

Positive-definite matrices and quadratic forms

Ten problems on positive-definite matrices: classifying 2×2 matrices by eigenvalues and by a test vector, the eigenvalue characterisation and what it says about determinants and inverses, Sylvester's criterion by completing the square, Gram matrices and ridge, a 3×3 Cholesky factorisation by hand with a solve, the gradient, Hessian and minimiser of a quadratic, Hessian PSD if and only if convex, the symmetric square root and the ellipsoid of a quadratic form, the step-size condition and convergence rate of gradient descent on a quadratic, and why a quadratic form sees only the symmetric part, with worked solutions and the mistakes that read definiteness off a determinant, trust the eigenvalues of a non-symmetric matrix, sample with the transposed Cholesky factor, assume the normal equations have a unique solution, or pick the step size from the smallest eigenvalue.

BEFORE YOU START

A positive-definite matrix is one whose quadratic form is a bowl: $x^\top Ax > 0$ in every direction. Every covariance matrix, every Gram matrix $X^\top X$ with independent columns, every Hessian at a strict minimum and every kernel matrix is one, and the properties that make them useful, a Cholesky factor, an inverse that is also positive definite, a square root, a condition number that sets the step size of gradient descent, all come from that one inequality. These ten problems classify small matrices by hand, prove the eigenvalue test and Sylvester's criterion, factor a 3×3 matrix by Cholesky and solve a system with it, minimise a quadratic by completing the square, show that a convex function is exactly one with a PSD Hessian, compute a symmetric square root, derive the step size and rate of gradient descent on a quadratic, and end with the fact that a quadratic form never sees the antisymmetric part of its matrix. The five mistakes at the end are the ones that survive a glance: a positive determinant read as positive definite, positive eigenvalues of a non-symmetric matrix, the Cholesky factor used transposed, normal equations assumed solvable, and a step size chosen from the wrong eigenvalue.

- A is a real symmetric $n \times n$ matrix unless a problem says otherwise, I is the identity, and vectors are columns. A is positive definite (PD, written $A \succ 0$) if $x^\top Ax > 0$ for every $x \neq 0$, positive semidefinite (PSD, $A \succeq 0$) if $x^\top Ax \geq 0$ for every x , and indefinite if $x^\top Ax$ takes both signs.
- A symmetric A has an orthonormal eigenbasis: $A = Q\Lambda Q^\top$ with $Q^\top Q = I$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$, and $Aq_i = \lambda_i q_i$ for the columns q_i of Q (the eigenvalues page). Equivalently $A = \sum_i \lambda_i q_i q_i^\top$. $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues and $\kappa = \lambda_{\max}/\lambda_{\min}$ is the condition number of a PD matrix.

- A Cholesky factorisation of a PD matrix is $A = LL^\top$ with L lower-triangular and $L_{jj} > 0$. Computed column by column: $L_{jj} = \sqrt{A_{jj} - \sum_{k < j} L_{jk}^2}$ and, for $i > j$, $L_{ij} = (A_{ij} - \sum_{k < j} L_{ik}L_{jk})/L_{jj}$.
- For $X \in \mathbb{R}^{N \times d}$ the Gram matrix is $X^\top X$; $\|v\|^2 = v^\top v$. The singular values of X are σ_i (the least-squares page).
- Gradients are columns: for scalar $f(x)$, $\nabla f \in \mathbb{R}^n$ and the Hessian $\nabla^2 f$ is the $n \times n$ matrix of second partials, so for symmetric A , $\nabla(x^\top Ax) = 2Ax$ and $\nabla(b^\top x) = b$ (the matrix-calculus page). A twice-differentiable f is convex when $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$ for all x, y and $\lambda \in [0, 1]$ (the Jensen page).
- Gradient descent with step η is $x_{k+1} = x_k - \eta \nabla f(x_k)$; $e_k = x_k - x^*$ is the error after k steps.

Builds on: [Eigenvalues and eigenvectors by hand](#), [Matrix calculus conventions](#)

Problems

1. Classify $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ and $C = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ as positive definite, difficulty ·
positive semidefinite or indefinite using their eigenvalues. For B , exhibit a vector v with $v^\top Bv < 0$; for C , a nonzero vector with $v^\top Cv = 0$.

2. For symmetric $A = \sum_i \lambda_i q_i q_i^\top$, show that $x^\top Ax = \sum_i \lambda_i (q_i^\top x)^2$ and difficulty · ·
hence that A is PD if and only if every $\lambda_i > 0$. Deduce that a PD matrix has $\det A > 0$ and $\text{tr } A > 0$, is invertible, and that A^{-1} is PD.

3. By completing the square in $x^\top Ax = ax^2 + 2bxy + cy^2$, show that the difficulty · ·
symmetric $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ is PD if and only if $a > 0$ and $ac - b^2 > 0$. For which t is $\begin{bmatrix} 2 & t \\ t & 2 \end{bmatrix}$
positive definite?

4. For $X \in \mathbb{R}^{N \times d}$, show that $X^\top X$ is PSD, that it is PD if and only if the difficulty · ·
columns of X are linearly independent, and that $X^\top X + \lambda I$ is PD for every $\lambda > 0$ with
eigenvalues $\sigma_i^2 + \lambda$, where the σ_i are the singular values of X (zero for $i > \text{rank } X$).

5. Compute the Cholesky factor L of $A = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 6 \end{bmatrix}$, verify $LL^\top = A$, difficulty $\dots$
and use it to solve $Ax = b$ for $b = (8, 10, 11)^\top$ by two triangular solves. What is $\det A$?

6. Let $f(x) = \frac{1}{2}x^\top Ax - b^\top x + c$ with A symmetric and PD. Compute ∇f difficulty $\dots$
and $\nabla^2 f$, show that $f(x) = \frac{1}{2}(x - x^*)^\top A(x - x^*) + f(x^*)$ with $x^* = A^{-1}b$ and $f(x^*) = c - \frac{1}{2}b^\top A^{-1}b$, and conclude that x^* is the unique minimiser. Evaluate for $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, $b = (1, 2)^\top$, $c = 0$. What happens if A has a negative eigenvalue?

