

Discussion 2



fig: a pelican on a bicycle
drawn by Qwen 3.8 27B

This week's resources:

- Matrix differentiation guide (written by Randal Barnes) link
- Mathematics for Machine Learning link
- 3Blue1Brown: Essence of Linear Algebra playlist
- Simon Willison's pelican on a bicycle benchmark for LLMs link

1 Proof of K-means Convergence

Consider the K-means algorithm applied to data points $\{x_n\}_{n=1}^N \subset \mathbb{R}^D$ with K clusters. Define binary indicator variables $r_{nk} \in \{0, 1\}$ with $\sum_{k=1}^K r_{nk} = 1$ for each n , and cluster centers $\{\mu_k\}_{k=1}^K$. The K-means objective is

$$J(\{r_{nk}\}, \{\mu_k\}) = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|x_n - \mu_k\|^2.$$

The K-means algorithm alternates between

1. **Update assignments:** Set $r_{nk} = 1$ for the k that minimizes $\|x_n - \mu_k\|^2$ and $r_{nj} = 0$ for all $j \neq k$.
2. **Update means:** Set the cluster centers to

$$\mu_k = \frac{\sum_{n=1}^N r_{nk} x_n}{\sum_{n=1}^N r_{nk}} \quad (\text{for clusters with } \sum_n r_{nk} > 0).$$

Prove that the K-means algorithm converges after a finite number of iterations. (Hint: It is sufficient to show that the objective is monotonically decreasing, and that there are a finite number of iterations).

Solution: We show two things: (A) each assignment or update step does not increase the objective J and decreases it unless a fixed point is reached; (B) there are only finitely many possible assignments $\{r_{nk}\}$, so decrease steps cannot occur forever — hence the algorithm must terminate after finitely many iterations.

(A) The objective never increases.

- *Assignment step decreases (or leaves unchanged) J :* Fix the current centers $\{\mu_k\}$. For each data point x_n , choosing $r_{nk} = 1$ for the nearest center minimizes its contribution to J . Hence J does not increase.
- *Update step decreases (or leaves unchanged) J :* Fix $\{r_{nk}\}$. The objective for each cluster k is

$$J_k(\mu) = \sum_{n:r_{nk}=1} \|x_n - \mu\|^2,$$

minimized uniquely by the mean of the assigned points as shown in lecture:

$$\mu_k^* = \frac{\sum_{n=1}^N r_{nk} x_n}{\sum_{n=1}^N r_{nk}}.$$

Hence updating μ_k to the mean does not increase J .

(B) Only finitely many assignments exist. Each point has K choices independently, so there are at most K^N assignment matrices.

Combine (A) and (B). Since J decreases strictly whenever assignments or centers change, and there are only finitely many assignments, the algorithm must eventually reach a state where neither step changes anything. At that point, K-means has converged after a finite number of iterations.

2 PC-yay!

(a) A warehouse robot records the following four two-dimensional displacement vectors:

$$\mathbf{x}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix} -1 \\ -2 \end{bmatrix}, \quad \mathbf{x}_4 = \begin{bmatrix} -2 \\ -1 \end{bmatrix}.$$

Each coordinate measures displacement in meters. We will use PCA to find the direction in which the robot's motion varies the most.

(i) Compute the sample mean

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i.$$

Solution: The sample mean is

$$\boldsymbol{\mu} = \frac{1}{4} \left(\begin{bmatrix} 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} -1 \\ -2 \end{bmatrix} + \begin{bmatrix} -2 \\ -1 \end{bmatrix} \right) = \boxed{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}.$$

(ii) Using the normalization $1/N$, compute the empirical covariance matrix

$$C = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top.$$

Solution: Since $\boldsymbol{\mu} = \mathbf{0}$,

$$\begin{aligned} C &= \frac{1}{4} \sum_{i=1}^4 \mathbf{x}_i \mathbf{x}_i^\top \\ &= \frac{1}{4} \left(\begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} + \begin{bmatrix} -1 \\ -2 \end{bmatrix} \begin{bmatrix} -1 & -2 \end{bmatrix} + \begin{bmatrix} -2 \\ -1 \end{bmatrix} \begin{bmatrix} -2 & -1 \end{bmatrix} \right) \\ &= \frac{1}{4} \left(\begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} + \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix} \right) \\ &= \boxed{\begin{bmatrix} 5/2 & 2 \\ 2 & 5/2 \end{bmatrix}}. \end{aligned}$$

- (iii) Find the eigenvalues of C and a corresponding unit eigenvector for each eigenvalue.

