

Partial derivatives, gradients and Hessians

Ten problems on functions of several variables: partial derivatives, the gradient as the direction of steepest ascent and the normal to a level set, directional derivatives, Hessians and the classification of critical points, the quadratic $\frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$ with a Newton step, log-sum-exp, second-order Taylor, the logistic-loss Hessian and convexity, with worked solutions and the mistakes that misread a Hessian.

BEFORE YOU START

A loss function takes many numbers and returns one. Its first derivatives, collected into the gradient, say which way is downhill and how steeply; its second derivatives, collected into the Hessian, say how the slope itself changes, which decides whether a flat point is a minimum, a maximum or a saddle, whether the function is convex, and how far a Newton step should go. These ten problems build both objects from ordinary partial derivatives, then use them on the functions machine learning keeps meeting: a quadratic, log-sum-exp and the logistic loss. The five mistakes at the end each misread one of the two.

- f maps $\mathbb{R}^n$ to $\mathbb{R}$. The partial derivative $\partial f / \partial x_i$ differentiates in x_i with every other variable held fixed. In two variables, $f_x = \partial f / \partial x$, and f_{xy} means differentiate in x , then in y .
- The conventions are those of the matrix-calculus page: vectors are columns, the row derivative $\partial f / \partial \mathbf{x}$ is $1 \times n$, and the gradient is its transpose, $\nabla f = (\partial f / \partial \mathbf{x})^\top$, an $n \times 1$ column. A point in the plane is written (x, y) ; a vector is written as a column, such as $(4, 8)^\top$.
- The Hessian $\nabla^2 f$ is the $n \times n$ matrix with $(\nabla^2 f)_{ij} = \partial^2 f / \partial x_i \partial x_j$; it is the Jacobian of ∇f . When the second partials are continuous, which holds for every function on this page, it is symmetric: $f_{xy} = f_{yx}$.
- For a unit vector u , the directional derivative is $D_u f = \left. \frac{d}{dt} f(\mathbf{x} + t\mathbf{u}) \right|_{t=0} = \nabla f(\mathbf{x})^\top u$, the rate of change of f per unit distance along u .
- A symmetric matrix H is positive semidefinite (PSD) if $\mathbf{v}^\top H \mathbf{v} \geq 0$ for every $\mathbf{v}$, and positive definite if $\mathbf{v}^\top H \mathbf{v} > 0$ for every $\mathbf{v} \neq 0$; equivalently, all its eigenvalues are ≥ 0 , or all > 0 .
- A critical point has $\nabla f = 0$. There, a positive definite Hessian means a strict local minimum, a negative definite one a strict local maximum, and one with eigenvalues of both signs a saddle; a zero eigenvalue leaves the test inconclusive. For 2×2 , $\det \nabla^2 f$ is the product of the two eigenvalues.
- A twice-differentiable f on $\mathbb{R}^n$ is convex exactly when $\nabla^2 f(\mathbf{x})$ is PSD at every $\mathbf{x}$.

- Newton's method steps from x_k to $x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$. $\sigma(t) = 1/(1 + e^{-t})$ is the logistic sigmoid, with $\sigma' = \sigma(1 - \sigma)$ (the differentiation-rules page).

Builds on: Differentiation rules: chain, product, quotient

Problems

1. Let $f(x, y) = x^2 y^3 + y e^{2x}$. Compute $\partial f / \partial x$, $\partial f / \partial y$ and $\nabla f(0, 1)$. difficulty ·

2. Let $f(x, y) = x e^y + y^2$. Find the rate of change of f at $(1, 0)$ in the direction of $v = (3, 4)^\top$. difficulty ·

3. Let $f(x, y) = x^2 + 4y^2$. At $(2, 1)$, find the unit direction of steepest ascent and the rate of increase along it. Then show that $\nabla f(2, 1)$ is perpendicular to the level curve of f through $(2, 1)$. difficulty ··

4. Let $f(x, y) = x^3 y^2 - 4xy + y^3$. Compute the Hessian $\nabla^2 f$ and confirm that $f_{xy} = f_{yx}$. difficulty ·

5. Find every critical point of $f(x, y) = x^3 + x^2 + 3xy + y^2$ and classify each as a local minimum, local maximum or saddle. difficulty ··