7. Let f be twice differentiable on $\mathbb{R}^n$ and, for fixed x and v , let $g(t) =$ difficulty $\dots$
 $f(x + tv)$. Show that $g''(t) = v^\top \nabla^2 f(x + tv) v$, and use it to prove that f is convex if and only if $\nabla^2 f(x) \succeq 0$ for every x . Conclude that $\frac{1}{2}x^\top Ax$ is convex exactly when $A \succeq 0$.

8. For PD $A = Q\Lambda Q^\top$ define $A^{1/2} = Q\Lambda^{1/2}Q^\top$. Show that $A^{1/2}$ is difficulty $\dots$
symmetric PD with $(A^{1/2})^2 = A$, and that $A^{-1/2}AA^{-1/2} = I$. Compute $A^{1/2}$ for $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$,
and describe the ellipse $\{x : x^\top Ax = 1\}$.

9. For $f(x) = \frac{1}{2}x^\top Ax - b^\top x$ with $A \succ 0$, show that gradient descent difficulty $\dots$
satisfies $e_{k+1} = (I - \eta A)e_k$, that it converges from every start if and only if $0 < \eta < 2/\lambda_{\max}$,
and that the error component along q_i shrinks by the factor $|1 - \eta\lambda_i|$ per step. Find the η
that minimises the worst factor and the resulting rate in terms of κ . Evaluate for $A = \text{diag}(1, 10)$.

10. For any square matrix M , not necessarily symmetric, show that difficulty $\dots$
 $x^\top Mx = x^\top Sx$ with $S = \frac{1}{2}(M + M^\top)$. Show that $M = \begin{bmatrix} 1 & 4 \\ 0 & 1 \end{bmatrix}$ has both eigenvalues equal
to 1 and yet $v^\top Mv < 0$ for some v . Why does the eigenvalue test of Problem 2 not apply?

Answers

1. $A \succ 0$ (eigenvalues 3, 1); B indefinite (eigenvalues 3, -1, $v = (1, -1)^\top$ gives -2); $C \succeq 0$ but not PD (eigenvalues 5, 0, $v = (2, -1)^\top$ gives 0)
2. $x^\top Ax = \sum_i \lambda_i (q_i^\top x)^2$; $A \succ 0 \iff \text{all } \lambda_i > 0$; then $\det A > 0$, $\text{tr } A > 0$, A is invertible and $A^{-1} \succ 0$
3. $\begin{bmatrix} a & b \\ b & c \end{bmatrix} \succ 0 \iff a > 0 \text{ and } ac - b^2 > 0$; $\begin{bmatrix} 2 & t \\ t & 2 \end{bmatrix} \succ 0 \iff |t| < 2$
4. $X^\top X \succeq 0$ always; $X^\top X \succ 0 \iff \text{rank } X = d$; $X^\top X + \lambda I \succ 0$ for $\lambda > 0$, with eigenvalues $\sigma_i^2 + \lambda$
5. $L = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 1 & 2 \end{bmatrix}$, $LL^\top = A$; $x = (1, 1, 1)^\top$; $\det A = 64$
6. $\nabla f = Ax - b$, $\nabla^2 f = A$; $f(x) = \frac{1}{2}(x - x^*)^\top A(x - x^*) + f(x^*)$ with $x^* = A^{-1}b$ and $f(x^*) = c - \frac{1}{2}b^\top A^{-1}b$, the unique minimiser; for the example $x^* = (0, 1)^\top$, $f(x^*) = -1$; with a negative eigenvalue f is unbounded below
7. $g''(t) = v^\top \nabla^2 f(x + tv) v$; f convex $\iff \nabla^2 f(x) \succeq 0$ for all x ; $\frac{1}{2}x^\top Ax$ convex $\iff A \succeq 0$
8. $A^{1/2} = Q\Lambda^{1/2}Q^\top \succ 0$ with $(A^{1/2})^2 = A$ and $A^{-1/2}AA^{-1/2} = I$; for the example $A^{1/2} = \frac{1}{2} \begin{bmatrix} \sqrt{3} + 1 & \sqrt{3} - 1 \\ \sqrt{3} - 1 & \sqrt{3} + 1 \end{bmatrix}$; $x^\top Ax = 1$ is an ellipse with semi-axes $1/\sqrt{3}$ along $(1, 1)^\top/\sqrt{2}$ and 1 along $(1, -1)^\top/\sqrt{2}$
9. $e_{k+1} = (I - \eta A)e_k$; convergence from every start $\iff 0 < \eta < 2/\lambda_{\max}$; the q_i component scales by $|1 - \eta\lambda_i|$; $\eta^* = \frac{2}{\lambda_{\min} + \lambda_{\max}}$ with rate $\frac{\kappa - 1}{\kappa + 1}$; for $\text{diag}(1, 10)$: $\eta^* = 2/11$, rate 9/11
10. $x^\top Mx = x^\top \frac{1}{2}(M + M^\top)x$; $M = \begin{bmatrix} 1 & 4 \\ 0 & 1 \end{bmatrix}$ has eigenvalues 1, 1 but $(1, -1)M(1, -1)^\top = -2$; the eigenvalue test needs a symmetric matrix

Worked solutions

Problem 1

Classify $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ and $C = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ as positive definite, positive semidefinite or indefinite using their eigenvalues. For B , exhibit a vector v with $v^\top Bv < 0$; for C , a nonzero vector with $v^\top Cv = 0$.