Solution: The characteristic equation is

$$\det(C - \lambda I) = \left(\frac{5}{2} - \lambda\right)^2 - 4 = 0,$$

so the eigenvalues are $\lambda_1 = \frac{9}{2}$ and $\lambda_2 = \frac{1}{2}$.

To find an eigenvector for an eigenvalue λ , we must find a nonzero vector in the nullspace of $C - \lambda I$ (this comes from rearranging the definition of an eigenvector $C\mathbf{v} = \lambda\mathbf{v} \longrightarrow (C - \lambda I)\mathbf{v} = \mathbf{0}$).

For $\lambda_1 = \frac{9}{2}$,

$$C - \lambda_1 I = \begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix}.$$

Let $\mathbf{v} = \begin{bmatrix} a & b \end{bmatrix}^\top$ and solve for the corresponding eigenvector:

$$\begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies -2a + 2b = 0 \implies b = a.$$

We may choose $a = 1$, which gives $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Its length is $\sqrt{2}$, so dividing by $\sqrt{2}$ gives the unit eigenvector

$$\mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

For $\lambda_2 = \frac{1}{2}$,

$$C - \lambda_2 I = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}.$$

In this case,

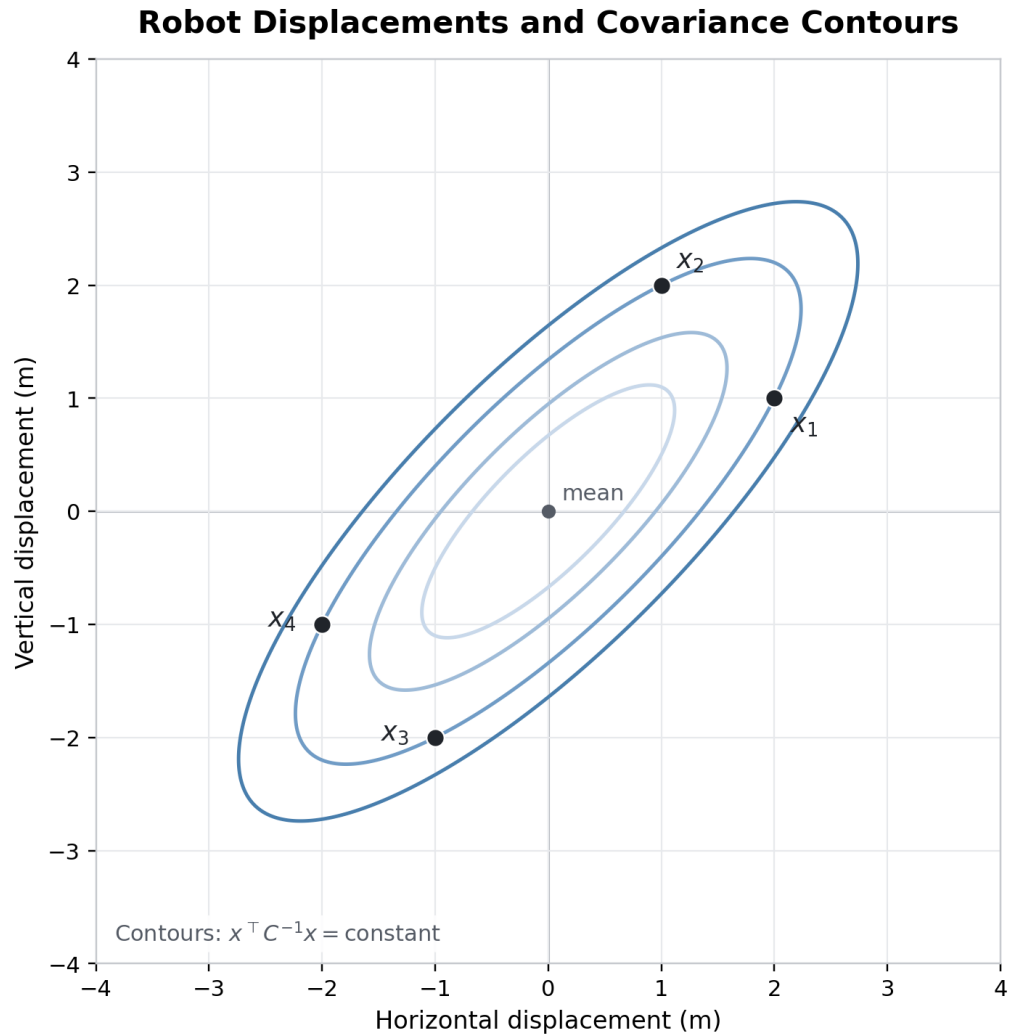
$$\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies 2a + 2b = 0 \implies b = -a.$$