6. Let $f(x) = \frac{1}{2} x^\top A x - b^\top x$ with $A \in \mathbb{R}^{n \times n}$, not necessarily symmetric, and $b \in \mathbb{R}^n$. Compute ∇f and $\nabla^2 f$. Assuming $\frac{1}{2}(A + A^\top)$ is invertible, take one Newton step from an arbitrary x_0 . Where does it land? difficulty ··

7. Let $f(x) = \log \sum_{j=1}^n e^{x_j}$ (log-sum-exp). Compute ∇f and $\nabla^2 f$. difficulty ···

8. Write the second-order Taylor polynomial of a function f of n variables around x , then apply it to log-sum-exp from Problem 7 around $x = 0$. Express the result in terms of $h \in \mathbb{R}^n$, the displacement from 0. difficulty ··

9. Logistic regression has data $X \in \mathbb{R}^{N \times d}$ with row n equal to $x^{(n)\top}$, difficulty $\dots$ labels $y_n \in \{0, 1\}$, weights $w \in \mathbb{R}^d$, probabilities $p = \sigma(Xw)$ with $p_n = \sigma(x^{(n)\top} w)$, and the mean binary cross-entropy $L(w)$, whose gradient is $\nabla_w L = \frac{1}{N} X^\top (p - y)$ (the regression page derives it). Show that $\nabla_w^2 L = \frac{1}{N} X^\top D X$ with $D = \text{diag}(p \odot (1 - p))$, and that it is positive semidefinite.

10. For which real numbers c is $f(x, y) = x^2 + cxy + y^2 + e^x$ convex on difficulty $\dots$ $\mathbb{R}^2$?

Answers

1. $\partial f / \partial x = 2xy^3 + 2ye^{2x}$, $\partial f / \partial y = 3x^2y^2 + e^{2x}$, $\nabla f(0, 1) = (2, 1)^\top$
2. $D_u f(1, 0) = \frac{7}{5}$, with $u = (\frac{3}{5}, \frac{4}{5})^\top$
3. $\nabla f(2, 1) = (4, 8)^\top$; steepest ascent along $(1, 2)^\top / \sqrt{5}$ at rate $4\sqrt{5}$; the level curve $x^2 + 4y^2 = 8$ has tangent $(-2, 1)^\top$ there, and $(4, 8)(-2, 1)^\top = -8 + 8 = 0$
4. $\nabla^2 f = \begin{pmatrix} 6xy^2 & 6x^2y - 4 \\ 6x^2y - 4 & 2x^3 + 6y \end{pmatrix}$, and $f_{xy} = f_{yx} = 6x^2y - 4$
5. $(0, 0)$ is a saddle: $\det = -5 < 0$, eigenvalues 5 and -1 . $(\frac{5}{6}, -\frac{5}{4})$ is a strict local minimum:
 $\det = 5 > 0$ and $f_{xx} = 7 > 0$
6. $\nabla f = \frac{1}{2}(A + A^\top)x - b$, $\nabla^2 f = \frac{1}{2}(A + A^\top)$; Newton's step gives $x_1 = (\frac{1}{2}(A + A^\top))^{-1}b$ from any x_0 , which is $A^{-1}b$ when A is symmetric
7. $\nabla f = s$ and $\nabla^2 f = \text{diag}(s) - ss^\top$, where $s_i = e^{x_i} / \sum_j e^{x_j}$
8. $\log \sum_i e^{h_i} \approx \log n + \frac{1}{n} \sum_i h_i + \frac{1}{2n} \sum_i h_i^2 - \frac{1}{2n^2} (\sum_i h_i)^2$
9. $\nabla_w^2 L = \frac{1}{N} X^\top D X$ with $D = \text{diag}(p \odot (1 - p))$, and $v^\top \nabla_w^2 L v = \frac{1}{N} \sum_n p_n(1 - p_n)(x^{(n)\top} v)^2 \geq 0$ for every v , so it is positive semidefinite
10. $\nabla^2 f = \begin{pmatrix} 2 + e^x & c \\ c & 2 \end{pmatrix}$, and f is convex exactly when $|c| \leq 2$

Worked solutions

Problem 1

Let $f(x, y) = x^2y^3 + ye^{2x}$. Compute $\partial f/\partial x$, $\partial f/\partial y$ and $\nabla f(0, 1)$.