A : $\det(A - \lambda I) = (2 - \lambda)^2 - 1 = 0$ gives $\lambda = 3, 1$, both positive: PD.	The eigenvalues page's Problem 1; Problem 2 below shows why positive eigenvalues mean PD.
B : $(1 - \lambda)^2 - 4 = 0$ gives $1 - \lambda = \pm 2$, $\lambda = 3, -1$: indefinite.	One eigenvalue of each sign.
$v = (1, -1)^\top$: $v^\top Bv = 1 - 2 - 2 + 1 = -2 < 0$.	v is the eigenvector for $\lambda = -1$, since $B(1, -1)^\top = (1 - 2, 2 - 1)^\top = -(1, -1)^\top$; its quadratic form is $\lambda\ v\ ^2 = -2$.
C : $(1 - \lambda)(4 - \lambda) - 4 = \lambda^2 - 5\lambda = 0$ gives $\lambda = 5, 0$: PSD but not PD.	A zero eigenvalue means $x^\top Cx = 0$ for a nonzero x , so the form is never negative but is not always positive.
$v = (2, -1)^\top$: $Cv = (2 - 2, 4 - 4)^\top = 0$, so $v^\top Cv = 0$.	The null vector of C ; C has rank one, $C = (1, 2)^\top(1, 2)$.
$A \succ 0$ (eigenvalues 3, 1); B indefinite (eigenvalues 3, -1, $v = (1, -1)^\top$ gives -2); $C \succeq 0$ but not PD (eigenvalues 5, 0, $v = (2, -1)^\top$ gives 0)	
All three have positive diagonal entries and positive trace, and B even has all entries positive, so none of those tests decides anything; B's determinant is -3 and C's is 0, which rule them out, but a positive determinant would not have ruled them in (Mistake 1). A is the covariance matrix of the Gaussian page's Problem 1, and C is what a covariance looks like when one coordinate is twice the other.	

Problem 2

For symmetric $A = \sum_i \lambda_i q_i q_i^\top$, show that $x^\top A x = \sum_i \lambda_i (q_i^\top x)^2$ and hence that A is PD if and only if every $\lambda_i > 0$. Deduce that a PD matrix has $\det A > 0$ and $\text{tr } A > 0$, is invertible, and that A^{-1} is PD.

$$\begin{aligned} x^\top A x &= x^\top \left(\sum_i \lambda_i q_i q_i^\top \right) x = \\ \sum_i \lambda_i (x^\top q_i)(q_i^\top x) &= \sum_i \lambda_i (q_i^\top x)^2. \end{aligned}$$

Distribute $x^\top$ and x over the sum;
 $x^\top q_i = q_i^\top x$ is a scalar.

If every $\lambda_i > 0$ and $x \neq 0$, then some $q_i^\top x \neq 0$, so $x^\top A x > 0$.

The q_i are an orthonormal basis, so $x = \sum_i (q_i^\top x) q_i$; if every coefficient were zero, x would be zero. All terms are ≥ 0 and at least one is > 0 .

If A is PD, then $\lambda_i = q_i^\top A q_i > 0$ for every i .

Take $x = q_i$ in the definition: $q_i^\top A q_i = q_i^\top (\lambda_i q_i) = \lambda_i \|q_i\|^2 = \lambda_i$.

$\det A = \prod_i \lambda_i > 0$ and $\text{tr } A = \sum_i \lambda_i > 0$; no eigenvalue is zero, so A is invertible.

The eigenvalues page's Problem 5, extended to $n \times n$ by $\det(Q\Lambda Q^\top) = \det \Lambda$ and $\text{tr}(Q\Lambda Q^\top) = \text{tr}(\Lambda Q^\top Q) = \text{tr } \Lambda$; a matrix is singular exactly when 0 is an eigenvalue.

$A^{-1} = Q\Lambda^{-1}Q^\top = \sum_i \frac{1}{\lambda_i} q_i q_i^\top$, with every $1/\lambda_i > 0$, so A^{-1} is PD by step 2.

$(Q\Lambda Q^\top)(Q\Lambda^{-1}Q^\top) = Q\Lambda\Lambda^{-1}Q^\top = I$ because $Q^\top Q = I$; the inverse has the same eigenvectors and reciprocal eigenvalues.

$x^\top Ax = \sum_i \lambda_i (q_i^\top x)^2$; $A \succ 0 \iff$ **all**
 $\lambda_i > 0$; **then** $\det A > 0$, $\text{tr } A > 0$, **A is**
invertible and $A^{-1} \succ 0$

In the eigenbasis the quadratic form is a weighted sum of squares, and definiteness is the sign pattern of the weights: all > 0 is PD, all ≥ 0 is PSD, mixed is indefinite. The converses of step 4 fail: $\det > 0$ and $\text{tr} > 0$ together do not give PD in three dimensions (diag(4, -1, -1)), and in two dimensions they do only because two eigenvalues with positive product and positive sum are both positive, which is Problem 3's criterion. Step 5 is why the precision matrix Σ^{-1} of a Gaussian is PD whenever the covariance is.

Problem 3

By completing the square in $x^\top Ax = ax^2 + 2bxy + cy^2$, show that the symmetric $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ is PD if and only if $a > 0$ and $ac - b^2 > 0$. For which t is $\begin{bmatrix} 2 & t \\ t & 2 \end{bmatrix}$ positive definite?

For $a \neq 0$: $ax^2 + 2bxy + cy^2 = a\left(x + \frac{b}{a}y\right)^2 + \left(c - \frac{b^2}{a}\right)y^2 = a\left(x + \frac{b}{a}y\right)^2 + \frac{ac - b^2}{a}y^2$.

Expand $a\left(x + \frac{b}{a}y\right)^2 = ax^2 + 2bxy + \frac{b^2}{a}y^2$ and subtract the extra $\frac{b^2}{a}y^2$.

If $a > 0$ and $ac - b^2 > 0$: both coefficients are positive, so the form is ≥ 0 , and it is 0 only when $y = 0$ and then $x + \frac{b}{a}y = x = 0$.

A positive combination of two squares vanishes only when both squares vanish.

If A is PD: $x = (1, 0)^\top$ gives $a > 0$, and $x = (-\frac{b}{a}, 1)^\top$ gives $\frac{ac - b^2}{a} > 0$, so $ac - b^2 > 0$.

Two test vectors, chosen to kill one square each in step 1; $a > 0$ lets the inequality be multiplied through.

$\begin{bmatrix} 2 & t \\ t & 2 \end{bmatrix} : a = 2 > 0$ and $ac - b^2 = 4 - t^2 > 0$ exactly when $|t| < 2$.

Step 2 with $a = c = 2, b = t$.