Choosing $a = 1$ gives $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$, which also has length $\sqrt{2}$. Therefore, the corresponding unit eigenvectors are

$$\boxed{\mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Any nonzero multiple of an eigenvector points in the same eigenvector direction. In particular, negating either eigenvector gives an equivalent answer; we normalize them to length one for PCA.

- (iv) The plot below shows the four robot displacements and several contours of equal covariance-scaled distance from the mean. Starting at the mean, draw and label the two eigenvector directions $\mathbf{v}_1$ and $\mathbf{v}_2$. Make the length of each arrow proportional to $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$, respectively, using the same scale factor for both arrows. Which eigenvector is the first principal component? Describe its physical meaning for the robot.



Solution: The first principal component is the eigenvector associated with the largest eigenvalue. Therefore,

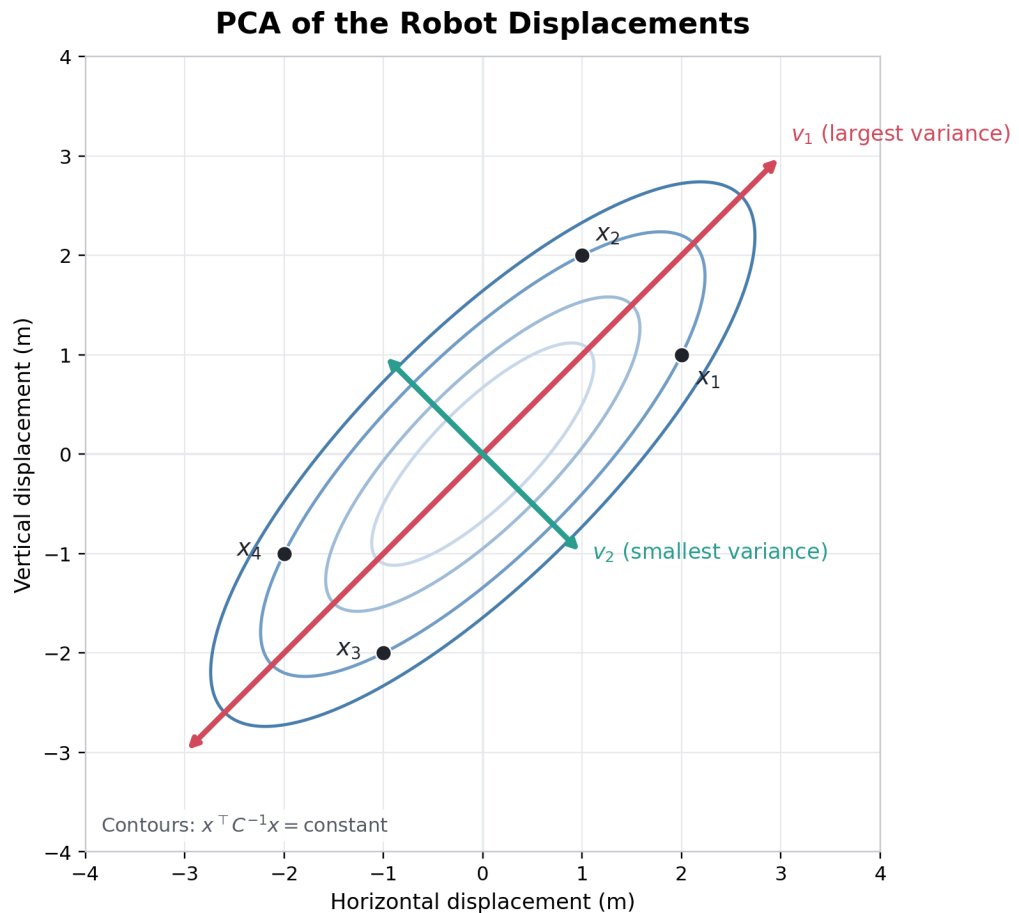
$$\mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