$\frac{\partial}{\partial x}(x^2y^3) = 2xy^3$ and $\frac{\partial}{\partial x}(ye^{2x}) = 2ye^{2x}$.	With y held fixed, y^3 and y are constant factors, and $(e^{2x})' = 2e^{2x}$ by the chain rule.
$\frac{\partial}{\partial y}(x^2y^3) = 3x^2y^2$ and $\frac{\partial}{\partial y}(ye^{2x}) = e^{2x}$.	Now x is held fixed, so x^2 and e^{2x} are the constant factors.
$\nabla f(0, 1) = (0 + 2 \cdot 1 \cdot e^0, 0 + e^0)^\top = (2, 1)^\top$.	The gradient stacks the two partials as a column; evaluate each at $x = 0, y = 1$.
$\partial f/\partial x = 2xy^3 + 2ye^{2x}, \partial f/\partial y = 3x^2y^2 + e^{2x}, \nabla f(0, 1) = (2, 1)^\top$	

Problem 2

Let $f(x, y) = xe^y + y^2$. Find the rate of change of f at $(1, 0)$ in the direction of $v = (3, 4)^\top$.

$\nabla f = (e^y, xe^y + 2y)^\top$.	Partial in x with y fixed, then in y with x fixed.
$\nabla f(1, 0) = (1, 1)^\top$.	$e^0 = 1$ and $2y = 0$ at $y = 0$.
$\ v\ = \sqrt{3^2 + 4^2} = 5$, so $u = v/\ v\ = (\frac{3}{5}, \frac{4}{5})^\top$.	A rate per unit distance needs a unit vector; v only names the direction.
$D_u f(1, 0) = \nabla f(1, 0)^\top u = \frac{3}{5} + \frac{4}{5}$.	The directional derivative is the gradient's inner product with the unit direction.

$$D_u f(1, 0) = \frac{7}{5}, \text{ with } u = \left(\frac{3}{5}, \frac{4}{5}\right)^\top$$

Sanity check: it cannot exceed $\|\nabla f(1, 0)\| = \sqrt{2} \approx 1.41$ (Problem 3), and $\frac{7}{5} = 1.4$, because v points almost along $(1, 1)^\top$.

Problem 3

Let $f(x, y) = x^2 + 4y^2$. At $(2, 1)$, find the unit direction of steepest ascent and the rate of increase along it. Then show that $\nabla f(2, 1)$ is perpendicular to the level curve of f through $(2, 1)$.

$$\nabla f = (2x, 8y)^\top, \text{ so } \nabla f(2, 1) = (4, 8)^\top.$$

The two partials of a sum of one-variable terms.

For a unit u , $D_u f = \nabla f^\top u = \|\nabla f\| \cos \theta$, with θ the angle between ∇f and u .

The inner product of two vectors is the product of their lengths and the cosine of the angle between them, and $\|u\| = 1$.

$D_u f$ is largest when $\theta = 0$: $u = \nabla f / \|\nabla f\| = (4, 8)^\top / \sqrt{80} = (1, 2)^\top / \sqrt{5}$, at rate $\|\nabla f\| = \sqrt{80} = 4\sqrt{5}$.

$\cos \theta \leq 1$, with equality only when u points along the gradient.

The level curve through $(2, 1)$ is $x^2 + 4y^2 = 8$.

$f(2, 1) = 4 + 4 = 8$, and a level curve is the set where f keeps that value.

Let $r(t)$ be any curve on it with $r(0) = (2, 1)$. Then $f(r(t)) = 8$ for all t , so $\nabla f(2, 1)^\top r'(0) = 0$.

Chain rule: the derivative of $f(r(t))$ is $\nabla f^\top r'(t)$, and the derivative of a constant is 0. Every tangent to the level curve is therefore perpendicular to the gradient.

Concretely, $r(t) = (2\sqrt{2} \cos t, \sqrt{2} \sin t)$ traces the curve and passes through $(2, 1)$ at $t = \pi/4$, where $r'(\pi/4) = (-2\sqrt{2} \sin \frac{\pi}{4}, \sqrt{2} \cos \frac{\pi}{4})^\top = (-2, 1)^\top$.

$(2\sqrt{2} \cos t)^2 + 4(\sqrt{2} \sin t)^2 = 8$, and $\sin \frac{\pi}{4} = \cos \frac{\pi}{4} = 1/\sqrt{2}$.