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix} \succ 0 \iff a > 0 \textbf{ and } ac - b^2 > 0;$$

$$\begin{bmatrix} 2 & t \\ t & 2 \end{bmatrix} \succ 0 \iff |t| < 2$$

This is Sylvester's criterion: all leading principal minors positive, the 1×1 minor a and the 2×2 minor $\det A$. Completing the square is the 2×2 Cholesky factorisation in disguise, $L = \begin{bmatrix} \sqrt{a} & 0 \\ b/\sqrt{a} & \sqrt{(ac - b^2)/a} \end{bmatrix}$, and the criterion fails exactly where a square root in Problem 5's algorithm would go negative. At $t = \pm 2$ the matrix is PSD with the null vector $(1, \mp 1)^\top$, and for $|t| > 2$ it is indefinite despite its positive diagonal and positive trace. For a covariance $\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$ the condition is $\rho^2 < 1$.

Problem 4

For $X \in \mathbb{R}^{N \times d}$, show that $X^\top X$ is PSD, that it is PD if and only if the columns of X are linearly independent, and that $X^\top X + \lambda I$ is PD for every $\lambda > 0$ with eigenvalues $\sigma_i^2 + \lambda$, where the σ_i are the singular values of X (zero for $i > \text{rank } X$).

$$v^\top X^\top X v = (Xv)^\top (Xv) = \|Xv\|^2 \geq 0.$$

Group the product as $(Xv)^\top (Xv)$; a squared norm is nonnegative.

$$v^\top X^\top X v = 0 \iff Xv = 0.$$

A norm is zero only for the zero vector.

$X^\top X \succ 0 \iff Xv \neq 0$ for every $v \neq 0 \iff$ the columns of X are linearly independent.

$Xv = \sum_j v_j x_j$ with x_j the columns, and $Xv = 0$ for some $v \neq 0$ is exactly a linear dependence among them. That needs $N \geq d$; with $N < d$ some nonzero v always has $Xv = 0$.

$$v^\top (X^\top X + \lambda I) v = \|Xv\|^2 + \lambda \|v\|^2 > 0 \text{ for } v \neq 0.$$

Step 1 plus $\lambda \|v\|^2 > 0$; the second term is positive even when the first is zero.

With $X = U\Sigma V^\top$: $X^\top X = V\Sigma^\top \Sigma V^\top$, so its eigenvalues are σ_i^2 with eigenvectors the right singular vectors v_i , and $(X^\top X + \lambda I)v_i = (\sigma_i^2 + \lambda)v_i$.

$U^\top U = I$ and $\Sigma^\top \Sigma$ is diagonal with entries σ_i^2 (the least-squares page); adding λI adds λ to every eigenvalue without moving the eigenvectors.

**$X^\top X \succeq 0$ always; $X^\top X \succ 0 \iff$
rank $X = d$; $X^\top X + \lambda I \succ 0$ for $\lambda > 0$,
with eigenvalues $\sigma_i^2 + \lambda$**

Every Gram matrix, covariance matrix ($\frac{1}{N-1}X^\top HX$ on the variance page) and kernel matrix is PSD for the reason in step 1, and PD only when nothing is collinear. The normal equations $X^\top Xw = X^\top y$ have a unique solution exactly in the PD case (Mistake 4); ridge regression's $X^\top X + \lambda I$ is always invertible, with condition number $(\sigma_{\max}^2 + \lambda)/(\sigma_{\min}^2 + \lambda)$, which λ pulls towards 1, and that is what makes ridge both solvable and well-conditioned. The bias–variance page prices the shrinkage this costs.

Problem 5

Compute the Cholesky factor L of $A = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 6 \end{bmatrix}$, verify $LL^\top = A$, and use it to solve $Ax = b$ for $b = (8, 10, 11)^\top$ by two triangular solves. What is $\det A$?

Column 1: $L_{11} = \sqrt{A_{11}} = 2$, $L_{21} = A_{21}/L_{11} = 1$, $L_{31} = A_{31}/L_{11} = 1$.

The algorithm with no earlier columns to subtract.

Column 2: $L_{22} = \sqrt{A_{22} - L_{21}^2} = \sqrt{5 - 1} = 2$, $L_{32} = (A_{32} - L_{31}L_{21})/L_{22} = (3 - 1)/2 = 1$.

Subtract the contribution of column 1 before taking the root and dividing.

Column 3: $L_{33} = \sqrt{A_{33} - L_{31}^2 - L_{32}^2} = \sqrt{6 - 1 - 1} = 2$.

Both earlier columns contribute to the (3, 3) entry.

$$L = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 1 & 2 \end{bmatrix} \text{ and } LL^\top = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 1+4 & 1+2 \\ 2 & 1+2 & 1+1+4 \end{bmatrix} = A.$$

Entry (i, j) of $LL^\top$ is the dot product of rows i and j of L .

Forward solve $Ly = b$: $y_1 = 8/2 = 4$, $y_2 = (10 - 1 \cdot 4)/2 = 3$, $y_3 = (11 - 1 \cdot 4 - 1 \cdot 3)/2 = 2$.

Row by row from the top; each row has one new unknown.

Back solve $L^\top x = y$: $x_3 = 2/2 = 1$, $x_2 = (3 - 1 \cdot 1)/2 = 1$, $x_1 = (4 - 1 \cdot 1 - 1 \cdot 1)/2 = 1$.

$L^\top$ is upper-triangular, so solve from the bottom; $Ax = LL^\top x = Ly = b$.

$$\det A = \det L \det L^\top = (\det L)^2 = (2 \cdot 2 \cdot 2)^2 = 64.$$

The determinant of a triangular matrix is the product of its diagonal, and $\det L^\top = \det L$.

$$L = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 1 & 2 \end{bmatrix}, LL^\top = A; x = (1, 1, 1)^\top; \det A = 64$$

Check: $A(1, 1, 1)^\top = (8, 10, 11)^\top$. The three square roots were of 4, 4 and 4, all positive, which is the proof that A is PD: Cholesky succeeds exactly when every pivot $A_{jj} - \sum_{k < j} L_{jk}^2$ is positive, and it is how software tests definiteness, at a third of the cost of an eigendecomposition. The Gaussian page uses L to sample, $x = \mu + L\varepsilon$ (and not $L^\top \varepsilon$, Mistake 3), and $\log \det \Sigma = 2 \sum_j \log L_{jj}$ is how its normaliser is computed without forming the determinant.