This vector points along the line $y = x$, so the robot's displacement varies most in the diagonal direction.

The contours below show points with equal covariance-scaled distance from the mean, $\mathbf{x}^\top \mathbf{C}^{-1} \mathbf{x} = \text{constant}$. Their long axis is parallel to $\mathbf{v}_1$, where the data has the most variance, while their short axis is parallel to $\mathbf{v}_2$, where the data has the least variance. Since arrow length represents standard deviation along each direction, the ratio of the two arrow lengths is

$$\frac{\sqrt{\lambda_1}}{\sqrt{\lambda_2}} = \sqrt{\frac{9/2}{1/2}} = 3.$$

Thus, the $\mathbf{v}_1$ arrow should be three times as long as the $\mathbf{v}_2$ arrow. Either orientation along an eigenvector direction is correct because negating an eigenvector gives an equivalent eigenvector.



- (v) Project each centered displacement onto the first principal component by computing $z_i = \mathbf{v}_1^\top (\mathbf{x}_i - \boldsymbol{\mu})$.

Solution: Since $\boldsymbol{\mu} = \mathbf{0}$,

$$z_1 = \frac{3}{\sqrt{2}}, \quad z_2 = \frac{3}{\sqrt{2}}, \quad z_3 = -\frac{3}{\sqrt{2}}, \quad z_4 = -\frac{3}{\sqrt{2}}.$$

- (vi) Suppose the robot stores only the one-dimensional values z_i . What information about its motion is lost? Give a specific example from the four recorded displacements.

Solution: Keeping only z_i discards motion in the direction of the second principal component, $\mathbf{v}_2$. For example, $\mathbf{x}_1 = (2, 1)^\top$ and $\mathbf{x}_2 = (1, 2)^\top$ both have projection $3/\sqrt{2}$, so they become indistinguishable after compression. The representation preserves their common diagonal motion but loses the difference between their horizontal and vertical displacements.

- (vii) The robot later records a new displacement

$$\mathbf{x}_{\text{new}} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}.$$

Project this displacement onto the first principal component, then reconstruct its two-dimensional approximation using

$$\hat{\mathbf{x}}_{\text{new}} = \boldsymbol{\mu} + \left(\mathbf{v}_1^\top (\mathbf{x}_{\text{new}} - \boldsymbol{\mu}) \right) \mathbf{v}_1.$$

In a safety-critical robot, why might replacing $\mathbf{x}_{\text{new}}$ with this approximation be problematic?

Solution: The one-dimensional projection is

$$z_{\text{new}} = \mathbf{v}_1^\top \mathbf{x}_{\text{new}} = \frac{1}{\sqrt{2}}(3 + 1) = 2\sqrt{2}.$$

Therefore, the reconstruction is

$$\hat{\mathbf{x}}_{\text{new}} = z_{\text{new}} \mathbf{v}_1 = 2\sqrt{2} \left(\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) = \begin{bmatrix} 2 \\ 2 \end{bmatrix}.$$

The discarded displacement is

$$\mathbf{x}_{\text{new}} - \hat{\mathbf{x}}_{\text{new}} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Although this direction has lower variance in the observed data, it may still contain important information. For example, losing a horizontal or vertical deviation could cause the robot to misjudge its position relative to an obstacle. Thus, low variance does not necessarily mean unimportant.

