta\lambda_i.$$

The worst-case contraction factor is therefore

$$\rho(\eta) = \max_{1\leq i\leq n} |1 - \eta\lambda_i|. \tag{86}$$

Thus, the optimal constant learning rate solves

$$\min_{\eta>0} \rho(\eta) = \min_{\eta>0} \max_{1\leq i\leq n} |1 - \eta\lambda_i|. \tag{87}$$

For fixed η, the function

$$\lambda \longmapsto |1 - \eta\lambda|$$

is convex. Its maximum over the interval $[\mu, L]$ is therefore attained at one of the endpoints. Since $\mu = \lambda_{\min}(Q)$ and $L = \lambda_{\max}(Q)$ are themselves eigenvalues of Q, it follows that

$$\rho(\eta) = \max\{|1 - \eta\mu|, |1 - \eta L|\}. \tag{88}$$

Consequently, the optimisation problem reduces to

$$\min_{\eta>0} \max\{|1 - \eta\mu|, |1 - \eta L|\}. \tag{89}$$

We first consider the nontrivial case $0 < \mu < L$. The behaviour of $\rho(\eta)$ can be understood by considering three ranges of the learning rate.

If

$$0 < \eta \leq \frac{1}{L},$$

then both endpoint factors are nonnegative:

$$1 - \eta\mu \geq 1 - \eta L \geq 0.$$

Hence,

$$\rho(\eta) = 1 - \eta\mu. \tag{90}$$

This expression decreases as η increases, so an optimal learning rate cannot lie in the interior of this interval.

If

$$\frac{1}{L} \leq \eta \leq \frac{1}{\mu},$$

then the two endpoint factors have opposite signs:

$$1 - \eta\mu \geq 0, \qquad 1 - \eta L \leq 0.$$

Their absolute values are therefore

$$|1 - \eta\mu| = 1 - \eta\mu, \qquad |1 - \eta L| = \eta L - 1,$$

and the contraction factor becomes

$$\rho(\eta) = \max\{1 - \eta\mu, \eta L - 1\}. \tag{91}$$

The first expression decreases with η, whereas the second increases with η. The maximum of the two is therefore minimised at the point where they are equal:

$$1 - \eta_{\text{opt}}\mu = \eta_{\text{opt}} L - 1. \tag{92}$$

Equivalently, the original endpoint factors have the same magnitude and opposite signs:

$$1 - \eta_{\text{opt}}\mu = -\big(1 - \eta_{\text{opt}} L\big). \tag{93}$$

Finally, if

$$\eta \geq \frac{1}{\mu},$$

then both endpoint factors are nonpositive:

$$1 - \eta L \leq 1 - \eta\mu \leq 0.$$

It follows that

$$\rho(\eta) = \eta L - 1, \tag{94}$$

which increases with η. Hence, no learning rate in the interior of this region can be optimal.

These three cases show that the optimum occurs in the middle region, where the two endpoint magnitudes are balanced. Intuitively, if one endpoint magnitude were strictly larger than the other, the learning rate could be adjusted slightly to reduce the larger one while it remained the worst-case factor. At the optimum, neither endpoint can be improved without worsening the other. This balancing property is sometimes referred to as an *equioscillation condition*.

Solving (93) gives

$$\begin{aligned} 1 - \eta_{\text{opt}}\mu &= -1 + \eta_{\text{opt}} L, \\ 2 &= \eta_{\text{opt}}(L + \mu). \end{aligned} \tag{95}$$

Therefore, the optimal constant learning rate is

$$\boxed{\eta_{\text{opt}} = \frac{2}{L + \mu}}. \tag{96}$$

Indeed, when $0 < \mu < L$,

$$\frac{1}{L} < \frac{2}{L + \mu} < \frac{1}{\mu},$$

so η_{opt} lies in the middle region, as required.