Problem 6

Let $f(x) = \frac{1}{2}x^\top Ax - b^\top x + c$ with A symmetric and PD. Compute ∇f and $\nabla^2 f$, show that $f(x) = \frac{1}{2}(x - x^*)^\top A(x - x^*) + f(x^*)$ with $x^* = A^{-1}b$ and $f(x^*) = c - \frac{1}{2}b^\top A^{-1}b$, and conclude

that x^* is the unique minimiser. Evaluate for $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, $b = (1, 2)^\top$, $c = 0$. What happens if A has a negative eigenvalue?

$$\nabla f = Ax - b \text{ and } \nabla^2 f = A.$$

$\nabla(\frac{1}{2}x^\top Ax) = Ax$ for symmetric A and $\nabla(b^\top x) = b$ (the matrix-calculus page); the Hessian is the Jacobian of $Ax - b$.

$$\frac{1}{2}(x - x^*)^\top A(x - x^*) = \frac{1}{2}x^\top Ax - x^\top Ax^* + \frac{1}{2}x^{*\top} Ax^*.$$

Expand; the two cross terms are equal because A is symmetric.

With $Ax^* = b$: $x^{*\top} Ax^* = b^\top x^*$ and

$$x^{*\top} Ax^* = b^\top A^{-1}b, \text{ so } \frac{1}{2}(x - x^*)^\top A(x - x^*) = \frac{1}{2}x^\top Ax - b^\top x + \frac{1}{2}b^\top A^{-1}b = f(x) - c + \frac{1}{2}b^\top A^{-1}b.$$

Substitute $Ax^* = b$ and $x^* = A^{-1}b$; compare with the definition of f .

$f(x) = \frac{1}{2}(x - x^*)^\top A(x - x^*) + f(x^*)$ with $f(x^*) = c - \frac{1}{2}b^\top A^{-1}b$; the first term is > 0 for $x \neq x^*$ and 0 at x^* .

Rearrange step 3; $A \succ 0$ applied to the vector $x - x^*$.

$$A^{-1} = \frac{1}{3} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, x^* = \frac{1}{3}(2 - 2, -1 + 4)^\top = (0, 1)^\top, f(x^*) = -\frac{1}{2}b^\top x^* = -\frac{1}{2}(0 + 2) = -1.$$

The 2×2 inverse with $\det A = 3$; $b^\top A^{-1}b = b^\top x^*$. Check: $f(0, 1) = \frac{1}{2} \cdot 2 - 2 = -1$ and $\nabla f(0, 1) = (1, 2)^\top - (1, 2)^\top = 0$.

If $Av = \lambda v$ with $\lambda < 0$: $f(tv) = \frac{1}{2}\lambda t^2 \|v\|^2 - t b^\top v + c \rightarrow -\infty$ as $t \rightarrow \infty$.

Along an eigenvector the quadratic term is $\frac{1}{2}\lambda t^2 \|v\|^2$, and a negative quadratic beats a linear term for large t .

$\nabla f = Ax - b$, $\nabla^2 f = A$; $f(x) = \frac{1}{2}(x - x^*)^\top A(x - x^*) + f(x^*)$ **with** $x^* = A^{-1}b$ **and** $f(x^*) = c - \frac{1}{2}b^\top A^{-1}b$, **the unique minimiser; for the example** $x^* = (0, 1)^\top$, $f(x^*) = -1$; **with a negative eigenvalue** f **is unbounded below**

Completing the square in n dimensions: the bowl is centred at x^* and shaped by A . Setting $\nabla f = 0$ finds the same x^* , but only step 4 says it is a minimum rather than a saddle, and only PD makes it unique: a PSD A with $Av = 0$ leaves f constant along v (a valley of minimisers if $b \perp v$, no minimiser at all otherwise), which is Mistake 4 in the least-squares setting. Newton's method is exactly this calculation applied to the quadratic model of a general f , which the Newton page takes up.

Problem 7

Let f be twice differentiable on $\mathbb{R}^n$ and, for fixed x and v , let $g(t) = f(x + tv)$. Show that $g''(t) = v^\top \nabla^2 f(x + tv) v$, and use it to prove that f is convex if and only if $\nabla^2 f(x) \succeq 0$ for every x . Conclude that $\frac{1}{2}x^\top Ax$ is convex exactly when $A \succeq 0$.

$$g'(t) = \nabla f(x + tv)^\top v.$$

Chain rule: the derivative of f along the curve $x + tv$, whose velocity is v (the Jacobians page).

$$g''(t) = v^\top \nabla^2 f(x + tv) v.$$

Differentiate $\sum_i \partial_i f(x + tv) v_i$ again: each $\partial_i f$ contributes $\sum_j \partial_{ij} f v_j$, giving $\sum_{ij} v_i \partial_{ij} f v_j$.

If $\nabla^2 f \succeq 0$ everywhere, then $g'' \geq 0$ for every x, v , so each g is convex: $f(\lambda y + (1 - \lambda)z) = g(\lambda)$ with $x = z, v = y - z$, and $g(\lambda) = g(\lambda \cdot 1 + (1 - \lambda) \cdot 0) \leq \lambda g(1) + (1 - \lambda)g(0) = \lambda f(y) + (1 - \lambda)f(z)$.

The Jensen page's Problem 2: a one-variable function with nonnegative second derivative is convex; the segment from z to y is the line $z + t(y - z)$ for $t \in [0, 1]$.

Conversely, if f is convex then every g is convex, so $g''(0) \geq 0$, that is $v^\top \nabla^2 f(x) v \geq 0$ for every v .

Restricting a convex function to a line keeps the chord inequality, and a convex twice-differentiable function of one variable has $g'' \geq 0$: if $g''(0) < 0$ then $g(h) + g(-h) - 2g(0) = g''(0)h^2 + o(h^2) < 0$ for small h , contradicting the midpoint chord inequality.

$f(x) = \frac{1}{2}x^\top Ax$ has $\nabla^2 f = A$ at every x , so f is convex exactly when $A \succeq 0$.

Problem 6, step 1; the Hessian is constant.