3 Calculus

(a) Consider the function

$$f(x, y) = \sigma(ax + by),$$

where $a, b \in \mathbb{R}$ and $\sigma(t) = \frac{1}{1 + e^{-t}}$ for $t \in \mathbb{R}$.

(i) Show that $\frac{d\sigma}{dt} = \sigma(t) (1 - \sigma(t))$.

Solution:

$$\frac{d}{dt}\sigma(t) = \frac{d}{dt} \left(\frac{1}{1 + e^{-t}} \right) \tag{1}$$

$$= \frac{e^{-t}}{(1 + e^{-t})^2} \tag{2}$$

$$= \left(\frac{1}{1 + e^{-t}} \right) \left(\frac{1 + e^{-t}}{1 + e^{-t}} - \frac{1}{1 + e^{-t}} \right) \tag{3}$$

$$= \boxed{\sigma(t) (1 - \sigma(t))} \tag{4}$$

(ii) Using the result you showed in part (i) and the chain rule, compute $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$.

Solution: First let's define a new variable $u = ax + by$ such that $f = \sigma(u)$. By chain rule, we have

$$\frac{\partial f}{\partial x} = \frac{\partial \sigma}{\partial u} \frac{\partial u}{\partial x}.$$

We can calculate $\frac{\partial u}{\partial x} = a$, giving us, in conjunction with the result from part (i),

$$\frac{\partial f}{\partial x} = \sigma(u) (1 - \sigma(u)) \cdot a.$$

Substituting $ax + by$ back in for u gives us

$$\frac{\partial f}{\partial x} = \boxed{a\sigma(ax + by) [1 - \sigma(ax + by)]}.$$

Following the same process for $\frac{\partial f}{\partial y}$, we get

$$\frac{\partial f}{\partial y} = \boxed{b\sigma(ax + by) [1 - \sigma(ax + by)]}.$$

(b) For $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$, define

$$g(\mathbf{x}) = \sum_{j=1}^n x_j^2.$$

Compute the partial derivative $\frac{\partial g}{\partial x_i}$ for a generic coordinate $i \in \{1, \dots, n\}$.

Solution: Since $g(\mathbf{x}) = \sum_{j=1}^n x_j^2$, we have

$$\frac{\partial g}{\partial x_i} = \boxed{2x_i}.$$

(c) Let $\mathbf{w} \in \mathbb{R}^n$ be a constant vector and define the scalar function

$$h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} = \sum_{j=1}^n w_j x_j.$$

Compute $\frac{\partial h}{\partial x_i}$ for a generic coordinate $i \in \{1, \dots, n\}$.

Solution: With $h(x) = \sum_{j=1}^n w_j x_j$, treating $\mathbf{w}$ as constant,

$$\frac{\partial h}{\partial x_i} = \boxed{w_i}.$$

(d) Let $\mathbf{w} \in \mathbb{R}^n$ be a constant vector and define the scalar function

$$\ell(\mathbf{x}) = \frac{1}{2} \sum_{j=1}^n x_j^2 - \mathbf{w}^\top \mathbf{x} = \frac{1}{2} g(\mathbf{x}) - h(\mathbf{x}).$$

In single-variable calculus, we can minimize/maximize a function $f(x)$ using its derivatives. Here ℓ takes a whole vector $\mathbf{x}$ as input instead of a single number, but we can reuse that same idea one coordinate at a time.

(i) Using your answers to parts (b) and (c), compute $\frac{\partial \ell}{\partial x_i}$ for a generic coordinate $i \in \{1, \dots, n\}$. Stacking these n partial derivatives into a vector gives the *gradient*

$$\nabla \ell(\mathbf{x}) = \begin{bmatrix} \frac{\partial \ell}{\partial x_1} \\ \dots \\ \frac{\partial \ell}{\partial x_n} \end{bmatrix}.$$