$\nabla f(2, 1) = (4, 8)^\top$; **steepest ascent**
along $(1, 2)^\top / \sqrt{5}$ **at rate** $4\sqrt{5}$; **the level**
curve $x^2 + 4y^2 = 8$ **has tangent**
 $(-2, 1)^\top$ **there, and** $(4, 8) \cdot (-2, 1)^\top =$
 $-8 + 8 = 0$

So the gradient points straight across the contours, uphill, and its length is the slope in that direction. Gradient descent steps along $-\nabla f$ for this reason.

Problem 4

Let $f(x, y) = x^3y^2 - 4xy + y^3$. Compute the Hessian $\nabla^2 f$ and confirm that $f_{xy} = f_{yx}$.

$$f_x = 3x^2y^2 - 4y \text{ and } f_y = 2x^3y - 4x + 3y^2.$$

First partials, each with the other variable held fixed.

$$f_{xx} = \frac{\partial}{\partial x}(3x^2y^2 - 4y) = 6xy^2.$$

Differentiate f_x in x again; $-4y$ is constant in x .

$$f_{xy} = \frac{\partial}{\partial y}(3x^2y^2 - 4y) = 6x^2y - 4.$$

Differentiate f_x in y .

$$f_{yx} = \frac{\partial}{\partial x}(2x^3y - 4x + 3y^2) = 6x^2y - 4.$$

Differentiate f_y in x : a different calculation that gives the same result, because the second partials of a polynomial are continuous.

$$f_{yy} = \frac{\partial}{\partial y}(2x^3y - 4x + 3y^2) = 2x^3 + 6y.$$

Differentiate f_y in y .

$$\nabla^2 f = \begin{pmatrix} 6xy^2 & 6x^2y - 4 \\ 6x^2y - 4 & 2x^3 + 6y \end{pmatrix}, \text{ and } f_{xy} = f_{yx} = 6x^2y - 4$$

Row i is the gradient of $\partial f / \partial x_i$ written as a row, so the Hessian is the Jacobian of ∇f ; symmetry means the off-diagonal entries need computing only once.

Problem 5

Find every critical point of $f(x, y) = x^3 + x^2 + 3xy + y^2$ and classify each as a local minimum, local maximum or saddle.

$$f_x = 3x^2 + 2x + 3y \text{ and } f_y = 3x + 2y.$$

First partials.

$f_y = 0$ gives $y = -\frac{3}{2}x$.	Solve the linear equation first, then substitute into the other.
$f_x = 3x^2 + 2x - \frac{9}{2}x = x(3x - \frac{5}{2}) = 0$, so $x = 0$ or $x = \frac{5}{6}$.	Substitute $y = -\frac{3}{2}x$; a product is zero when a factor is.
The critical points are $(0, 0)$ and $(\frac{5}{6}, -\frac{5}{4})$.	$y = -\frac{3}{2}x$ at each root.
$\nabla^2 f = \begin{pmatrix} 6x+2 & 3 \\ 3 & 2 \end{pmatrix}$.	$f_{xx} = 6x + 2, f_{xy} = f_{yx} = 3, f_{yy} = 2$.
At $(0, 0)$: $\nabla^2 f = \begin{pmatrix} 2 & 3 \\ 3 & 2 \end{pmatrix}$ with $\det = 4 - 9 = -5$; its eigenvalues are 5 and -1 , along $(1, 1)^\top$ and $(1, -1)^\top$.	A negative determinant is a negative product of the two eigenvalues, so they have opposite signs; f curves up along one eigenvector and down along the other. Along $(1, -1)^\top$, $f(t, -t) = t^3 - t^2$, which is below $f(0, 0) = 0$ for small $t \neq 0$.
At $(\frac{5}{6}, -\frac{5}{4})$: $f_{xx} = 6 \cdot \frac{5}{6} + 2 = 7$, and $\nabla^2 f = \begin{pmatrix} 7 & 3 \\ 3 & 2 \end{pmatrix}$ with $\det = 14 - 9 = 5$.	A positive determinant means the eigenvalues share a sign, and the trace $7 + 2 = 9$, their sum, makes that sign positive: the Hessian is positive definite.
$(0, 0)$ is a saddle: $\det = -5 < 0$, eigenvalues 5 and -1. $(\frac{5}{6}, -\frac{5}{4})$ is a strict local minimum: $\det = 5 > 0$ and $f_{xx} = 7 > 0$	The minimum is only local: $f(x, 0) = x^3 + x^2$ goes to $-\infty$ as $x \rightarrow -\infty$.