At this learning rate, the endpoint contraction factors are

$$\begin{aligned} 1 - \eta_{\text{opt}}\mu &= 1 - \frac{2\mu}{L + \mu} \\ &= \frac{L - \mu}{L + \mu}, \end{aligned} \tag{97}$$

and

$$\begin{aligned} 1 - \eta_{\text{opt}} L &= 1 - \frac{2L}{L + \mu} \\ &= -\frac{L - \mu}{L + \mu}. \end{aligned} \tag{98}$$

Thus,

$$\boxed{1 - \eta_{\text{opt}}\mu = -\big(1 - \eta_{\text{opt}} L\big)}, \tag{99}$$

confirming that the endpoint factors have equal magnitude and opposite signs. The eigendirection associated with μ contracts without changing sign, while the eigendirection associated with L contracts and changes sign at every iteration.

The resulting optimal worst-case contraction factor is

$$\boxed{\rho_{\text{opt}} = \rho(\eta_{\text{opt}}) = \frac{L - \mu}{L + \mu}}. \tag{100}$$

Introducing the condition number

$$\kappa = \frac{L}{\mu},$$

we can equivalently write

$$\boxed{\rho_{\text{opt}} = \frac{\kappa - 1}{\kappa + 1}}. \tag{101}$$

There is also a useful geometric interpretation. The affine map

$$\lambda \longmapsto 1 - \eta\lambda$$

maps the eigenvalue interval $[\mu, L]$ to

$$[1 - \eta L,\, 1 - \eta\mu].$$

The contraction factor $\rho(\eta)$ is the greatest distance from zero of a point in this image interval. At the optimal learning rate,

$$[1 - \eta_{\text{opt}} L,\, 1 - \eta_{\text{opt}}\mu] = \left[-\frac{L-\mu}{L+\mu}, \frac{L-\mu}{L+\mu}\right]. \tag{102}$$

The image interval is therefore centred at zero. Its two endpoints are equally far from zero, which explains geometrically why the endpoint contraction factors have equal magnitude and opposite signs.

In the special case $L = \mu$, all eigenvalues of Q are equal. Since Q is symmetric, this implies

$$Q = \mu I.$$

The same formula gives

$$\eta_{\text{opt}} = \frac{2}{L+\mu} = \frac{1}{\mu},$$

and hence

$$I - \eta_{\text{opt}} Q = I - \frac{1}{\mu}\mu I = 0.$$

Thus,

$$\rho_{\text{opt}} = 0, \tag{103}$$

and gradient descent reaches the minimiser in a single iteration.

For the general case, gradient descent with the optimal constant learning rate satisfies

$$\|x_k\|_2 \le \rho_{\text{opt}}^k \|x_0\|_2 = \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x_0\|_2. \tag{104}$$

Moreover, because each squared eigencomponent contracts by at most ρ_{opt}^2 per iteration,

$$f(x_k) \le \rho_{\text{opt}}^{2k} f(x_0) = \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} f(x_0). \tag{105}$$

This is the best worst-case rate obtainable when the same constant scalar learning rate must be used at every iteration and in every eigendirection.

1.16 Why ill-conditioning causes slow convergence

In diagonal coordinates, the objective is

$$g(y) = \frac{1}{2}\sum_{i=1}^{n} \lambda_i y_i^2.$$

An eigenvalue λ_i represents the curvature of the function in the direction u_i.

The learning rate must be small enough to ensure stability in the largest-curvature direction:

$$\eta < \frac{2}{L}.$$

On the other hand, progress in the smallest-curvature direction is controlled by

$$1 - \eta\mu.$$

If $\mu \ll L$, every learning rate that is stable in the largest-curvature direction produces only a small change in the smallest-curvature direction.

Thus a scalar learning rate must compromise between:

- stability in directions of large curvature;
- rapid progress in directions of small curvature.

The condition number

$$\kappa = \frac{L}{\mu}$$

quantifies the disparity between these curvatures. A large condition number means that the level sets of the quadratic are highly elongated and that ordinary gradient descent converges slowly.