Solution: By linearity, $\frac{\partial \ell}{\partial x_i} = \frac{1}{2} \frac{\partial g}{\partial x_i} - \frac{\partial h}{\partial x_i}$. Substituting the results from parts (b) and (c),

$$\frac{\partial \ell}{\partial x_i} = \frac{1}{2} (2x_i) - w_i = \boxed{x_i - w_i}.$$

This one formula holds for every coordinate i . Stacking all n of them into a vector gives the gradient

$$\nabla \ell(\mathbf{x}) = \begin{bmatrix} x_1 - w_1 \\ \vdots \\ x_n - w_n \end{bmatrix} = \mathbf{x} - \mathbf{w}.$$

- (ii) Recall from single-variable calculus that we find values of x that minimize/maximize $f(x)$ by solving $\frac{df}{dx} = 0$; the solutions are called *critical points*. In vector-calculus, a critical point of ℓ is a vector $\mathbf{x}$ where the partials $\frac{\partial \ell}{\partial x_i} = 0$ for all i components of $\mathbf{x}$ (equivalently $\nabla \ell(\mathbf{x}) = \mathbf{0}$).

Set all n partial derivatives equal to zero and solve for the minimizer $\mathbf{x}^*$. (*Hint: the i -th component of $\mathbf{x}^*$ is the value of x_i that makes $\frac{\partial \ell}{\partial x_i} = 0$; solve for one coordinate and then stack each component into the vector $\mathbf{x}^*$.)*

Solution: A critical point needs every partial derivative to equal 0. Because the gradient is component-wise, this splits into n independent single-variable equations:

$$\frac{\partial \ell}{\partial x_i} = x_i - w_i = 0 \implies x_i = w_i, \quad i = 1, \dots, n.$$

Each coordinate is pinned down on its own, and assembling them gives

$$\boxed{\mathbf{x}^* = \mathbf{w}.$$

- (iii) In single-variable calculus, once we have a critical point we check the sign of the second derivative: $f''(x) > 0$ means the function f is concave up, so the point is a minimum. Use this same second-derivative idea to argue that $\mathbf{x}^*$ is a minimum, and not a maximum or a saddle point.

Solution: Component-wise (single-variable) view. Group ℓ term by term. Since $g(\mathbf{x}) = \sum_j x_j^2$ and $h(\mathbf{x}) = \sum_j w_j x_j$,

$$\ell(\mathbf{x}) = \sum_{j=1}^n \left(\frac{1}{2} x_j^2 - w_j x_j \right).$$

Notice that each term in the sum only involves one coordinate of $\mathbf{x}$. So minimizing ℓ is equivalent to minimizing each term in the sum independently. Each term $\frac{1}{2} x_j^2 - w_j x_j$ is an upward-opening parabola whose second derivative is $\frac{\partial^2 \ell}{\partial x_j^2} = 1 > 0$, so by the single-variable second-derivative test, each critical point $x_j = w_j$ is a minimizer. Minimizing each term also minimizes the entire sum because we add all the terms together. Hence $\mathbf{x}^* = \mathbf{w}$ is a minimum.

Vector view. The second-derivative test generalizes by collecting all the second partials into the *Hessian* matrix $H_{ij} = \frac{\partial^2 \ell}{\partial x_i \partial x_j}$. Here

$$\frac{\partial^2 \ell}{\partial x_i^2} = 1, \quad \frac{\partial^2 \ell}{\partial x_i \partial x_j} = 0 \quad (i \neq j),$$

so $H = I$, the identity matrix. This Hessian is positive definite ($\mathbf{v}^\top H \mathbf{v} > 0$ for $\forall \mathbf{v} \neq \mathbf{0}$), meaning that it "curves" upwards in every direction. If the Hessian is positive definite at $\mathbf{x}^*$, then the function is convex and $\ell(\mathbf{x}^*)$ is a minimum.

Completing the square. Completing the square gives $\ell(\mathbf{x}) = \frac{1}{2} \|\mathbf{x} - \mathbf{w}\|^2 - \frac{1}{2} \|\mathbf{w}\|^2$: a squared distance (never negative, zero only at $\mathbf{x} = \mathbf{w}$) minus a constant. The smallest possible value this can take on is when the first term is zero ($\mathbf{x} = \mathbf{w}$).