Problem 6

Let $f(x) = \frac{1}{2}x^\top Ax - b^\top x$ with $A \in \mathbb{R}^{n \times n}$, not necessarily symmetric, and $b \in \mathbb{R}^n$. Compute ∇f and $\nabla^2 f$. Assuming $\frac{1}{2}(A + A^\top)$ is invertible, take one Newton step from an arbitrary x_0 . Where does it land?

$$\nabla(x^\top Ax) = (A + A^\top)x \text{ and } \nabla(b^\top x) = b.$$

Problems 4 and 3 of the matrix-calculus page: x appears on both sides of A , so both A and $A^\top$ contribute.

$\nabla f = \frac{1}{2}(A + A^\top)x - b.$	The gradient is linear: scale the first term by $\frac{1}{2}$ and subtract the second.
$\nabla^2 f = \frac{1}{2}(A + A^\top).$	The Hessian is the Jacobian of ∇f , and the Jacobian of Mx is M ; the constant b drops out. It is symmetric, as a Hessian must be, even when A is not.
Write $H = \frac{1}{2}(A + A^\top)$. Then $x^\top Ax = x^\top Hx$.	$A = H + K$ with $K = \frac{1}{2}(A - A^\top)$, and $x^\top Kx$ is a scalar equal to its own transpose $x^\top K^\top x = -x^\top Kx$, so it is 0. Only the symmetric part of A is ever seen by f .
$x_1 = x_0 - H^{-1}(Hx_0 - b) = x_0 - x_0 + H^{-1}b = H^{-1}b.$	Newton's step with $\nabla f(x_0) = Hx_0 - b$ and $\nabla^2 f = H$; $H^{-1}H = I$.
$\nabla f(x_1) = HH^{-1}b - b = 0.$	So x_1 is the critical point, reached in one step from any x_0 : Newton's step jumps to the critical point of the second-order Taylor model, and for a quadratic that model is f itself.
$\nabla f = \frac{1}{2}(A + A^\top)x - b$, $\nabla^2 f = \frac{1}{2}(A + A^\top)$; Newton's step gives $x_1 = (\frac{1}{2}(A + A^\top))^{-1}b$ from any x_0, which is $A^{-1}b$ when A is symmetric	When $\frac{1}{2}(A + A^\top)$ is also positive definite, f is strictly convex and x_1 is its global minimum.

Problem 7

Let $f(x) = \log \sum_{j=1}^n e^{x_j}$ (log-sum-exp). Compute ∇f and $\nabla^2 f$.

Let $Z = \sum_j e^{x_j}$, so $f = \log Z$ and $\partial Z / \partial x_i = e^{x_i}$.	Only the $j = i$ term of the sum depends on x_i .
$\frac{\partial f}{\partial x_i} = \frac{1}{Z} \cdot e^{x_i} = s_i$, where $s_i = e^{x_i} / \sum_j e^{x_j}$.	Chain rule: the derivative of $\log$ at Z is $1/Z$, times $\partial Z / \partial x_i$. The vector s is the softmax of x .

$\frac{\partial s_i}{\partial x_k} = \frac{\delta_{ik} e^{x_i} Z - e^{x_i} e^{x_k}}{Z^2} = \delta_{ik} s_i - s_i s_k,$ with $\delta_{ik} = 1$ if $i = k$ and 0 otherwise.	Quotient rule on e^{x_i}/Z: the numerator depends on x_k only when $k = i$, and $\partial Z/\partial x_k = e^{x_k}$.
$\nabla^2 f = \text{diag}(s) - ss^\top.$	Entry (i, k) of the Hessian is $\partial s_i/\partial x_k$; the $\delta_{ik} s_i$ terms form the diagonal matrix $\text{diag}(s)$ and the $s_i s_k$ terms form the outer product $ss^\top$.
$\nabla f = s$ and $\nabla^2 f = \text{diag}(s) - ss^\top$, where $s_i = e^{x_i} / \sum_j e^{x_j}$	Two checks: the Hessian is symmetric, and $\nabla^2 f \mathbf{1} = s - s(s^\top \mathbf{1}) = 0$ because $\sum_i s_i = 1$, matching $f(x + c\mathbf{1}) = f(x) + c$, which is linear along $\mathbf{1}$ and so has no curvature there.