$g''(t) = v^\top \nabla^2 f(x + tv) v$; f **convex**
 $\iff \nabla^2 f(x) \succeq 0$ **for all** x ; $\frac{1}{2}x^\top Ax$
convex $\iff A \succeq 0$

Convexity in n dimensions is convexity along every line, and the second derivative along a line is the Hessian's quadratic form in that direction, so PSD everywhere is the whole story. It has to be everywhere (the Jensen page's Mistake 5) and PSD is enough: PD gives strict convexity, but the converse fails (x^4 is strictly convex with $f''(0) = 0$). This is the test that certifies least squares (Hessian $2X^\top X$, Problem 4), ridge, logistic regression (Hessian $\frac{1}{N}X^\top DX$ on the Hessians page) and cross-entropy with softmax (the Jensen page's Problem 7) as convex, and it is what a neural network with a hidden layer fails.

Problem 8

For PD $A = Q\Lambda Q^\top$ define $A^{1/2} = Q\Lambda^{1/2}Q^\top$. Show that $A^{1/2}$ is symmetric PD with $(A^{1/2})^2 = A$, and that $A^{-1/2}AA^{-1/2} = I$. Compute $A^{1/2}$ for $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, and describe the ellipse $\{x : x^\top Ax = 1\}$.

$(A^{1/2})^\top = Q\Lambda^{1/2}Q^\top = A^{1/2}$, and its eigenvalues $\sqrt{\lambda_i}$ are positive, so $A^{1/2} \succ 0$.

$\Lambda^{1/2}$ is diagonal, hence symmetric; Problem 2's test.

$$(A^{1/2})^2 = Q\Lambda^{1/2}Q^\top Q\Lambda^{1/2}Q^\top = Q\Lambda Q^\top = A.$$

$$Q^\top Q = I \text{ in the middle and } \Lambda^{1/2}\Lambda^{1/2} = \Lambda.$$

$$A^{-1/2} = Q\Lambda^{-1/2}Q^\top \text{ and } A^{-1/2}AA^{-1/2} = Q\Lambda^{-1/2}\Lambda\Lambda^{-1/2}Q^\top = QQ^\top = I.$$

The same cancellation; $\Lambda^{-1/2}\Lambda\Lambda^{-1/2} = I$ entrywise, and $QQ^\top = I$ for a square orthogonal matrix.

For the example, $Q = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$, $\Lambda = \text{diag}(3, 1)$, so $A^{1/2} = \sqrt{3}q_1q_1^\top + 1 \cdot q_2q_2^\top = \frac{\sqrt{3}}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \sqrt{3}+1 & \sqrt{3}-1 \\ \sqrt{3}-1 & \sqrt{3}+1 \end{bmatrix}.$

The eigenvalues page's Problem 8 for Q and Λ ; $A^{1/2} = \sum_i \sqrt{\lambda_i} q_i q_i^\top$ with $q_1 = (1, 1)^\top / \sqrt{2}$ and $q_2 = (1, -1)^\top / \sqrt{2}$.

Check: $(A^{1/2})^2$ has diagonal $\frac{1}{4}((\sqrt{3}+1)^2 + (\sqrt{3}-1)^2) = \frac{1}{4}(8) = 2$ and off-diagonal $\frac{1}{4} \cdot 2(\sqrt{3}+1)(\sqrt{3}-1) = \frac{1}{4} \cdot 4 = 1$.

$$(\sqrt{3} \pm 1)^2 = 4 \pm 2\sqrt{3} \text{ and } (\sqrt{3}+1)(\sqrt{3}-1) = 2.$$

With $y = Q^\top x$: $x^\top Ax = y^\top \Lambda y = 3y_1^2 + y_2^2 = 1$, an ellipse with semi-axis $1/\sqrt{3}$ along q_1 and 1 along q_2 .

Problem 2's sum of squares in the eigenbasis; $\lambda_i y_i^2 = 1$ on the axis gives $y_i = 1/\sqrt{\lambda_i}$.

$A^{1/2} = Q\Lambda^{1/2}Q^\top \succ 0$ **with** $(A^{1/2})^2 = A$
and $A^{-1/2}AA^{-1/2} = I$; **for the example**
 $A^{1/2} = \frac{1}{2} \begin{bmatrix} \sqrt{3} + 1 & \sqrt{3} - 1 \\ \sqrt{3} - 1 & \sqrt{3} + 1 \end{bmatrix}$; $x^\top Ax = 1$ **is**
an ellipse with semi-axes $1/\sqrt{3}$
along $(1, 1)^\top / \sqrt{2}$ **and** 1 **along**
 $(1, -1)^\top / \sqrt{2}$

A PD matrix has two useful square roots: the symmetric $A^{1/2}$ here and the triangular Cholesky L of Problem 5, with $A = A^{1/2}A^{1/2} = LL^\top$. Either one turns $\mathcal{N}(0, I)$ into $\mathcal{N}(0, A)$ by multiplication, and either inverse whitens, which is the variance page's $W = \Lambda^{-1/2}Q^\top$ up to a rotation. The ellipse is the unit ball of the norm $\|x\|_A = \sqrt{x^\top Ax}$ and the level set of the Gaussian with precision A : long axes where the eigenvalue is small, and Problem 9 shows that the ratio of its axes, $\sqrt{\kappa}$, is what slows gradient descent.

Problem 9

For $f(x) = \frac{1}{2}x^\top Ax - b^\top x$ with $A \succ 0$, show that gradient descent satisfies $e_{k+1} = (I - \eta A)e_k$, that it converges from every start if and only if $0 < \eta < 2/\lambda_{\max}$, and that the error component along q_i shrinks by the factor $|1 - \eta\lambda_i|$ per step. Find the η that minimises the worst factor and the resulting rate in terms of κ . Evaluate for $A = \text{diag}(1, 10)$.

$$x_{k+1} - x^* = x_k - \eta(Ax_k - b) - x^* = (x_k - x^*) - \eta A(x_k - x^*), \text{ so } e_{k+1} = (I - \eta A)e_k.$$

$\nabla f = Ax - b$ (Problem 6) and $b = Ax^*$, so $Ax_k - b = A(x_k - x^*)$.

With $e_k = \sum_i c_i^{(k)} q_i$: $c_i^{(k+1)} = (1 - \eta\lambda_i)c_i^{(k)}$, so $c_i^{(k)} = (1 - \eta\lambda_i)^k c_i^{(0)}$.

$(I - \eta A)q_i = q_i - \eta\lambda_i q_i$; the eigenvectors decouple the iteration into n scalar recurrences.