1.17 Different learning rates in different directions

The diagonal representation suggests assigning a separate learning rate to each eigendirection. Consider

$$y_{i,k+1} = y_{i,k} - \eta_i \lambda_i y_{i,k}. \tag{106}$$

Equivalently,

$$y_{i,k+1} = (1 - \eta_i \lambda_i) y_{i,k}.$$

The i-th coordinate converges if and only if

$$0 < \eta_i < \frac{2}{\lambda_i}.$$

If the eigenvalues are known, the choice

$$\eta_i = \frac{1}{\lambda_i} \tag{107}$$

gives

$$y_{i,k+1} = 0.$$

Thus every coordinate reaches the minimiser in one iteration.

Let

$$D = \operatorname{diag}\left(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n}\right) = \Lambda^{-1}.$$

The update in diagonal coordinates is then

$$y_{k+1} = y_k - D\Lambda y_k = 0.$$

Transforming this update back to the original coordinates gives

$$x_{k+1} = x_k - U\Lambda^{-1}U^\top Q x_k. \tag{108}$$

Since

$$Q^{-1} = U\Lambda^{-1}U^\top,$$

we obtain

$$x_{k+1} = x_k - Q^{-1}Q x_k = 0. \tag{109}$$

This is Newton's method applied to the quadratic. Since the Hessian is the constant matrix Q, Newton's method reaches the minimiser in one iteration.

The same observation motivates preconditioning. Instead of the scalar learning rate η, one may use a positive definite matrix P and consider

$$x_{k+1} = x_k - P\nabla f(x_k) = (I - PQ)x_k. \tag{110}$$

The ideal choice is $P = Q^{-1}$. In practice, one seeks a matrix P that approximates Q^{-1} but is cheaper to compute and apply.

1.18 Summary

Starting from the pure quadratic

$$f(x) = \frac{1}{2}x^\top Ax,$$

the analysis proceeds as follows.

1. Decompose A into symmetric and skew-symmetric parts:

   $$A = \frac{A + A^\top}{2} + \frac{A - A^\top}{2}.$$

   The skew-symmetric part makes no contribution because

   $$x^\top(A - A^\top)x = 0.$$

   Thus the function depends only on

   $$Q = \frac{A + A^\top}{2}.$$

2. Assume that $Q \succ 0$. Then

   $$f(x) > 0 \quad \text{for every } x \neq 0, \qquad f(0) = 0,$$

   so the origin is the unique global minimiser.

3. By the spectral theorem,

   $$Q = U\Lambda U^\top,$$

   where U is orthogonal and

   $$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \qquad \lambda_i > 0.$$

4. The coordinate transformation

   $$y = U^\top x, \qquad x = Uy,$$

   is bijective and preserves Euclidean lengths.

5. The transformed objective is

   $$g(y) = f(Uy) = \frac{1}{2}\sum_{i=1}^{n} \lambda_i y_i^2.$$

   Convexity is preserved, and corresponding points have the same objective values. Hence the original and transformed minimisation problems are equivalent.

6. Gradient descent becomes
$$y_{i,k+1} = (1 - \eta\lambda_i)y_{i,k}.$$
The coordinates evolve independently.

7. Gradient descent converges from every initial point if and only if
$$0 < \eta < \frac{2}{L}, \qquad L = \lambda_{\max}(Q).$$

8. The convergence factor is
$$\rho(\eta) = \max_i |1 - \eta\lambda_i|.$$

9. The optimal constant learning rate is
$$\eta_{\text{opt}} = \frac{2}{L + \mu}, \qquad \mu = \lambda_{\min}(Q),$$
and its contraction factor is
$$\rho_{\text{opt}} = \frac{L - \mu}{L + \mu} = \frac{\kappa - 1}{\kappa + 1}.$$

10. Direction-dependent learning rates
$$\eta_i = \frac{1}{\lambda_i}$$
eliminate every coordinate in one step. In the original coordinates, this corresponds to multiplication by Q^{-1}, as in Newton's method.

The diagonalisation does not change the objective or the meaning of its minimum. It expresses the same quadratic optimisation problem in the orthonormal coordinate system formed by its principal curvature directions.