Problem 8

Write the second-order Taylor polynomial of a function f of n variables around x , then apply it to log-sum-exp from Problem 7 around $x = 0$. Express the result in terms of $h \in \mathbb{R}^n$, the displacement from 0.

Let $g(t) = f(x + th)$; then $g'(0) = \nabla f(x)^\top h$ and $g''(0) = h^\top \nabla^2 f(x) h$.	Chain rule: $g'(t) = \nabla f(x + th)^\top h = \sum_i h_i \partial f/\partial x_i$, and differentiating each $\partial f/\partial x_i$ once more along h gives $\sum_{i,j} h_i h_j \partial^2 f/\partial x_i \partial x_j$.
$f(x + h) = g(1) \approx g(0) + g'(0) + \frac{1}{2} g''(0) = f(x) + \nabla f(x)^\top h + \frac{1}{2} h^\top \nabla^2 f(x) h.$	The one-variable Taylor polynomial of g at 0, evaluated at $t = 1$; the error is third order in $\ h\$.
At $x = 0$: $f(0) = \log n$ and $s = \frac{1}{n} \mathbf{1}$.	Every $e^0 = 1$, so the sum is n and each softmax entry is $1/n$.
$\nabla f(0)^\top h = \frac{1}{n} \sum_i h_i.$	$\nabla f = s$ (Problem 7).
$\nabla^2 f(0) = \frac{1}{n} I - \frac{1}{n^2} \mathbf{1} \mathbf{1}^\top, \text{ so } h^\top \nabla^2 f(0) h = \frac{1}{n} \sum_i h_i^2 - \frac{1}{n^2} \left(\sum_i h_i \right)^2.$	$\nabla^2 f = \text{diag}(s) - ss^\top$ (Problem 7) with $s = \frac{1}{n} \mathbf{1}$, and $h^\top \mathbf{1} = \sum_i h_i$.

$$\log \sum_i e^{h_i} \approx \log n + \frac{1}{n} \sum_i h_i + \frac{1}{2n} \sum_i h_i^2 - \frac{1}{2n^2} \left(\sum_i h_i \right)^2$$

Steps 2 to 5. Read with $\bar{h} = \frac{1}{n} \sum_i h_i$, it is $\log n + \bar{h} + \frac{1}{2} \cdot \frac{1}{n} \sum_i (h_i - \bar{h})^2$: the mean of the h_i plus half their variance, so near 0 log-sum-exp sits above the mean by an amount set by the spread.

Problem 9

Logistic regression has data $X \in \mathbb{R}^{N \times d}$ with row n equal to $x^{(n)\top}$, labels $y_n \in \{0, 1\}$, weights $w \in \mathbb{R}^d$, probabilities $p = \sigma(Xw)$ with $p_n = \sigma(x^{(n)\top} w)$, and the mean binary cross-entropy $L(w)$, whose gradient is $\nabla_w L = \frac{1}{N} X^\top (p - y)$ (the regression page derives it). Show that $\nabla_w^2 L = \frac{1}{N} X^\top D X$ with $D = \text{diag}(p \odot (1 - p))$, and that it is positive semidefinite.

$$\nabla_w L = \frac{1}{N} \sum_n (p_n - y_n) x^{(n)}.$$

$X^\top$ has columns $x^{(n)}$, so $X^\top (p - y)$ is the sum of those columns weighted by the entries of $p - y$.

$$\frac{\partial p_n}{\partial w} = p_n(1 - p_n) x^{(n)\top}, \text{ a } 1 \times d \text{ row.}$$

Chain rule: $\sigma' = \sigma(1 - \sigma)$ evaluated at $x^{(n)\top} w$, times the row derivative of $x^{(n)\top} w$, which is $x^{(n)\top}$.

The Jacobian of $(p_n - y_n) x^{(n)}$ with respect to w is $x^{(n)} \frac{\partial p_n}{\partial w} = p_n(1 - p_n) x^{(n)} x^{(n)\top}$, a $d \times d$ matrix.