$e_k \rightarrow 0$ for every $e_0 \iff |1 - \eta\lambda_i| < 1$ for every $i \iff 0 < \eta\lambda_i < 2$ for every $i \iff 0 < \eta < 2/\lambda_{\max}$.

A geometric sequence tends to zero exactly when its ratio has modulus below 1; the binding constraint is the largest eigenvalue, since $\eta\lambda_i < 2$ for all i is $\eta\lambda_{\max} < 2$, and $\eta > 0$ is needed for $\lambda_{\min}$.

The worst factor is $\rho(\eta) = \max_i |1 - \eta\lambda_i| = \max(1 - \eta\lambda_{\min}, \eta\lambda_{\max} - 1)$.

For $0 < \eta < 2/\lambda_{\max}$ the values $1 - \eta\lambda_i$ lie between $1 - \eta\lambda_{\max}$ and $1 - \eta\lambda_{\min}$, and the largest modulus is at one of the two ends.

ρ is minimised where the two branches are equal: $1 - \eta\lambda_{\min} = \eta\lambda_{\max} - 1$, so $\eta^* = \frac{2}{\lambda_{\min} + \lambda_{\max}}$ and $\rho(\eta^*) = 1 - \frac{2\lambda_{\min}}{\lambda_{\min} + \lambda_{\max}} = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\kappa - 1}{\kappa + 1}$.

One branch decreases in η and the other increases, so the maximum of the two is smallest where they cross; divide numerator and denominator by $\lambda_{\min}$.

$A = \text{diag}(1, 10)$: $\kappa = 10$, $\eta^* = 2/11$, rate $9/11 \approx 0.818$; with $\eta = 1/\lambda_{\max} = 0.1$ the factors are 0.9 and 0, rate 0.9; with $\eta = 0.3 > 2/10$ the stiff component grows by $|1 - 3| = 2$ per step.

Step 5 with $\lambda_{\min} = 1$, $\lambda_{\max} = 10$; the other two step sizes from step 4.

$e_{k+1} = (I - \eta A)e_k$; **convergence from every start** $\iff 0 < \eta < 2/\lambda_{\max}$; **the q_i component scales by $|1 - \eta\lambda_i|$; $\eta^* = \frac{2}{\lambda_{\min} + \lambda_{\max}}$ with rate $\frac{\kappa - 1}{\kappa + 1}$; for $\text{diag}(1, 10)$: $\eta^* = 2/11$, rate $9/11$**

The step size is set by the stiffest direction and the speed by the flattest: at η^* the error shrinks by $(\kappa - 1)/(\kappa + 1) \approx 1 - 2/\kappa$ per step, so reaching accuracy ϵ takes about $\frac{\kappa}{2} \log(1/\epsilon)$ steps, linear in the condition number. For a general smooth f the same analysis applies to the Hessian at the minimum, which is why conditioning is the quantity that normalisation, preconditioning and Adam's per-coordinate scaling attack, and why Newton's method, which multiplies by A^{-1} and makes every λ_i effectively 1, converges in one step on a quadratic. Mistake 5 chooses η from the wrong eigenvalue.

Problem 10

For any square matrix M , not necessarily symmetric, show that $x^\top Mx = x^\top Sx$ with $S = \frac{1}{2}(M + M^\top)$. Show that $M = \begin{bmatrix} 1 & 4 \\ 0 & 1 \end{bmatrix}$ has both eigenvalues equal to 1 and yet $v^\top Mv < 0$ for

some v . Why does the eigenvalue test of Problem 2 not apply?

$x^\top Mx = (x^\top Mx)^\top = x^\top M^\top x.$	A 1×1 matrix equals its transpose, and $(x^\top Mx)^\top = x^\top M^\top x.$
$x^\top Mx = \frac{1}{2}(x^\top Mx + x^\top M^\top x) = x^\top \frac{1}{2}(M + M^\top)x = x^\top Sx.$	Average the two equal expressions and factor out $x^\top$ and x .
M is upper-triangular with diagonal $(1, 1)$, so its eigenvalues are $1, 1$.	The eigenvalues of a triangular matrix are its diagonal entries (the eigenvalues page's Problem 4).
$S = \frac{1}{2} \left(\begin{bmatrix} 1 & 4 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 4 & 1 \end{bmatrix} \right) = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix},$ Problem 1's B , indefinite with eigenvalues $3, -1$.	Add M to its transpose and halve.
$v = (1, -1)^\top: v^\top Mv = 1 \cdot 1 + 4 \cdot 1 \cdot (-1) + 0 + 1 \cdot 1 = -2.$	Expand $x^\top Mx = M_{11}x_1^2 + M_{12}x_1x_2 + M_{21}x_2x_1 + M_{22}x_2^2$ with $x = (1, -1)^\top$; it agrees with $v^\top Sv = -2$ from Problem 1.
$x^\top Mx = x^\top \frac{1}{2}(M + M^\top)x; M = \begin{bmatrix} 1 & 4 \\ 0 & 1 \end{bmatrix}$ has eigenvalues $1, 1$ but $(1, -1)M(1, -1)^\top = -2$; the eigenvalue test needs a symmetric matrix	A quadratic form sees only the symmetric part of its matrix; the antisymmetric part $\frac{1}{2}(M - M^\top)$ contributes $x^\top Kx = 0$ because $x^\top Kx = -x^\top K^\top x = -x^\top Kx$. Problem 2 used $M = Q\Lambda Q^\top$ with orthonormal Q , which non-symmetric matrices do not have: M here has a single eigenvector direction (it is a shear) and no orthogonal eigenbasis, so its eigenvalues say nothing about its quadratic form (Mistake 2). This is also why the matrix-calculus page's $\nabla(x^\top Mx) = (M + M^\top)x$ has the symmetrised matrix in it, and why the Hessians page symmetrises A before reading off definiteness.

Where this goes wrong

1. Reading positive definite off a positive determinant

Sylvester's criterion is remembered as “the determinant is positive”, which is the last of its conditions and not the first.