$x^{(n)}$ and y_n do not depend on w ; only the scalar p_n does, and a constant column times a scalar's row derivative is a column times a row.

$$\nabla_w^2 L = \frac{1}{N} \sum_n p_n(1 - p_n) x^{(n)} x^{(n)\top} = \frac{1}{N} X^\top D X.$$

The Hessian is the Jacobian of the gradient, so sum step 3 over n ; a weighted sum of outer products of the rows of X is $X^\top \text{diag}(\text{weights}) X$, since $X^\top D X = \sum_n D_{nn} x^{(n)} x^{(n)\top}$.

For any $v \in \mathbb{R}^d$: $v^\top \nabla_w^2 L v = \frac{1}{N} \sum_n p_n(1 - p_n) (x^{(n)\top} v)^2$.

$v^\top x^{(n)} x^{(n)\top} v = (x^{(n)\top} v)^2$, a square.

$\nabla_w^2 L = \frac{1}{N} X^\top D X$ **with** $D = \text{diag}(p \odot (1 - p))$, **and** $v^\top \nabla_w^2 L v = \frac{1}{N} \sum_n p_n (1 - p_n) (x^{(n)\top} v)^2 \geq 0$ **for every** v , **so it is positive semidefinite**

Every p_n lies strictly between 0 and 1, so every weight $p_n(1 - p_n)$ is positive and every term is ≥ 0 . A PSD Hessian at every w makes L convex, so any critical point is a global minimum.

Problem 10

For which real numbers c is $f(x, y) = x^2 + cxy + y^2 + e^x$ convex on $\mathbb{R}^2$?

$$f_x = 2x + cy + e^x \text{ and } f_y = cx + 2y.$$

First partials.

$$\nabla^2 f = \begin{pmatrix} 2 + e^x & c \\ c & 2 \end{pmatrix}.$$

$f_{xx} = 2 + e^x$, $f_{xy} = f_{yx} = c$, $f_{yy} = 2$. It depends on x but not on y .

A symmetric 2×2 matrix with positive diagonal is PSD exactly when its determinant is ≥ 0 .

The determinant is the product of the eigenvalues and the trace their sum; a positive trace rules out two negative eigenvalues, so $\det \geq 0$ leaves both ≥ 0 , while $\det < 0$ forces one negative.

$$\det \nabla^2 f = 2(2 + e^x) - c^2 = 4 + 2e^x - c^2.$$

Product of the diagonal minus the square of the off-diagonal entry.

If $c^2 \leq 4$, then $\det \nabla^2 f \geq 2e^x > 0$ at every point, so the Hessian is positive definite everywhere and f is convex.

$4 - c^2 \geq 0$, and $e^x > 0$ for every x .

If $c^2 > 4$, then $\det \nabla^2 f < 0$ wherever $e^x < \frac{1}{2}(c^2 - 4)$, so the Hessian has a negative eigenvalue there and f is not convex.

e^x takes every positive value, so such x always exist; at those points f curves down along an eigenvector, which a convex function never does.

$\nabla^2 f = \begin{pmatrix} 2 + e^x & c \\ c & 2 \end{pmatrix}$, and f is convex exactly when $|c| \leq 2$

The boundary case $|c| = 2$ is still convex because e^x keeps the determinant positive; without the e^x term, $x^2 \pm 2xy + y^2 = (x \pm y)^2$ would be only just convex, flat along one line.

Where this goes wrong

1. Directional derivative along a non-unit vector

The formula $\nabla f^\top u$ is easy to remember and easy to use with whatever vector the problem gives.

$$\nabla f(1, 0) = (1, 1)^\top \text{ for } f = xe^y + y^2$$

Right so far: this is step 2 of Problem 2.

“The directional derivative is the gradient dotted with the direction.”

The shortcut that causes the mistake: true only when the direction is a unit vector.

$$D_v f(1, 0) = (1, 1) (3, 4)^\top = 7 \text{ ✗}$$

$\nabla f^\top v$ scales with the length of v : it is the rate along v per unit of the parameter t in $f(x + tv)$, moving at speed $\|v\| = 5$. The rate per unit distance divides by that: $7/5$ (Problem 2). The giveaway is that 7 exceeds $\|\nabla f(1, 0)\| = \sqrt{2}$, the steepest rate in any direction (Problem 3).

2. Calling a saddle a minimum because $f_{xx} > 0$

In one variable a positive second derivative at a critical point settles it, and f_{xx} looks like that second derivative.