$\begin{bmatrix} a & b \\ b & c \end{bmatrix} \succ 0 \iff a > 0 \text{ and } ac - b^2 > 0$	Right so far: Problem 3.
“The determinant is what decides it: positive determinant, positive definite.”	The shortcut that causes the mistake: Problem 3's proof needed the leading entry $a > 0$ first, and the determinant only afterwards.
$A = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$ has $\det A = 1 > 0$, so $A \succ 0$ ✗	$x^\top Ax = -\ x\ ^2 < 0$ for every $x \neq 0$: negative definite. The determinant is the product of the eigenvalues, and two negative eigenvalues multiply to a positive number; $\text{diag}(4, -1, -1)$ does the same in three dimensions with a positive trace too. The criterion needs every leading principal minor positive, a and then $ac - b^2$, which for $\text{diag}(-1, -1)$ fails at the first. The cheapest decisive test is Problem 5's: attempt the Cholesky factorisation and watch for a non-positive pivot, which here fails at $\sqrt{-1}$.

2. Trusting the eigenvalues of a non-symmetric matrix

Positive eigenvalues mean positive definite, and the matrix at hand is not checked for symmetry before the test is applied.

$A = Q\Lambda Q^\top \succ 0 \iff \text{every } \lambda_i > 0$	Right so far: Problem 2, for symmetric A .
“Eigenvalues are eigenvalues; compute them and look at the signs.”	The habit that causes the mistake: Problem 2's proof wrote $x^\top Ax = \sum_i \lambda_i (q_i^\top x)^2$, which needs an orthonormal eigenbasis, and only symmetric matrices are guaranteed one.

$M = \begin{bmatrix} 1 & 4 \\ 0 & 1 \end{bmatrix}$ has eigenvalues 1, 1 > 0, so $x^\top M x > 0$ for all $x \neq 0$ ✗

$(1, -1)M(1, -1)^\top = -2$ (Problem 10).
The quadratic form of M is the quadratic form of $\frac{1}{2}(M + M^\top) = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$, whose eigenvalues are 3 and -1, so the form is indefinite. The right procedure for a non-symmetric matrix is to symmetrise first and then test; a Jacobian $\partial f / \partial x$ of a non-gradient vector field is the usual place this comes up, and the Hessians page's $\frac{1}{2}(A + A^\top)$ is the same step taken for the Hessian of $\frac{1}{2}x^\top A x$.

3. Sampling with the transposed Cholesky factor

$A = LL^\top$ and $A = L^\top L$ look alike, and whichever factor is to hand gets multiplied onto the noise.

$A = LL^\top$ with $L = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 1 & 2 \end{bmatrix}$, and $x = \mu + L\varepsilon$ has covariance $LL^\top = A$ for $\varepsilon \sim \mathcal{N}(0, I)$

Right so far: Problem 5 and the Gaussian page's Problem 7.

“ L and $L^\top$ are the same factor, just written the other way.”

The habit that causes the mistake: the covariance of $M\varepsilon$ is $MM^\top$, and the order of the two factors is not optional.

$x = \mu + L^\top \varepsilon$ has covariance A ✗

Its covariance is $L^\top (L^\top)^\top = L^\top L = \begin{bmatrix} 6 & 3 & 2 \\ 3 & 5 & 2 \\ 2 & 2 & 4 \end{bmatrix} \neq A$: the first coordinate now has variance 6 instead of 4. $L^\top L$ and $LL^\top$ share eigenvalues and determinant (64 in both cases), so a check on those would not catch it; a check on $\text{Var}(x_1)$ would. The same slip in the other direction happens in the whitening step, $L^{-1}(x - \mu)$ against $L^{-\top}(x - \mu)$, and in software it is the difference between a library returning the lower factor and one returning the upper. Problem 8's symmetric root $A^{1/2}$ has no such ambiguity, which is one reason to prefer it when cost is not the issue.

4. Assuming the normal equations have a unique solution

$X^\top X$ is positive semidefinite and usually positive definite, and the usual case becomes the only case.

$X^\top X \succeq 0$ for every X , with $v^\top X^\top X v = \|Xv\|^2$

Right so far: Problem 4, step 1.

“A Gram matrix is positive definite, so $(X^\top X)^{-1}$ exists and $w = (X^\top X)^{-1} X^\top y$.”

The habit that causes the mistake: Problem 4 made PD conditional on linearly independent columns, and nothing has been said about the columns.

$w^* = (X^\top X)^{-1} X^\top y$ is the unique least-squares solution for any $X \nprec$

If two features are collinear, or there are more features than examples ($d > N$), some $v \neq 0$ has $Xv = 0$, $X^\top X$ has a zero eigenvalue and is singular, and every $w^* + tv$ fits equally well: a whole line (or more) of minimisers, as in Problem 6's PSD case. Problem 1's $C = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ is $X^\top X$ for the single row $X = (1, 2)$, and its null vector $(2, -1)^\top$ is the direction along which the fit cannot tell the features apart. Ridge adds λI and restores a unique solution (Problem 4), and the least-squares page's pseudoinverse picks the minimiser of smallest norm; both are choices, not consequences of the normal equations.

5. Choosing the gradient-descent step from the smallest eigenvalue

The slow direction is the one with the small eigenvalue, and the step is enlarged to speed it up.

The component of the error along q_i scales by $1 - \eta\lambda_i$ per step

Right so far: Problem 9, step 2.

“The slowest direction is $\lambda_{\min}$, so pick η to make $1 - \eta\lambda_{\min}$ small.”

The habit that causes the mistake: Problem 9's convergence condition is on $\lambda_{\max}$; the step that is right for the flattest direction is far too long for the stiffest.

For $A = \text{diag}(1, 10)$, take $\eta = 1 = 1/\lambda_{\min}$, which kills the slow component in one step ✗

The stiff component scales by $1 - 10 = -9$: after k steps it is $(-9)^k$ times its start, so the iteration diverges, oscillating across the valley in ever larger swings. Any $\eta \geq 2/\lambda_{\max} = 0.2$ does this. The admissible range is $0 < \eta < 0.2$, the best choice is $\eta^* = 2/11$ (Problem 9), and the slow direction then shrinks by only $9/11$ per step: with a single step size the slow direction cannot be sped up without breaking the fast one, and that gap, κ , is what momentum (the optimisers page) and preconditioning exist to close.

eulerpad.app/practice/positive-definite-matrices-and-quadratic-forms