At $(0, 0)$, $f = x^3 + x^2 + 3xy + y^2$ has

$$\nabla^2 f = \begin{pmatrix} 2 & 3 \\ 3 & 2 \end{pmatrix} \text{ and } \det = -5$$

Right so far: these are the numbers in step 6 of Problem 5.

“ $f'' > 0$ means a minimum; the determinant only has to be nonzero for the test to apply.”

The analogy that causes the mistake: the one-variable test, with the determinant treated as a validity check rather than a sign to read.

$f_{xx}(0,0) = 2 > 0$ and $\det \neq 0$, so $(0,0)$ is a local minimum ✗

The sign of the determinant is the content of the test: $\det$ is the product of the eigenvalues, and -5 means one is negative. $f_{xx} > 0$ is the curvature along the x axis only. Along $(1, -1)^\top$, $f(t, -t) = t^3 - t^2 < 0 = f(0,0)$ for small $t \neq 0$, so $(0,0)$ is a saddle. The f_{xx} sign matters only once $\det > 0$.

3. Reading convexity off the diagonal of the Hessian

For a diagonal matrix the diagonal entries are the eigenvalues, and it is tempting to read every Hessian that way.

$$\nabla^2 f = \begin{pmatrix} 2 + e^x & c \\ c & 2 \end{pmatrix} \text{ for } f = x^2 + cxy + y^2 + e^x$$

Right so far: this is step 2 of Problem 10.

“A matrix with positive entries on its diagonal is positive definite.”

The shortcut that causes the mistake: true for diagonal matrices, where the diagonal entries are the eigenvalues.

Both diagonal entries, $2 + e^x$ and 2 , are positive for every c , so f is convex for every c ✗

The diagonal entries are $u^\top \nabla^2 f u$ for u along the axes, the curvature in two directions out of all of them. With $c = 3$ at $x = 0$ the Hessian is $\begin{pmatrix} 3 & 3 \\ 3 & 2 \end{pmatrix}$, and along $u = (1, -1)^\top$ the curvature is $3 - 6 + 2 = -1$. The off-diagonal entries count; the determinant brings them in, and convexity needs $|c| \leq 2$ (Problem 10).

4. Using A where its symmetric part belongs

The one-variable rule $\frac{d}{dx}(\frac{1}{2}ax^2 - bx) = ax - b$ suggests an answer that holds only for symmetric A .

$f(x) = \frac{1}{2}x^\top Ax - b^\top x$ with A not symmetric

Right so far: the function is stated correctly, and nothing has been differentiated yet.

“ $\frac{1}{2}x^\top Ax - b^\top x$ is the matrix version of $\frac{1}{2}ax^2 - bx$.”

The analogy that causes the mistake.

$\nabla f = Ax - b$, so Newton's step lands at $A^{-1}b$ ✗

The two occurrences of x give Ax and $A^\top x$, so $\nabla f = \frac{1}{2}(A + A^\top)x - b$ (Problem 6), and f never sees the antisymmetric part of A . With $A = \begin{pmatrix} 2 & 2 \\ 0 & 2 \end{pmatrix}$ and $b = (3, 0)^\top$, $A^{-1}b = (\frac{3}{2}, 0)^\top$, where the true gradient is $(0, \frac{3}{2})^\top$, not 0. The two answers agree only when $A = A^\top$, which is why examples built on symmetric matrices never expose it.

5. Gradient in the place of the row derivative

In one variable the linear term of Taylor's formula is $f'(x)h$, and ∇f is called “the derivative” often enough to be written in its place.

The quadratic term of the second-order Taylor polynomial is $\frac{1}{2}h^\top \nabla^2 f(x) h$

Right so far: it is 1×1 , as a term of a scalar must be (Problem 8).

“ $f(x + h) \approx f(x) + f'(x)h$, and ∇f is the derivative.”

The analogy that causes the mistake: the one-variable formula with ∇f substituted for f' .

$f(x + h) \approx f(x) + \nabla f(x)h + \frac{1}{2}h^\top \nabla^2 f(x)h$ ✗

$\nabla f(x)$ is $n \times 1$ and h is $n \times 1$, so the product does not exist. The linear term uses the row derivative $\partial f / \partial x = \nabla f^\top$, the $1 \times n$ Jacobian of a scalar: $\nabla f(x)^\top h$. In code with 1-D arrays, `g * h` multiplies entry by entry and returns a vector instead of failing, so the “approximation” of a scalar comes out as a vector and nothing raises an error.